Analyse de texte et acqusition des connaissances : aspects méthodologiques

Louis-Claude PAQUIN
Luc DUPUY
Centre d'Analyse de Textes par Ordinateur
Université du Québec à Montréal

Yves ROCHON
Environnement Québec

Résumé

Dans le domaine des évaluations environnementales, les problèmes qui se posent avec une importance particulière sont ceux du stockage, de l'indexation conceptuelle et l'exploitation d'immenses bases de données textuelles en format libre. Ces organisations sont appelées à devenir des utilisateurs importants de systèmes et elles demandent certainement toute la pertinence des systèmes experts. Un des problèmes nodaux de ce secteur du traitement des connaissances est l'analyse des demandes de projets. Dans ce qui suit nous illustrons une série de traitements dont l'objectif est d'automatiser, dans le cadre d'un projet de système expert, certaines des dimensions de l'analyse textuelle de ces demandes. Nous présentons quelques-unes des utilisations du logiciel SATO en espérant montrer la pertinence de celui-ci pour la constitution de dictionnaires de connaissances qui sont des composant socio-terminologiques importants d'un système expert.

0. Introduction

Le texte qui suit illustre les éléments d'une méthode pour l'acquisition des connaissances à partir de l'analyse de données textuelles, méthode qui n'en est qu'à ses premières ébauches. Une méthode unifiée demeure, jusqu'à plus ample informé, hors de portée. Toutefois, les résultats obtenus, même à titre exploratoire, nous semblent assez intéressants pour être soumis à un public plus large. Afin de contextualiser les opérations d'analyse, le projet SAGÉE est présenté, ainsi que quelques considérations sur l'acquisition des connaissances. C'est ainsi qu'une série de traitements seront un à un présentés, commentés et illustrés. Le propos qui suit est un compte-rendu ayant un caractère plus empirique que théorique. L'objectif visé est d'abord et avant tout d'illustrer certains aspects d'une méthode d'analyse encore en voie de développement. Le lecteur trouvera ailleurs (Paquin, L.-C. et Dupuy, L. 1989) l'encadrement qui sert de fond théorique à la démarche analytique présentée ici.

1. Le projet SAGÉE

À la Direction des évaluations environnementales du Ministère de l'Environnement du Québec les auteurs travaillent depuis plus de trois ans à un projet de système expert SAGÉE (Système d'Aide à la Gestion des Évaluations Environnementales). SAGÉE vise à supporter 25 spécialistes issus de diverses disciplines, ayant des points de vue différents sur l'environnement et sur le sens à donner à leur tâche dans l'application de la Loi sur la qualité de l'environnement du Québec. Ce support ne consiste pas en l'automatisation de tout le processus d'évaluation de projets. Il vise plutôt l'amélioration de l'accès aux informations pertinentes dans l'accomplissement des tâches reliées à cet exercice très complexe qu'est l'évaluation de projets ayant des incidences sur l'environnement comme, par exemple, un projet de marina. Cette

complexité est attribuable à divers paramètres : l'ampleur et à la diversité des champs d'activités couverts {projet industriel, aménagement portuaire, infrastructure routière, etc.}, la multiidisciplinarité des connaissances requises {écologie, hydrologie, sociologie, ingénierie, etc.}, la variété des données, le caractère ponctuel des impacts, etc. Cette situation se complique à cause de l'absence de cadre de référence solidement établi. En effet, le caractère récent et pluri-culturel des sciences environnementales fait qu'il existe peu d'approches méthodologiques, théoriques, normatives ou standardisées permettant de valider les modélisations produites.

SAGÉE poursuit un quadruple objectif. Le premier vise un aspect informationnel : «identifier et récupérer les données et les connaissances pertinentes au traitement d'un dossier parmi l'ensemble des données et connaissances en circulation au Ministère et au Gouvernement». Le deuxième concerne la vigilance en matière d'analyse environnementale: assurer un bon enchaînement des «gestes administratifs associés à la procédure d'examen et d'évaluation des impacts sur l'environnement des grands projets en gestation au Québec». Le troisième objectif touche une dimension qualitative : aider les chargés de projet à mieux percevoir les enjeux liés aux dossiers à traiter, leur éviter d'ignorer certains aspects ou de traiter certains autres trop en profondeur. Le quatrième objectif concerne le processus administratif : on voudrait profiter de la rétroaction sur la procédure du traitement des dossiers dans le cadre de l'administration du règlement. Au total, en tant que «projet pilote de système expert gouvernemental», SAGÉE explore la possibilité de «lier les informations contenues dans les bases de données gouvernementales et les connaissances propres au système expert» et surtout de mettre au point une méthodologie d'acquisition des connaissances qui tienne compte de la spécificité du milieu des analystes du Ministère de l'Environnement.

Dès le début du projet, on a su que ce type de développement informatique nécessitait le recours à la technologie des systèmes experts (SE). En effet, les premiers essais de structuration du domaine d'expertise issus des méthodologies de programmation usuelle (lire ici les stratégies de développement de bases de données...) avec déclaration des variables et élaboration de modèles se sont avérés infructueux, principalement à cause de la faiblesse de ces outils pour structurer des ensembles complexes de variables. Par ailleurs, la technique des SE n'épuisait pas l'ensemble des fonctionnalités requises par la tâche à accomplir, elle devait être couplée à la gestion de bases de données (GBD). L'architecture de SAGÉE était conçue comme l'adjonction d'un module d'inférences à une base de donnés. Il s'agissait alors de déterminer la place de chacune de ces deux technologies (SE et GBD) et les modalités de leurs échanges d'information. Une modélisation commune semblait nécessaire, nous avons donc par la suite eu recours au formalisme entité-relation utilisé en développement de base de données (Valiquette, L. et Béland, R. 1988). À l'usage, ce formalisme s'est avéré lourd pour représenter l'enchevêtrement complexe des concepts relatifs à l'environnement; il a quand même permis de faire un certain découpage de cette connaissance.

Suite à l'analyse de la retranscription d'une quinzaine d'entrevues réalisées avec les chargés de projet, nous avons constaté que leur apport cognitif est certes important puisque c'est là que résident les heuristiques, mais qu'il est loin d'épuiser le domaine. Les entrevues nous ont plutôt servi à dresser l'inventaire des sources d'information : lois, règlements, directives, manuel du chargé de projet de la Direction des Évaluations Environnementales (DÉE) et les documents afférents au traitement des dossiers : rapport d'analyse, proposition de décrets et correspondance. Toutes ces sources sont de nature textuelle et sont abondantes; nous utilisons actuellement un ensemble de textes totalisant plus de 16 méga-octets, soit approximativement 5 000 pages de texte. En fait, la plupart des données manipulées par SAGÉE sont de type textuel. Dans le cas de l'avis de projet qui initie le processus d'évaluation, la proportion d'informations en texte libre représente neuf-dixième du volume des données. En voici quelques sections:

- Objectifs et justification du projet
- Description du projet
- Description du milieu et des principales contraintes
- Principales répercussions appréhendées
- Remarques

En fait, le processus de gestion d'un projet utilise et génère une grande quantité de textes. Pour traiter ces informations, il faut en effectuer manuellement la mise en forme. Ceci implique l'établissement d'une correspondance entre les formulations trouvées dans la demande de projet et celle prévues dans les limites du modèle conceptuel. Cette normalisation comporte le danger que le dossier traité ne soit pas conforme à celui qui a été soumis.

La définition des concepts et de leurs relations requiert un effort, souvent considérable, de structuration et de standardisation. Ces mêmes opérations doivent être effectuées sur les textes mais selon des modalités différentes. Il ne faudrait pas, cela serait une erreur irréparable, découper les textes pour en faire des données contraintes. Il faut plutôt utiliser les techniques d'indexation qui servent habituellement à l'élaboration d'un thesaurus pour une base de données documentaire : les locutions, les relations de synonymie, d'association, de généralisation et de spécification, etc. Les développeurs de SAGÉE se sont trouvés, en résolvant leur besoin cognitif au moyen de textes, à tirer bénéfice des connaissances contenues dans la production textuelle de l'organisation. Incidemment, ce type de stratégie analytique permet la valorisation de la production textuelle de l'organisation, production qui est généralement reléguée aux oubliettes institutionnelles.

Les procédures d'analyse des textes (ATO) que nous avons mis au point visent à réaliser un dictionnaire de concepts. Ce dictionnaire représente la base de notre travail de modélisation du flux des connaissances. Il est important autant pour la formulation des règles de production que pour l'enchaînement et le contenu des panoramas de la base de données. Cet outil de schématisation nous donne une vue d'ensemble du domaine de concepts à modéliser. Il n'est pas constitué à partir de l'idée que l'on se fait des concepts mais par la dynamique propre du discours des analystes du Ministère de l'Environnement. La matière textuelle du discours des experts sert de substrat pour l'élaboration d'un schéma conceptuel et son opérationnalisation dans le contexte de la maquette SAGÉE. Ce dictionnaire nous a ainsi permis de valider la modélisation conceptuelle des données. Il nous aide à avoir une meilleure connaissance des traitements administratifs et intellectuels à l'oeuvre dans le processus d'analyse de dossiers.

Actuellement, nous sommes à étendre la structuration de nos connaissances à ses proportions définitives : l'ensemble des connaissances utilisées dans le traitement de l'ensemble des types de projets. Le but est de définir les lien entre les items d'un type de projet et les impacts appréhendés. L'analyse des rapports produits par l'organisation nous a permis de décomposer tout projet, projet de route par exemple, au niveau des infrastructures nécessaires et des activités à réaliser. Ces éléments conjugués aux caractéristiques du milieu nous permettent de prédire, au moyen des règles d'inférences appropriées, l'évolution de chaque élément saillant vers des impacts attendus.

2. L'acquisition des connaissances

Le processus d'acquisition des connaissances est d'emblée reconnu par la majorité des textes de la littérature (Boose,J. et Gaines, B. 1990; Tourigny, N. et Simian, G. 1990, pour ne donner que les cas les plus récents) comme un processus très complexe qui dépasse largement le fait d'«extraire» de la tête d'un expert les connaissances de celui-ci. Se posent notamment d'importantes difficultés comme :

- le manque d'histoires de cas détaillées, la difficulté d'avoir accès à des outils manipulant des données normés et standardisés et la diversité, quelque fois hétéroclite, des modes de représentation;
- les problèmes de portabilité posés par le développement d'applications trop spécifiques,
- la trop grande diversité des modèles ergonomiques;
- les contraintes «machines» qui sectorisent les types et genres de développements.

