

1.1.2 Univers industriels

Avec l'implantation micro de la technologie, la notion de système expert s'applique de plus en plus à des univers variés. Depuis le milieu des années 80, il y a une demande grandissante pour des applications dans les sphères de l'activité industrielle. Dans ce contexte, le problème de la représentation des connaissances d'un domaine d'expertise se double de celui de l'acquisition des connaissances. L'acquisition des connaissances propres à un domaine présente deux sortes de difficultés.

S'il est relativement facile, dans de tels univers, de faire l'inventaire des unités cognitives pertinentes, le problème est de discerner les différents rôles qu'elles sont appelées à jouer. Plus problématique encore est l'identification des fonctions cognitives devant être remplies et les moments où celles-ci doivent stratégiquement être réalisées pour simuler le raisonnement de l'expert. Par ailleurs, il faut que le système d'acquisition des connaissances puisse assurer de manière continue l'assimilation des nouvelles "entrées".

La variété des applications de type S-E pose aussi le problème plus général de la délimitation des savoirs. En effet, une des premières tâches dans la conception et l'implantation de systèmes experts est d'identifier les fonctions (Marcus, S. 1988) du domaine de compétence qu'il faut structurer. Or comme chaque domaine est spécifique, la structuration des concepts et des stratégies doit être effectuée spécifiquement. Autrement dit, chaque domaine présente ses idiosyncrasies épistémologiques, théoriques et méthodologiques devant être intégralement prises en compte. On doit reconnaître que les stratégies de résolution ou de calcul des solutions des problèmes symboliques sont encore mal identifiées; il manque à ce jour une classification claire des méthodes utilisées par les différents types de systèmes experts (McDermot, 1988). On ne peut que constater la montée des expériences de type "méthodologiques" avec tout le foisonnement terminologique que cela suppose.

La conception de S-E utilisés dans un univers industriel ajoute au problème de la représentation de la connaissance celui du transfert des expertises. L'utilité d'un système expert réside dans le fait de pouvoir refléter et conserver ce que les experts d'un domaine connaissent. Ceci pose le problème de l'efficacité communicationnelle dans la mesure où les messages produits par le S-E doivent être conformes à ceux que produirait un expert. Un S-E doit pouvoir expliquer, justifier et préciser les parcours qu'il emprunte; à la limite, un S-E doit offrir un support de type pédagogique à ses utilisateurs (Barr, A. & Feigenbaum, E. 1982). L'utilisation du S-E pose également le problème de la mise-à-jour des connaissances.

1.1.3 Univers administratifs

Dans ce type d'univers, les préoccupations des utilisateurs et des concepteurs concernent principalement la manipulation (production, analyse, gestion, etc.) de données textuelles. Il existe des problèmes qui se posent en amont de leur simulation dans le cadre d'une application de type S-E. On pense ici notamment aux difficultés que posent les activités (Gingras, M. 1988) de stockage, d'indexation conceptuelle et d'exploitation d'immenses bases de données textuelles en format libre. Le terme de format libre désigne un format qui ressemble à celui d'un livre ou d'une revue. Ce format, à géométrie variable (le paragraphe, la page ou le chapitre), est le véhicule principal de la connaissance propre à cet univers : les textes de jurisprudence, les répertoires de politiques et de décrets, etc.

1.1.4 Problèmes de méthodes

Les modèles d'analyse utilisés par les approches de type "sciences et techniques" sont fortement inspirés des analyses en psychologie cognitive, centrées sur la genèse et la production des concepts (Eliot, L. B. 1987; Reitman Olson, J. & Rueter, H. H., 1987). On doit reconnaître que la structuration sociale de la pratique scientifique et technique a tendance à favoriser des foyers bien identifiés : pour les domaines scientifiques les experts sont généralement connus, ils ont une notoriété particulière qui les désigne spontanément comme points d'origine lors de la constitution d'un domaine

d'expertise. Dans le cas des secteurs où le savoir est stocké dans des archives textuelles, les experts ne sont pas aussi aisément identifiables. Bien souvent, on constate que l'expert est un groupe dont chacun des membres contrôle une zone d'expertise bien spécifique. On a qu'à penser au secteur des politiques administratives des institutions gouvernementales pour se convaincre des différences qui caractérisent l'opposition entre "sciences dures" et "sciences humaines". Dans le second cas, le savoir est structuré comme un univers polysémique où les raisonnements présentent des aspects divergents pour ne pas dire contradictoire. On voit bien alors que les expertises ne peuvent être appréhendées qu'en fonction des caractéristiques socio-cognitives des groupes d'intervenants.

Les méthodes strictement cognitivistes laissent dans l'ombre la dimension sociale de l'expertise. Le savoir est construit par et dans l'activité discursive et les interactions micro-sociales. Il faut alors tenir compte des multiples perspectives surtout lorsque le savoir est afférent à des domaines moins bien définis comme ceux que l'on retrouve dans les organisations publiques et para-publiques. Par exemple, une notion telle que "l'intention de frauder" s'analyse difficilement in abstracto; le recours au contexte fourni par les archives s'avère indispensable. L'utilisation efficace du potentiel offert par les archives textuelles requiert une adaptation des méthodes d'acquisition des connaissances aux réalités quotidiennes de la pratique administrative. Ceci nous semble important parce que les archives textuelles représentent le mode dominant de conservation des structures socio-cognitives que sont les savoirs afférents à telle ou telle pratique sociale. En dernière analyse, il faut bien reconnaître que toute forme de savoir dépend en tout ou en partie d'une structure socio-linguistique qui permet le stockage, la manipulation et la transmission des éléments d'un domaine de savoir.

