Une approche au transfert d'expertise, l'Analyse de Textes par Ordinateur

Louis-Claude PAQUIN et Luc DUPUY, chercheurs au

Centre d'Analyse de Textes par Ordinateur
Université du Québec à Montréal

0. Contexte

Les systèmes experts permettent de s'attaquer à une classe de problèmes pour lesquels les solutions informatiques conventionnelles s'avèrent peu efficaces en raison des caractéristiques socio-cognitives du savoir des experts. Les données manipulées par les experts se présentent souvent de manière complexe et incomplète; de plus, la structuration de ces données est sujette à de fréquentes révisions. Le principe à la base de ces systèmes est la séparation entre le savoir propre au domaine d'application et les instructions informatiques qui en invoquent les unités pertinentes au moment opportun.

On peut dire aujourd'hui que les progiciels générateurs de systèmes experts ont atteint un niveau de sophistication et de stabilité satisfaisant. On doit souligner qu'ils imposent en revanche des coûts élevés quant aux opérations de repérage et de saisie des connaissances. Ceci constitue l'un des principaux obstacles à la pénétration de cette technique informatique dans les organisations.

Ce texte[1] participe aux efforts de recherche pour automatiser ou assister le transfert du savoir-faire humain vers le système expert. Plusieurs applications s'inspirent de la psychologie cognitive: les unes se basent sur la théorie du *personal construct* (littéralement "construit personnel")[2] où les objets sont évalués en triade par l'expert pour générer une matrice de distance de laquelle sera tirée une hiérarchie conceptuelle, les autres sur l'analyse de protocoles[3], le commentaire à haute voix produit au cours de la résolution de problèmes. Inspirée par la neurophysiologie du système nerveux, une approche connexionniste[4] permettant un apprentissage "littéral" du système par accumulation de stimuli est en développement dans les laboratoires. Une autre voie se dessine du côté de l'analyse morpho-syntaxique des textes[5] pour dépister les concepts pertinents dans un corpus de grande taille. L'approche que nous proposons s'inscrit dans cette veine. Elle s'inspire de l'analyse du discours et elle est basée sur une perception unifiée et "naturelle" du savoir. Elle permet un dépistage constant, objectif, reproductible et indépendant des problématiques définies dans les textes.

La genèse de cette contribution est double. D'une part il y a notre intervention en ingénierie cognitive au sein d'une organisation gouvernementale sur un projet pilote[6] qui a pour objet l'instanciation d'un système expert (SE) de gestion à la Direction des évaluations environnementales. Les connaissances pertinentes qui sont presqu'entièrement de nature textuelle occupent un volume considérable[7], il

s'agit de textes de lois, de règlements, de directives, de décrets, de correspondance. En outre, treize entrevues avec les chargés de projet d'une longueur d'une heure et demie ont été effectuées et saisies. D'autre part, nous participons à un projet de recherche ayant pour objectif l'élaboration d'un Système d'Analyse de Contenu Assistée par Ordinateur[8]. L'orientation de cette recherche est pragmatique: la valorisation des données textuelles. L'option théorique retenue privilégie l'analyse morphologique du discours (présentée en détail plus loin), plutôt qu'une approche strictement syntaxique ou sémantique de la représentation des données textuelles.

Entrepris il y a quelques mois, notre travail a surtout consisté à préciser les prémisses théoriques de notre approche, de même qu'à s'assurer de sa faisabilité à grande échelle et en temps réel. En raison de sa complexité, la tâche a été décomposée en deux phases consécutives. La première étape qui est présentée ici consiste en la mise à jour des structures cognitives du domaine d'expertise. Le résultat prend la forme d'un dictionnaire d'objets valués reconstitués par l'analyse des substantifs et des configurations de mots qui leur sont associées[9]. La seconde étape nous fournira les informations nécessaires à la rédaction de règles d'inférences et à la mise en ordre des opérations à accomplir au moyen de celles-ci. Cette étape ultérieure s'appuiera sur l'analyse des verbes d'actions impliquant les objets structurés précédemment isolés, et sur un découpage du corpus en segments homogènes pour constituer des sous-textes décrivant chacune des étapes de la solution du problème posé.

Dans les sections qui suivent, nous traitons d'abord de la saisie d'expertise (§ 1-2). Ensuite nous caractérisons notre intervention à partir des textes (§ 3-4). Puis, à la suite d'une définition du savoir, nous proposons la logique naturelle comme cadre de référence (§ 5- 6); cette discussion débouche sur la présentation de l'analyse morphologique du discours. Une attention particulière est accordée aux objets de la schématisation (§ 7-8). Ces précisions théoriques nous amènent ensuite à décrire les modalités de notre analyse (§ 9). Nous décrivons enfin dans l'ordre chacune des trois étapes de la constitution du dictionnaire des objets valués (§ 10-12).

1. Reconstituer une expertise

A moins qu'il ne se double d'un pédagogue, l'expert possède habituellement une connaissance analytique ou semi-implicite plutôt que synthétique de son domaine de compétence. Transposer son savoir-faire en un discours pédagogique lui demandera un entraînement particulier. De cette incontournable difficulté nous proposons l'explication suivante : le savoir-faire d'un expert peut être considéré comme une famille de réponses à des questions qui n'ont plus à être posées. Les opérations cognitives sont caractérisées, entre autres par Jean Piaget[10], comme des manipulations s'appliquant sur des schèmes mémorisés. C'est probablement au cours du processus de mémorisation que les questions à l'origine des opérations cognitives se sont dissipées, leur travail de support étant terminé[11].

