Le passage des termes aux concepts[1]

Louis-Claude Paquin[2]
Centre d'analyse de textes par ordinateur
Université du Québec à Montréal

INTRODUCTION

Dans les années 1990, la pénétration des systèmes à base de connaissances dans les organisations tant publiques que privées dépend en grande partie de l'efficacité et du coût de la phase du transfert d'expertise. Pour la plupart des applications, le repérage et l'organisation des connaissances passe par une lecture analytique de textes. Qu'il s'agisse de textes réglementaires, de cahiers de procédures, de mémos ou encore de retranscriptions d'entrevues avec les experts du domaine, la connaissance relative à l'expertise à modéliser réside en grande partie dans des textes. Même lorsqu'elle n'atteint qu'une envergure limitée, une expertise repose la plupart du temps sur un large corpus dont les éléments sont fortement interreliés: par exemple d'un article de loi découlent des règlements et une procédure dont l'application donne lieu à des guides et à des interprétations. Pour augmenter l'efficacité du processus, le recours à des outils informatiques et à des chercheurs en analyse de textes s'impose.

Il s'avère cependant que le transfert d'expertise à partir des textes pose plusieurs difficultés qui sont loin d'être résolues à l'heure actuelle; en voici un inventaire partiel. D'abord il s'agit d'un domaine de recherches potentiellement lucratif sur le plan des redevances ou des subventions, ce qui impose des contraintes sur le mode d'intervention. Ensuite, il n'y a pas de théorie générale et unifiée du texte, mais un foisonnement d'approches parfois contradictoires. Enfin le transfert d'expertise est une activité trop nouvelle pour être bien spécifiée. Pour composer avec ces difficultés, nous optons pour une approche de type recherche-action orientée vers la méthodologie.

La chaîne de traitements ayant déjà fait l'objet d'une présentation détaillée[3], dans ce texte je présenterai les réflexions et les prises de position qui ont accompagné l'élaboration d'une méthodologie pour constituer un dictionnaire des concepts à partir de l'analyse des textes d'un domaine d'expertise donné. Les thèmes abordés sont les suivants: la nécessité de la prise en compte de la textualité pour le dépistage des concepts, la logique naturelle comme cadre théorique, le concept de concept, les différentes stratégies pour le repérage des termes et enfin le passage des termes aux concepts. En terminant, les prochaines étapes sont esquissées.

---

1 Ce texte est présenté au Colloque international "Les industries de la langue. Perspectives des années 1990" Thème I : Aspects technologiques: Représentation des connaissances dans le traitement des langues naturelles, Montréal le 22 novembre 1990.

2 Je tiens à saluer mon collègue Luc Dupuy en compagnie duquel j'ai écrit plusieurs textes dont je m'inspire ici largement.

3 L.-C. Paquin, L. Dupuy, Y. Rochon, "Analyse de texte et acquisition des connaissances : aspects méthodologiques" (à paraître dans *Informatique cognitive des Organisations*, numéro spécial sur l'analyse documentaire automatique).

INVENTAIRE DES DIFFICULTES

La première source de problème touche aussi les autres thèmes de recherches en intelligence artificielle et, par delà, toute la recherche appliquée. Il s'agit de domaines de recherches dont les résultats sont susceptibles de s'appliquer à de vastes marchés et, par le fait même, de s'avérer très lucratifs. La structure de financement de la recherche étant de plus en plus assujettie à l'obtention de redevances ou de contrats de services de la part d'organisations privées ou publiques, l'objectif des chercheurs s'est déplacé de l'élaboration et la validation empirique de modèles théoriques desquels découleront des applications, vers la mise au point rapide de produits ou biens livrables. C'est ainsi que pour se procurer un avantage sur la concurrence, les centres de recherches construisent des progiciels ou des applications de type clé en main dont les stratégies computationnelles sont inaccessibles à l'utilisateur. Il est alors difficile de vérifier la concordance entre les résultats obtenus et les principes théoriques annoncés. De plus, la solution apportée au problème est la plupart du temps trop générale pour les besoins exprimés par les utilisateurs. Enfin le passage d'une théorie à l'application robuste et efficace entraîne parfois des compromis ou des incomplétudes qui peuvent s'avérer "troublants" sur le plan théorique. Ainsi, par exemple, les interfaces en langue naturelle pour l'interrogation des bases de données qui sont dérivées de théories linguistiques qui, lors de l'analyse d'une question, ignorent toutes les formes fonctionnelles pour plus d'efficacité.

Un deuxième obstacle provient de la divergence des multiples approches en analyse de textes par ordinateur (ATO) qui parfois, à tort ou à raison, s'opposent. Il suffit de rappeler ici les débats où sont opposés le quantitatif et le qualitatif, le statistique au syntaxique, le structuralisme au fonctionnalisme, le macroscopique au microscopique, etc. Autant de points de vue sur le texte, autant de contextes théoriques différents, de modèles particuliers: linguistiques, cognitifs, psychologiques, documentaires, anthropologiques, sociologiques, etc. Ce foisonnement reflète l'absence d'une théorie générale et unifiée du texte de laquelle découlerait, au stade opérationnel, une chaîne de traitements garantissant un dépistage consistant et complet des connaissances qui y sont à l'oeuvre. Une telle théorie qui rendrait entièrement explicite ses composants, règles de formation, principes, niveaux, etc., est-elle possible? Le principe qui veut que plus une théorie est spécifique, moins la couverture de son domaine d'application est large se voit encore une fois confirmé. Qui plus est, à l'intérieur des disciplines qui, à des degrés divers, ont le texte pour objet, les divergences abondent. Les études de P. Bourdieu[4]. suggèrent que celles-ci ne sont pas attribuables qu'a des aspects intrinsèquement "théoriques". Il a bien montré que les divergences épistémologiques sont en fait les indices révélant les divisions politiques à l'intérieur du dommaine de la pratique scientifique. L'enjeu de ces luttes pour le monopole de l'autorité scientifique c'est l'obtention de subventions, de contrats, d'un avantage sur la "concurrence", etc. Par ailleurs, les chercheurs sont reconnus pour élaborer en vase clos des nomenclatures et des modèles complexes dont l'utilisateur ne sait trop que faire pour résoudre son problème.