1.2.1 L'analyse du discours

L'analyse du discours (telle qu'elle se pratique en sociologie) (Ghiglione, R., & Matalon, B., 1985; Lecomte A., 1985-1988) repose en grande partie sur l'hypothèse suivante: les énoncés d'un discours se présentent comme des formes d'objets-noyaux, des

faisceaux, aux configurations régulières. Analyser la morphologie d'un discours revient à construire un modèle du texte. Une telle construction d'effectue en répertoriant à travers les strates de la matière syntaxique des objets conceptuels et en reconstituant les itinéraires sémantiques que ces objets empruntent. Ce type d'analyse du discours exploite la particularité du langage naturel d'être à lui-même son propre métalangage, c'est-à-dire qu'il sert à la fois à représenter la réalité et à représenter la représentation de la réalité. Il s'agit en somme d'une lecture par extraction et échantillonnage de segments de texte représentant les enjeux importants du discours. Ces segments, articulés les uns aux autres, forment un nouveau texte se donnant comme résultat de l'acte d'interprétation. La construction des séquences d'un texte s'effectue selon l'axe nominal et selon l'axe verbal. Dans le premier cas, le fil du texte se dessine à partir des relations qu'organisent les formes nominales. Ce sera, par exemple, la reprise systématique d'une catégorie sémantique, au moyen de différentes expressions nominales ou pronominales. Dans le second cas, les formes verbales et les formes déverbales (nom formé par dérivation d'un verbe) instaurent une logique de l'action en orientant les parcours empruntés par les sujets des énoncés. Ainsi, certaines formes verbales seront utilisées pour marquer les oppositions entre le continu et le discontinu, entre le potentiel et l'actuel, etc. Dans une telle perspective, la logique naturelle guide notre examen de la structuration des objets; la grammaire (sémantique et syntaxe) nous sert à isoler les régularités matérielles de la langue qui les représente.

L'analyse de discours représente donc une modalité d'intervention sur les textes qui peut prendre en compte les éléments conceptuels et l'organisation que ces derniers supposent. Jusqu'à maintenant le bénéfice méthodologique du lien entre analyse de discours n'as pas, dans le domaine de l'acquisition des connaissances (transfert d'expertise), été apprécié à sa juste valeur. par les cognitivistes qui pourtant reconnaissent l'importance de l'analyse des protocoles, pour ne nommer que cette modalité particulière de l'archive textuelle.

2. Le projet ACTE

FRANÇOIS DAoust, LUC DUPUY, LOUIS-CLAUDE PAQUIN
CENTRE D'ANALYSE DE TEXTES PAR ORDINATEUR
UNIVERSITÉ DU QUÉBEC À MONTRÉAL

2.1. Contexte du projet

Dans les grands organismes, dont ceux de l'appareil gouvernemental, la production textuelle - faite de rapports, de directives, de projets, de correspondance, etc. - connaît un volume grandissant qui rend de plus en plus difficile leur exploitation. Ainsi, les "travailleurs du texte", chercheurs, gestionnaires, décideurs, etc. dont la lecture et l'analyse de texte constituent la principale activité, sont débordés par une masse de documents qu'ils doivent analyser en fonction d'objectifs qui leur sont propres: accumulation de faits, d'événements ou de connaissances, interprétation, élaboration de stratégies, prise de décision, etc.

D'un autre côté, les outils et les méthodes informatiques pour la compréhension des textes n'ont cessé de se diversifier et de se perfectionner dans les laboratoires de recherche, tant sur le plan de la performance que sur celui de la validité théorique. De plus, l'utilisation généralisée des programmes de "traitement de textes" et celle des bases de données textuelles ont rendu possible l'exploitation de cette ressource non-négligeable de savoir et d'expertise que représente l'ensemble des textes produits par une organisation.

Les conditions sont donc favorables pour un transfert d'expertise des laboratoires vers les organisations.

1.2 L'expertise du Centre d'ATO

Le Centre d'ATO possède une expérience de plus d'une décennie en développement d'outils informatiques en analyse de texte par ordinateur (ATO). Citons entre autres des logiciels, tels SATO, Déredec, FX et D_expert. Les chercheurs du Centre d'Analyse de Textes par Ordinateur possèdent une expérience d'intervention dans les organisations publiques et para-publiques, en particulier comme agent de formation en ingénierie cognitive et en analyse de textes par ordinateur. Voici une liste approximative des projets du Centre d'ATO avec le Gouvernement du Québec:

- Au ministère des Affaires municipales, utilisation de SATO pour l'analyse des lois municipales;

FRANÇOIS DAoust, LUC DUPUY, LOUIS-CLAUDE PAQUIN
CENTRE D'ANALYSE DE TEXTES PAR ORDINATEUR
UNIVERSITÉ DU QUÉBEC À MONTRÉAL

- Au ministère de l'Agriculture, utilisation du D_expert et de Faisceaux dans le contexte de systèmes de diagnostic;
- Au secrétariat du Conseil du Trésor, utilisation de SATO pour le traitement du répertoire des politiques administratives, utilisation du D_expert pour la mise au point d'un système expert sur l'attribution d'une certaine classe de contrats de service (SAGAC);
- Au Comité consultatif en gestion du personnel (CCGP), utilisation de SATO pour l'indexation assistée de conventions collectives et autres textes; projet analogue avec le Conseil du Trésor;
- Au ministère de l'Education, utilisation de SATO pour repérer le vocabulaire du domaine de l'évaluation, pour élaborer une stratégie de calibrage des textes, utilisation du D_expert dans le cadre d'un système expert pour guider le choix de stratégies statistiques;
- Au ministère de l'Environnement, utilisation de SATO et du D_expert dans le projet SAGEE (Système d'analyse pour la gestion des évaluations environnementales);
- Au ministère du Revenu, projet d'utilisation conjointe de SATO et du D_expert dans le cadre d'un programme de formation des vérificateurs fiscaux.