Il en résulte que l'accès au savoir d'un expert passe par la recherche des questions à ces réponses. Retrouver le savoir d'un expert, c'est refaire le chemin de cette découverte, c'est reconstituer le contexte dans lequel s'est produit la genèse d'une découverte ou le moment de consolidation d'informations éparses, c'est-à-dire l'apprentissage. Voilà pourquoi l'entrevue avec l'expert est tenue pour essentielle[12]. Au moyen de l'entrevue avec l'expert, l'ingénieur de la connaissance se doit de simuler un rapport question-réponse structuralement analogue à celui qui prévalait au moment où l'expert fit la découverte initiale ou l'apprentissage. Le lieu de la reconstitution de l'expertise est le discours oral ou écrit, voilà pourquoi la prise en compte de son expression est si importante. Pour que la part implicite du savoir soit réduite à un minimum lors du questionnement, la reconstitution doit se faire par l'intermédiaire d'un processus discursif explicite, c'est-à-dire une "opération intellectuelle qui s'effectue par une suite d'opérations élémentaires partielles et successives"[13].

Dans une perspective élargie au-delà de la problématique exposée ici, reconstituer une expertise consiste à mettre en oeuvre un ensemble de procédures génétiques, cognitives et historiques. Les procédures génétiques visent à faire apparaître la dynamique de l'élaboration des concepts (par ex.: le passage du concret vers le formel). Les procédures cognitives concernent le type et la nature des opérations spécifiant le travail d'un concept donné. Les procédures historiques permettent de mettre en relief la part de la dynamique sociale (commande sociale, tendances paradigmatiques, modes, etc.) qui contraint la production d'un concept ou d'une famille de concepts.

2. Enregistrer au lieu de noter

Instancier un système expert nous met tôt ou tard face à une divergence des vues quant à l'expertise concernée. N'utiliser qu'une seule source d'expertise, une série d'entrevues avec un expert par exemple, ne fait que repousser ce constat et rend plus difficile l'intégration des vues divergentes après coup. De ce constat nous tirons la proposition suivante: une expertise sera d'autant mieux cernée qu'elle sera saisie à partir de multiples sources, les accords entre les différentes sources indiquant des îlots très robustes et les désaccords, des îlots faibles. Par ailleurs, ne considérer que les sources d'expertise humaines c'est négliger une très grande partie du matériau disponible. Une grande partie de la mémoire d'une organisation est de nature textuelle et réside dans ses archives; elle est composée de rapports, de mémos, de lettres, etc.

Qu'il s'agisse d'entrevues avec les agents d'une organisation ou des documents de cette organisation, la prise de notes est un mode d'extraction de données qui s'avère inadéquat en raison de son caractère subjectif et des problèmes posés par la validation. Pour les entrevues, une prise de notes sélective ou une paraphrase rend impossible de distinguer après coup les inférences de l'analyste des verbalisations originales. Dans le cas du matériau textuel, un très grand corpus empêche la perception de régularités dont les occurrences sont trop éloignées les unes des autres. En somme, la perte de contact avec la formulation initiale rend difficile la validation des structures cognitives extraites et impossible

la vérification de leur exhaustivité. Si la transcription sur support magnétique des entrevues entraîne des coûts de saisie et de révision, une économie du temps des experts est réalisée. Quant à la révision, elle est facilitée par l'intégration des lexiques techniques produits par l'approche dans le traitement de texte utilisé.

3. Une approche plutôt qu'un logiciel

Notre analyse de la problématique du transfert d'expertise à grande échelle nous conduit à la mise au point d'une approche plutôt qu'à celle d'un logiciel. Afin de parer aux inévitables déperditions qu'entraîne cette opération de réduction, parmi l'arsenal du raisonnement scientifique, il nous a semblé plus sûr d'adopter une ligne de conduite principalement inductive. L'induction[14] a pour point de départ des objets particuliers connus: les textes, tels qu'ils se présentent. Si les textes sont les objets qui nous sont immédiatement accessibles, les connaissances qu'ils renferment sont obscures, difficiles à appréhender parce que mêlées au discours. Puis, au moyen de mécanismes de dépistage propres au texte, une remontée est effectuée pour arriver aux structures cognitives générales du corpus. Cette démarche s'oppose à la déduction qui pose par un filtre, une idée pré-conçue de la configuration finale qui sera fouillée de façon déterministe.

La plupart des logiciels dont nous avons pu prendre connaissance[15], projettent des grilles de codification qui, à notre avis, font abstraction du processus de fabrication de la connaissance dans le discours pour ne considérer qu'un ensemble d'unités discrètes. Le biais est direct lorsque la grille d'analyse est fixée dans les procédures, indirect lorsque, dans un premier temps, l'utilisateur est appelé à construire une grille qui sera projetée sur le texte. Dans les deux cas, cette façon de faire, commode d'un point de vue opérationnel, privilégie une description statique de la connaissance.

Nous proposons une approche interactive où la dimension heuristique prime: des manipulations standardisées et sans a-priori quant au contenu, révèlent partiellement l'expertise et guident la suite des explorations. Cette démarche cyclique est composée d'autant de boucles exploration (extraction) /validation que jugées nécessaires. Ainsi, non seulement la validation, mais aussi la gouverne (contrôle) des opérations est laissée aux membres de l'organisation chargés de la construction du système expert. A la limite, notre contribution consiste à leur faire prendre conscience de l'existence et de l'utilité d'outils informatiques qui sont en mesure d'augmenter l'efficience du processus. Notre approche ne tient donc pas de la moulinette fermée où la confiance dans l'outil prime, mais de la calculette où les manipulations répétées, libres et variées augmentent la créativité de l'usager.

4. L'Analyse de Texte par Ordinateur (ATO)

Le principal obstacle à l'utilisation intégrale des textes pour le transfert d'expertise réside selon nous dans une méconnaissance des difficultés inhérentes

au texte[16]. Les mots qui le composent ne renvoient pas toujours au réel via les concepts, mais peuvent servir à une re-catégorisation de ceux-ci. La modification continuelle du cadre référentiel du texte amène le lecteur à produire des inférences multiples. C'est pourquoi on dit que son fonctionnement est polysémique et que les interprétations qu'on peut en faire ne l'épuisent jamais. Notre maîtrise des langues naturelles est en devenir, il reste des problèmes qui ne sont pas encore résolus de façon satisfaisante tels, l'anaphore[17], la coordination, les formulations incomplètes, etc.