---

4 P. Bourdieu, "La spécificité du champ scientifique et les conditions sociales du progrès de la raison", *Sociologie et sociétés*, 7 (1), mai 1975, pp.91-118

En somme, jusqu'à présent, les secteurs particuliers d'application de l'ATO ont donné lieu à des modèles de traitements qui s'avèrent difficilement composables entre eux et transposables dans le secteur du dépistage de la connaissance, même s'ils semblent connexes a priori.

La troisième source de difficulté tient dans la nouveauté et la complexité de la tâche elle-même. D'abord, la notion d'expertise est mal circonscrite. Dans le contexte de l'intelligence artificielle, elle renvoie habituellement à l'ensemble suffisant des connaissances mises en oeuvre pour accomplir une tâche donnée. Toujours dans le même contexte, la notion de connaissance est hétérogène, elle recouvre des concepts et leurs relations, des actions, des transformations, des évaluations, des stratégies, etc. Ensuite, l'opération de transfert effectuée sur cette expertise qui est habituellement itérative (essais et erreurs), peut être décomposée en plusieurs étapes, chacune comportant ses difficultés: la phase de dépistage où la connaissance pertinente est mise à jour, la phase de modélisation où une représentation est construite à partir des résultats de l'étape précédente et une phase de formalisation de la représentation dans les termes du paradigme opérationnel retenu (prédicats du premier ordre, objets valués, règles d'inférences, etc.).

Il ressort de l'analyse qui précède que nous en sommes encore au stade du bricolage. Il ne doit pas en découler un constat d'impossibilité, mais une grande prudence au niveau des objectifs et des attentes suscitées, autant du côté des chercheurs que de celui des utilisateurs de l'ATO. Pour contourner au moins partiellement ces difficultés nous avons opté pour le développement d'une méthodologie par phases de traitements interactifs"sur mesure" à partir de cas concrets.[5]. Celle-ci, à l'aide de certains outils généraux, offre une assistance à l'utilisateur dans l'accomplissement de sa tâche, ici le transfert d'expertise. Il s'agit d'une intervention au sein même de l'organisation qui pourrait être appelée recherche-action. Une formation de base est d'abord donnée. Vient ensuite une phase d'analyse-conseil où la méthodologie est adaptée à la situation spécifique. Par la suite, l'intervention se transforme en une supervision. La plupart des utilisateurs engagés dans ce type d'intervention manifestent un haut taux de satisfaction, notamment parce qu'elle leur parmet d'acquérir graduellement une autonomie dans la solution de leur problème et surtout en raison de sa transparence ils comprennent ce qu'ils font. Cependant, les profits suffisants pour soutenir un niveau d'activité scientifique qui maintienne la méthodologie à jour sont difficilement générés.

UNE APPROCHE MÉTHODOLOGIQUE

La stratégie d'analyse proposée peut être qualifiée de mixte et surtout pragmatique car elle ne découle pas d'un assujettissement théorique qui force une redéfinition de la tâche et conditionne le stade opérationnel. Les cadres théoriques utilisés sont empruntés à la science documentaire, à la linguistique, à la sociologie, à la philosophie et à l'anthropologie. Ils ont été sélectionnés en fonction d'une évaluation empirique de la pertinence des résultats obtenus lors de leur opérationnalisation. La délimitation de l'extension des concepts prime sur l'étude des modalités linguistiques de leur expression (terminologie). Ainsi, non seulement l'enchaînement des mots dans les phrases est pris en compte, mais

---

5 Notamment au Ministère de l'Environnement et au Secrétariat du Conseil du Trésor du Québec.

aussi les conditions de production du texte, le projet communicationnel sous-jacent et l'insertion du texte dans la trame des ses voisins (intertextualité).

La stratégie d'analyse proposée peut aussi être qualifiée de constructiviste à l'instar de la démarche analytique humaine qu'elle vient appuyer[6]. En effet, nous voyons la lecture humaine comme un processus de filtrage et d'annotation de certains éléments en fonction d'hypothèses et d'objectifs spécifiques. C'est pourquoi le sens des termes n'est que rarement perçu dès la première lecture; il est plutôt acquis par saturation de contextes accumulés et condensés (synthèse) suite à l'accomplissement successif de plusieurs cycles. Chacun des cycles comporte minimalement les étapes suivantes: la formulation d'hypothèses, la description des documents, l'extraction des données et l'analyse proprement dite. Les résultats d'une précédente analyse participent à la reformulation des hypothèses.

De plus, notre méthodologie nécessite l'implication des utilisateurs à tous les stades du traitement afin qu'ils deviennent les véritables développeurs du système. Ainsi, les utilisateurs ne font pas qu'exprimer leurs besoins et valider le système reçu, mais assument la gouverne (contrôle) des opérations de transfert d'expertise. L'implication des utilisateurs facilite l'implantation du système parce qu'elle garantit la reconnaissance de la spécificité culturelle du matériau textuel, c'est-à-dire les schèmes socio-culturels régissant l'acte de lecture/écriture du corpus de textes considéré. En effet, la plupart du temps, pour que la performance d'un système à base de connaissances soit jugée adéquate, il faut non seulement que les réponses obtenues s'avérent justes, mais que ses dialogues et son "raisonnement" reflètent la culture de l'organisation et s'intégrent à l'ergonomie de l'environnement informatique.