Constitué à l'origine autour de quelques chercheurs, le centre d'ATO a d'abord été un regroupement de chercheurs (1983), puis un organisme de service à la recherche (FCAR, en 1984) et un laboratoire de recherche (UQAM, en 1986). Il compte maintenant dix chercheurs à temps complet et son budget approche le demi-million de dollars.

1.3 Le projet d'un atelier logiciel cognitif et textuel

Le projet d'un atelier cognitif et textuel (ACTE) est né d'un double besoin. D'abord, il est nécessaire d'augmenter la robustesse et la convivialité des systèmes existants. C'est le cas en particulier du

D_expert, prototype fonctionnel en LISP dont on voudrait extraire un module exécutable et portable sur les micro-ordinateurs en usage.

L'association intime des deux systèmes (SATO et D_expert) semble apporter une réponse appropriée au développement et à la généralisation des projets en cours. En effet, la méthodologie développée pour l'extraction des connaissances repose en bonne partie sur l'analyse de données textuelles. De plus, la manipulation des textes avec des outils informatiques pourrait profiter grandement de la méthodologie des systèmes experts qui permettent la construction graduelle d'algorithmes complexes.

Schématiquement, la pertinence du projet ACTE se résume aux arguments suivants:

- Premièrement, la faisabilité d'un tel outil repose sur l'existence de logiciels dont on a vérifié la pertinence et le potentiel.
- Deuxièmement, le projet permettra de développer et de consolider les logiciels existants afin d'en augmenter la performance et l'efficacité.
- Troisièmement, le projet ACTE sera l'occasion d'améliorer et d'uniformiser l'ergonomie des outils existants en vue de leur union dans le module intégré.
- Quatrièmement, ce projet sera l'occasion d'augmenter la connectivité bureautique du système par l'adjonction d'un protocole de communication permettant de programmer l'accès à des banques de données.
- Finalement, en mariant ainsi deux technologies majeures pour le travail dit de bureau, on se donne une plateforme solide et générale pour développer des applications particulières susceptibles de répondre à une variété de besoins.

Le projet est divisé en deux phases consécutives:

D'abord l'intégration et l'optimisation des outils existants: SATO un analyseur lexico-textuel et D_expert un générateur de systèmes experts. En parallèle, la section de linguistique computationnelle du Centre d'ATO poursuit les recherches en vue du développement d'un analyseur lexico-syntaxique du français, ALSF, en

collaboration avec l'Institut national de la langue française (CNRS, France). Ensuite l'installation dans l'atelier d'un certain nombre d'applications linguistiques valorisant les descriptions potentielles de l'analyseur, telles le dépouillement des termes, la construction automatique d'index structurés, la production de résumés automatiques, etc.

Cet atelier, sans compromis avec la convivialité, pourra s'attaquer en temps réel à de très grandes masses de textes, tout en n'imposant aucune préconstruction théorique quant aux modalités du traitement des textes. Son architecture favorisera l'intégration éventuelle de modules ou procédures de provenance diverse et les raccords avec des serveurs aux spécifications différentes.

3 Description de SATO et D_expert

3.1 SATO

3.1.1 Les caractéristiques de SATO

SATO (Daoust, F., 1985-1989) est donc un système d'analyse de texte par ordinateur destiné à faciliter les analyses de contenu. Il s'agit, en quelque sorte, d'un système de base de données textuelles qui permet d'annoter des textes multilingues et de les manipuler de diverses façons: repérage de concordances, construction de lexiques, catégorisation des mots, dénombrements de tout ordre et analyseurs lexicométriques. SATO fonctionne en mode menu ou en mode commande et dispose d'un mécanisme d'aide en contexte.

3.1.2 Description générale

Le système SATO est conçu pour permettre à l'utilisateur d'interroger son texte et de contrôler pas à pas les diverses étapes de traitement. Pour ce faire, cependant, il faut d'abord que SATO lise le texte et en reconnaisse les multiples composantes: mots, ponctuations, paragraphes, références de pagination, etc. Ce traitement constitue une étape préalable dont les résultats sont

gardés en permanence dans des fichiers d'un format spécifique à SATO.

C'est au cours de ce pré-traitement que SATO construit, à partir des mots du texte, le lexique, c'est-à-dire le catalogue de l'ensemble des formes (mots, ponctuations, nombres...) contenues dans le texte.

Une des caractéristiques intéressantes du système SATO, c'est qu'il est possible d'associer aux mots du texte, ou aux formes, des propriétés à valeur numérique ou symbolique. C'est un peu comme si l'on pouvait rajouter des dimensions au texte afin de l'annoter.

On peut imaginer par exemple que définir une propriété pour le lexique revient à rajouter une colonne au catalogue des formes du texte. De même, on peut imaginer que définir une propriété pour le texte revient à rajouter au texte original une ligne sur laquelle il est possible d'inscrire une annotation pour chacun des mots du texte.

3.1.3 Les outils d'analyse de SATO

Les outils d'analyse fournis dans SATO peuvent être regroupés en six catégories.