Il existe cependant des outils informatiques, des logiciels[18] et des progiciels[19], développés au centre d'ATO de l'UQAM pour analyser les textes qui, utilisés conjointement, permettent un dépistage satisfaisant de concepts potentiels dans un vaste corpus de textes. La liste validée, des objets structurés seront constitués autour des concepts retenus par le biais de leurs configurations syntagmatiques. Des fouilles effectuées sur un grand nombre de segments textuels à partir d'une description morphologique garantissent une réduction constante, objective, reproductible et indépendante des problématiques définies dans les textes. Pour parvenir à nos fins, nous intégrons dans notre approche les deux méthodes d'analyse des textes qui sont habituellement tenues pour opposées: quatitative, basée sur les statistiques de co-occurrence, et qualitative, basée sur une description exhaustive des structures syntaxiques.

Un dépistage assisté des concepts basé sur des patrons morpho-syntaxiques augmente l'exhaustivité, l'envergure des sources utilisées, et la rigueur terminologique. L'enquête est menée entièrement à partir des données et non pas de l'idée que l'on s'en fait ou à partir des analogies constatées avec d'autres domaines. Tout au long de la procédure, des sous-produits utiles pour l'organisation sont générés: des lexiques permettant l'unification terminologique, des bases de données textuelles consultables par le système expert ou autrement et surtout une sensibilisation de l'organisation à la richesse des textes qu'elle produit.

En résumé, dans le cadre du transfert d'expertise, notre intervention, basée sur l'analyse du discours, vient combler l'absence de méthodologie pour transformer le discours en données cognitives. Pour ce faire, nous comptons utiliser le métalangage inhérent au texte lui-même pour isoler par leur récurrence les invariants organisés et hiérarchisés. Avant de présenter les traitements informatiques qui opèrent la transformation en objets structurés assujettis à la logique formelle des systèmes experts, il nous a semblé essentiel de caractériser la logique "naturelle" du discours de même que les objets "de schématisation" qu'elle régit. Auparavant une mise au point quant à la perception du savoir s'impose.

5. Perception unifiée du savoir

La plupart des contributions au transfert d'expertise présentent une perception éclatée sinon tronquée du savoir. Dans une perspective ontologique, on parlera des "primitifs" d'un savoir et des méta-connaissances, un savoir du savoir sur le

réel. D'un point de vue typologique, on dira d'un savoir qu'il est procédural, analogique, probabiliste, etc. Ou encore on l'identifiera en regard des objets auquel il se rattache: substances, causalité, temporalité, etc. Pour dépister les savoirs des textes il nous faut recourir, non pas à une vision analytique partielle et construite du savoir, mais à un cadre plus général qui incorpore les dimensions positives de la cognition mentionnées plus haut et les aspects conjoncturaux inhérents aux contextes de son développement, en d'autres mots la discursivité de son exposition.

Un tel cadre nous est proposé par Michel Foucault qui définit le savoir comme un "ensemble d'éléments, formés de manière régulière par une pratique discursive"[20]. Cette définition contient deux idées cardinales: celle de régularité et celle de pratique discursive. Les éléments d'un savoir sont différents et variés, mais dotés d'une stabilité cognitive et temporelle qui rend possible la tenue d'activités symboliques, tels le raisonnement (par l'humain ou le système expert), l'argumentation, la démonstration, etc. Cette stabilité n'est pas intrinsèque aux objets, qui ne sont que des constructions, mais maintenue par une pratique discursive orientée en vertu d'une finalité donnée. Une expertise peut donc être considérée comme une formation discursive "où entre les objets, les types d'énonciation, les concepts, les choix thématiques on pourrait définir une régularité (un ordre, des corrélations, des positions et des fonctionnements, des transformations)"[21].

Il apparaît donc clairement que les objets d'un savoir sont irréductibles aux concepts d'un secteur de la connaissance positive[22]. De plus, les savoirs sont multiples et ne se recouvrent pas tout à fait, car dans toute organisation, à la division du travail correspond une division du savoir, des savoirs succèdent à d'autres savoirs[23]. En d'autres mots, des discours différents simultanés ou consécutifs sont tenus au moyen des mêmes concepts par différents agents ou groupes d'agents à l'intérieur de l'organisation. Dans une perspective plus large que le dépistage des unités cognitives, il faudrait tenir compte de la dimension sociale du savoir et, en plus des concepts qui ne sont souvent que les ingrédients d'une heuristique[24], considérer d'autres notions, tels le sens commun, les croyances, les processus cognitifs spécifiques à une "école" de savoir, etc.

6. La logique naturelle comme cadre de référence

Contrairement aux formulations logiques manipulées par les systèmes experts, les formulations linguistiques véhiculant le savoir ne sont pas assujetties à des règles d'exposition très strictes. Ainsi l'expression des mêmes objets peut varier considérablement. On peut bien sûr tirer directement du discours des formalisations partielles: des taxonomies, des arbres de décisions sectoriels, des procédures algorithmiques, etc. Mais, ce faisant, on néglige un résiduel important: le processus d'assimilation symbolique du réel, les transactions et les tractations au milieu desquelles se construisent les concepts.

