ACTE : l'ingénierie cognitive et textuelle pour l'indexation hypertextuelle

François Daoust, Luc Dupuy et Louis-Claude Paquin

Centre d'Analyse de Textes par Ordinateur
Université du Québec à Montréal
Case postale 8888, Succursale A
Montréal, P.Q.
Canada H3C 3P8
Tél.: (514) 987-8256
Fax.: (514) 987-8538
S640@UQAM.Bitnet

Introduction

L'hypertexte est une (nouvelle) technologie de l'information qui s'inscrit dans la thématique de notre centre de recherches: l'analyse de textes par ordinateur. Notre contribution consiste d'une part à mettre en relief le caractère "textuel" des documents mis en réseaux et, d'autre part, à proposer une (autre?) approche computationnelle à l'analyse des textes. Cette approche ou «ingénierie» est duelle: textuelle et cognitive. Elle est textuelle en ce que l'analyse prend non seulement en compte l'enchaînement des mots dans les phrases, mais aussi les conditions de production, le projet communicationnel sous-jacent, l'insertion du texte dans la trame des voisins <référence à SACAO>. Elle est cognitive parce que la délimitation de l'extension des concepts et leurs manipulations prime sur la causalité et les modalités linguistiques. La méthodologie d'analyse qui découle de cette approche nous semble offrir le cadre d'un passage plus efficace et viable du texte à l'hypertexte.

De cette approche computationnelle découle un projet d'atelier cognitif et textuel (ACTE) commandité par un consortium inter-ministériel du Gouvernement du Québec <note> Cet atelier est essentiellement le lieu de l'intégration de systèmes (experts) à base de connaissance et d'un "concordancier". L'objectif poursuivi est l'accomplissement efficace en temps réel et sur de très larges corpus de textes des opérations fondamentales de l'analyse de texte. Celles-ci sont limitées: le tri, l'étiquettage, le filtrage (pattern matching) et le stockage. Ces opérations fondamentales peuvent être combinées en des opérations

de plus haut niveau et contrôlées par des règles d'inférences. Pour le passage du texte à l'hypertexte l'opération importante est l'établissement de relations multiples mais sélectives entre n'importe quel terme ou segment et n'importe quel autre terme ou segment, à l'intérieur comme à l'extérieur de l'espace textuel, à partir de n'importe quel point du texte analysé. Pour être pertiente, la sélection des relations doit être faite à partir d'un modèle riche du texte.

Cette contribution vise à montrer à partir des caractéristiques de l'hypertextualité et de celles de la textualité que le passage du texte à l'hypertexte doit être inséré dans une démarche "analytique". Le cadre d'analyse de textes proposé pour l'indexation hypertextuelle est la logique naturelle qui s'applique aux schématisations. ACTE est ensuite présenté comme un environnement qui permet d'opérationnaliser ce type d'analyse.

L'hypertextualité

Dans la perspective de la documentation logicielle, hypertexte (Barret, E. 1988,1989; McCarty, W.,1988) est défini comme l'ensemble des ressources documentaires “on-line” auxquelles il est possible d’accéder de multiples façons. Le stockage et l'exploitation de documents structurés en réseau de références s'est généralisé et répandu. Le noeuds du réseau sont des segments de textes à géométrie variable; leurs liens sont les chemins possible d’exploration. Ce dispositif dégage le texte de ses contraintes linéaire et séquentielle imposées par le format du support “livre” pour l'inscrire dans un espace multidimentionnel où la recherche n'est plus filtrage mais navigation. L'hypertexte séduit parce qu'il va à l'encontre de la thèse skinérienne de la formation aux activités documentaires: l’utilisation de la documentation sous format textuel qui nous condamne à des parcours «labyrintesques» <note>. Le modèle proposé est plutôt celui de la co-opération: le parcours et la production de liens. Cette architecture permet une production (écriture) et une consommation.(lecture) de textes non plus linéaires mais associatives..Le concept de navigation, typique de l’environnement des SGBD, permet présentement l'accès le plus “intelligent” aux archives ou banques documentaires.