La génération de lexiques, avec le nombre d'occurrences de chaque forme dans le texte ou une partie quelconque du texte.

Le repérage de concordances, c'est-à-dire de segments de texte dans lesquels apparaissent un ou plusieurs mots.

La production de dénombrements, c'est-à-dire de décomptes de mots dans une suite de segments de texte : phrases, paragraphes, etc.

L'application d'analyseurs automatiques destinés à produire des avis d'interprétation sur le texte :

LISIBILITE fournit divers indices de difficulté/facilité de lecture du texte.

PARTICIPATION permet d'évaluer la part relative d'une classe quelconque de mots dans un ensemble de sous-textes.

DISTANCE fournit une mesure de distanciation, ou de différenciation lexicale, entre divers textes. L'algorithme permet aussi de repérer les mots ou classes de mots qui contribuent le plus à distinguer ces textes.

DECRIRE permet de calculer la distribution d'une propriété sur le texte ou le lexique.

Finalement, l'utilisateur peut élaborer ses propres analyseurs en combinant les outils primitifs déjà décrits. Le schéma d'analyse ainsi produit pourra prendre la forme d'une procédure qu'on peut appeler par la commande

EXECUTER.

Des commandes de catégorisation permettent d'affecter des valeurs, généralement des catégories, aux mots du texte ou aux formes :

PROPRIETE permet de définir une nouvelle propriété, d'effacer ou de modifier une propriété existante. SATO est doté d'un mécanisme d'héritage: une nouvelle propriété peut hériter des valeurs d'une propriété mère; une propriété lexicale peut être "projetée" sur le texte et une propriété textuelle (symbolique) peut être "résumée" au niveau du lexique.

VALEUR permet d'affecter une valeur à un ou à plusieurs mots ou formes.

DICTIONNAIRE permet de consulter un dictionnaire préalablement inscrit sur fichier.

CONCORDANCE permet aussi d'affecter des valeurs à des mots dans un contexte de locutions ou cooccurrences.

Enfin, il y a un catégorisateur plein écran qui permet d'annoter très facilement tout mot ou forme lexicale. Cette annotation est modifiable tout aussi facilement et n'altère d'aucune façon le texte lui-même dont l'intégralité est garantie.

Des commandes de service qui permettent de contrôler à loisir l'impression et l'affichage des résultats. Ainsi on peut décider de masquer lors de l'affichage certains mots ou annotations de propriétés. On peut souligner (ou afficher en différentes couleurs) des mots qui possèdent une morphologie donnée ou des valeurs de

propriétés particulières : par exemple des mots d'une certaine longueur ou fréquence, des mots annotés selon une certaine catégorie grammaticale ou sémantique, etc.

L'intérêt de ces techniques provient surtout du fait qu'on peut les utiliser dans le cadre d'analyses comparatives. Un lexique par exemple, considéré de façon isolé, a un intérêt limité. Mais, quand on compare plusieurs lexiques, provenant de plusieurs textes ou parties distinctes d'un même texte, on peut voir des régularités ou irrégularités significatives. Dans SATO, la commande **DOMAINE** permet de définir un sous-ensemble quelconque des mots du texte, de manière à y opérer de telles analyses comparatives.

Ainsi, on pourrait comparer des chapitres. Il est aussi possible de définir un domaine qui serait composé des phrases où apparaissent un ou plusieurs mots déterminés. On pourrait aussi comparer des sous-textes formés de mots ayant reçu une catégorie particulière. Imaginons par exemple que, dans le texte d'une pièce de théâtre, on ait identifié les réparties de chacun des personnages. On pourrait alors définir comme autant de sous-textes les interventions de chacun d'eux.

3.1.4 Une syntaxe simple et efficace

Le système SATO est construit autour d'une syntaxe qui permet de décrire très facilement, et avec beaucoup de flexibilité, les objets primitifs du texte, à savoir les mots. Cette description représente en effet un véritable patron de fouille qui couvre tant les caractères d'édition du mot que ses valeurs de propriétés.

Voici quelques exemples:

parle	le mot "parle";
parle\$	tous les mots débutant par "parle";
p ent	tous les mots débutant par "p" et se terminant par "ent";
p_rle	tous les mots débutant par "p" suivi d'un caractère
"parle" ou	quelconque et se terminant par "rle" comme

"perle";
 parl(e,ent,ure) "parle", "parlent", "parlure";
 |ent*freq=5,>5 tous les mots se terminant par "ent" et dont la
 fréquence
 est plus grande ou égale à 5;
 ab\$*ALP=(fr,an)tous les mots débutant par "ab" et provenant des
 alphabets
 français ou anglais.
 \$*ALP~fr tous les mots qui ne sont pas en français.

Cette syntaxe de description des mots, combinée à une structure de
 commandes également très générale, confère à SATO une grande
 souplesse. C'est ainsi qu'on dispose d'une base solide pour
 implanter des analyseurs et assurer une communication efficace
 entre l'utilisateur-trice et le texte informatisé. En fait, les
 possibilités d'utilisation de SATO sont presque illimitées, d'autant
 plus que le système est capable de manipuler des centaines de
 pages de façon très efficace.

3.2 D_expert

D_expert (auparavant appelé Déredéc-EXPERT*) (Paquin, L.-C.,
 1986-1989) est un générateur de systèmes experts (GSE) développé
 en français au Québec. La présente version, écrite en Le_LISP,
 fonctionne sur Macintosh avec un minimum de 2 mo de mémoire,
 sur les IBM compatibles sous OS2 avec 3 mo de mémoire et sur
 VAX/VMS.