Les logiques formelles[25] font généralement abstraction de la nature des objets qu'elles manipulent. Dans les situations de production des savoirs, les objets

manipulés sont rarement quelconques. Ils ne peuvent certainement pas se réduire à des variables valuées de façon binaire (vrai ou faux). Le plus souvent, ces objets symboliques sont dotés de plusieurs variables dont les valeurs ne sont pas boléennes mais scalaires, c'est-à-dire qui appartiennent à des classes méréologiques. La notion de classe collective (ou méréologique) se distingue de celle de classe distributive (ou ensembliste) comme le continu s'oppose au discontinu. Cette opposition peut être marquée formellement de la manière suivante: la notion de classe distributive est basée sur la relation être_élément_de qui est irréflexive, asymétrique et intransitive; la notion de classe méréologique est basée sur la relation être_partie_de qui est réflexive, symétrique et transitive. En résumé une classe méréologique ou une méréonomie est une hiérarchie partie-tout; en voici un exemple simple: la main et ses doigts. Chacun des doigts n'est pas tant une partie de la main que son prolongement.

Dans les classes méréologiques prédominent des complexes de relations entre un tout et ses parties, entre les parties de parties, etc. Ces manipulations quotidiennes des objets symboliques semblent relever d'autres systèmes d'opérations que ceux thématisés par les logiques formelles (l'inférence par calcul simple de la vérité). C'est pourquoi nous nous écartons des formalismes déterministes (logique des prédicats, etc.), pour utiliser un cadre de référence qui tienne compte de la dynamique topologique de l'énonciation du savoir. En effet, le problème que l'on cherche à résoudre n'est pas de nature "prédictive", mais qualitative, spatiale et temporelle. Le raisonnement quotidien (scientifique ou non...) "se présente comme un enchaînement, une combinaison ou une confrontation d'énoncés ou de représentations, respectant des contraintes internes susceptibles d'être explicitées, conduit en fonction d'un but"[26]. On a affaire à des schémas topologiques plutôt qu'à des parcours strictement déterministes.
Les circonstances du raisonnement quotidien sont importantes, car les objets signifiants ne sont pas seulement manipulés à des fins de démonstration[27].
Quatre postulats caractérisent cette approche:

1) Chaque fois qu'un locuteur A fait un discours, il propose une schématisation à un interlocuteur B.
2) Les activités logico-discursives de A s'exercent dans une situation d'interlocution déterminée.
3) La schématisation que A propose à B est fonction de la finalité de A mais aussi des représentations qu'il se fait de B, de la relation qu'il soutient avec B et de ce dont il est question, c'est-à-dire du thème T.
4) La schématisation comporte des images de A, de B et de T. Elle contient aussi des marques de son élaboration.

En plus des activités déductives, parmi les manipulations on trouve des validations, des inductions, le développement d'hypothèses, des analogies, etc. La description et la définition des termes d'un savoir est faite dans les termes de la langue naturelle qui est à elle-meme son propre métalangage[28]. Le sens des énoncés décrivant les termes d'un savoir est fonction des opérations et des tractations inhérentes aux finalités du rapport communicationnel qui gouvere la validité et la fidélité des énoncés. Ainsi nous avons recours à la logique naturelle

parce qu'elle s'intéresse aux opérations de schématisation mises en jeu par les locuteurs impliqués dans une pratique discursive. Les schématisations opérent en structurant les objets cognitifs et les articulant dans l'espace d'un savoir. Ces opérations sont toujours tributaires de circonstances spécifiques, soit la pratique sociale qui en détermine les conditions de possibilité. Dans cette perspective, les schématisations se caractérisent surtout par leur variété[29].

7. Les objets de la schématisation

Les objets de la schématisation, comme les objets des logiques formelles, sont les noyaux autour desquels se construit le raisonnement. Comme nous avons vu précédemment, ils présentent une différence importante: ils peuvent être l'instanciation de classes-objets ou de classes méréologiques (cf. supra); des parties de parties (par ex.: "Les justifications c'est souvent à cause d'un problème de courbe, un problème de mauvais drainage de la route, la chaussée est toute désuette, des problèmes d'accidents. Tout ça c'est la justification."[30]); des archétypes d'une classe particulière, des métonymies (par ex.: "Le ministère des Transports connaît les règlements autant que nous autres" [ici le tout représente les parties, soit les agents du ministère des transports]). Dans le cadre de la logique naturelle, les propriétés des objets d'une schématisation, de même que les relations qui peuvent exister entre eux, sont représentées par des prédicats. En plus des relations utilisées dans le cadre des logiques formelles (implication, relation de contraire, d'équivalence, etc.), on retrouve des relations de transformation d'objets, des relations méta-fonctionnelles (l'introduction d'un texte, d'un auteur, etc.), etc.

L'opération d'ancrage est le processus par lequel l'unité sémantico-cognitive vient prendre place dans un processus de schématisation; elles sont stabilisées à l'intérieur des formes linguistiques soit nominales soit verbales. Dans les deux cas, ces formes sont appelées substantif en raison de leur référence au réel. Les ancrages nominaux matérialisent au sein du discours des classes méréologiques d'objets. On comprendra qu'une notion comme celle de projet n'a pas en soi de "sens"; elle trouve son sens seulement à partir des éléments (ingrédients) qui en précisent les limites (par ex.: "Le projet à l'étude consiste en la réfection de l'émissaion d'eaux usées de l'usine de pâtes et papier C."). Les ancrages verbaux fournissent les éléments de la dynamique des objets: les propriétés et les relations (par ex.: "Le projet a pour objectif d'améliorer la production de sauvagine du marais Lac Noir (comté de l'Islet) qui a une superficie de 47 hectares").

Rappelons que les objets d'une schématisation sont récurrents, étant constamment repris et reformulés par les interlocuteurs tout au long du processus discursif. Plusieurs substantifs nominaux peuvent référer successivement au même objet (synonymie). Notons enfin qu'à une désignation nominale donnée semble correspondre une manière particulière de structurer la référence à l'objet. Ainsi, par ex.:

[question] "Au niveau des répercussions analysées, comment appréhender quand vous analysez ce point?"
[réponse] "Il y a des résidences d'affectées, des problèmes de bruit." (...) Il y a quelques lacs affectés. (...) un terrain de golf d'affecté.