Nous croyons qu'une suite de traitements interactifs permet aux développeurs de systèmes â base de connaissances, de mettre progressivement à jour, non seulement les connaissances à l'oeuvre dans les textes, mais le processus de leur élaboration dans le discours. Ainsi, si une grande place est accordée à la dimension heuristique, c'est qu'il nous semble important que les utilisateurs prennent conscience de la connaissance que renferment leurs textes, qu'ils la reconnaissent comme étant leur et se l'approprient pour la réinjecter dans leurs activités quotidiennes. C'est ainsi que tout au long de la démarche proposée, des sous-produits utiles pour l'organisation sont générés: des lexiques permettant le contrôle terminologique, des bases de données textuelles interrogeables par le système à base de connaissances ou autrement et surtout une sensibilisation de l'organisation à la richesse des textes qu'elle produit.

Voyons maintenant les caractéristiques logicielles requises pour opérationnaliser cette méthodologie. D'une part, les opérations fondamentales de l'analyse de textes doivent être accomplies sur de très larges corpus de textes. Cees opérations sont en nombre limité: la segmentation et le découpage, l'étiquetage et le filtrage (pattern matching), le groupement et les comparaisons. Pour accomplir le transfert d'expertise, elles doivent être combinées en des opérations de plus haut niveau, telles la concordance (extraction par patron de fouille de mots qui apparaissent dans un contexte réduit), la projection de

---

6 Voir J. Duschastel, L. Dupuy, F. Daoust, "Système d'analyse du contenu assisté par ordinateur (SACAO)", *Actes du Colloque La description des langues naturelles en vue d'applications linguistiques*, Québec: Centre international de recherche sur le bilinguisme, 1989, pp. 197-210.

dictionnaires (liste de mots déjà catégorisés), catégorisation manuelle, etc. Ces dernières opérations doivent de plus être effectuées rapidement de façon à soutenir l'intérêt de l'utilisateur et à encourager son interaction avec le corpus de textes étudié. D'autre part, les ressources matérielles nécessaires doivent être accessibles et connues des utilisateurs-développeurs. La qualification requise pour manipuler l'environnement informatique ne doit pas être trop élevée car elle nécessiterait une formation et un entraînement trop long. Cet environnement doit être convivial et robuste pour ne pas engendrer des frustrations. Le logiciel SATO[7] rassemble la plupart de ces caractéristiques.

Plutôt que de proposer un modèle de traitement pour «automatiser» le transfert d'expertise à partir des textes qui ne s'appliquerait qu'à un micro-monde ne correspondant que rarement aux situations rencontrées dans les organisations, étant donné aussi l'ampleur des difficultés et notre mode d'intervention dans les organisations, nous avons choisi de découper la tâche pour offrir immédiatement des services. Dans un premier temps, nous nous sommes intéressés aux problèmes soulevés par le dépistage des connaissances dans les textes. Nous nous sommes concentrés sur les «objets», c'est-à-dire les concepts sont recherchés.

PRISE EN COMPTE DE LA TEXTUALITÉ

La recherche de concepts qui seraient formulés clairement et explicitement dans les textes est habituellement une expérience fort décevante. Les formulations sont partielles, trop contextualisées ou encore entremêlées à d'autres concepts, et par là même peu utilisables directement parce que locales. Découper les textes afin d'en sélectionner des segments jugés significatifs sur la base de critères linguistiques ou autres pour en faire des données contraintes, les concepts, admissibles aux systèmes à base de connaissances manifeste autant une incompréhension de la spécificité du texte que de la nature des concepts. L'expérience montre que les concepts ne sont pas déposés dans les textes mais littéralement construits par le lecteur en interaction avec le texte[8]. Une maîtrise suffisante du domaine, ainsi qu'une connaissance des conditions de production, s'avèrent préalables à la reconstitution des concepts qui sont à l'oeuvre. En effet, des modifications du cadre référentiel du texte peuvent amener le lecteur à produire des inférences. Dans un texte, tous les mots ne réfèrent pas toujours au monde extérieur; ils peuvent servir à une re-catégorisation des concepts pour constituer des paradigmes, c'est-à-dire des classes d'équivalences contextuelles, soit à constituer des méréonomies. Il s'agit de hiérarchies régissant des complexes de relations entre un tout et ses parties, entre les parties de parties, etc[9].

---

7 Système d'Analyse de Textes par Ordinateur, développé par François Daoust du Centre d'Analyse de Texte par Ordinateur de l'Université du Québec à Montréal. SATO est utilisé sur du matériel de type IBM-PC-AT™ et tourne sous la version 3.3 de DOS™. Il peut avantageusement être utilisé avec des logiciels multitâches comme WINDOWS™ (version 3.0) ou DESQVIEW™, ce qui permet d'augmenter la productivité et l'interactivité du travail d'analyse.

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

9 "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,

Dans les textes, l'effet de référence est largement tributaire des formes nominales auxquelles on associe le processus de dénomination[10]. Ces formes nominales sont appelées termes. Toutefois l'effet de référence n'est que rarement le fait du terme isolé; il est habituellement consolidé, spécifié, qualifié, élaboré par d'autres références {épithète, complément du nom, proposition relative}. Ainsi, des marques référentielles[11] proviennent de configurations d'énoncés et des transformations linéaires engendrant la dynamique textuelle. Ces marques sont identifiables linguistiquement à partir de stratégies discursives[12], notamment le choix des thèmes et des primitifs sémantiques, qui confèrent à certaines expressions nominales une fonction de régie textuelle sur d'autres pour les caractériser; par exemple, la catégorie de longueur qui joue le rôle de foncteur formatuer de nom dans l'expression «la longueur du quai». Ce terme ne se comprend quà l'intérieur du complexe des relations (syntaxiques) qu'il entretient avec les mots qui le précèdent et ceux qui le suivent.