Il n'est pas de notre compétence de régler ici ces problèmes. En fait, nous voulons en ajouter un à la liste : celui des archives textuelle. En effet, dans plusieurs des secteurs on retrouve une quantité importante d'archives textuelles contenant un volume important de connaissances. Et si, paradoxalement, le texte est le principal véhicule des connaissances scientifiques ou techniques, il est un des «experts» les moins étudiés du point de vu des processus socio-cognitifs, exception

faite du domaine juridique où le raisonnement «textuel» est de plus en plus modélisé (pour une étude récente Moulin, B. 1990). Paradoxalement, depuis Ericsson et Simon en 1984 (Ericsson, K. A. et Simon, H. A. 1984) on parle d'analyse de protocoles et d'entrevue effectuées auprès des experts alors qu'on néglige de généraliser ce type d'analyse au cas des archives textuelles des organisations qui pourtant non seulement fondent celles-ci «légalement» , mais en constituent les fondations procédurales. Il faut toutefois reconnaître que la saisie des documents, dans la plupart des domaines administratifs, demeure encore une des principales difficultés d'accès aux archives textuelles.

Dans le cadre du projet SAGÉE, le terme acquisition de la connaissance est conçu, pour la phase courante, comme un processus de traduction des éléments socio-cognitifs contenus dans les archives textuelles. Il s'agit de traduire les éléments de la procédure d'examen et d'évaluation des impacts sur l'environnement pour l'adapter à chacune des demandes de projet. Cette adaptation représente une contrainte importante : il faut la réaliser en tenant compte du fait que 25 spécialistes sont impliqués dans le processus d'évaluation. Il faut donc s'assurer que le langage de description utilisé fasse consensus. Il faut s'assurer également que ce langage soit élaboré à partir des documents existants, compte tenu de la dimension «historique» que l'on y retrouve : histoire du développement de la pratique d'analyse des dossiers environnementaux, histoire des changements législatifs, etc. L'expertise dans le secteur des évaluations environnementales n'est pas localisée de manière spécifique chez un expert ou un groupe d'experts. Il s'agit d'une famille de points de vue et de perspectives qui visent des objectifs communs mais qui sont singulièrement différents. Les archives textuelles représentent donc un matériaux très important qu'il faut exploiter au maximum avant d'envisager de nouvelles manières d'«extraire» la connaissance.

La traduction doit produire une description formelle des concepts qui permette l'accomplissement d'une chaîne d'inférences. Une telle description s'avère problématique dans la mesure où nous nous confrontons à la fluidité des concepts qui ne sont pas, à l'instar des formes solides, des entités ayant des limites tangibles et concrètes. Prenons, par exemple, le concept d'«impact appréhendé majeur» que l'analyste doit identifier dans le texte de la demande. Une telle notion est d'autant plus problématique que dans les textes, on y accède rarement de façon directe : l'accès au concept est médiatisé par les termes spécifiques de la demande. Autrement dit, les termes désignent contextuellement une instance particulière du concept. En effet, dans les textes, l'effet de référence est largement tributaire des formes nominales auxquelles on associe le processus de dénomination (Rey-Debove, J. 1976). L'effet de référence n'est que rarement le fait du terme isolé; il est habituellement consolidé, spécifié, qualifié, élaboré par d'autres références {épithète, complément du nom, proposition relative}. Certaines expressions nominales exercent une fonction de régie sur d'autres formes et les caractérisent ou les spécifient; par exemple, la notion de longueur dans l'expressions «la longueur du quai».

Quiconque veut faire l'acquisition de la connaissance dans les textes doit, même de façon intuitive, partir des termes dépistés dans les textes pour "remonter" vers les concepts. La recherche de concepts qui seraient formulés clairement et explicitement dans les textes est habituellement une première expérience fort décevante. Les définitions sont partielles, contextualisées ou relatives à d'autres concepts, par là même peu utilisables directement parce que locales. Les concepts ne sont pas déposés dans les textes, ils sont fabriqués par l'interaction du lecteur et du texte (Deschênes, A.-J. 1988). Un passage systématique et rigoureux des termes aux concepts nécessite un cadre ou modèle de «concept» qui tienne autant compte du point de départ, les contextes, que du point d'arrivée, les concepts utilisés pour la construction des règles d'inférences. Les concepts ne sont pas seulement des prédicats qui régissent des arguments (leurs caractéristiques), mais des objets valués dynamiquement c'est-à-dire dotés de caractéristiques qui les définissent comme centres de régie dans les processus de raisonnement. Le modèle des concepts comme combinaison de «primitives» décrivant la réalité s'avère le plus intéressant parce qu'il est autant efficace dans le cadre logique des systèmes experts que dans le cadre morpho-syntaxique des textes. Les objets valués conviennent autant à la rédaction des règles d'inférences qu'à la description du groupe nominal. Ce modèle est cependant réducteur parce qu'il ne permet pas la prise en compte aisée du phénomène textuel de la re-catégorisation, soit lorsqu'un terme est mis à la place d'un autre, plus général, plus spécifique.

Dans les textes les concepts ne sont donc pas directement accessibles. Un terme, c'est-à-dire l'expression linguistique du concept, pris dans un contexte donné, est accompagné d'une consolidation sémantique particulière de caractéristiques alors que le concept est une forme schématique qui «encapsule» les consolidations possibles. Les quelques éléments de méthode que nous proposons pour constituer un dictionnaire de concept permettent d'isoler, à partir de leur récurrence dans des corpus donnés, les régularités socio-terminologiques (Lerat, P. 1989). Il s'agit de mettre progressivement à jour leur organisation en terme de configuration et de les inscrire dans la hiérarchie congnitive globale pressentie.

Pour reconstituer ces configurations conceptuelles, nous utilisons des patrons morpho-syntaxiques, ce qui garantit un dépistage indépendant des problématiques définies dans les textes. Le recours au jugement des experts intervient après coup pour valider et réduire le matériau cognitif recueilli. Au développement d'une application ou à la mise au point d'une grilles de codification, nous avons privilégié une méthodologique par phases de traitements interactifs. Ceci pour deux raisons. D'une part, la logique du langage symbolique de l'application impose souvent des contraintes et des artifices techniques dans la production de descriptions du savoir du domaine. D'autre part, la projection de grilles de codification sur le texte entraîne une déstructuration de l'énonciation de la réalité. Nous croyons qu'une suite de traitements interactifs permet aux développeurs de systèmes experts de mettre progressivement à jour le processus d'élaboration de la connaissance dans le discours.

L'analyse des textes requiert une stratégie de type constructiviste. Le sens des termes ne se donne pas dès la première lecture. La lecture humaine est un processus séquentiel (non-indexé...) et sélectif; il fonctionne par cycles successifs et il est largement tributaire d'une perspective qu'il faut sans cesse ré-investir dans le travail de lecture. L'analyse de texte possède elle aussi son cycle de vie et de développement. Le cycle nous conduit d'une première lecture à une phase d'annotation ou de mise en relief de certains éléments vers une phase de synthèse et ceci récursivement jusqu'à saturation des effets de sens recherchés. Le «sens» d'un texte se construit donc au fur et à mesure que le lecteur investit le texte de ses questions. Autrement dit, le sens des composants du texte ne se donne pas immédiatement : il est la résultante des cycles question-réponse où le lecteur sélectionne autant de parcours textuels susceptibles d'être des éléments de réponse. Dans cette optique, un logiciel d'analyse de texte devient particulièrement utile car il permet à un analyste de conserver les traces de ses opérations de lecture obtenant ainsi autant de versions du textes que le requièrent les dimensions de la problématique d'analyse.

3. Aspects de l'analyse de texte par ordinateur

3.1 Gestion du flux documentaire et organisation des corpus d'analyse

Deux éléments ont entraîné une augmentation de l'importance accordée aux archives textuelles sur support informatique. La prise en compte de la capacité des nouvelles technologies informatiques à exploiter les connaissances des textes et l'intérêt porté au contenu de ces documents par le personnel du projet SAGÉE. Peu à peu, les fonctions de production et gestion de données textuelles s'ajoutent à la déjà traditionnelle fonction de production de texte papier qui est généralement remplie par la micro-informatique. Ce besoin ce faisant encore plus pressant puisque même au niveau de la production textuelle, les gens du secrétariat du projet éprouvent des difficultés à retrouver ces textes, difficulté qui tend à s'accroître avec le temps car le professionnel est amené à devenir de plus en plus un producteur de textes autonome grâce aux ressources de la micro-informatique. De plus, l'analyse des textes nécessite une structuration de ces derniers pour permettre une meilleure manipulation des documents.

En regard de ces besoins généraux et de nos besoins spécifiques, nous avons travaillé à l'élaboration d'un mécanisme facilitant la transformation du texte «document-papier» en texte «donnée». Ce mécanisme qui est présentement en voie d'installation, consiste lier le texte produit à une base de données pour en facilité la gestion documentaire. Pour ce faire, nous avons instauré une nomenclature des noms des fichiers informatiques selon les principes suivants:

Un documents correspond à un fichier;

L'unité de base de l'identification d'un fichier est le numéro de dossier auquel il appartient et sa version;

Chaque responsable des dossiers a son propre répertoire;

Celui-ci se divise en sous-répertoire correspondant à des type de document (e.i. DIR pour directive) et à l'intérieur desquels se trouve les documents (fichiers) identifiés par le numéro de dossier;

Suite à ceci nous avons centralisé les documents produits sur un serveur via un réseau local de micro-ordinateur puis nous avons associé dans une base de données les nom des fichiers (document) aux informations relatives aux dossiers.

Cette association générée automatiquement permet de définir l'auteur du document, sa nature et son objet. Le nom du fichiers informatique est considéré comme une cote à l'image de la cote documentaire d'une bibliothèque. La gestion des fichiers par une base de données offre des possibilités très intéressantes comme : repérer des documents par diverses façons pour les lire, les archiver ou en faire l'analyse ou encore étudier la production d'une unité pour une période donnée.

La gestion du flux documentaire représente un moment important de la constitution des archives textuelles. Ceci nous permet de faire une analyse pour un type de document, pour une période de temps donnée, ou encore pour un type de projet. Cette façon de faire offre une grande souplesse pour permettre l'analyse de corpus en fonction des problématiques spécifiques (par exemple : on peut étudier comment s'organise le concept d'impact dans les rapports d'analyse environnementale des projets d'aménagement routier). Une gestion documentaire efficace et adaptée aux réalités de l'institution permet de faire un premier dégrillage de documents potentiellement pertinents. Il demeure cependant que la gestion des documents ne suffit pas à elle seule au travail d'analyse de contenu des textes. Il faut ensuite enrichir le traitement à l'aide de techniques d'analyse de contenu des textes (Bardin, L. 1989).