La notion de "répercussion" est analysée ici à partir de différentes perspectives. Ceci illustre le fait que le "sens" d'un mot ou d'un terme est la manière de faire correspondre une signification et la réalité extra-linguistique. Pour clarifier, voici un exemple tiré de l'arithmétique. Les énoncés 2 + 2 et 1 + 3 sont équivalents dans la mesure où 2 + 2 = 1 + 3 = 4; ainsi, c'est seulement en termes du résultat qu'ils sont équivalents, mais la manière d'obtenir ce résultat est tout à fait différente. On peut vraisemblablement penser que l'acte de référencer est tout à fait comparable, que les énoncés portent sur des éléments arithmétiques ou discursifs.

8. L'analyse morphologique du discours

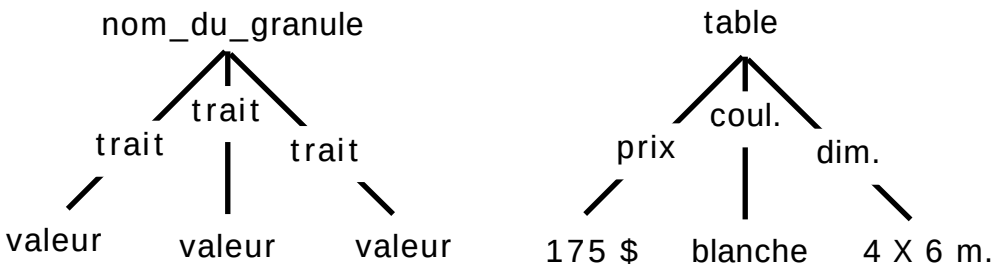
Les objets de schématisation ne s'enchaînent pas comme les arguments d'une démonstration formelle; ils se profilent suivant des trajets qu'il nous faut reconstituer et des relations de voisinage qu'il nous faut repérer. Cette terminologie réfère au noyau théorique qui tend à se constituer autour de la topologie. La propriété qu'ont les objets d'être construits, transformés, modifiés, etc. est illustrée par la variété des ancrages d'objets de schématisation, de même que la diversité des prédicats. Dans ce contexte, les rapports des objets ne peuvent être appréhendés qu'au moyen de la catégorie d'espace, c'est pourquoi nous parlons de topologie où des rapports de proximité et d'éloignement marquent les différentes parties d'un discours ou d'un texte (l'introduction doit se différencier de la conclusion et du développement; la problématique doit être distincte de l'analyse; les arguments doivent démarquer l'attitude du sujet par rapport aux objets dont il est question dans le discours ou le texte, etc.).

L'analyse morphologique[31] du discours, de laquelle notre approche s'inspire, 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 général du texte en répertoriant à travers les strates de la manifestation syntaxique des objets de schématisation et, au-delà des limites strictes de la phrase, 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. Ceci justifie une lecture par extraction et échantillonnage de segments de texte (en termes technique on parle de "thématisation par spécification"), tenus pour représentation canonique des 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 séquentialité[32] est le processus par lequel les objets sont mis en discours et subséquemment reconstitués par la lecture (un parcours discursif particulier). 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 (anaphoriques). 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.

9. Notre analyse

Dans un premier temps, nous cherchons à repérer les concepts pour constituer un dictionnaire qui pourra être transposé en structures cognitives du domaine d'expertise. Par structures cognitives, nous entendons un répertoire d'unités cognitives suffisantes pour définir l'espace et les opérateurs du problème à résoudre. Dans le cadre de notre intervention, ces primitives de la réalité à représenter sont des unités d'organisation qui ne se décomposent pas en prédicats mais en caractéristiques intrinsèques, les traits qui portent une valeur[33]; ils sont dotés de la forme suivante:



D'un point de vue formel, il s'agit d'arbres finis définis par la valeur des étiquettes de leur noeuds. L'expérience nous a démontré que ce mode de représentation de la connaissance convient à la plupart des univers cognitifs présents dans les organisations; il offre modularité, flexibilité et lisibilité. Ces structures cognitives définissent l'extension maximale de chacun des concepts manipulés par le système. Elles régissent la construction des règles d'inférences (les conditions de la prémisse autant que les inférences), de même que celle des faits (les données du problème).

Pour constituer ces structures cognitives, nous analysons les substantifs nominaux en termes des configurations de mots, appelés ingrédients, qui leur sont associés. Ainsi, par exemple pour le substantif "projet" on aura des configurations telles, l'assujettissement d'un projet, la pertinence d'un projet, etc. Cette analyse tire sa justification de ce que l'effet de référence au réel (les

concepts dans le cas qui nous occupe ici) dans un discours donné est tributaire de formes nominales qui consolident d'autres formes nominales en classes-objets. Ainsi, les marques référentielles[34] proviennent des configurations d'énoncés et des transformations linéaires engendrant la dynamique textuelle. Ces marques sont identifiables linguistiquement à partir des stratégies discursives[35] qui confèrent à certaines formes nominales une fonction de régie textuelle. À titre d'illustration, nous mentionnerons les trois configurations suivantes:

Description définie + chaînes de pronoms anaphoriques
Par ex.: Objet(x) (...) il,il,etc.
Suite de GP + déictique | substantif anaphorique
Par ex.: le projet n'est pas complet (...) le projet retarde (...)
c'est un cas de rejet (...)
Suite de GP + nominalisation
Par ex.: gp1,gp2,gp3... la demande est complète
Présentatif + (gn | gp) + que + gp*
Par ex.: Il y a des critères qu'il faut respecter : a,b,c, (...)

Voilà qui illustre comment peuvent s'ordonner linéairement les différentes configurations d'énoncés suivant les finalités du raisonnement. Une approche "logique naturelle" met en évidence les rapports de transition qui existent entre les énoncés d'un discours.