Le texte n'est donc pas constitué d'un ensemble de données discrètes; il ne se réduit pas plus à l'ensemble des mots qui le composent qu'aux relations réunissant ceux-ci en un contenu (la signification). Il est d'abord et avant tout un acte de langage sur des savoirs (discours)[13]. Le modèle du texte préconisé ici est un ensemble des systèmes interreliés. Le terme ensemble est employé au lieu de hiérarchie à dessein parce que les systèmes entretiennent entre eux de multiples relations de dépendance parfois mutuelle. La perception de la parole relève du système phonologique; celle du texte repose sur un ou plusieurs systèmes typographiques. La référence des mots au monde par le dictionnaire constitue le système lexical; le rôle de chacun des mots dans l'énoncé est le fait de deux systèmes: morphologique, le système des marques que portent les mots et syntaxique, celui qui régit la combinatoire des mots dans les énoncés. Les autres systèmes sont moins bien définis. Le système sémantique est souvent pensé comme une sorte de calcul sur les propriétés lexicales des mots en fonction de leur position syntaxique dans un segment donné. Dans les conditions normales de lecture il nous faut composer avec l'intrication des systèmes. Il est par exemple virtuellement impossible de choisir automatiquement entre plusieurs catégorisations de surface potentiellement contradictoires sans une analyse de la structure profonde de l'énoncé.

Ces systèmes linguistiques constituent la micro-structure du texte, dont l'éventuelle maîtrise n'est pas une condition suffisante pour accomplir la tâche analytique du dépistage des concepts. Des niveaux de description

---

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. La version anglaise de ce texte est à paraître sous le titre L. Dupuy, L.-C. Paquin "An approach to Expertise Transfer : Computer-Assisted Text Analysis" *Advances in Computing and the Humanities : Content, Concepts,Meaning.* J A I Press, Greenwich, Connecticut, vol 3-4.

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

11 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.

12 M. Foucault, *L'Archéologie du savoir*, Éditions Gallimard, 1969., pp. 85-93.

13 M. Foucault, *op. cit,* p. 238.

supplémentaires, proprement textuels, appelés macro-structures sont requis. Parmi ces derniers, mentionnons le réseau d'argumentation, l'environnement communicationnel, l'organisation thématique, les figures de style, etc. Ces systèmes s'appliquent à des segments d'une autre nature que la phrase, des segments à géométrie variable tels le paragraphe ou tout autre découpage arbitrairement construits par les exigences de cohérence interne. La transformation du matériau langagier en matériau formel apte au calcul logique nécessite une prise en compte de l'ensemble de ces systèmes, c'est-à-dire la «textualité».

LES SCHÉMATISATIONS

Nous avons vu que dans les situations discursives, contrairement aux logiques formelles qui font en grande partie abstraction de la nature des objets qu'elles manipulent, les termes ne sont jamais quelconques. Ils réfèrent au réel et sont toujours spécifiés (mis en contexte) à un certain degré. Dans les textes, les concepts ne sont pas manipulés à des fins de démonstration[14] mais de schématisation. Les schématisations sont les opérations discursives structurant des objets cognitifs pour les articuler dans l'espace d'un savoir. La logique naturelle s'intéresse à de telles opérations, qualifiées de logico-discursives[15], mises en jeu par les locuteurs impliqués dans une pratique discursive. Cette théorie présuppose que le discours est une organisation de signes verbaux qui portent la marque d'activités intellectuelles par lesquelles les individus analysent et interprètent les mondes qui sont offerts[16]. 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.[17].

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,

---

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

15 Pour une courte introduction à la logique naturelle voir Y. Chang, "Une conception des opérations logico-discursives", *Sociologie et intelligence artificielle* A. Turmel ed. Québec, Laboratoire de recherches sociologiques, Département de sociologie, Université Laval, Collection Séminaire de recherche 1988 N° 1, pp.105-129.

16 M.-J. Borel. et al.,. *op. cit*.: p 41.

17 M.-J. Borel et al.,. *op. cit*.: 99-146; 241.

etc.). 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). De plus, à une désignation nominale donnée semble correspondre une manière particulière de structurer la référence au réel.

L'opération d'ancrage est le processus par lequel l'unité sémantico-cognitive vient prendre place dans un processus de schématisation. Les unités se trouvent à être stabilisées à l'intérieur des formes linguistiques soit nominales soit verbales. 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'émissaire d'eaux usées de l'usine de pâtes et papier"). 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").

Dans la perspective où la langue naturelle est à elle-même son propre métalangage[18], c'est-à-dire qu'elle porte des marques de son organisation qui sont dissociées du contenu véhiculé, la stratégie d'analyse consiste à utiliser ce métalangage pour isoler, par leur configuration et leur récurrence, des noyaux de référence dont les propriétés associatives sont appréhendées avec profit au moyen de la catégorie d'espace (position, adjacence, proximité et éloignement). Le cadre théorique que nous retenons pour l'analyse de textes appliquée au dépistage des concepts est la logique naturelle appliquée aux ancrages nominaux des schématisations. Au lieu d'une description arborescente de la hiérarchie de relations des mots de chacune des phrases, qui s'avère lourde et difficile à valoriser, l'analyse par la logique naturelle produit des inventaires, des classifications, des topographies ou encore des partitions du texte; cela facilite la constitution d'un dictionnaire de concepts.