Étant donné que le projet SAGÉE dispose d'un mécanisme de gestion du flux documentaire, la constitution des corpus (regroupement de textes pour une analyse donnée) est une opération très simple. Il suffit à l'analyste de sélectionner dans la base de donnée documentaire les éléments appropriés et de constituer un fichier-maître qui pourra être soumis au logiciel d'analyse de texte SATO (Daoust, F. 1989).

3.2 Étapes préparatoires de l'analyse de contenu des textes

3.2.1 Le logiciel SATO (Système d'Analyse de Textes par Ordinateur)

La quantité de données textuelles utilisée par les analystes du projet SAGÉE les oblige à utiliser des techniques informatiques d'analyse de contenu des documents. La masse des données se caractérise par le fait qu'elle représente un volume important de documents (au moment de la rédaction de ce texte : 30 méga-octets). Elle se caractérise aussi par le fait que la base documentaire est constamment enrichie et mise à jour. Aussi, pour développer le dictionnaire des connaissances du projet SAGÉE, nous avons utilisé les ressources d'un logiciel. Il s'agit de SATO, un Système d'Analyse de Textes par Ordinateur développé par François Daoust du Centre d'Analyse de Textes par Ordinateur de l'Université du Québec à Montréal. SATO est utilisé sur du matériel de type IBM-PC-AT™ et tourne sous la version 3.3 de DOS™. Il peut avantageusement être utilisé avec des logiciels multitâches comme WINDOWS™ (version 3.0) ou DESQVIEW™, ce qui permet d'augmenter la productivité et l'interactivité du travail d'analyse.

Ce logiciel interactif permet à un usager d'analyser les dimensions lexicale et textuelle de ses documents. Dans le premier cas, l'usager dispose d'un index alphabétique exhaustif des mots d'un document; ici on a accès aux éléments de manière hors-contexte. Dans le second cas, l'usager peut définir et consulter différents segments de son document. En plus de pouvoir traiter des collections multilingues (4 alphabets différents par corpus) SATO offre la possibilité d'annoter les mots ou segments de mots, c'est-à-dire de leur adjoindre des descripteurs numériques (des chiffres comme dans le cas de dénombrements) ou symboliques (des étiquettes

alpha-numériques). Par exemple, le mot «impact» pourrait avoir comme descripteur une fréquence de «32» et l'étiquette grammaticale «nomcommun». Les annotations s'appliquent aussi bien au lexique (liste hors-contexte des mots) que dans le texte (le mot et ses contextes d'occurrence). L'usager peut annoter manuellement ou automatiquement les éléments de ses documents en ayant le recours à des dictionnaires ou par des stratégies de fouilles contextuelles, par exemple associer l'étiquette «rôle professionnel» à la chaîne "chargé de projet" dans tout le texte.

SATO met à la disposition de l'usager un langage de requête simple et efficace. Celui-ci admet comme élément de recherche soit l'expression littérale des éléments ou une combinaison de caractères et de caractères de remplacement permettant notamment des jeux de troncation des parties gauche ou droite des chaînes de caractères. Il faut remarquer que la syntaxe s'applique autant aux mots du texte ou du lexique qu'aux descripteurs utilisés pour l'indexation des mots ou segments de mots. Par exemple, le patron de fouille <chargé> dépistera la chaîne «chargé» alors que le patron <|ent> dépistera tous les mots qui se terminent par <ent> comme «parlent», «comment», etc. Les requêtes peuvent être une combinaison de mots et de descripteurs; en effet, un patron de fouille comme

«|ent*type=bio-physique*fréquence>34»

dépistera toutes les chaînes se terminant par <ent>, dont le type est bio-physique et dont la fréquence dans le corpus est plus grande que 34.

Dans l'environnement SATO, une requête produit soit des listes lexicales soit des concordances, c'est-à-dire les mots de la requête et le contexte (segment de texte) où ils occurrent, celui-ci étant déterminé par l'analyste (un groupe de 10 mots, une phrase, un paragraphe, etc.). Le logiciel permet également la construction de lexiques (liste de mots triés sur une clé alphabétique ou numérique). Pour les fins de l'analyse, le texte peut être divisé en domaines, c'est-à-dire des sous-textes obtenus à partir des patrons de fouille des documents ou parties de documents. SATO permet également des dénombrements; plusieurs analyseurs lexicométriques y sont incorporés tels, la distance (chi carré entre 2 lexiques), la lisibilité, la participation d'un ou plusieurs sous-texte à l'ensemble du corpus, etc. SATO fonctionne en mode menu et dispose d'un mécanisme d'aide en contexte. Il convient autant aux aspects exploratoires de l'analyse de documents qu'aux stratégies systématiques de l'analyse de contenu (Bardin, L. 1989).

3.2.2 Blocage des locutions grammaticales

Pour faciliter la réduction du bruit (résultats non-désirés) lorsque l'usager soumet des requêtes, SATO offre la possibilité de «bloquer» ou «figer» automatiquement les locutions grammaticales comme les locutions conjonctives (afin que), prépositionnelles (à la faveur de), adverbiales (à peu près), etc. Ceci présente un double avantage : a) inscrire au lexique du corpus les formes «fonctionnelles» (ou mots-outils); b) distinguer les formes nominales apparaissant dans ces locutions des autres formes nominales du corpus (par exemple la forme «mesure» peut représenter une mesure ou apparaître dans l'expression «à mesure que»).

Pour ce faire nous utilisons une procédure SATO (locutions) qui permet d'indexer les expressions à bloquer. Une fois ce travail d'indexation effectué, le texte est redonné à SATO qui inscrit au lexique les locutions grammaticales comme des entrées lexicales distinctes.

3.2.3 Catégorisation des parties du discours : l'utilisation d'une base de données lexicales (BDL)

La description grammaticale des éléments lexicaux s'avère être une phase importante de l'analyse de contenu des documents. Elle permet d'augmenter considérablement le pouvoir sélectif des requêtes. Par exemple, le lexique de la plupart des textes administratifs de SAGÉE est composé de 25% de noms communs. Une fois ces éléments indexés, l'analyste n'a plus qu'à consulter un mot sur quatre pour retrouver l'information pertinente. Bien entendu, ce type d'indexation n'est pas fait manuellement. La description morpho-grammaticale comporte une phase automatique permettant de catégoriser entre 85 et 90% des entrées d'un lexique.

Ce type de description se fait à l'aide d'une base de données lexicales. Dans le contexte du projet SAGÉE, nous avons utilisé la base BDL-SATO (Base de Données Lexicales) développée par Luc Dupuy du Centre d'Analyse de Textes par Ordinateur. BDL regroupe une quinzaine de collections d'unités lexicales organisées sous forme de dictionnaires SATO (au total BDL-SATO regroupe 358 820 entrées lexicales ou mots du français écrit). L'algorithme de BDL-SATO est très simple. Il s'agit de comparer les chaînes du lexique aux chaînes contenues dans les dictionnaires. Si les chaînes sont identiques, la chaîne du lexique reçoit la catégorie associée à la chaîne du dictionnaire. Les collections d'unités lexicales sont regroupées de la façon suivante :

- délimiteurs (signes de ponctuation, 26 entrées);
- interjections (57 entrées);
- conjonctions et locutions conjonctives (275 entrées);
- prépositions et locutions prépositives (356 entrées);
- déterminants (adjectifs numériques, indéfinis, etc.) (171 entrées);
- adjectifs qualificatifs (26 099 entrées);
- verbes à l'infinitif (8 384 entrées);
- pronoms (pronoms personnels, relatifs, etc.) (120 entrées);
- verbes conjugués (222 658 entrées);
- participes passés (33 447 entrées);
- participes présents (8 343 entrées);
- adverbes et locutions adverbiales (1 582 entrées);
- noms communs (55 653 entrées);
- noms propres (toponymes, etc.) (1 649 entrées).

La structure d'une collection lexicale de BDL est fort simple. Il s'agit d'une liste des formes du français écrit. Par exemple le dictionnaire «noms communs» comprend plus de 50 000 formes nominales dont l'extrait suivant montre la structure :

```
abaissement*gramr=nomcommun
abaissements*gramr=nomcommun
abaisseur*gramr=nomcommun
abaisseure*gramr=nomcommun
abaisseures*gramr=nomcommun
abaisseurs*gramr=nomcommun
abandon*gramr=nomcommun
abandonnateur*gramr=nomcommun
abandonnateure*gramr=nomcommun
abandonnateures*gramr=nomcommun
abandonnateurs*gramr=nomcommun
abandonné*gramr=nomcommun
abandonnés*gramr=nomcommun
abandons*gramr=nomcommun
abasourdissement*gramr=nomcommun
abasourdissements*gramr=nomcommun
abatis*gramr=nomcommun
abattage*gramr=nomcommun
abattages*gramr=nomcommun
abattement*gramr=nomcommun
abattements*gramr=nomcommun
abatteur*gramr=nomcommun
abatteurs*gramr=nomcommun
abattis*gramr=nomcommun
abattoir*gramr=nomcommun
abattoirs*gramr=nomcommun
```

Ainsi se structure chacun des dictionnaires de BDL-SATO. Une fois sous SATO, on utilise la fonction de consultation d'un dictionnaire pour indexer les mots du lexique. Il en va de même pour tous les autres dictionnaires. Évidemment, SATO offre une procédure automatique effectuant ce genre de travail. La procédure DOGRAMR (Daoust, F., Fiches d'utilisation, 1989) a été développée à cette fin. Le temps requis pour indexer un corpus de 100 000 mots est d'environ 15 minutes (temps utilisateur) pour un IBM-PC de type 80386 (16 Mz). Ces résultats sont présentés ici à titre indicatifs.

3.2.4 Catégorisation manuelle

Après la phase d'indexation automatique, il faut, au besoin, compléter la catégorisation morpho-grammaticale des formes n'apparaissant pas aux dictionnaires. Le logiciel SATO facilite grandement cette tâche. Une fois consultés les dictionnaires de BDL, il suffit de soumettre une requête permettant de dépister les entrées lexicales n'ayant pas reçu de catégorie à la propriété GRAMR. Lorsque l'inventaire est affiché par SATO, l'analyste n'a qu'à pointer la forme désirée et lui associer la catégorie appropriée. Qui plus est, une fois les nouvelles formes indexées on peut les intégrer au dictionnaire approprié. Il va sans dire que cette façon de procéder permet une mise-à-jour systématique et régulière. Dans l'exemple qui suit, on voit que la forme accotement sera indexée comme «nomcommun», pour être ultérieurement ajoutée au dictionnaire des noms communs.

fréq	GRAMR	
1	nil	abbord
2	nil	abi
1	nil	abiotiques
1	nil	accostant
1	nil	accoste
2	nil	*accotement*

valeur : ... adverbe article conjonction *nomcommun* ...