Les trois étapes de la constitution d'un dictionnaire des objets valués sont les suivantes:

i) Constitution d'une liste de concepts
Catégorisation morphologique
Blocage des locutions
Epuration de la liste des mots
Etablissement de la synonymie
ii) Dépistage des traits
iii) Dépistage des valeurs

Le repérage des règles d'inférences viendra dans une étape ultérieure où nous analyserons la détermination, définie comme les modalités des rapports entre les objets et les opérations susceptibles de leur être appliquées. Un lexique des opérations sera constitué des verbes transitifs descripteurs des transformations subies par les objets valués. À chacune des opérations, les arguments sont rattachés en termes d'objets valués. Les verbes d'action ou d'état seront de plus analysés en termes de modulation (actif, passif, nécessaire, facultatif, etc.), de localisation et de temporalité.

À plus long terme, les segments significatifs seront découpés en vertu de l'homogénéité de leur contenu et catégorisés selon une grille dont voici une liste provisoire: définitions, descriptions, lois, axiomes et stratégies. Ces derniers segments indiquent comment réordonner les autres segments pour respecter l'ordre des opérations pour résoudre les problèmes. De plus, nous prêterons attention à l'énonciation pour elle-même, soit les modalités d'inscription des déterminations dans un rapport (social) de communication par le sujet et les modulations des objets des schématisations (par ex.: soumettre une demande,

proposer un projet, etc.). L'énonciation nous intéresse en ce qu'elle gouverne les aspects dynamiques de la production des schématisations (suivant les phases de son raisonnement, l'expert peut passer du dire au faire, du certain au probable, etc.; observer la différence entre On dit que et x dit que). Les énoncés seront structurés en segments de textes significatifs de manière plus ou moins complexe (configurations à 1, 2, ... , n énoncés). Ils se présentent comme des conjonctions, des concessions, des restrictions, des transitions, etc. (par ex.: la demande de projet doit passer par l'étape a et b et c, etc.), des hypothèses ou des conséquences. Cette analyse nous permettra de mettre à jour les transitions d'états du problème et les stratégies de contrôle qui mènent à sa solution.

10. Constitution d'une liste de concepts

Cette première étape consiste à passer du lexique du corpus analysé, c'est-à-dire l'ensemble des mots qui le composent avec leur fréquence d'appatition, aux concepts ou termes, entendus comme noyaux des ancrages nominaux. Ce passage se décompose en plusieurs opérations que nous commentons dans l'ordre.

D'abord nous procédons à l'étiquetage morphologique du lexique. Pour ce faire, nous utilisons présentement un algorithme[36] composé de règles filtrant les flexions qui sont appliquées après la projection de dictionnaires d'exceptions. Ensuite, nous bloquons locutions terminologiques, c'est-à-dire les termes qui dépassent l'unité lexicale. Dans les domaines techniques une grande partie des termes est composée de plusieurs mots qui, pris séparément, ont chacun une signification (par ex.: traitement de texte). Leur construction autour d'une tête nominale semble être tributaire de patrons morpho-syntaxiques. Cette constatation nous amème à extraire des segments de textes (concordances) à partir de la co-occurrence de patrons morphologiques[37]. A titre d'illustration, voici trois schémas de patrons:

[nom + de + nom]
  par ex.: avis de projet, centre de documentation, chargé de projet
[nom + préposition + verbe infinitif]
  par ex.: matériel à draguer
[nom + adjectif] En raison de sa généralité, ce schéma de patron est étendu à
[nom + adjectif + adjectif]
  par ex.: réunion générale annuelle.

Tous les segments obtenus ne sont pas des locutions terminologiques; un expert doit procéder à l'épuration de la liste des éléments lexicaux. Cette opération repose sur des jugements de valeur et nécessite rigueur et constance. Une fréquence d'apparition d'une suite de mots plus grande que la fréquence de leur tête nominale employée séparément est un indice habituellement sûr. Il n'y a cependant pas de critères de sélection absolus mais relatifs au domaine d'expertise à couvrir et en fonction de la configuration prévue des unités cognitives. D'une part, il serait erroné de lier les mots étude de répercussion parce qu'il y a aussi étude d'impact, ce qui permet la construction de l'unité cognitive [étude -> type (répercussion / impact)]. D'autre part, la liaison serait

justifiée si l'unité cognitive entrevue était plus générale: [étape -> type (étude_de_répercussion / étude_d'impact)]. En somme, il faut faire attention de ne pas lier, sur la foi de co-occurrences nombreuses, les concepts et leurs valeurs si un trait est implicite. Il ne saurait y avoir de locutions terminologiques en soi mais par rapport à l'usage que l'on entend en faire. Les locutions terminologiques recevront l'étiquette nominale.

Les locutions terminologiques ayant été bloquées, le lexique des mots du corpus est restreint aux nominaux, puis une réduction manuelle est effectuée par un expert afin de ne retenir que les concepts afférents au domaine d'expertise. Voici quelques mots qui feront partie du lexique des concepts de SAGÉE: avis_de_projet, dragage, sédiment, etc.

Les concepts équivalents du lexique sont ensuite ramenés manuellement à une forme canonique, partagée par tous. Ainsi, par exemple, problème de bruit et pollution sonore sont synonymes. Il faut auparavant que la synonymie soit validée par l'expert. S'il s'agit de deux états différents, consécutifs dans le déroulement d'un processus, il n'y a pas d'équivalence. Cette opération permet de réduire le nombre de concepts et, lors d'un balayage du corpus, d'accéder au maximum d'occurrences d'un terme standard à partir de ses synonymes.

Les deux prochaines sections décrivent le passage des concepts aux objets valués. Leur qualité et leur quantité déterminent la richesse de l'inférence du SE. Pour utiliser le matériau textuel, il faut arrimer les groupes nominaux qui ont pour tête les concepts du lexique à la structure objet-trait-valeur.