Ce GSE a été développé de façon à ce que des experts non-
 informaticiens puissent générer des systèmes experts par eux-
 mêmes. Nous préconisons en effet une intervention externe en
 ingénierie cognitive de type décroissante afin que l'organisation
 devienne autonome dans l'instanciation et la maintenance du SE qui
 lui convient. Le D_expert se caractérise donc en tout premier lieu
 par son aisance d'utilisation. Il n'y a pas de commandes à
 apprendre, seule une compréhension minimale de son architecture
 est nécessaire pour en exploiter toutes les possibilités. Une
 formation de 6 heures est habituellement suffisante pour s'y
 initier.

3.2.1 Représentation de la connaissance

La clarté de la représentation n'a pas été sacrifiée pour la performance du programme. La longueur des chaînes de caractères (256 au maximum) favorise l'expression naturelle de la terminologie du domaine d'expertise où les termes sont, la plupart du temps, composés de plusieurs mots. La structuration de la connaissance proposée est celle de l'objet valué (appelé "granule") qui offre modularité, flexibilité et lisibilité. L'objet valué est défini par la valeur de ses caractéristiques (appelés "traits"). Il permet la réduction d'une multiplicité de termes divers mais apparentés en des systèmes cognitifs. Les valeurs des traits ne sont jamais des procédures, ce qui assure l'indépendance des granules.

Pour en faciliter la gestion, les granules sont regroupés en vertu de critères taxinomiques quelconques en des ensembles appelés bases. Les granules peuvent être inscrits dans un tableau de dépendance cognitive; un procédé d'héritage sélectif multiple des propriétés permet d'établir avec les autres granules autant de liens que nécessaire et de gérer la transmission des valeurs. La sémantique des liens est laissée à l'utilisateur. Les traits et les liens sont gérés de façon indépendante des granules.

La règle d'inférences est la seule structure de contrôle: en plus de produire des faits (inférer), elle permet:

- de poser des questions à l'utilisateur en lui faisant instancier la valeur d'un trait (sur le Macintosh, le choix peut être fait par la sélection d'une zone dans une image);
- d'effectuer des quêtes dans des bases de données;
- d'effectuer des calculs;
- de transférer des valeurs d'un fait à un autre;
- de gérer l'itérativité;
- d'émettre des messages;
- d'évaluer des expressions LISP;
- transmettre des commandes au systèmes d'exploitation: DOS, OS2 ou VMS(VAX).

Le registre des actions disponibles ne cesse de s'étendre selon les besoins des utilisateurs.

Les faits relatifs à un problème donné sont regroupés en requêtes. Ceci permet de soumettre un problème et d'obtenir un résultat provisoire, de soumettre un autre problème, puis de compléter ou de réviser le premier problème pour le traiter de nouveau et ainsi de suite.

3.2.2 Éditeur de la connaissance

L'utilisateur construit et révisé son système expert par des sélections dans des menus emboîtés. Avec le D_expert, le système expert est construit et révisé par des sélections dans des menus emboîtés. Tous les granules disponibles sont décrits avec toutes les valeurs admissibles de chacun de leurs traits dans un dictionnaire (appelé ici "structures cognitives"). Les filtres et les inférences des règles, de même que les faits, sont construits en choisissant dans le dictionnaire un granule, puis en sélectionnant une valeur pour chacun des traits du granule retenu.

En tout temps, on peut ajouter ou retirer dans le dictionnaire:

- des granules
- des traits aux granules déjà déclarés
- des valeurs à des traits de granules déjà déclarés

Un propagateur prend en charge la mise à niveau des règles déjà construites avec les réaménagements apportés au dictionnaire. Un correcteur orthographique permet de modifier toute chaîne de caractères désignant une base, un granule, un trait ou une valeur à partir de l'affichage d'une base de granule, règle ou requête. Toutes les occurrences de cette chaîne à la même position, tant dans les structures cognitives que dans les règles et les requêtes seront alors remplacées.

De nombreux liens import-export ont été développés via fichiers-texte:

- toute chaîne de caractères peut être fournie à partir d'un fichier en format texte en la sélectionnant et en actionnant une clé
- lexiques de termes pour constituer des registres de valeur
- import-export des structures cognitives en format MORE 1.1 et Thinktank

- import-export des structures cognitives et des requêtes au moteur d'inférences en format base de données matricielles

3.2.3 Moteur d'inférences:

Le moteur d'inférences compare la prémisse des règles d'inférences aux faits (chaînage avant); à chacun des cycles, toutes les règles pertinentes sont invoquées (l'arbre de recherche est parcouru en largeur); le conflit entre plusieurs règles pertinentes est résolu par une mise en ordre croissante selon le nombre de filtres contenus par leur prémisse; le traitement de l'incertitude se fait par combinaison de coefficients de confiance (Mycin); le traceur est multi-niveau, il est possible de préciser quelles informations quant au déroulement d'un traitement sont souhaitées: l'identité des règles, le résultat du filtrage, le cumul des coefficients et des statistiques.

Les valeurs d'un trait faisant l'objet d'un questionnement peuvent être documentées d'un texte en format libre. Un mécanisme permet la navigation dans une hiérarchie de question en sélectionnant l'option par défaut "Aucun(e)". Les faits construits sur un même granule sont unifiés si la valeur de leurs traits ne présentent aucune contradiction de valeur. De plus, les faits qui présentent des liens hiérarchiques sont généralisés: des faits parents sont générés avec les valeurs de l'enfant. Les résultats obtenus, de même que les réponses fournies peuvent être conservés. Il est possible de générer un rapport en fusionnant certains résultats avec le fichier primaire d'un traitement de textes.