Une première revue de littérature nous montre cependant l’existence d’une opposition importante entre la définition de l’hypertexte comme

stratégie d'écriture et celle où l'on insiste plutôt sur le processus de lecture comme stratégie de navigation dans le contenu des textes. Cette opposition est-elle attribuable aux caractéristiques intrinsèques du médium hypertexte plutôt qu'à des pratiques différentes? En effet, la dimension de la vitesse d'accès à l'information n'est pas le seul critère démarquant les documents de la génération "hypertexte" des documents de la génération "Gutenberg". Si la notion d'hypertexte est prometteuse du point de vue d'une nouvelle structuration de l'espace de lecture/écriture, elle se fonde paradoxalement sur des applications logicielles, certes rapides, mais qui sont loin d'offrir les avantages du format "livre" dans un espace à trois dimensions (je pense ici aux différents PIMs - personal information management systems - qui ont certes leur pertinence mais qui reposent largement sur la notion plutôt limitée de fiche ou d'article de base de donnée). Pour satisfaire les besoins de la lecture comme acte d'interprétation des textes (vision du lecteur "humaniste"), l'hypertexte doit être assujetti à une méthodologie qui respecte les schèmes socio-culturels régissant l'acte de lecture/écriture.

Si la critique que les humanistes (McCarty, W., 1988) font des présupposés "modernistes" (instantanéité et multi-accès) est pertinente (le livre n'est jamais lu strictement de façon linéaire; l'acte narratologique est lui aussi un dialogisme), il reste qu'elle met justement en lumière une des forces de l'hypertextualité : l'accès aux documents hypertextuels est supporté précisément par des procédures d'accès plus efficaces que le recours aux tables onomastiques ou index livresques. Les entités identifiées dans la structure de définition de l'espace textuel deviennent les portes d'entrée pour avoir accès au texte. D'une certaine manière, le modèle entités-relations (re)donne au données textuelles leur troisième dimension, voire leur nième dimension. À la différence des index référentiels (onomastiques, conceptuels) fondés largement sur la notion de concordance (ou adjacences explicitement réalisées dans le texte) les index hypertextuels reposent sur la notion d'accès transversal aux documents. Cet accès se fait selon une logique de type associative.

Toutefois le modèle hypertexte ne fournit pas de cadre pour "bien" délimiter les noeuds/segments_de textes ou encore pour "bien" établir et classer les liens. Nous proposons le recours au modèle textuel. Les partitions (ou domaines de textes) sont alors définies à partir des points de convergences thématiques, argumentatifs, rhétoriques, etc. Celles-ci peuvent être "intra-textuelles" ou "extra-textuelles". Ces

deux dimensions organisent en quelque sorte les itinéraires textuels; elles structurent la discursivité textuelle. Donc, jusqu'à un certain point, le succès de l'indexation hypertexte repose sur la richesse des relations conceptuelles explicitées dans le discours mais définies pour et dans le domaine de savoir (Foucault, M., 1969). Pour permettre l'explicitation des éléments d'un savoir, hypertextualité doit refléter la structure d'un discours socialement construit et partagé par un ensemble d'utilisateurs et non pas d'un discours issu d'une socialisation solitaire de l'acte d'écriture.

La navigation dans les archives textuelles suppose la clarification/explicitation d'un projet spécifique de lecture/écriture partagé par un groupe donné de personnes. L'indexation hypertextuelle nécessite d'une part l'analyse des textes et de l'autre la délimitation. le réseau social de l'acte de lecture.

l'hypertexte pose le problème fondamental d'une organisation sémantico-pragmatique de l'information; on rétablit en fait les causes d'un malaise autour de l'insaisissable de l'information que Wiener définissait comme n'étant ni matière ni énergie. L'information est un rapport problématologique qui s'est matérialisé sous forme de texte; c'est en fait le principe même du cycle de développement et de vue du texte, cette idée de mise en forme définie non comme état mais comme processus (reprendre l'étymologie in-formation);

La textualité

Dans les grands organismes, dont ceux de l'appareil gouvernemental, la production textuelle - faite de rapports, de directives, de projets, de correspondance, etc. - connaît un volume grandissant qui rend de plus en plus difficile leur exploitation. Ainsi, les "travailleurs du texte", chercheurs, gestionnaires, décideurs, etc. dont l'analyse de données textuelles (lecture d'archives et de documents, rédaction de rapports, etc.) constitue l'activité principale, sont débordés par une masse de documents qu'ils doivent analyser en fonction d'objectifs qui leur sont spécifiques: accumulation de faits, d'événements ou de connaissances, interprétation, élaboration de stratégies, prise de décision, etc. Le texte prend de multiples formes en fonction du projet communicationnel qui lui est assigné: études, rapports, directives,

décrets, réponse en format libre à des questionnaires, retranscription d'entrevues, etc.