11. Rattachement des traits

L'extension maximale de chacun des concepts/objets est isolée en superposant tous leurs contextes d'occurrence (concordances) dans le corpus. Plusieurs traits peuvent être repérés au moyen de patrons simples; comme par exemple le patron [nom_du_trait + préposition + nom_du_concept]. Pour le concept quai on trouve les segments suivants:

La longueur totale du quai (...)
L'emplacement du quai (...)
Si le quai est situé à (...)
La largeur du quai (...)

Nous travaillons, dans la perspective de l'analyse morphologique 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 de (gp | gn) + 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.

2) Suite de GP + 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 nominalisation 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 au déverbal "préparation" un trait anaphorique.

Lors de cette analyse, les informations quant à l'inscription des concepts dans une structure hiérarchique seront dépistées autour d'expressions partitives, telles une_espèce_de, une_partie_de, etc. Ces liens doivent être précieusement recueillis et adjoints aux concepts dans le dictionnaire.

12. Dépistage des valeurs

Une fois dépistées les configurations nominales qui fournissent les traits des objets, on poursuit les recherches en examinant les adjectifs présents dans les contextes dépistés. Parmi les formes adjectivales recherchées, il y a les quantificateurs (les numéraux, les cardinaux et les ordinaux), celles qui instancient des positions sur des échelles (par ex.: froid, tiède, chaud, brulant, bouillant, etc.), ou des partitions de masses (par ex.: demi, trois-quart, etc.). Ces formes adjectivales font apparaître les échelles argumentatives qui positionnent virtuellement les autres valeurs qualitatives ou quantitatives possibles[38].

13. Conclusion

Suite à l'analyse des sustantifs nominaux, nous constituons un dictionnaire de concepts/objets décrits de manière extensive dans les termes de leurs caractéristiques (les traits) pour lesquelles l'admissibilité de la valeur est spécifiée en terme d'échelle ou de contrainte. Ce dictionnaire peut être entre autre utilisé pour générer les structures cognitives d'un système expert. Il nous reste à tester la validité sur plusieurs corpus et à produire une évaluation des résultats obtenus.

Nous venons de proposer une approche s'inspirant de l'analyse du discours basée sur une perception unifiée et "naturelle" du savoir. Son dépistage des concepts décrivant l'espace d'un problème répond à des critères de constance,

d'objectivité, de reproductibilité et d'indépendance quant aux problématiques définies dans les textes. Cette approche interactive qui convient à une application en temps réel sur un très grand corpus a pour but de susciter la créativité de l'utilisateur.

Face à la masse sans cesse croissante de textes produits par les organisations qui dépasse de loin leur capacité de lecture, nous prônons l'utilisation de progiciels pour le traitement de la connaissance consécutive au traitement de textes afin de valoriser cette source d'expertise inestimable.

---

[1] Nous tenons à remercier chaleureusement Jules Duchastel pour ses précieux conseils, Pierre Mackay, directeur du Centre d'ATO et Diane Lessard, notre collègue, pour leur lecture attentive de ce texte, ainsi que leurs suggestions pertinentes. Nous tenons également à souligner la contribution importante de messieurs Alain Lecomte (GRAD, Grenoble) et Jean-Marie Marandin (L.I.S.H.) au domaine de l'analyse du discours et spécialement du développement des hypothèses discutées dans ces lignes.

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

[3] Ericcson K. A.; Simon H. A. Protocol Analysis: Verbal reports as data, MIT Press, Cambridge Mass., 1984.

[4] Grosberg, S. Neural Network and Natural Intelligence, MIT Press, Cambridge Mass., 1988.

[5] Voir: Frey, W.; Reyle, U. & Rohrer, C. "Automatic construction of a knowledge base by analysing texts in natural language". International Joint Conference on Artificial Intelligence, 1983, 727-729.

Nishida, T. Kosaka, A. & Doshita, S. Towards knowledge acquisition from natural language documents-automatic model construction from hardware manuals. International Joint Conference on Artificial Intelligence, 1983, 482-486.

Un système comporte cette fonctionnalité pour la langue allemande: Diederich J.; Ruhmann I. & May M., "KRITON: a knowledge-acquisition tool for expert systems", International Journal of Man Machine Studies;1987; 26; 29-40

[6] Le Système d'Aide à la Gestion des Evaluations Environnementale (SAGÉE) a déjà fait l'objet d'une présentation: Actes du premier colloque québécois en Informatique cognitive des organisations, Juin 1987: section 2, 21-28.

[7] On évalue la taille présente du corpus à envion 3 millions de mots (15 meg.)

[8] Ce projet de recherches, dirigé par Jules Duchastel et financé par le Fonds FCAR du Québec dans le cadre du programme "actions spontanées" a commencé ses activités en janvier 1988.

[9] Un dictionnaire des concepts semble nécessaire dans plusieurs approches au transfert d'expertise. Voir Hart A., Knowledge Acquisition for Expert Systems: McGraw-Hill; 1986: 65-66.

[10] L'abstraction réfléchissante "consiste à tirer d'un système d'actions ou d'opérations de niveau inférieur certains caractères dont elle assure la réflexion (...) sur des actions ou opérations de niveau supérieur, car il n'est possible de prendre conscience des processus d'une construction antérieure qu'au moyen d'une reconstruction sur un nouveau plan." Jean Piaget, Études d'épisétmologie génétique, XIV, Presses Universitaires de France, 1961, p. 203.

[11] Michel Meyer, Découverte et justification en science, Paris, Éditions Klincksieck, 1979, Chap. IX "La conception problématologique de la science", pp. 289-353.

[12] "A systems analyst can fact-find by questionnaire, by sampling records or by observing poeple at work, but no analysis is complete without face-to-face discussions witch users." A. Hart, op. cit: p. 49.

[13] André Lalande, Vocabulaire technique et critique de la philosophie, Paris, Presses Universitaires de France, 1976, p. 237.