4.- L'analyse des groupes nominaux

4.1 Une hypothèse de travail

Sur un corpus de textes, représentatif de la production d'une organisation, on recherche tous les groupes nominaux construits à partir de termes désignant des concepts préalablement reconnus comme pertinents dans le domaine spécifique de l'expertise que l'on souhaite représenter. Ceux-ci devraient nous donner d'une part tous contextes où apparaissent les traits ou caractéristiques pertinents du concepts et, d'autre part, pour chacun des traits, l'ensemble des valeurs possibles ou tout au moins une contrainte sur leur admissibilité. Il s'agit d'un passage du lexique du corpus analysé, c'est-à-dire l'ensemble des mots qui le composent avec leur fréquence d'apparition aux termes (expressions socio-terminologiques), entendus comme noyaux potentiels de concepts ou des têtes nominales. Le groupe nominal est entendu ici dans un sens assez large : il regroupe les relatives et les attributs via les verbes d'états. Pour la forme nominale «projet» on aura par exemple des configurations du type :

«l'assujettissement d'un projet»
«la pertinence du projet»
etc.

4.2 Blocage des multitermes

4.2.1 Les multitermes (ou locutions terminologiques)?

Le terme représente lexicalement une unité cognitive : il est l'expression socio-terminologique du concept. La plupart du temps, un terme appartenant à un domaine d'expertise est composé de plusieurs mots qui, pris séparément, ont chacun une signification différente de celle de leur réunion (concaténation). On a ici affaire à des cas de micro-syntaxe du groupe nominal (Benveniste, É. 1966). La construction de ces multitermes autour d'une tête nominale semble se faire conformément à des patrons morpho-syntaxiques qui traduisent un acte de dénomination catégorielle. Quand un analyste parle par exemple d'«avis de projet», l'expression renvoie à une des dimensions de la notion de projet. Ainsi, la construction de la morphologie nominale [NOM de NOM : AVIS de PROJET] se trouve à produire un acte classificatoire différent de l'utilisation des seuls termes «avis» ou «projet».

La méthode d'analyse proposée ici est encore rudimentaire. Pour l'instant, l'environnement SATO ne dispose pas encore d'algorithmes permettant de dépister automatiquement les termes à l'intérieur d'un texte. Nous disposons toutefois de mécanismes d'analyse qui permettent de dresser automatiquement l'inventaire d'éventuels candidats. Pour l'essentiel, la procédure dépiste des chaînes nominales relativement «figées» à partir de certaines règles de composition morphologique (Gros, M. 1989).

4.2.2 Dépistage des multitermes

Les multitermes sont dépistés à l'aide de concordances résultant de la projection de patrons comportant un ou plusieurs filtres SATO. Une concordance est un segment de texte arbitrairement long contenant le ou les mots spécifiés comme filtre. Les concordances peuvent tenir compte de l'ordre des filtres ou pas. Le cas échéant elles sont soit ordonnées, si dans le segment les filtres respectent l'ordre de la requête, soit strictes, si les filtres sont adjacents. Dans un tel cadre, il y a deux types de stratégies pour le dépistage des multitermes : a) les fouilles sur des candidats pré-déterminés;
b) les fouilles basées sur la description morphologique du texte.

4.2.3 Les patrons de fouille

Les fouilles sur des têtes candidates pré-sélectionnées peuvent être effectuées une à une ou en lot. La première option consiste en une commande de concordance sur le tête choisie avec un rappel de contexte assez large pour dépister le ou les termes associés. Pour dépister un plus grand nombre d'occurrences de la tête choisie, l'analyste peut utiliser les ressources de troncation offertes par SATO. Cela permet de dépister toutes les formes de la tête, le singulier et le pluriel dans le cas des noms : par exemple une requête (dans un contexte numérique de +/- 3 mots) sur la chaîne «projet$» donnera les résultats suivants :

```
 #1    *PAGE=1/1/4/8 ... *PAGE=1/1/5/5
de la réception de l'avis de projet.
#2     *PAGE=1/1/8/5 ... *PAGE=1/1/8/12
plus le chargé de projet qui décide,
#3     *PAGE=1/1/9/8 ... *PAGE=1/1/10
L'avis de projet arrive toujours au
```

4.2.4 Stratégie de dépistage des multitermes par sélection des candidats

Les multitermes sont validés par un expert du domaine en faisant l'analyse des concordances obtenues; dans l'exemple précédent, les multitermes avis_de_projet et chargés_de_projet sont pertinents. Il existe des méthodes (Salton, G. 1983) basées sur diverses mesures d'occurrences pour décider s'il s'agit de multitermes ou de simples groupes nominaux. La configuration morphologique n'est pas un critère suffisant comme l'illustrent les chaînes suivantes : «l'avis de Pierre», «les chargés de mission».

La sélection d'une liste de têtes candidates à former des multitermes peut se faire en ajoutant une propriété qui pourrait avoir pour nom l'étiquette «tête» avec oui et nil pour valeur. D'abord on fait écrire le lexique de tous les mots qui ont reçu l'étiquette «nomcommun» (en SATO ceci se traduit par «$*gramr=nomcommun») comme catégorie morphologique. Ensuite on étiquette les têtes candidates jugées pertinentes:

Ecrire Lexique $*gramr=nom$

fréq	GRAMR	TETE	
1	nomcommun	nil	abénaquie
2	nomcommun	nil	abstraction
1	nomcommun	nil	accent
3	nomcommun	nil	acceptabilité
1	nomcommun	nil	acceptation
1	nomcommun	nil	acceptabilité
13	nomcommun	nil	accès
1	nomcommun	nil	accessibilité
1	nomcommun	nil	accident
1	nomcommun	nil	accidents
55	nomcommun	nil	accord
4	nomcommun	nil	accorde
1	nomcommun	nil	accordes
2	nomcommun	nil	**accotements**
2	nomcommun	nil	accumulation
1	nomcommun	nil	achat

propriété : Gramr **Tête**
valeur : Nil **Oui**

Par exemple, le mot «accotements» est jugé pertinent comme candidat pour être la tête d'un multiterme. Par la suite, les concordances seront effectuées à partir de filtres SATO qui permettent de dépister les segments textuels contenant les mots indexés oui à la propriété «tête» .

Cette stratégie à partir de têtes candidates pré-sélectionnées suppose une bonne connaissance de la terminologie du secteur de la spécialité, sinon des omissions sont à prévoir. Elle n'offre en effet aucune garantie d'exhaustivité.

4.2.5 Stratégie de dépistage des multitermes par patrons morphologiques

Les fouilles, basées sur la description morphologique du texte, se font sans présupposés sémantiques quant au contenu des textes. Elles portent sur tous les mots ayant une catégorie morphologique donnée, le nom en l'occurrence. Le dépistage des multitermes se fait alors par extraction de segments (concordances) à partir de la co-occurrence de patrons morphologiques. Les patrons présentés ici sont au nombre de trois, respectivement [nom + de + nom], [nom + préposition + verbe infinitif] et [nom + adjectif]. Ainsi un filtre SATO du type :

[nom + préposition + nom]

donnera les extraits suivants :

1 *PAGE=1/1/5 ... *PAGE=1/1/5/8
l'**avis de projet**. Comment faites
2 *PAGE=1/1/8/5 ... *PAGE=1/1/8/12
plus le **chargé de projet** qui décide,
3 *PAGE=1/1/9/8 ... *PAGE=1/1/10
L'**avis de projet** arrive toujours au
4 *PAGE=1/1/18/4 ... *PAGE=1/1/18/11
Un **dossier de route** par exemple,

Remarquons au passage que ceci illustre une des capacités d'analyse grammaticale simple de SATO. Il est en effet possible de réaliser des micro-grammaires en chaînes (Salkoff, M. 1979; Harris, Z. et al 1989). Ces grammaires peuvent être programmées par l'analyste sous forme de concordances strictes SATO. La chaîne [Nom + préposition + Nom] n'est qu'un des possibles modèles de chaîne. On peut aisément penser des modèles incorporant des adjectifs, verbes infinitifs, participes passés, etc. Ces modèles demeurent très rudimentaires dans la mesure où nous sommes guidés par une linguistique spontanée ou naïve, i.e. celle que nous maîtrisons comme utilisateurs de la langue. Les procédures que nous testons actuellement demanderont certainement à être validées par des équipes de linguistes ou de socio-linguistes. C'est une carence que nous ne sous-estimons pas et accueillerons toute collaboration constructive.

4.2.6 Blocage des locutions terminologiques

En SATO, cette opération est très simple. Elle consiste à faire une copie du texte où se retrouveront toutes les occurrences dépistées par les précédents patrons. Cette copie sera à son tour re-soumise à SATO pour qu'apparaissent au lexique du texte les locutions terminologiques candidates du domaine SAGÉE. Sur cette version du texte se fera la validation définitive des multitermes. Notons au passage que ces opérations peuvent au besoin être automatisées. En plus de ses caractéristiques interactives, SATO offre un langage de programmation simple permettant à l'usager d'automatiser les séquences d'opérations jugées pertinentes et valides.

4.2.7 Épuration de la liste des candidats termes

Quelle que soit la stratégie de dépistage utilisée, la liste obtenue comporte un certain nombre (parfois assez élevé) d'occurrences indésirables (bruit). Il en est ainsi parce que la réalisation d'un patron morphologique est un critère de dépistage de candidats qui en aucun cas devrait primer sur les autres critères, d'ordre plus qualitatifs (reposant sur une dimension sémantique ou thématique). Voici un début de lexique de termes candidats dépistés par les patrons morphologiques du type groupe nominal ([Nom de Nom]) et certaines de ses variantes :

fréq

1 abaissement_du_lac
1 abaissement_du_niveau
1 abaissement_du_niveau_de_trois
1 abandon_du_projet
1 abandon_du_projet_de_la_sainte
1 abords_des_rives
1 abords_du_parc
1 abords_du_secteur_d'_aménagement
1 abords_du_site
1 abri_des_inondations
1 absence_d'_accumulation
1 absence_de_bonnes
1 absence_de_concentration

Comme on peut le constater, il ne comporte pas que des termes. L'expérience nous démontre qu'il est difficile de trancher sur le niveau de «fixité» d'un segment répété sans connaissance du domaine. C'est pourquoi le recours aux experts pour la collation des termes est indispensable. Cette opération, effectuée de façon manuelle, repose sur des critères sémantiques et pragmatiques; les critères employés ne sont pas toujours objectifs. Ces critères doivent cependant être clairement explicités et partagés par les membres de l'équipe.