LE CONCEPT DE CONCEPT

D'un point de vue philosophique, le concept est à la fois acte de pensée et objet de pensée[19]. Il est acte de pensée en ce qu'il permet de découper dans un univers d'objets la classe de ceux qui possèdent telle ou telle propriété. Il est objet de pensée par son aptitude à la dissociation en éléments ou parties. Des procédures réglées d'analyse et de composition peuvent lui être appliquées. Ainsi, un ensemble d'objets peut être ordonné en vertu de concepts; un concept se différencie par un ensemble de caractéristiques. Sur le continuum entre l'universel et le particulier, le concept est plus près du premier pôle. Il est induit à partir des instances particulières d'objets, tout en permettant de les distinguer. Il en va de même en neurobiologie où le concept est défini comme «un objet de mémoire qui ne possède qu'une faible composante sensorielle, voire pas du

---

18 A. 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.

19 G.-G. Granger, "Catégories et raison" *Encyclopédie de la philosophie*, Paris, P.U.F., p. 530.

tout»[20]; cet objet de mémoire sert par la suite à identifier les objets perçus. Pour ce qui est de la délimitation des concepts, deux perceptions s'opposent. La première, positiviste, considère les concepts comme des formes solides, des entités ayant des limites tangibles et concrètes. Beaucoup plus commode pour l'informatisation, cette perception ne tient cependant pas la route; les concepts débordent les uns sur les autres, s'amalgament, se modifient, etc. Depuis Bergson, la perception qui prévaut est que les concepts s'assimilent plutôt à des fluides:

> "Je fais, refais et défais mes concepts à partir d'un horizon mouvant, d'un centre toujours décentré d'une périphérie toujours déplacée qui les déplace et les différencie"[21]

Même si les distinctions ne sont pas toujours étanches, les concepts sont empiriquement découpés en fonction de caractères morphologiques ou utilitaires. Ce jeu de caractéristiques ou propriétés qui peut varier selon les différentes perspectives adoptées permet en retour de constituer les objets en classes.

Dans le cadre plus étroit des systèmes à base de connaissances, le concept est un primitif, l'unité de la connaissance. Il est considéré unitaire parce que sa décomposition en unités plus fines n'est pas jugée nécessaire pour l'utilisation qui lui est réservée. Les concepts pressentis lors de l'analyse seront reconstitués en configurations de caractéristiques dont la valeur est contrainte. Ici le principe aristotélicien d'identité doit jouer à fond. Dans un système formel il est impossible qu'un même attribut appartienne et n'appartienne pas en même temps au même sujet sous le même rapport (le principe du tiers exclu). Les concepts seront ensuite traduits en descriptions formelles appelées à être mises en relation du type implication (règles d'inférences), ces énoncés logiques qui font porter la valeur de vérité de certaines instances de concepts (prémisse) sur d'autres instances de concepts.

Les concepts en tant que modèles construits peuvent être représentés sous la forme de prédicats qui ont pour arguments leurs caractéristiques. Cependant ils sont plus efficacement représentés en objets symboliques (arbres finis définis par la valeur des étiquettes de leur noeuds) dotés de variables dont les valeurs ne sont pas exclusivement booléennes mais scalaires. Par exemple le concept de température pourrait être vu comme un doublet de variables. La première variable serait la mesure en degré Celcius avec comme valeur un nombre d'une précision de deux chiffres. La seconde variable serait l'appréciation avec comme valeur un élément de l'ensemble suivant: {bouillant, chaud, tiède, froid, glacé}. Comme on peut le voir, la variable ne s'évalue pas de façon booléenne (par oui ou par non), mais par la sélection d'un élément dans un ensemble. La notion ensembliste d'éléments discrets n'est pas tout à fait adéquate pour décrire l'attribution d'une position sur une échelle parce qu'il s'agit de la graduation continue d'une qualité. Il vaut mieux considérer l'échelle comme une espèce très contrainte de classe méréonomique (cf. supra). Cette représentation des concepts en objets valués s'avère intéressante parce qu'elle est aussi efficace dans le cadre logique des systèmes à base de connaissances, que dans le cadre

---

[20] J.-P. Changeux, *L'homme neuronal*, Paris, Fayard, 1983, p.174.

[21] G. Deleuze, *Différence et répétition*, Paris, P.U.F. 1969 p. 4.

morpho-syntaxique des textes, le «point de départ» de l'enquête. En effet, ce formalisme convient autant à la rédaction des règles d'inférences qu'à la description du groupe nominal.

Nous verrons qu'on peut reconstituer les concepts par classification de l'ensemble des contextes (spécifications) entourant leur expression linguistique, appelée terme. La relation qui unit le terme au concept en est une d'instance au générique. Le terme est l'expression d'un concept dans un contexte donné, c'est-à-dire accompagné d'une consolidation particulière de caractéristiques, alors que le concept est une forme schématique qui «encapsule» les consolidations possibles. Les concepts existent hors du discours, mais ne se révèlent que par son intermédiaire. Parfois l'instance générale, le concept, est désignée par les mêmes mots que ceux qui en désignent les instances particulières, les termes; il en résulte une grande confusion. En effet, il y a une différence majeure entre le discours sur un concept et l'utilisation de ce concept dans un discours.

LE REPÉRAGE DES TERMES

Les termes sont plus que des suites de caractères séparées par des blancs (mots graphiques). Ils forment une classe particulière de mots à l'intérieur de celle des substantifs parce qu'ils représentent un concept (une notion[22]). En tant que représentation, les termes ne sont pas les concepts eux-mêmes dans leur universalité, mais le résultat d'un filtrage particulier, la matérialisation linguistique d'une partie de leurs caractéristiques sélectionnées en fonction d'un projet discursif particulier. Dans le cadre de la logique naturelle on parle d'ancrage nominal. Parfois les termes sont des mots mais la plupart du temps il s'agit de groupes de mots appelés termes complexes ou composés; par exemple, les termes suivants: «traitement de texte», «oedème interstitiel sérofibrineux avec granylocytes». Parfois ils se trouvent répertoriés dans les dictionnaires, mais la plupart du temps ils ne le sont pas. Comme on peut le constater, il s'agit de syntagmes plus ou moins figés qui gardent la trace de la structure et du contenu de leur contexte d'énonciation mais qui se comportent sémantiquement comme des mots simples.