* Le Déredéc-EXPERT s'est mérité un prix au congrès d'Avignon: Les systèmes experts et leurs applications en 1987.

4. ACTE: devis général

4.1 Présentation

L'Atelier cognitif et textuel (ACTE) sera donc constitué par l'intégration de SATO et du D_expert. Au niveau méthodologique, cette intégration est motivée par l'intérêt de faire appel à la technologie combinée de l'analyse de texte et des systèmes experts.

FRANÇOIS DAoust, LUC DUPUY, LOUIS-CLAUDE PAQUIN
CENTRE D'ANALYSE DE TEXTES PAR ORDINATEUR
UNIVERSITÉ DU QUÉBEC À MONTRÉAL

Les domaines d'application d'un tel atelier sont très nombreux. Citons, entre autres,

- la gestion documentaire;
- la construction de systèmes d'analyse des textes consacrés à des besoins précis;
- l'ingénierie cognitive à partir du matériau textuel;
- l'accès à l'information sur différents serveurs sans apprentissage de langages d'interrogation;
- la génération de systèmes experts ou à base de connaissance; etc.

Au niveau informatique, l'intégration de SATO et du D_expert dans un même environnement de programmation présente des avantages importants. Les deux modules partageront au maximum des bibliothèques de fonctions communes (gestion d'écran, accès au disque, etc.). Cette mise en commun entraîne une économie de code et donc de temps requis pour son écriture et son épuration. Le passage physique des fonctionnalités d'un module à celles de l'autre se trouvera facilité et accéléré. Au lieu d'un échange d'informations par voie de fichier, les deux modules seront dotés d'un espace mémoire commun (tableau noir).

Actuellement, le D_expert est à l'état d'un prototype fonctionnel; son cycle de développement est terminé, ses fonctionnalités ont été validées sur le terrain. Présentement en LISP et avec toutes les structures en mémoire vive, l'inévitable accroissement du dictionnaire de la connaissance et de la base de règles d'inférences qui accompagne le passage des systèmes experts développés de l'état de maquettes à celui de prototype, fait apparaître les limites de l'état présent du système. L'encombrement de la mémoire vive nécessite une configuration de plus en plus grande et rend impossible en même temps la cohabitation avec une autre application.

Cet état de fait amène une double nécessité. D'une part, on devra dégager les fonctionnalités propres au moteur d'inférences de celles qui relèvent de la gestion des données (dictionnaire de la connaissance, règles d'inférences et requêtes). Ces données, dont la structure est présentement arborescente, prendront la forme de fiches et seront gardées sur disque; seuls les index permettant leur

accès seront gardés en mémoire vive. D'autre part, on devra passer du code LISP interprété à du code compilé (C ou Pascal) pour plus de robustesse et d'efficacité.

Aussi, l'intégration du D_expert dans l'environnement logiciel de SATO fera en sorte que les règles d'inférences pourront utiliser directement les patrons de fouille de SATO, exécuter des commandes SATO, accéder à des serveurs d'information, etc.

A l'inverse, SATO profitera directement des capacités du module d'inférences. Comme nous l'avons vu dans la section précédente, SATO peut être vu comme une boîte à outils offrant un ensemble d'instruments performants qui sont manipulés interactivement pour "disséquer" un texte, ou vérifier des hypothèses de lecture. Des fichiers comportant des commandes peuvent être exécutés en lots donnant lieu ainsi à des macro-commandes. Cependant nous aurions besoin de construire des analyseurs particuliers plus sophistiqués incorporant des stratégies de contrôle plus complexes, telles l'exécution conditionnelle d'actions, des boucles avec échappement. Pour construire ces analyseurs, l'utilisation de règles d'inférences, un mode de modélisation qui se répand de plus en plus, semble plus opportun que le développement d'un langage ad hoc. De plus, l'adjonction d'un gestionnaire de base de données permettra d'instancier des propriétés autant au lexique qu'au texte par des valeurs en format libre (chaîne de caractères).

L'atelier cognitif et textuel comportera un moteur d'analyse lexico-textuel fourni par SATO et un moteur d'inférences fourni par le D_expert. Son interface proviendra en grande partie de SATO alors que la gestion des données sera confiée à un module spécialisé.

4.2 Une ergonomie intégrée

SATO et D_expert partageront un même interface-usager où l'écran (25 lignes de 80 caractères) est composé de trois zones: un éditeur de texte en haut, un éditeur d'objets au centre et une barre de menu en bas. Une commande permettra de passer d'un écran à l'autre. Le principe directeur est d'éviter les superpositions de fenêtres, tel les menus déroulants, car des délais de restauration

d'écran sont encourus; délais inacceptables à des vitesses de télécommunication de 2400 bauds.

La barre de menu des commandes, sauf sur Macintosh, est empruntée à la librairie SATO. Elle est situé au bas de l'écran et son défilement est accompli au moyen des curseurs; l'accès direct à une commande est possible en tapant la première lettre; la sélection courante est affichée en vidéo inverse.