Les critères doivent être précis. Dans une perspective de transfert d'expertise, il faut faire attention de ne pas lier sur la foi d'une co-occurrence nombreuse des concepts avec leurs valeurs lorsque le trait est implicite (cf la description de l'objet valué). Cependant les critères de liaison sont toujours relatifs au domaine d'expertise à couvrir. D'un certain point de vue, il serait erroné de lier les mots «étude de répercussion» parce qu'il y a aussi «étude d'impact», ce qui permet la construction du granule [étude-> type (répercussion / impact)]; la position contraire se trouve tout aussi justifiée si le granule entrevu est considéré comme étant plus général : [étape -> type (étude de répercussion / étude d'impact)]. Ceci n'est pas sans rappeler que c'est d'abord la question de lecture qui construit l'objet; il faut également prendre en compte qu'un texte ou un corpus est susceptible de plusieurs types de lectures, et donc de plusieurs types de constructions terminologiques.

4.2.8 Réduction d'une liste de candidats à l'aide de SATO

La réduction de la liste de candidats se fait en SATO de la façon suivante. D'abord une propriété est temporairement adjointe au lexique aura pour valeurs OUI ou NIL. Puis la catégorisation se fait à OUI lorsque le mot est considéré un terme par un expert. Pour que la réduction soit quelque peu valide, il faudrait qu'elle soit faite séparément par plusieurs experts de points de vue différents. Il est important que la propriété SATO identifie chacun des intervenants par un nom différent.

L'acceptation ou le rejet de l'entrée lexicale sont inscrits au lexique (catégorisation) par chacun des experts du domaine de façon individuelle (dans notre exemple, les étiquettes lcp (Louis-Claude Paquin), ld (Luc Dupuy) et yr (Yves Rochon) représentent les initiales des analystes).

La fusion de chacune des catégorisations de provenance diverse se fait de la façon suivante : un lexique de toutes les entrées lexicales acceptées est déposé dans un fichier. Ce fichier est ensuite transformé en dictionnaire SATO pour être projeté sur une version qui deviendra le journal de bord du processus de catégorisation:

fréq	lcp	ld	yr	
1	nil	nil	nil	abaissement_du_lac
1	nil	nil	nil	abandon_du_projet
1	nil	oui	oui	abords_des_rives
1	nil	nil	nil	abords_du_parc
1	nil	nil	nil	abords_du_site
1	oui	nil	nil	abri_des_inondations
1	nil	nil	nil	absence_d'_accumulation
1	nil	nil	nil	absence_de_concentration
1	nil	nil	nil	absence_de_contamination
1	nil	nil	nil	absence_de_marina
1	nil	nil	oui	absence_de_suivi
1	nil	nil	nil	abstraction_du_règlement
1	oui	oui	oui	abus_d'_alcool
1	oui	oui	oui	abus_de_drogues

Les données de ce lexique peuvent être exportées en format tabulaire c'est-à-dire en format admissible à des chiffriers, à des bases de données ou des logiciels de traitements statistiques. Ces traitements, d'ordre statistique, peuvent être organisés en dictionnaire SATO et intégrés ultérieurement au corpus.

Un nombre optimal d'experts doit être déterminé par les analystes du domaine. Comme tout travail de définition d'une politique terminologique, la nature et la quantité des «décideurs» doit être établie d'une manière qui tienne compte des caractéristiques du milieu. Un nombre impair de juges peut être important car il assure la possibilité d'utiliser le critère ultime de sélection : la majorité simple. Un autre critère serait qu'il suffise du tiers pour qu'une intégration des vues plus

poussée soit entreprise en demandant aux répondants dans une intervention séparée, pourquoi le terme a été sélectionné ou encore pourquoi il a été rejeté. Un tel questionnaire a pour effet de dépister les imperfections (tels les trous) dans la conception de la politique terminologique au sein d'un groupe d'experts qui entretiennent des relations discursives (telles la production et la lecture de rapports).

4.2.9 Constitution et projection de dictionnaires de multitermes

Il est possible de conserver ces multitermes, sous forme de dictionnaire SATO, pour projection ultérieure sur d'autres textes. Pour ce faire, il suffit simplement d'utiliser la ressource SATO pour la création des dictionnaires. Celle-ci est simple. À l'aide d'un index (comme ceux que l'on vient de décrire) on sélectionne au lexique les éléments terminologiques pour ensuite les acheminer vers un fichier dictionnaire où ils seront stockés en ordre alphabétique. Ces dictionnaires peuvent par la suite être consultés à volonté.

Pour élargir le filtrage aux flexions féminine et plurielle, lorsque c'est pertinent, on peut utiliser un opérateur de troncation à droite (dans l'environnement SATO une chaîne peut être tronquée en la suffixant du caractère «$») :

accord de principe	problème$ de bruit
activités de circulation	pylône$ de réseau$
analyse de recevabilité	qualité des sédiments
analyse environnementale	recalibrage de cours d'eau
avis de projet$	rejet d'environnement
centre$ de documentation	réflexion$ de route
chargé$ de projet$	réunion$ de direction
coupe de roc	tableau$ de synthèse
coupe de roc de dynamitage	technologie$ de traitement
description du trajet	territoir$ de chasse
dossier$ de route	toile de polythène
eau de pluie	visite$ de terrain
étude d'impact	voie$ de contournement
étude de répercussion	volée$ de plomb
filière de documentation	
massif de béton	matériel à draguer
ministère de l'Agriculture	matériaux à draguer
ministère de l'Environnement	
outils de travail	réunion générale
période de consultation publique	nord américain
plan de surveillance spécial	

Pour faire de la liste de mots obtenus un dictionnaire de locutions, il faut lui donner un format approprié au moyen de l'utilitaire «locution» de l'environnement SATO. Celui-ci assure la conversion des chaînes retenues en fichiers d'opérations SATO (qui ressemble d'assez près à un "batch file" ou une macro WordPerfect™).

4.2.10 Projection du dictionnaire

Le dictionnaire constitué, on revient à une version du texte antérieure aux patrons morphologiques pour y projeter le dictionnaire.

Ces dictionnaires de multitermes doivent être validés avec les procédures prévues par la politique terminologique de l'institution. Après une validation auprès des experts les ayant utilisé dans le cadre d'une description de leur champ d'expertise, ces dictionnaires peuvent êtres présentés à des spécialistes de la terminologie pour une phase de validation ou de correction.

Cette façon de faire permet de diminuer le risque de redites ou la répétition d'une description du domaine.

4.2.11 Exemple d'un texte généré avec les multitermes

Voici, à titre d'illustration, un texte regénéré avec les multitermes bloqués:

*PAGE=dirt/1/3
analyste # 101
projet_d'_aménagement_du_littoral de la rivière Ristigouche
Promoteur : Le conseil_de_Bande de la réserve de Restigouche
Dossier #XYZ Janvier 1986
Une importance_particulière doit être apportée aux points suivants :
camionnage_des_matériaux_de_remblayage;
impacts des travaux_de_dragage et de disposition_des_matériaux_dragués sur la qualité de l'eau, la flore et la faune (site_des_travaux et cône_de_diffusion);
les impacts associés à la modification de la section_hydraulique de la rivière Restigouche à l'emplacement_du_projet;
les impacts_du_projet sur la zone d'herbaçaie à spartine;
les aspects_visuels et esthétiques liés aux ouvrages;
la circulation_maritime et la sécurité du public;
impacts liés à la phase exploitation du port_de_plaisance et des aménagements sur le remblayage.

4.3 Lexique des candidats termes

4.3.1 Présentation

Cette étape a pour but d'opérer le passage du lexique des mots étiquetés nominaux, conformément à l'effet linguistique de référence au réel et non au discours lui-même.

4.3.2 Catégorisation manuelle

La catégorisation manuelle se fait principalement sur le lexique de tous les noms ou candidats nominaux du lexique. Les multitermes doivent au préalable être catégorisés morphologiquement «nomcommun» pour qu'ils soient intégrés. Un, préférablement plusieurs experts, sont alors appelés à sélectionner parmi toutes les formes nominales, celles qui représentent des réalités jugées pertinentes pour résoudre une tâche dans un domaine particulier. De même, un dictionnaire des termes est constitué. Celui-ci servira de support aux phases subséquentes du traitement.

4.3.3 Synonymie

L'objectif ici visé est de ramener à une seule forme canonique les termes qui ont une même signification. Il faut auparavant s'assurer que l'équivalence sémantique est juste ou s'il s'agit de deux états différents, consécutifs dans le déroulement de la procédure comme par exemple les termes «problème_de_bruit» et «pollution_sonore». Pour traiter la synonymie, il s'agira de construire une propriété SYNONYME et d'y indexer les termes équivalents, mais moins fréquents dans le corpus.

4.3.4 Classification des termes

Cette opération a pour but de situer les termes, les uns par rapport aux autres. Comme la complexité de la tâche croit avec le nombre de termes pertinents retenus. Deux mille termes n'est pas exceptionnel sur un corpus d'envergure moyenne (entre 700 et 1 000 pages de textes). La stratégie proposée ici en est une de réductions successives. Il y a là une trop grande masse d'information pour pouvoir procéder à une classification efficace.

Il n'y a pas a priori de bonne classification; une classification repose sur des critères moins discutables que d'autres. Une bonne politique d'indexation et le consensus du groupe d'analyste assurent un longue vie à un système de catégorie qui ordonne la connaissance d'un domaine.

4.3.5 Première réduction

La première réduction est une classification des données lexicales en domaines sémantiques (bases). Cette première division peut ne préfigurer en rien de la classification “définitive”; elle est effectuée par pure commodité de traitement de l'information. Ces organisations sont tributaires soit de la logique d'apprentissage, soit de l'organisation-même du domaine. En fait, cette division exige la formulation préalable d'une hypothèse ou encore à l'adhésion à l'une ou l'autre école de pensée qui s'intéressent au domaine choisi et qui s'accordent avec le point de vue qui sera celui de la tâche à accomplir.