[14] Cette définition de l'induction s'inspire directement de la description aristotélicienne (Physicorum I, 184a)

[15] Voir entre autres:

Bennet, J. S. "ROGET: A knowledge-based system for acquiring the conceptual structure of a diagnostic expert system". Journal of Automated Reasoning; 1985; 1: 49-74;

Boose, J. H. "A Knowledge Acquisition Program for Expert System Based On Personnal Construct Psychology". International Journal of Man Machine Studies; 1985; 23: 495-525;

Bradshaw J. M. "Expertise transfer and complex problems: using AQUINAS as a knowledge-acquisition workbench for knowledge-based systems". International Journal of Man Machine Studies;1987; 26; 3-28;

Eshelman, L.; Ehret, D.; McDermott J. & Tang M., "MOLE: a tenacious knowledge-acquisition tool".International Journal of Man Machine Studies;1987; 26; 41-54;

Klinger, G. ; Bentolina, J. ; Genetet, S. ; Grimes, M. ; Mac Dermott, J. "KNACK Report-driven knowledge acquisition". International Journal of Man Machine Studies;1987; 26; 65-79;

Aussenac, N. ; Michez, B. "M.A.C.A.O.: Application d'un modèle psychologique à la réalisation d'un outil d'aide à l'acquisition des connaissances. Actes du colloque Représentation du réel et informatisation; 26 et 27 mai 1988 tenu à Saint Etienne (France).

[16] L'analyse du discours est la discipline qui s'intéresse à ces problèmes. Voir pour une vue d'ensemble D. Maingueneau, Initiation aux méthodes de l'analyse du discours, Paris, Hachette, 1976.

[17] L'anaphore (du grec anaphora : référence, rappel, recours) désigne la caractéristique qu'ont certains mots (généralement des pronoms) de mettre en relation (de référencer) des mots ou des termes avec des éléments énoncés antérieurement. Par exemple : "il" dans "Pierre mange une pomme, il aime les fruits". Dans cet exemple, le pronom il fait référence a Pierre. Les relations anaphoriques sont de plusieurs types: pronom utilisé comme substitut, termes qui consolident une suite d'énoncés : le dragage, l'excavation, le pavage, ... ces travaux seront requis, etc.

[18] SATO, Système d'Analyse de Textes par Ordinateur par F. Daoust et Déredec, Atelier pour l'analyse et la construction de systèmes cognitifs par P. Plante.

[19] CBSF, Catégorisation de base du français et LCMF, Lemmatisation et Catégorisation Morphologique du Français par L. Dumas; ALSF, Analyseur Lexico-Syntaxique du Français par J. M. Marandin.

[20] Michel Foucault, L'Archéologie du savoir, Éditions Gallimard, 1969, p. 238.

[21] Michel Foucault, op. cit., p. 53.

[22] Pierre Bourdieu, "La spécificité du champ scientifique et les conditions sociales du progrès de la raison", Sociologie et Société, Science et structure sociale, Vol. 7, N° 1, pp. 91-118.

[23] J. P. Poitou, "The Expert ans the System" projet d'article fourni par l'auteur en mai 1987.

[24] G. Polya, How to Solve It. A New Aspect of Mathematical Method, Princeton, New Jersey, Princeton University Press,1973.

[25] Pour une classification voir Susan Haack, Philosophy of Logics. Great Britain: Cambridge University Press; 1978; 276.

[26] Pierre Oléron, Le raisonnement, Paris, Presses Universitaires de France, Que sais-je #1671, 1977, p.9.

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

[28] Antoine Culioli, "La formalisation en linguistique", dans : Culioli, A., Fuchs, C. et Pêcheux, M., Considérations théoriques à propos du traitement formel du langage, Documents de linguistique quantitative, N° 7, Dunod, 1970, pp. 1-13.

[29] Borel, M.-J.; Grize, J.-B.; Miéville, D. op. cit.: 99-146; 241.

[30] Les exemples qui suivent sont tirés de retranscriptions d'entrevues réalisés dans le cadre du projet SAGÉE; il s'agit de discours oral ce qui explique le relâchement de la syntaxe. C'est nous qui soulignons.

[31] Lecomte, A.; Marandin, J.-M. "Analyse de discours et morphologie discursive", Document de travail, 1984 , 67 pages.

[32] Lecomte, A. "Espace des séquences : approche topologique et informatique de la séquence," Maldidier, D. et al., "Analyse de discours: nouveaux parcours", Langages, Mars 1986, Larousse, p. 93.

[33] Le progiciel générateur de systèmes experts développé au centre d'ATO de l'UQAM par Louis-Claude PAQUIN, appelé Déredec-EXPERT, utilise ce mode de représentation de la connaissance qu'il baptise "granule". Le granule n'est pas d'une structuration en trois niveaux, mais l'organisation interne d'un objet cognitif qui peut être inscrit dans une hiérarchie arbitrairement complexe.

[34] Nous reprenons en l'élargissant l'exposé d'Alain Lecomte "Algorithmes de la séquence", Exposé présenté le 27 janvier 1983 dans le séminaire de la ACP "ADELA", 24 pages.

[35] Le choix des thèmes et l'instanciation nominale des primitifs sémantiques; Michel Foucault, op.cit., pp. 85-93.

[36] CBSF, Catégorisation de Base Syntaxique du Français par L. Dumas. La couverture de cet algorithme est d'environ 80% des mots d'un texte; en cas d'ambiguité, l'étiquette accolée désigne toutes les possibilités.

[37] Pour ce faire, nous utilisons SATO par F. Daoust qui est doté des capacités d'une part de bloquer les concordances dépistées et d'en faire un lexique avec fréquences et d'autre part de constituer des listes de locutions qui peuvent être projetées sur d'autres textes.

[38] Oswald Ducrot, Les échelles argumentatives, Paris, Les Éditions de Minuit,1980.