L'éditeur d'objets est un tampon (page virtuelle) qui admet un certain nombre d'objets, les mots du lexique et du texte en SATO et les éléments des objets valués en D_expert. En SATO, cet éditeur permet d'annoter manuellement les mots en fournissant une valeur à une de leurs propriétés. En D_expert, il permet de sélectionner une partie de structure à réviser, par exemple un trait d'un granule; les commandes au niveau d'emboîtement relatif à la sélection sont alors proposées; pour le trait on trouverait "Effacer le trait" et "Ajouter une valeur". L'éditeur d'objets est tiré de l'actuel écran d'annotation SATO; cependant des modifications lui seront apportées pour élargir la typologie des objets admissibles à ceux du D_expert et permettre un défilement arrière limité.

L'éditeur de texte, un nouveau développement, est doté d'un tampon en mémoire vive d'une centaine de lignes. Il présente un minimum de fonctionnalités qui sont offertes dans la barre de menu et permet le chargement et la sauvegarde de "petits-fichiers". Les commandes composées à l'aide des menus emboîtés y sont affichées (journal des commandes). Il sera possible, le cas échéant, d'apporter des modifications aux commandes déjà effectuées, puis de les sélectionner et de les soumettre à nouveau. Ce dispositif facilite la mise au point de fichiers exécutables et permet d'introduire des commandes SATO dans les règles d'inférences. Au moyen de l'écran D_expert, la rédaction d'une règle d'inférence dont les actions sont des commandes SATO est entreprise; la prémisse complétée, on passe à l'écran SATO et, par tentative, les actions SATO sont mises au point sur un texte. Lorsque le résultat est satisfaisant, les commandes sont sélectionnées, un retour à l'écran D_expert est effectué et les commandes sélectionnées viennent compléter la règle en cours d'écriture. L'éditeur de texte

sert aussi à la saisie et à la révision des chaînes de caractères qui constituent les objets cognitifs et les propriétés symboliques en format libre pour les mots.

L'aide contextuelle, couplée à la librairie de menus, est affichée en boucle par dessus l'éditeur de texte. Présentement, le texte réside sur un fichier externe, mais est mis en mémoire en début d'exécution. L'atelier utilisera un fichier indexé pour limiter l'utilisation de la mémoire. Seront aussi affichés à cet endroit: la documentation adjointe aux traits et aux valeurs par le constructeur du système expert de même que les avertissements et les messages d'erreur.

En plus du mode interactif, l'atelier disposera d'un langage de commandes. SATO possède déjà un tel langage et nous devons en définir un pour D_expert. Finalement, une syntaxe unique permettra d'unifier les deux langages. Une boîte à outils d'interface sera fournie pour que le développeur puisse, selon ses besoins, définir l'ergonomie des écrans de saisie et de choix multiples qui seront offerts à l'utilisateur.

4.3 Une refonte des structures internes

Le D_expert est composé de deux modules: un éditeur de la connaissance et un moteur d'inférences. Le recodage du D_expert implique deux types de tâches: d'une part le transfert des fonctions d'interface et de rétention des données cognitives vers les librairies d'interface SATO et un SGBD interne à ACTE; et d'autre part, un recodage en PASCAL à partir du code LISP des fonctions propres au générateur de systèmes experts. Ainsi, lors d'un cycle d'inférences, la restriction, c'est-à-dire le repérage de toutes les règles dont la prémisse est susceptible d'appariement avec la conjoncture factuelle (base de faits), se fera selon une stratégie de base de données. L'accès aux règles d'inférences résidentes sur disque sera direct, leur adresse étant répertoriée dans un index résidant en mémoire vive.

Ce passage sera accompagné d'une transformation de la structure de données, présentement arborescente en fiches standardisées.

Un seul modèle de fiche pourra représenter toutes les données cognitives: le dictionnaire de la connaissance, les règles d'inférences (les filtres de la prémisse et les actions de la conclusion) de même que les requêtes (ensembles de faits). Ce modèle de fiche permettra une meilleure documentation des données cognitives: dates de création, de révision, et commentaires en plein texte. La clé d'accès aux données cognitives sera composée du triplet base-granule-trait. Le granule ne sera plus représenté par un arbre, mais par un ensemble de fiches qui auront en commun leurs deux premiers identificateurs: base-granule. Le statut de l'objet sera explicite et non plus déduit de la position qu'il occupe dans une arborescence donnée. Les opérateurs de filtrage affecteront les inférences et les faits; leur gamme sera étendue à la nécessité, l'obligation, la facultativité, etc. La valeur pourra être un patron.

L'intégration du D_expert et de SATO nécessitera enfin des adaptations à ce dernier module. D'abord, l'ingénierie cognitive des résultats SATO devra être faite; une base de granules SATO sera mise sur pied, de même que des prédicats pour tester les résultats générés par SATO en position de prémisse dans les règles d'inférences. Chacun des résultats sera décomposé en trait/valeur; par exemple la description d'une propriété numérique aura pour traits i) la classe de mots décrite (patron de mot); ii) le nombre de mots qu'il comporte; iii) la fréquence moyenne; iv) l'écart type. Ensuite, les résultats de SATO, présentement affichés ou déposés dans un fichier unique, doivent être retenus afin qu'ils puissent être filtrés par les règles d'inférences. Cette rétention nécessitera l'instanciation de fiches de dépendances des commandes SATO pour conserver et expliciter les contextes des résultats.

5. ACTE: Un outil au service des organisations

L'atelier proposé est conçu de façon à répondre aux besoins actuels des organisations en bases de données textuelles, principalement l'accès sélectif aux connaissances contenues dans les textes. Cet accès pourra être direct en temps réel au moyen de patrons de fouille; par exemple trouver dans toutes les conventions collectives les passages où il est question de congés de maternité. Si, par contre, les textes résident dans une base de données

conventionnelle et que leur accès doit se faire par le biais d'un thésaurus, l'atelier intervient dans l'étape d'indexation des textes, soit le choix de termes représentatifs du contenu accompagnés de leur référence.