Cette réduction peut être effectuée en SATO en ajoutant au lexique du corpus une propriété «domaine» pour laquelle la liste des valeurs est constituée du système de catégories sélectionné précédemment. L'illustration est une catégorisation effectuée sur le «bainstorming» d'un chargé de projet effectué avec un idéateur comme le logiciel MORE™. Les catégories sont tirées du modèle conceptuel des données:

DOMAINE----------	DESC----------
bio-physique	abiotiques
site	abords_du_cours_d'_eau
aménagement	abri
aménagement	camionnage_des_matériaux_de_remblayage
domaine_spatial	camions
aménagement	canal_de_fuite
exploitation	canal_de_navigation
aménagement	canalisation
bio-physique	canards
socio-économique	occupation_résidentielle
impact	odeur
bio-physique	oiseaux_migrateurs
enjeux	opinions_des_groupes_d'_intérêt
enjeux	opposition_significative
exploitation	opération_des_batardeaux
intervenant	ordre_des_ingénieurs
communauté	organisation_du_territoire
socio-économique	organisation_sociale
socio-économique	territoire_agricole
site	topographie_du_site
aménagement	tracés_de_routes_possibles
aménagement	tranchée
exploitation	transbordement
exploitation	transformation_du_bois

4.3.6 Validation de la première distribution du lexique des concepts

Le lexique des termes de chacune des catégories permet de revoir les concepts qui ont reçu une catégorie donnée. S'opère alors une suite d'opérations de transfert de concepts d'une catégorie à une autre. Il y a ensuite le problème des concepts qui appartiennent à deux catégories de façon non-exclusive. Un nombre élevé de ces concepts polysémiques (pouvant appartenir à plusieurs catégories) est un indicateur du fait que la grille de classification retenue entre en contradiction avec la distribution des caractéristiques. Il y a peut-être aussi lieu d'intégrer des hiérarchies intermédiaires de catégories. En somme, le but de cette réduction est de gérer les concepts qui seront constitués à partir des termes et de leur contexte immédiat.

4.4 Agrégation

4.4.1 L'agrégation des caractéristiques aux termes

L'agrégation est l'opération du rattachement au terme des caractéristiques pertinentes de l'objet représenté. L'objet dont il est ici question peut être autant concret qu'abstrait. Il s'agit d'opérer un passage, cette fois du terme au concept. Pour ce faire, on doit trouver un petit nombre de caractéristiques à partir de l'ensemble des formes observées de l'objet. Ces traits distinctifs standardisés adjoints aux termes portent une valeur.

Le modèle épistémologique sous-jacent tient les objets, les concepts et les situations pour des principes organisateurs. Autour de ces noyaux gravitent des catégories au sens aristotélicien du terme:

substance	QUOI?
qualité	COMMENT
quantité	COMBIEN?
relation	AVEC QUOI?
lieu	OU?
temps	QUAND?
position	
possession	
action	
passion	

La démarche de questionnement sur les caractéristiques pertinentes des concepts ne débouche pas sur un modèle global sans rupture ou contradiction. Elle sert à trouver des critères qui permettent d'isoler une série d'îlots structurés de connaissance dans le discours scientifique ou technique d'un domaine. Comme ces critères permettent de regrouper ou de séparer des individus entre eux, ils doivent être documentés. Pour ce faire, les individus sont triés par affinités et comparés tous à tous en vertu des précédents critères.

Le passage des termes aux concepts comporte 2 opérations qui ne sont pas consécutives mais simultanées:

1) le rattachement des traits;

2) la spécification de la valeur.

Les traits ne constituent pas en eux-mêmes des unités cognitives. Ils servent plutôt à les caractériser. Les traits sont dépistés dans les textes tels quels au moyen, soit d'une lecture de séries de contextes, soit au moyen de patrons de fouilles basés sur la morphologie, soit inférés parce que des valeurs sont considérées apparentée. Il est important que le nom des traits représente le critère. Les valeurs servent à spécifier le concept dans ses caractéristiques jugées pertinentes. Les valeurs peuvent être des quantificateurs (numéraux, les cardinaux et les ordinaux) ou encore une suite de positions sur une échelle thématique, par exemple : froid, tiède, chaud, brûlant, bouillant, etc.

Il y a plusieurs stratégies pour l'agrégation, cependant il n'y en a pas une meilleure que les autres, elles sont plutôt complémentaires. Les secteurs laissés pour compte dans une méthode sont dans d'autres mieux traités.

4.4.2 Agrégation quantitative

La première opération manuelle est l'inscription de chacun des termes retenus dans une fiche standard. On peut la structurer soit en remplissant les champs d'une base de données soit en catégorisant le lexique avec SATO. Les champs sont les suivants : le champ «domaine» est consacré aux catégories de termes précédemment attribuées. Le champ «terme_lié» est un terme considéré comme étant équivalent mais plus général d'emploi ou situé immédiatement au niveau supérieur dans la hiérarchie conceptuelle du domaine. Le champ «trait» sera rempli si le terme est considéré comme une valeur du «terme_lié».

DOMAINE	activité_de_la_DEE
TERME_LIÉ	action
TRAIT	type
TERME	**actions_légales**

DOMAINE	activité_de_la_DEE
TERME_LIÉ	critère_d'_analyse
TRAIT	nom
TERME	**exigences_de_construction**

DOMAINE	aménagement
TERME_LIÉ	accès
TRAIT	qualité
TERME	**accès_difficile**

DOMAINE	aménagement
TERME_LIÉ	activité
TRAIT	description
TERME	**marnage**

DOMAINE	aménagement
TERME_LIÉ	travaux
TRAIT	nom
TERME	**camionnage_des_matériaux_de_remblayage**

DOMAINE	bio-physique
TERME_LIÉ	caractistique_pédologique
TRAIT	type
TERME	**nature_des_sédiments**

Ces fiches seront par la suite analysées sous l'angle des relations terme -> terme_lié, terme -> domaine afin de dépister les éventuels bouclages. Puis tous les traits assignés au terme_lié seront regroupés avec leur valeur en arbres, pour dépister les redoublements de traits avec des registres de valeurs différentes. Cette opération n'est pas encore prototypée. Les fiches seront par la suite fusionnées. Nous proposons pour les fiches résultantes une structure de fiche qui semble complète sans être trop lourde.

Cette fiche ne peut être complétée pour tous les termes, certains seront plus rapidement et plus complètement décrits que d'autres; il s'agit sans doute de concepts fondamentaux pour la description du domaine. Tout au long du développement des schémas d'inférences suivis, il est toujours possible de compléter les fiches pour leur adjoindre une valeur à un champ fixe ou encore pour ajouter de nouvelles caractéristiques devenues pertinentes.

4.4.3 Agrégation par patrons morphologiques

Cette stratégie est utilisée pour isoler la régularité structurelle des concepts en superposant tous leurs contextes d'occurrences dans le corpus. Elle est en fait complémentaire à la précédente. La complétion des fiches pour chacun des concepts est précédée d'une recherche de l'extension structurelle maximale des concepts. La recherche se faisant à partir de patrons morphologiques, il s'agit de procéder à un arrimage des groupes nominaux à la structure terme-trait-valeur.

Un grand nombre de traits peuvent être repérés au moyen de patrons simples; par ex. : [nom_du_trait + de + nom_du_concept]. Ainsi, par exemple, pour le concept de quai on trouve les segments suivants:

... la longueur totale du quai ...
... l'emplacement du quai ...
... la largeur du quai ...

Nous travaillons, dans la perspective de l'analyse morphologique (Lecomte, A. 1978, 1984, 1986, 1988) du discours, à trouver des patrons plus complexes pour dépister des configurations qui échappent aux concordances. A titre d'illustration, voici deux exemples de patrons complexes.

1) Suite d'énoncés nominaux + substantif anaphorique :

Par ex. : «Moi, ce que je considère le plus important, c'est une bonne description du projet, un bon inventaire de la zone d'étude puis une bonne évaluation des impacts de son projet sur le milieu récepteur de cette eh... Pour moi c'est les 3 **points** les plus importants.»

Cet exemple illustre bien comment le substantif POINTS se trouve à définir les ingrédients d'un projet acceptable, i.e. «une bonne description», «un bon inventaire» et «une bonne évaluation des impacts».

2) Suite d'énoncés nominaux + nominalisation:

Par ex. : «La préparation ça peut-être différent d'un dossier à l'autre. Mais la façon dont ça se prépare. On essaie de voir quelles sont les questions qui vont venir à ça. D'abord il y a une présentation du ministère, ce que le ministère a fait dans le dossier, parce qu'il a le droit de parole au début des audiences. Le promoteur a droit de parole et ensuite le ministère de l'Environnement et le demandeur aussi l'explique. On explique le projet, on explique nous autres les raisons pourquoi on est dans le dossier, comprends-tu? Alors, il y a cette **préparation-là** et aussi l'opération : quelles sont les questions qui peuvent venir de la part de l'assistance ou de la part des commissaires.»

Dans cet exemple, la forme «préparation» est le vecteur sémantique qui se trouve à organiser les énoncés. On remarquera particulièrement l'utilisation de «-là» dans l'expression «préparation-là» qui adjoint à la forme déverbale (forme nominale dérivée du verbe correspondant : ici préparation dérive de préparer) «préparation» un trait déictique qui précise que la notion de préparation doit être comprise dans le contexte précis de la préparation générale des dossiers pour l'analyse auxquelles s'ajoutent d'autres opérations.

Lorsque pour les concepts les traits ont été isolés, on procède à la délimitation du domaine de valeur. Cette opération se fait notamment par l'examen des adjectifs présents dans les configurations nominales dépistées.