La fabrication de termes est essentielle à la tenue de discours de spécialité, qu'il soit scientifique, technique, administratif ou autre. Il s'agit d'une activité très importante qui, sauf dans certains secteurs très limités comme la physique, n'est pas assujettie à des normes explicites. D'un point de vue systèmique, la langue est instable, de sorte qu'il est virtuellement impossible de prédire toutes les constructions possibles. On peut cependant décomposer morphologiquement les termes complexes en deux composants distincts. D'une part, il y a le composant nominal qui désigne la classe conceptuelle générale, appelé tête et d'autre part, un ou plusieurs composants servant à restreindre l'étendue sémantique de la tête, appelé modificateur[23].

Produire une liste exhaustive de termes s'avère presque impossible, car les critères de reconnaissance par les locuteurs d'une langue de spécialité sont mal connus. D'une part, un consensus autour de régularités

---

22 AFNOR Pr. 240-000, p.85.

23 *Glossaire de termes sur l'analyse documentaire*, définitions recueillies par S. Bertrand Gastaldy, texte fourni par l'auteure.

socio-terminologiques[24] tend à s'établir. D'autre part, la performance diffère selon l'individu en fonction de ses structures cognitives, de ses intentions et enfin de sa finalité. De plus, comme leur pénétration est souvent progressive, il peut y avoir plusieurs termes concurrents pour désigner un même concept. Le repérage des termes dans de très grands corpus de textes par la lecture humaine s'avère trop coûteuse, trop lente et peu fiable en raison de l'attention soutenue qu'il faut maintenir. Le recours à l'ordinateur s'impose, surtout s'il peut produire un repérage constant, objectif et reproductible des termes dans les textes. Il existe plusieurs stratégies de repérage automatique des termes complexes.

La première se fonde uniquement sur la redondance de segments de textes; le seul critère discriminant est d'ordre statistique[25]. Le bruit engendré par cette méthode est très grand car les segments fréquents peuvent n'avoir aucun statut linguistique et le silence peut être important; des termes peuvent n'apparaître que quelque fois dans un corpus et s'avérer pertinents. D'une part, il semble clair qu'un savoir linguistique est nécessaire pour améliorer la performance de l'ordinateur dans le repérage des termes. Toute la question est d'en déterminer le type et la portée, étant donné que le temps de calcul tend à augmenter de façon logarithmique avec la profondeur de description linguistique. D'autre part on constate que si le critère de la fréquence d'apparition élevée d'un segment s'avère discriminant, il n'est pas le seul à prendre en compte.

La deuxième méthode est basée empiriquement sur des régularités dans la structure de surface des textes. Ces régularités sont exprimées sous la forme de patrons de fouille à partir de catégories morpho-syntaxiques, tels {nom + préposition + nom} pour «traitement de texte» ou {nom + préposition + verbe à l'infinitif} pour «machine à coudre», etc. Cette approche, qui demande des moyens restreints (un logiciel produisant des concordances et un dictionnaire de catégories morphologiques[26]), est rapide et fournit des résultats somme toute satisfaisants pour le transfert d'expertise. Le bruit provient du fait que les catégories morphologiques sont assignées aux mots hors contexte; elles peuvent donc être plurielles: par exemple le mot «été» peut être un substantif ou un participe passé. Le bruit provient aussi de la généralité des patrons demandés. Le silence est minime puisqu'il suffit que le terme apparaisse une fois dans son intégralité pour qu'il soit dépisté.

La troisième méthode consiste en une prise de décision sur la base d'un traitement automatique de la langue naturelle. Le problème est que ces systèmes manipulent des porteurs de sens et non pas le sens lui-même. J. Sowa, un

---

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

25 L. Lebart, A. Salem, *Analyse statistique des données textuelles,* Paris, Dunod, Bordas, 1988.

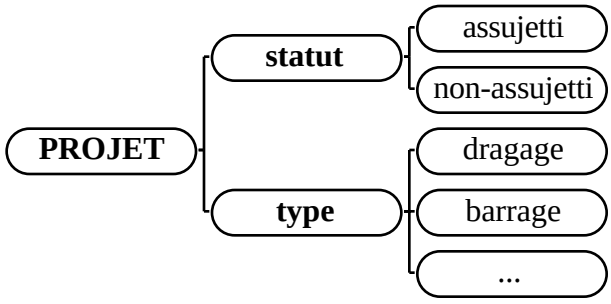
26 Nous avons opérationnalisé cette méthode dans l'environnement SATO de la façon suivante: d'abord une base de données lexicale (BDL) est projetée sur les textes. Puis un fichier de commandes SATO intitulé MARQUELO est exécuté. Les commandes consistent à dépister des concordances à partir de patrons de fouille sur les informations morpho-syntaxiques fournies hors-contexte par BDL puis à lier typographiquement les segments dépistés. Les résultats sont validés par l'utilisateur et un blocage physique des termes composés retenus peut être par la suite effectué automatiquement. BDL est développée par Luc Dupuy; elle regroupe présentement 358,820 mots du français écrit en une quinzaine de collections d'unités lexicales, telles substantifs, verbes à l'infinitif, verbes conjugués, adjectifs qualificatifs, pronoms, conjonctions, prépositions, adverbes, déterminants. Le temps requis pour indexer un corpus de 100 000 mots est d'environ 15 minutes (temps utilisateur) pour un IBM-PC de type 80386 (16 Mz).

membre de l'équipe de recherche en systèmes de IBM, affirme qu'après plus de trente ans de recherche, les succès sur un domaine restreint et l'échec lorsque le domaine est sans restriction s'explique par la nature fondamentale du langage. Une grammaire volumineuse ne suffit pas à étendre la couverture d'un petit système en un traitement de la langue naturelle sans aucune restriction[27]. Toutefois, un progiciel pour le dépistage des termes dans les textes, TERMINO™, a été mis au point dernièrement[28]. Les termes sont repérés sur la base d'une description arborescente, fondée sur la théorie X barre, de la syntaxe réalisée dans les phrases et d'un certain nombre d'heuristiques. Il reste à mesurer pour ce type d'approche le taux de bruit et de silence.