De plus, l'atelier vient enrichir les méthodologies pour le transfert d'expertise humaine contenue dans les textes vers les systèmes à base de connaissance. Ses fonctionnalités d'inscription de propriétés autant au lexique qu'au texte qui peuvent ensuite être questionnées permet un dépistage des concepts constant, objectif, reproductible et indépendant des problématiques définies dans les textes. Il s'agit donc d'utiliser le métalangage inhérent au texte lui-même pour isoler par leur récurrence les invariants organisés et hiérarchisés. Après une catégorisation morphologique des mots du texte, un dépistage des locutions terminologiques peut être effectué à partir de la co-occurrence de catégories; par exemple: traitement de textes {[nom] de [nom]}

D'une part, l'analyse des groupes nominaux d'un corpus de textes permet le dépistage d'unités cognitives et leur structuration en objets valués. Une fois que, parmi tous les substantifs, les concepts pertinents ont été retenus, les configurations nominales, appelés ingrédients, qui leur sont associés sont recherchées. Ainsi, par exemple pour le substantif "projet" on aura des configurations telles, l'assujettissement d'un projet, la pertinence d'un projet, etc. Les formes adjectivales présentes dans les contextes dépistés font apparaître les quantifications et les échelles argumentatives qui positionnent virtuellement les autres valeurs qualitatives ou quantitatives possibles.

D'autre part, l'analyse des groupes verbaux assiste la rédaction des règles d'inférences. En effet, l'examen des verbes d'action permet le dépistage des opérations définies sur les objets. Leurs flexions et leur contexte en fournissent la modulation (actif, passif, nécessaire, facultatif, etc.), la localisation et la temporalité.

En plus de permettre l'accès sélectif aux textes par une recherche de contenu, en plus de fournir une assistance à la conversion des objets du discours en objets valués, l'atelier est aussi conçu comme un outil général pour l'analyse du contenu des textes. La neutralité

de l'instrument, qui permet la coexistence de plusieurs niveaux d'analyse potentiellement contradictoires, favorise une démarche d'aller-retour entre la constitution de modèles sur les textes et leur validation empirique.

Il faut voir en effet, qu'il n'y a pas dans ACTE de projection déterministe d'un modèle pré-construit sur le texte. Le savoir sémantique et procédural appartient à l'usager. L'approche privilégiée par l'atelier est donc la mise à jour de l'organisation du texte par l'ajout de descriptions successives du texte en alternance avec l'exploration de résultats provisoires. Grâce à l'analyseur lexico-textuel, l'usager peut très facilement projeter sur le texte ses propres systèmes de catégories issus d'hypothèses explicites quant à l'interprétation du texte. Ainsi, les dénombrements pourront être effectués sur les catégories tout autant que sur les mots. Cette façon de faire amène le lecteur à expliciter les éléments textuels susceptibles d'être porteurs de sens et à arrêter les critères à partir desquels ceux-ci seront retenus et comptabilisés.

Voilà, brièvement décrites, les caractéristiques de la composante d'analyse lexico-textuelle de l'atelier. Par ailleurs, le générateur de systèmes experts permet, au moyen d'un dictionnaire de connaissances et de règles d'inférences exploitant cette connaissance, de modéliser et d'encapsuler des expertises qui, par la suite, seront exploitées par des utilisateurs.

Cette composante de l'atelier permettra la mise sur pied d'expertises incorporant la connaissance de langages d'interrogation de bases de données, textuelles ou non. Ainsi, par le biais de l'atelier, l'accès à des informations contenues sur différents supports, ayant chacun leurs particularités, n'exigera plus la maîtrise de multiples langages d'interrogation.

Le mécanisme du moteur d'inférences, qui déclenche des actions lorsque des conjonctures factuelles sont réalisées, convient non seulement à l'analyse de textes ou de scénarios de décisions, mais aussi à la génération de textes (lettres, rapports, etc...) Une conjoncture de faits donnée peut provoquer, en plus de l'adjonction d'inférences aux faits, l'émission de segments textuels décrivant et commentant la situation rencontrée. Ces usages sont

loin d'épuiser les possibilités de l'atelier cognitif et textuel. D'une part, il a le potentiel pour changer notre rapport au texte jusqu'ici linéaire, fortement marqué par les dimensions du temps (vitesse de lecture) et d'espace (de la première à la dernière page), pour un rapport instantané où les questions qui sont adressées aux textes sont immédiatement répondues et les réponses obtenues peuvent être à leur tour questionnées, etc. La créativité de l'utilisateur ainsi stimulée encourage le recours systématique aux textes. D'autre part il s'agit d'un puissant outil de modélisation des tâches textuelles permettant de construire de façon autonome des "applications textuelles" sur mesure intégrées par la suite aux procédures quotidiennes de l'organisation.

Bibliographie

Cicourel, A.V., (1986) "Social Measurement as the Creation of Expert Systems", Dans : Metatheory in Social Science. Pluralisms and Subjectivities, Chicago, The University of Chicago Press, 1986, pp. 246-270, 390 pages.

Paquin, L., D-EXPERT (Version 2.0), Université du Québec à Montréal, Centre d'Analyse de Textes par Ordinateur, 1987, 119 pages.

Politakis, P.G., Empirical Analysis for Expert Systems, Massachusetts, Pitman Publishing Inc., 1985, Research Notes in Artificial Intelligence, #6, 181 pages.