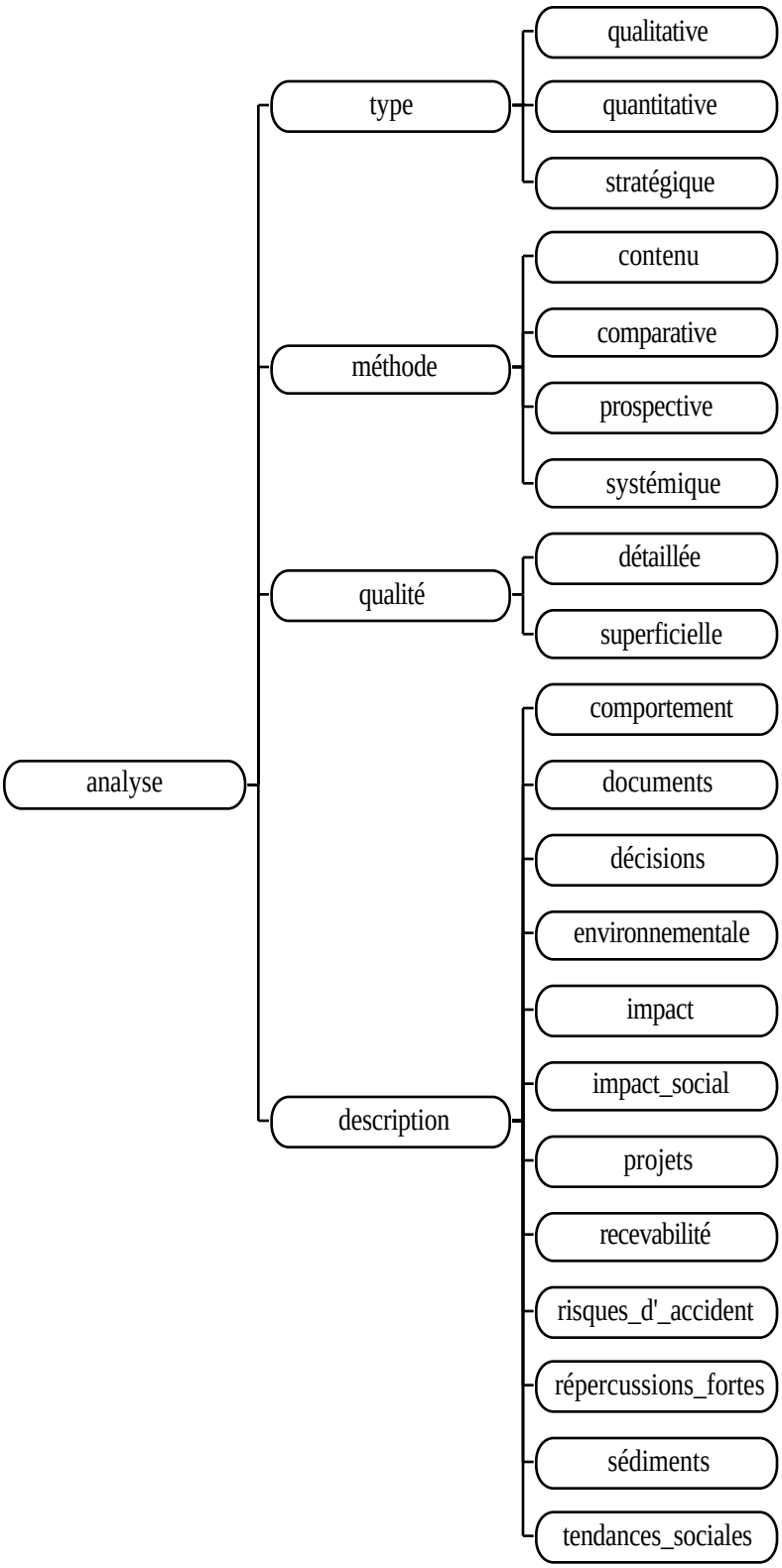
4.4.4 Agrégation des multitermes

Dans certaines conditions d'absence de caractérisation explicite des contextes, surtout quand le niveau d'abstraction de l'objet représenté par le terme est élevé, l'examen du lexique des multitermes peut nous permettre d'opérer l'agrégation sur les séries de termes qui partagent la même tête nominale. La partie gauche du multiterme est alors structurante, il s'agit d'un concept de niveau intermédiaire, alors que la partie de droite tient lieu de valeur. Voici par exemple tous les multitermes construits autour de la tête nominale «analyse»:

analyse_comparative	analyse_de_risques_d'_accident
analyse_d'_impact	analyse_du_comportement
analyse_d'_impact_social	analyse_du_contenu

analyse_des_impacts	analyse_environnementale
analyse_des_impacts_sociaux	analyse_habituelle
analyse_des_répercussions_fortes	analyse_nécessaire
analyse_des_sédiments	analyse_prospective
analyse_des_tendances_sociales	analyse_qualitative
analyse_détaillée	analyse_quantitative
analyse_de_décisions	analyse_stratégique
analyse_de_documents	analyse_superficielle
analyse_de_projets	analyse_systémique
analyse_de_recevabilité	

Dans ce cas, pour procéder à l'agrégation on doit comparer chacune des valeurs entre elles pour constituer des sous-groupes homogènes. Les critères utilisés peuvent être divers, mais doivent toujours être explicités. Inférer un trait consiste alors à sélectionner une étiquette qui nomme des groupes de valeurs apparentées. Entre autres critères de regroupement : lorsque les valeurs appartiennent à une même échelle, le trait inféré portera le nom de l'échelle. Cependant, comme un trait ne peut avoir qu'une et une seule valeur à la fois, il faudra répartir les valeurs dans des traits différents afin de les utiliser conjointement. Voici, à titre d'illustration une agrégation possible du précédent concept «analyse»:



.c2. 4.5 Quelques critères d'agrégation

Le concept construit à partir d'un terme doit pouvoir servir à réaliser par des valeurs différentes plusieurs autres termes apparentés. Il faut donc rechercher la configuration de traits la plus générale possible.

Le trait ne doit pas lui-même désigner un concept, mais être le descripteur d'une dimension du concept; par exemple, la grandeur, la hauteur ou la couleur de la table. Dans la mesure où l'analyse révèle une dénivellation de la hiérarchie, il devient nécessaire d'effectuer une opération de promotion : il s'agit de promouvoir le trait au granule et de promouvoir le granule à la base. Dans bien des cas, ceci suppose un retour au texte pour désambiguïser le terme. SATO permet dans ce cas de produire des fiches à partir de filtres définis par l'analyste, comme l'illustrent les extraits suivants produits à partir d'une requête comportant la chaîne «analyse_stratégique$» comme unique filtre :

1 *PAGE=mhumain/7/24/4 ...*PAGE=mhumain/7/25/5
Application de l'analyse_stratégique à la formulation de la directive...

2 *PAGE=mhumain/53/12/3 ... *PAGE=mhumain/53/14/7
Elles peuvent être rattachées à trois thèmes fondamentaux qui consistent à donner plus d'ampleur à l'analyse_stratégique, à modifier notre comportement au niveau_institutionnel et à accroître le niveau d'excellence de notre expertise professionnelle.

3 *PAGE=mhumain/60/40/3 ... *PAGE=mhumain/60/47/32 solutions_possibles
Pour Une Prise En Considération Plus Adéquate Des impacts_sociaux Dans La Procédure; Les solutions_possibles pour améliorer la prise en compte des impacts_sociaux dans le cadre de la procédure_d'_évaluation et d'examen_des_impacts sur l'environnement peuvent être ramenées à trois thèmes fondamentaux qui consistent à donner plus d'ampleur à l'analyse_stratégique, à modifier nos modalités_institutionnelles et à accroître le niveau d'excellence de notre expertise professionnelle.

4 *PAGE=mhumain/60/49/5 ... *PAGE=mhumain/60/50/61
Donner plus d'ampleur à l'analyse_stratégique Les problèmes que soulèvent les impacts_sociaux sont très variés.

5 *PAGE=mhumain/60/50/7 ... *PAGE=mhumain/61/4/7
Une solution judicieuse pour améliorer la prise en considération_des_impacts_sociaux dans le cadre de la procédure consiste à donner plus d'ampleur à l'analyse_stratégique dès les premières phases_de_planification et de conception_des_projets et dans tout le processus afférent à la prise de décision.

6 *PAGE=mhumain/61/6/12 ... *PAGE=mhumain/61/9/5
Du même coup, l'analyse_stratégique permet dès le début de repérer et d'évaluer adéquatement les effets_imprévus ou incertains des projets sur les gens et les collectivités, et conséquemment, de favoriser une prise de décision mieux éclairée et plus équitable.

7 *PAGE=mhumain/61/11 ... *PAGE=mhumain/61/13/2
Cette section-ci vise donc l'application de l'analyse_stratégique à la première étape de la procédure_administrative, soit la formulation de la directive_ministérielle.

#8 *PAGE=mhumain/61/21 ... *PAGE=mhumain/61/24/2
Tableau 9 Application De L'analyse_stratégique à La Formulation De La Directive Démarche 1.

9 *PAGE=mhumain/65/48/17 ... *PAGE=mhumain/65/49/41
du Manuel du chargé_de_projet intitulé «analyse_stratégique», 1985, pp.

10 *PAGE=mhumain/68/31/5 ... *PAGE=mhumain/68/33/25
Les relations entre intervenants s'intensifieront et occuperont ainsi une place_importante dans la gestion de nos dossiers, d'où l'importance également de mettre plus d'ampleur sur l'analyse_stratégique.

11 *PAGE=mhumain/69/10 ... *PAGE=mhumain/69/10/22
Le premier, et sans conteste le plus important des trois, consiste à donner plus d'ampleur à l'analyse_stratégique.

5. Perspectives de développement de l'analyse des archives textuelles de SAGÉE

Le travail en cours a mis en évidence trois aspects qui méritent d'être approfondis. La hiérarchisation des objets textuels, l'analyse des processus de raisonnement permettant à partir des textes d'apporter une aide à la rédaction des règles de production et la mise au point de mécanismes d'indexation textuelle intelligemment assisté par ordinateur.

La hiérarchisation est une opération qui consiste à relier des concepts/objets à des concepts/objets plus génériques. Elle sert à isoler des régularités pour des regroupements en des concepts/objets plus généraux ou mieux équilibrés. Elle servira par la suite à inscrire des liens et l'héritage. On pense que l'intégration de SATO et du D_expert au sein de l'ACTE (Atelier Cognitif et Textuel, Daoust, F., Dupuy, L. et Paquin, L.-C., 1989) nous permettra d'élaborer des stratégies d'analyse textuelle pour faciliter la traduction des hiérarchies cognitives textuelles vers les structures cognitives du système-expert. Il s'agit là d'un problème extrêmement complexe que nous ne pourrons certainement pas complètement solutionner mais auquel on pense pouvoir apporter un éclairage pertinent.

L'assistance à la rédaction des règles d'inférences représente une deuxième perspective de développement. Nous analyserons plus spécifiquement la relation de détermination nominale, définie comme les modalités des rapports entre les objets et les opérations susceptibles de leur être appliquées.

Finalement, la mise au point de mécanismes d'indexation textuelle intelligemment assisté par ordinateur constitue une troisième nécessité sur le plan du développement. Le flux documentaire exige à lui seul que l'on se penche sur l'automatisation des procédures. Nous pensons pouvoir ré-investir l'acquis des structures cognitives dans le processus d'indexation automatique des documents du projet SAGÉE. Si cette voie s'avère réalisable, on aurait possiblement une façon de résoudre partiellement un autre problème complexe : celui de la mise à jour des bases de faits.