Le texte n'est pas un univers de données discrètes ou numériques est d'abord et avant tout discours et acte de langage.sur des savoir (Foucault, M., 1969). Le texte est, au-delà de son apparence première, un objet stratifié qui ne se réduit pas plus à l'ensemble des mots qui le composent qu'aux relations réunissant ceux-ci en énoncés ou encore à un contenu pur et simple. 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 compréhension de la parole relève du système phonologique; celle du texte repose sur le système typographique. 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 (complémentaires?): 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 définis. Le système sémantique est pensé comme une sorte de calcul sur les propriétés lexicales des mots et leur position morpho-syntaxique dans un segment donné. Il est à noter que la complexité graduelle est du à l'exposé. 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 deux ou plus de catégorisation de surface potentiellement contradictoire sans une description de la structure profonde de l'énoncé. Telle est la conception linguistique du texte.

Au niveau informatique nous disposons depuis longtemps de systèmes fabriquant des index (tri et comptage de formes) et des concordances (extraction par patron de fouille de mots qui apparaîssent dans un contexte réduit) <note sur SATO et sur JEUDEMO et celui de Oxford> . Ces opérations ont rendu possible une analyse de type quantitative qui a essuyé des critiques justifiées en raison du peu de sensibilité linguistique de la définition des formes à compter: «suite de caractères délimitée par un séparateur (le blanc)».. On a très tôt pensé que les dénombrements analysés devraient être préalablement syntaxiquement décrits pour que tant la catégorisation que les résultats obtenus tiennent compte du contexte des populations de mots filtrées. Il y a cependant peu de programmes qui effectuent la description arborescente de la syntaxe réalisée dans des énoncés parcourus ou parsés. Après plus de trente ans de recherches dans le

domaine du traitement automatique de la langue naturelle, J. Sowa un membre de l'équipe de recherche en systèmes de IBM affirme que les succès sur un domaine restreint et leur échec lorsque le domaine est sans restrcictions s'explique par la nature fondamentale du langage. Une grammaire volumineuse se suffit pas à étendre la couverture d'un petit système en un traitement de la langue naturelle sans aucune restriction[1]. Depuis plus de douze ans, P. Plante du Centre d'ATO s'intéresse aux problèmes du parsage; il collabore présentement à la confection d'un analyseur lexical et syntaxique du français (ALSF) qui est basé sur un modèle théorique qui lui permet souplesse et transparence pour résoudre les principaux problèmes reliés au parsage.<note sur les travaux de P. Plante>.

Il n'a jusqu'à présent été question que des micro-structures du texte. Le texte est analysé sous l'angle de la distribution de fréquences d'apparition des mots désignant des concepts dans diverses parties d'un corpus de textes. Avec le premier type d'outil, les fréquences retenues sont brutes; dans le second, elles sont qualifiées à la syntaxe, c'est-à-dire elles gardent la trace de la structure et du contenu de leur contexte d'énonciation. Une analyse de texte, telle que pratiquée dans les sciences humaines demande des niveaux de description supplémentaires, proprement textuels appelées macro-structures <note>. Parmi ces derniers mentionnons les figures de style ou de pensées, la logique du réseau d'argumentation, l'environnement communicationnel, la thématique, etc. Ces systèmes s'appliquent à des unités d'une autre nature: à géométrie variable tels la phrase, le paragraphe ou encore tout autre découpage arbitraire justifié par une grille ou d'autres critères. Leur description ne semble pas uniquement dépendre de la structure arborescente de la description lexico-syntaxique. Dans la plupart des cas, les indices ne sont pas assez nombreux pour qu'une analyse puisse se faire. Par contre, lorsqu'il y a des indices, un filtrage basé sur des séquences de patrons morphologiques semble suffisant. Une connaissance du cadre formel propre au type de texte est de plus requise. Est-ce une lettre, un mémo, une documentation, un règlement, un article de loi, le résumé d'un texte, etc?

[1] "... 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 are not sufficient to scale up a small system to an unrestricted natural language processor"
in SOWA, J. F. "Multi-Domain Semantic Theory" copie de travail fourni par l'auteur lors d'une conférence à Montréal et daté du 28 novembre 1988.

Qui plus est, non seulement une connaissance de l'univers particulier du texte est requise, mais le lecture doit être informée des conventions sociales qui ont présidées à l'émergence du texte. Cette dimension, appelée intertextualité, situe le texte à dé-coder au-delà les systèmes linguistiques. La seule façon de contourner l'incertitude quant aux indices nécessaires et fournir quant même un cadre computationnel utile, c'est d'inclure le lecteur dans le processus de la fabrication du sens. Cette intuition est confirmée autant par les dernières théories de la psychologie <note> que de la sociologie <note> affirmant que les textes n'ont pas un sens univoque; le sens est plutôt construit par le lecteur au travers ses structures cognitives et culturelles résultantes de sa socialisation. Dans cette perspective, l'expertise de la lecture doit être prise en compte par le système. Pour ce faire sans introduire de biais, les techniques d'ingénierie cognitive utilisées pour construire les sytèmes (experts) à base de connaissance: l'entrevue, l'analyse de protocoles (verbalisations durant l'accomplissement d'une action); ou d'interruption (questionnement).