Ces trois stratégies de repérage se heurtent aux mêmes obstacles: la dispersion de l'information qui nécessite parfois la prise en compte de relations supra-phrastiques; les élisions, par exemple lorsque la tête du terme complexe est remplacée par un pronom, et les autres obscurités qui nécessitent une connaissance du monde de référence. Ainsi, quelle que soit la stratégie pratiquée, un tri manuel de contextes non pertinents repérés (le bruit) s'avère nécessaire. En effet, si l'ordinateur s'acquitte mieux que l'humain de l'aspect mécanique du balayage des textes, l'acquisition du niveau requis de familiarité avec le contenu pour prendre les bonnes décisions demeure problématique. Quoi qu'il en soit, il est important de rappeler que le dépouillement terminologique nécessaire pour induire les concepts à l'oeuvre dans un texte dans le but de construire un système à base de connaissances présente des différences appréciables sur le plan des modalités avec l'observation des formes et usages d'un terme pour fin de documentation. Dans le premier cas, les termes qui sont appelés à devenir des concepts doivent être validés par les experts du domaine, dans le second cas la visée est la normalisation de l'usage des experts.

LE PASSAGE DES TERMES AUX CONCEPTS

Un concept est doté d'un identificateur qui en désigne toutes les instances dans tous les contextes et un ensemble d'identificateurs de caractéristiques. Il sera global si les caractéristiques sont dotées d'un domaine de valeur, par exemple:



27 "... the successes of language processors on small domains and their failure on unrestricted domains result from the fundamental nature of language. In particular, a large grammar is not sufficient to scale up a small system to an unrestricted natural language processor" J.-F. SOWA, "Multi-Domain Semantic Theory" copie de travail fourni par l'auteur lors d'une conférence à Montréal datée du 28 novembre 1988.

28 S. David, P. Plante, *De la nécessité d'une approche morpho-syntaxique en analyse de textes*, texte fourni par les auteurs.

La sélection d'une valeur pour chacune des caractéristiques donnera un concept local ou instancié. Les règles d'inférences sont construites sur des concepts instanciés; par exemple:

Si le projet de dragage est assujetti, **alors** (...)

Voyons maintenant comment, à partir des termes on parvient à de tels concepts. Il s'agit de mettre progressivement à jour leur organisation en terme de configuration. Cette étape est cruciale; de la précision et de la cohérence des représentations construites dépend en partie la validité du système à base de connaissances.

Une fois que, parmi tous les substantifs, des termes ont été dépistés et que cette sélection a été validée, nous ne sommes pas pour autant en présence des concepts à l'oeuvre dans les textes analysés. Les concepts désignés par les termes s'avèrent difficiles à appréhender parce que mêlés au discours. A partir des termes intégrés à un contexte, seuls objets particuliers immédiatement accessibles, une démarche inductive doit être entreprise. Au moyen d'opérations de décontextualisation, de standardisation, d'explicitation, de condensation, de classification et de structuration, une remontée est effectuée pour arriver aux concepts[29]. Cette démarche s'oppose à la déduction qui, à partir d'une idée préconçue, pose une configuration finale qui sera fouillée de façon déterministe. Il va sans dire que face à la diversité et à la complexité des opérations mentionnées plus haut, l'intervention humaine est indispensable. De plus, un projet précis doit avoir été au préalable formulé de façon à restreindre le nombre illimité de caractéristiques pouvant décrire un objet. Enfin, des habiletés de toutes sortes sont mises à contribution parmi lesquelles figure une connaissance étendue du monde de référence. L'ordinateur nous offre une grande souplesse dans la visualisation de l'information. en terme de présentation de l'information. Des index permutés (Key Word In Context KWIC), qui sur une ligne affichent un certain nombre de caractères de chaque côté du mot en vedette, permettent de se faire rapidement une idée des configurations en présence autour d'un concept donné:

l'assujettissement d'un	projet, il faut avoir
de la pertinence d'un	projet, des études
recevabilité d'un [avis de	projet], permet de
dépend l'évaluation d'un	projet de dragage

Pour valider des regroupements, des contextes plus ou moins grands peuvent être affichés. Quant aux procédures de classification automatique, elles nous semblent peu utiles à ce stade-ci de notre expérimentation.