En guise de conclusion

Dans le contexte du projet SAGÉE on attend d'un système-expert qu'il soit adapté à la culture de l'organisation. Il ne doit pas seulement produire des résultats similaires à ceux d'un analyste. Il doit le faire en utilisant les mêmes structures socio-terminologiques, c'est-à-dire les procédés de dénomination utilisés pour classifier les objets de l'analyse, les différents types d'agents, les structures conceptuelles, etc. Et si l'acquisition des connaissances est un processus nodal dans l'élaboration d'un système expert, il ne doit jamais laisser dans l'ombre celui du transfert des connaissances, c'est-à-dire l'itinéraire inverse qui conduit du savoir de l'expert vers l'utilisateur du système. C'est en fonction de ces considérations que nous avons approfondi l'analyse des groupes nominaux, ceux-ci étant les premiers objets permettant de construire les schémas cognitifs utilisés dans le contexte du projet SAGÉE.

Cette stratégie a donné d'intéressants résultats. D'abord, l'élaboration d'un dictionnaire de concepts force l'organisation à uniformiser et à structurer les concepts qu'elle manipule.

Ensuite, la valorisation des archives textuelles a permis la sensibilisation de l'organisation à la richesse cognitive des textes qu'elle produit.

Finalement, la démarche d'analyse a permis la génération de sous-produits textuels utiles pour l'organisation, notamment :

- des bases de données lexicales pour unification terminologique;
- la constitution de bases de données textuelles;
- le recours systématique à l'histoire «textuelle» de l'organisation pour enrichir ses prises de positions présentes.

Mais ce ne sont là que des résultats préliminaires. La lecture de ce texte un peu scolaire montre à l'évidence

a) que le texte en format libre est une source facilement manipulable qui n'exige pas pour son utilisation que soit atteint une perfection linguistique;

b) que l'on peut disposer de moyens simples et efficaces pour préparer les données textuelles en format libre;

c) que l'interaction avec le texte des archives textuelles peut être un sol très fertile pour l'acquisition des connaissances;

d) que le texte est un excellent moyen pour le cogniticien de se «socialiser» au domainde de spécialité et qu'il reste un des meilleurs environnements pour valider les intuitions formelles de l'analyste, et ce, à partir de l'idiosyncrasie du point de vue des experts et des autres agents de l'organisation.

Somme toute, l'archive textuelle doit être considérée comme une source non-négligeable d'expertise. Le problème qui se pose ici est de trouver des moyens pour implanter dans la culture de l'organisation les habitudes de production textuelles de manière à s'assurer d'un renouvellement de cette forme d'énergie. Une des conclusions importantes qui se dégage est finalement que le développement d'un système expert ne doit pas être considéré comme une panacée mais plutôt comme une bonne occasion de construire interactivement des procédures systématiques et vérifiables de raisonnement intelligemment assistée par ordinateur.

Bibliographie

Bardin, L., L'Analyse de contenu, Presses Universitaires de France, Paris, 1989, 291 pages.

Benveniste, É., "Fondements syntaxiques de la composition nominale", in Problèmes de linguistique générale II, Éditions Gallimard, 1974, pp. 145-162

Boose, J. et Gaines, B., The Foundations of Knowledge Acquisition, Academic Press, New York, 1990, 385 pages.

Daoust, F., SATO (Système d'analyse de textes par ordinateur), Manuel de référence pour les micro-ordinateurs PC, PC compatibles et VAX/VMS, Centre d'Analyse de Textes par Ordinateur, Université du Québec à Montréal, 1989, 102 pages.

Daoust, F., SATO (Système d'analyse de textes par ordinateur), Fiches d'utilisation, Centre d'Analyse de Textes par Ordinateur, Université du Québec à Montréal, 1989, 27 pages.

Daoust, F., Dupuy, L., Paquin, L.-C., "ACTE : Workbench for Knowledge Engineering and Textual Data Analysis in the Social Sciences", in Proceedings of the Fourth International Conference on Symbolic and Logical Computing, Dakota State University, 1989, pp. 122-136.

Deschênes, A.-J., La compréhension et la production de textes, Presses de l'Université du Québec, Québec, 1988, 136 pages.

Ericsson, K. A., Simon, H. A., Protocol Analysis : Verbal Reports as Data, The MIT Press, Cambridge, Massachussetts, 1984, 426 pages.

Gros, M. "Degré de figement des noms composés", in Langage, #90, Les expressions figées, Larousse, Paris, 1989, pp. 57-72.

Harris, Z., The Form of Information in Science, Kluwer Academic Publishers, Dordrecht, 1989, 586 pages.

Lecomte, A., (1978) La thématisation. Quelques remarques linguistiques et discursives sur son fonctionnement, Dans : Lecomte, A., Paraphrase et thématisation. Essais d'analyse logique., Neuchâtel, Centre de recherches sémiologiques, Université de Neuchâtel,1978, Décembre, 32, pp. 69-82, 95 pages.

Lecomte, A., (1978) L'homme hilare ou vers une théorie logico-discursive de la paraphrase, Dans : Lecomte, A., Paraphrase et thématisation. Essais d'analyse logique., Neuchâtel, Centre de recherches sémiologiques, Université de Neuchâtel,1978, Décembre, 32, pp. 1-67, 95 pages.

Lecomte, A., Marandin, J.-M., (1984), "Analyse de discours et morphologie discursive", Montréal, Centre d'Analyse de Textes par Ordinateur, Université du Québec à Montréal, 67 pages. (Draft)

Lecomte, A. (1986), "Espace des séquences; approche topologique et informatique de la séquence", Dans : Maldidier, Denise et al. Langages, #81, 1986, Analyse de discours, nouveaux parcours. Hommage à Michel Pêcheux, pp. 91-110.

Lecomte, A., (1988) Le marmot et la mamelle, critique des représentations du raisonnement, Centre de Coordination pour la Recherche et l'Enseignement en Informatique et Société (CREIS), Représentation du réel et informatisation, Saint-Étienne, I.U.T. de Saint-Étienne,1988, 21 pages.

Lerat, P. "Lexicologie des institutions", in Lexique 3, Lexique et institutions, Presses Universitaires de Lille, 1989, pp. 159-165.

Moulin, B., "Un outil pour l'acquisition des connaissances à partir de textes prescriptifs", in L'acquisition des connaissances, Revue ICO, Février 1990, pp. 27-42.

Paquin, L.-C., D-EXPERT, Manuel de l'utilisateur, (Version 2.0), Centre d'Analyse de Textes par Ordinateur, Université du Québec à Montréal, 1990, 140 pages.

Paquin, L.-C. et Dupuy, L. "An approach to Expertise Transfer : Computer-Assisted Text Analysis" Advances in Computing and the Humanities : Content, Concepts,Meaning. Advances in Computing and the Humanities, J A I Press, Greenwich, Connecticut, vol 3-4.

Rey-Debove, J., "Problèmes de sémantique lexicale", in Sémantique et logique, Jean-Pierre Delarge, Éditeur, Paris, 1976, pp.167-180.

Salkoff, M. Analyse syntaxique du Français : Grammaire en chaîne, John Benjamins B.V., Amsterdam, 1979, 334 pages.

Salton, G. Introduction to Modern Information Retrieval New York, 1983.

Tourigny, N. et Simian, G., "Méthodes, techniques et outils d'acquisition des connaissances", in L'acquisition des connaissances, Revue ICO, Février 1990, pp. 9-26

Valiquette, L. et Béland, R. SAGÉE: projet de système d'aide à le gestion en évaluation environnementale, Actes du premier coloque québécois en Informatique cognitive des organisations, Québec, GIRICO 1988, pp. 21-28.

Notes biographiques

Louis-Claude Paquin, Section Ingénierie Textuelle et Cognitive, Centre d'Analyse de Textes par Ordinateur, Université du Québec à Montréal.
Louis-Claude PAQUIN est chercheur au Centre d'Analyse de textes par ordinateur de l'Université du Québec à Montréal depuis juin 1986. Il a développé un générateur de systèmes experts, le D_expert. Ce logiciel est utilisé dans le développement de plusieurs projets de sustème expert au sein de l'administration publique. Il oeuvre à l'implantation des nouvelles technologies dans les organisations, principalement la valorisation de leurs textes et le traitement de leurs connaissances. Sa principale contribution au domaine est d'ordre méthodologique. Il s'intéresse aussi à l'analyse du discours par ordinateur. Docteur en philologie médiévale, il a établi le texte d'un traité alchimique jusqu'alors inconnu, le Liber secretorum.
Yves Rochon, Responsable du secteur bureautique, Direction des évaluations environnementales Ministère de l'Environnement du Québec.
Yves Rochon fait partie de l'équipe de développement du projet SAGÉE. Il est président du comité des usagers du projet DELTA et membre du comité d'experts sur l'analyse des SGBD textuels pour le Ministère des Communications. Il a suivi des études de baccalauréat en science biologique et une maîtrise à l'Institut national de la recherche scientifique portant sur le développement de systèmes d'information pour les activités d'évaluation et d'examen en environnement. Il a réalisé la base de données sur les polluants toxiques industriels (BTI) et la base de données sur les critères de qualité du milieu aquatique (CQED) pour la Direction de la qualité du milieu aquatique du Ministère de l'Environnement du Québec.
Luc Dupuy, Section Ingénierie Textuelle et Cognitive, Centre d'Analyse de Textes par Ordinateur, Université du Québec à Montréal.
Luc Dupuy est agent de recherche pour la section Ingénierie Textuelle et Cognitive du Centre d'Analyse de Textes par Ordinateur de l'Université du Québec à Montréal depuis avril 1985. Il a complété (1986) une maîtrise en communications au Département de communication de l'UQÀM. Il travaille au développement d'une base de données lexicales (BDL-SATO) pour l'analyse socio-terminologique. Il assure avec Louis-Claude Paquin le développement du D_expert sur les matériels IBM PC et VAX780. Il s'intéresse principalement aux aspects méthodologiques et socio-terminologiques de l'analyse de texte par ordinateur. Il s'intéresse au développement d'une approche socio-cognitive de l'analyse de la représentation des connaissances. Il prépare un doctorat au Département de sociologie de l'Université du Québec à Montréal (analyse la dynamique des représentations sociales au sein des institutions administratives). Il adore la planche à voile et la science-fiction.