Les systèmes à base de connaissance résolvent des problèmes en parcourant une chaîne d'informations générée à partir des faits de l'espace de problème (base de faits). Cette façon de faire nous a inspiré un renversement d'approche computationnelle, le passage d'une stratégie de parsage déterministe pour une sémantique procédurale. Par ce terme nous entendons la reconstruction de la signification au moyen d'une chaîne inférentielle dirigée par un but particulier: l'hypothèse de lecture. Cette chaîne est faite par le déclenchement de règles d'interprétation ou de recatégorisation des segments à partir des faits dont on dispose. D'un côté, la configuration d'indices relevés dans les descriptions disponibles (morphologiques, syntaxiques, sémantiques) et les heuristiques du lecteur (sens commun). Cette stratégie présente l'attrait de respecter le caractère hautement associatif des propriétés associées aux unités lexicales.

A base de connaissance ou non, l'analyse de texte par ordinateur doit être encadrée par une méthodologie fiable et appropriée. On doit y retrouver minimalement les étapes suivantes: la formulation d'hypothèses, la description des documents, l'extraction des données et l'analyse proprement dite. La plupart du temps l'analyse de texte est un processus cyclique où les résultats d'une précédente analyse participent à la reformulation des hypothèses.

Du texte à l'hypertexte

La tâche de construire un hypertexte à partir d'un corpus de textes se nomme l'indexation hypertextuelle. Elle comporte deux opérations fondamentales: découper le texte en segments et établir des liens entre ceux-ci. Le découpage des segments destinés à devenir des unités hypertextuelles, appelées noeuds de base, doit répondre aux deux critères suivants: d'une part leur format doit pouvoir s'afficher à l'écran et/ou être imprimé sur une page; d'autre part il doit être cohérent.et complet c'est-à-dire compréhensible par lui-même. Une bonne façon de respecter le contenu du texte serait de faire de chacun des paragraphes du texte un noeud hypertextuel. Cependant comme les paragraphes sont contigus, une économie de dénotation est établie, les concepts en présence dans l'énoncé sont remplacés par des mots-outils (anaphores) ou des concepts associés (re-catégorisation.). Il n'existe pas, à notre connaissance, de programmes d'analyse syntaxique qui puisse dé-anaphoriser le texte. Les noeuds doivent donc être revus par un lecteur/auteur qui "intervient" dans le texte "officiel". L'intervention est double: rétablir la référence au contexte en désambiguant l'anaphorisation et en restituant la chaîne de re-catégorisation; indiquer pour la recherche la place du segment dans la structure du document par l'assigantion d'une étiquette unique. Il est à noter que le principe-même de découpage implique un démembrement des macro-structure du texte (cf. supra). Pour pallier à cette perte, nous proposons le recours à l'analyse de textes.

L'autre opération de l'indexation hypertextuelle est le liage des noeuds. Un lien est un chemin possible d'exploration entre un noeud de départ et un noeud d'arrivée. A chacun de ces types, une réthorique, c'est-à-dire un ensemble de règles ou critères régissant soit l'émission d'un lien, soit sa réception. Les critères d'association des liens hypertextuels ne reposent plus que sur leur contiguité. Les noeuds sont assemblés en "modules"; il s'agit d'une activité de classification.qui demande de l'évaluation, donc difficilement transposable en algorithme. La typologie des liens hypertextuels n'est pas figée: deux classes tendent à s'établir: les liens par la référence, par la hiérarchie. Les liens par la référence sont explicites ou implicites c'est-à-dire que dans le premier cas il existe dans le texte lui-même des indications facilitant leur repérage automatique alors que dans le second il n'y en a pas. Les références sont explicites soit typographiquement (par ex.: entourés de parenthèses «(Paquin 1988, p. 37) soit par des suites de caractères isolés des autres mots par un

alinéa (par ex.: l'article de la loi de la Protection de l'environnement «Q q a 4»), soit au moyen d'expressions (par ex.: «comme le dit Minsky:».

Les références explicites peuvent avoir pour destination: soi-même (par ex.: une autre partie du texte où le concept employé est défini); d'autres textes à l'intérieur du corpus. Si les références sont précédées du texte, il s'agit d'une citation. Les références implicites ne présentent pas comme tel de marques textuelles, ce qui les rend très difficle à dépister automatiquement. Leur dépistage requiert le jugement et l'intuition d'un lecteur/auteur hypertextuel. A titre d'illustration deux types de références implicites. Au niveau micro-textuel le regroupement des synonymes: la tâche consiste à reconnaître, à faire valider et à documenter la similitude observée, à sélectionner un terme qui sera préféré et à le relier aux autres termes qui se trouvent à être non préférés. Au niveau macro-textuel le discours indirect libre: une citation sans référence. Le dépistage et le marquage du discours indirect libre apporte à l'hypertexte une dimension importante. Les passages du texte qui sont rapportés sont re-liés aux textes-sources.