Nous avons vu que les termes inscrivent une instance particulière d'un concept dans une trame discursive donnée. Cette instance est le sujet ou l'objet d'une action, le résultat d'une transformation, elle est dotée de circonstances

29 Cette définition de l'induction s'inspire de la description aristotélicienne (*Physicorum* I, 184a)

décrivant le temps, le lieu, la condition, la conséquence, la cause, et ainsi de suite:

> Les circonstances disent la multiplicité irréductible à l'unité: non pas en nombre seulement, mais en site, forme, temps, couleur ou nuance, matière, phase, voisinage... contingence.[30]

Dans un premier temps, les relations syntaxiques autres que celles de la détermination sont écartées. Les configurations nominales associées aux termes (soulignées dans l'exemple plus haut) sont recherchées pour être regroupées, organisées et condensées. On appelle agrégation l'opération par laquelle les caractéristiques ou traits sont adjointes aux concepts. Il est à noter que les traits ne renvoient pas à des concepts, mais constituent des contraintes à leur extension. Parfois les traits sont explicites et se présentent dans les textes sous la forme {trait + de + concept}: «la longueur du quai». Cependant, dans la plupart des cas, les traits sont implicites, l'analyste, suite à l'examen de l'ensemble des termes et de leur contexte immédiat, doit en assigner un petit nombre. C'est notamment le cas des termes composés qui peuvent à ce stade-ci être décomposés. Vous trouverez à la fin du texte un exemple d'agrégation par décomposition des termes.

Cette décomposition est discrétionnaire; elle doit être opérée en fonction des unités cognitives prévues. Ainsi, par exemple, dans certains cas, le terme «étude d'impact» constituera l'identificateur d'un concept; alors que dans d'autres, «impact» sera une valeur du type d'«étude». Quant à elles, les formes adjectivales présentes dans les contextes dépistés font apparaître les quantifications (numérales, les cardinales et les ordinales) et les échelles qui positionnent virtuellement les autres valeurs qualitatives ou quantitatives possibles (par exemple : froid, tiède, chaud, brûlant, bouillant, etc.). Par ailleurs la mise en concept à partir d'un grand nombre de termes demandera une classification intermédiaire au moyen d'une grille de codification par domaine ou tout autre critère pertinent.

A l'instar des catégories, il ne semble pas y avoir de traits fondamentaux et universels qui s'appliqueraient avec profit à toutes les situations. Leur sélection est d'une part fonction du but spécifique poursuivi et d'autre part le résultat d'a priori sémantiques et d'un consensus. Les contextes sont classifiés par similitude et différence en des groupes homogènes et mutuellement exclusifs que les traits viennent étiqueter. Les principes de classification doivent être clairs, connus et cohérents; les niveaux d'abstraction explicités. Il ne faudra pas s'étonner de l'hétérogénéité des traits quant à leur nombre et à leur nature. Il ne faudra pas s'étonner non plus si cette démarche ne débouche pas sur un modèle global sans rupture ou contradiction. Le découpage de la réalité en catégories étanches est difficile; il semble assujetti à la deuxième loi de la thermodynamique qui veut que lorsqu'en quelque part on fait croître l'ordre, le désordre croît ailleurs un peu plus:

> As far as we can detect, the second law of thermodynamics is ultimately inexorable, but allows various fluctuations. "Order" can

---

30 M. Serres, *Les cinq sens, philosophie des corps mêlés -1*, Paris Grasset 1985, p.332.

> rise somewhere if "disorder" rises somewhere else a little more. (...) The second law of thermodynamics can be taken to imply the epistemological principle that indeterminacy is the base state of the universe. Indeterminacy can be constrained, but such constraints cannot last indefinately.[31]

Etant donné ces considérations épistémologiques, l'objectif visé ne devrait pas être de construire un modèle conceptuel complet, mais plutôt, là où cela s'avère possible, d'isoler une série d'îlots structurés de connaissance dans les productions discursives d'un domaine.

CONCLUSION

Le passage des termes du texte aux concepts, dans le cadre d'une analyse des schématisations, est constitué d'une suite d'opérations: une description morpho-syntaxique, le dépistage et la validation de termes pertinents, la structuration des concepts par une classification des contextes dépistés. Les concepts reconstitués, il reste à les inscrire dans le réseau d'associations pressenti. S'il y a lieu, une hiérarchie cognitive pourra être structurée selon certaines relations dépistées dans les contextes d'apparition des termes.

Afin de compléter cette méthodologie pour un transfert d'expertise à partir de l'analyse de textes par ordinateur, il nous reste à procéder à l'analyse des ancrages verbaux qui permette de dépister les transitions d'état (opérations) définies sur les concepts. Suite à un examen des séquences où apparaissent les concepts, ceux-ci seront insérés dans des transitions d'état telles la modification, l'accroissement, l'intervention, etc. Leurs flexions et leurs contextes fournissent la modulation (actif, passif, nécessaire, facultatif, etc.), la localisation et la temporalité de l'énonciation du processus en cours. Une attention particulière doit être accordée aux connecteurs, dont voici une liste partielle: conjonctions, concessions, restrictions, transitions, etc. Cette étape pose le problème de la transformation de la structure de cas à celle des règles d'inférences et de la structuration de ces dernières pour former une expertise.

Rappelons en terminant que plutôt que de mettre au point un logiciel offrant un ensemble de fonctions permettant d'appliquer de façon déterministe un modèle théorique donné, nous avons privilégié une approche composite interactive où la dimension heuristique prime. La prise en charge par l'organisation de la connaissance contenue dans ses textes pour la réinjecter dans sa pratique, notamment dans la production d'autres textes, nous semble aussi importante sinon plus que la réalisation du système à base de connaissances ayant motivé le recours aux textes.

---

31 R. de Beaugrande, "Systemic versus contextual aspects of terminology", *Proceedings International Congress on Terminology and knowledge Engineering, Trier (RFA)*, Frankfurt/M, INDEKS Verlag, 1988, pp. 11-12.

analyse_comparative
analyse_d'_impact
analyse_des_sédiments
analyse_des_tendances_sociales
analyse_détaillée
analyse_de_décisions
analyse_de_documents
analyse_de_projets
analyse_de_recevabilité
analyse_de_risques_d'_accident
analyse_du_comportement
analyse_du_contenu
analyse_environnementale
analyse_prospective
analyse_qualitative
analyse_quantitative
analyse_stratégique
analyse_superficielle
analyse_systémique