Les liens hiérarchiques servent à rattacher les concepts manipulés dans le discours à une structure conceptuelle, la plupart du temps hiérarchique. Ainsi un objet se trouve rattaché à son générique (sous-classe) et cette dernière à l'instance (classe). Ce type de lien est pertient dans les domaines déjà structurés: la description de méchanismes par exemple. Par ailleurs, il y a des domaines où le concept de hiérarchie n'est pas pertinent: les décrets tombent souvent dans cette catégorie. La description d'autres domaines demandera l'entrecroisement de plusieurs hiérarchies <trouver un exemple>. Dans les textes, les liens hiérarchiques établis entre les concept sont, la plupart du temps, dénoncés par des marques linguistiques, telles: est un (le moineau EST UN oiseau), partie (le fer fait partie de l'acier), etc. Dans tous les cas, il s'agit d'une structure externe au texte lui-même que l'hypertextualité permet de rajouter. D'autres peuvent être établis à la discrétion du lecteur/auteur; on les dira procéduraux lorsque leur destination sera déterminée par l'exécution d'une fonction avec des paramètres donnés.

Le lecteur/auteur qui transforme le texte en l'hypertexte, revise la formulation du texte des noeuds, lie les textes entre eux en opérationnalisant le jeu des références croisées, rattache les concepts en présence à leurs synonymes, à leur générique ou encore à leur définition. De plus, celui-ci peut annoter les textes, c'est-à-dire lier un

commentaire à un passage du texte. Il va sans dire que le passage du textuel à l'hypertextuel requiert une planification basée sur une étude be besoin. Cette activité exige de la part du lecteur/auteur outre des compétences en analyse de textes, les connaissances suivantes: les dimensions critiques de l'espace d'idée concernée; les caractéristiques qui distinguent une idée d'une autre et enfin les schèmes de nomination appropriés. De plus, le réseau construit doit être validé par une représentation graphique.

La logique naturelle comme cadre d'analyse

Le concept d'hypertexte met en évidence le fait fondamental que tout texte est un parcours organisé, une suite "régie" d'éléments qui trouvent leur identité non seulement à l'intérieur de ce parcours, le contexte, mais qui sont également déterminés par le co-texte, l'ensemble de tous les textes avoisinnants. Avoisinnant car c'est d'un espace qu'il s'agit. La lecture et l'écriture sont toujours topiques, localisés dans un espace textuel en fonction de la structure pressentie. Cet idée d'espace textuel est très près de celle d'un réseau hypertextuel; la métaphore spatiale constitue un point de contact, de passage de l'un à l'autre. Notre modèle du texte (cf. supra) présente le texte comme un révélateur des rapports sociaux de production de l'information; le cadre de notre analyse doit donc mettre à jour les transactions et les tractations au milieu desquelles se construisent les concepts au travers un corpus.

Les logiques formelles[2] font généralement abstraction de la nature des objets qu'elles manipulent. Dans les situations discursives, les objets manipulés ne sont jamais quelconques, ils sont toujours spécifiés (mis en contexte) à un certain degré. Les concepts sont efficacement représentés comme des objets symboliques dotés de plusieurs variables dont les valeurs ne sont pas boléennes mais scalaires. Par expemple le concept de température pourrait être vu comme un doublet de variables. La première variable serait la mesure en degré centigrades avec comme valeur un nombre avec 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 boléenne (par oui ou par non), mais par la sélection d'un élément dans un

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

ensemble. La notion ensembliste d'élément discrets n'est pas adéquate pour décrire l'attribution d'une position sur une échelle parce qu'il s'agit de la graduation coninue d'une qualité. On appelle classe méréologique le complexe de relations entre un tout et ses parties, entre les parties de parties; voici un exemple simple: la main et ses doigts; chacun des doigts n'est pas tant une partie de la main que son prolongement. L'échelle est une espèce très contrainte de classe méréonomique.

Dans les textes les concepts ne sont pas manipulés à des fins de démonstration[3]. mais de schématisation. Les schématisations sont des opérations discursives structurant des objets cognitifs et les articulant dans l'espace d'un savoir. Ainsi nous avons recours à la logique naturelle parce qu'elle s'intéresse à de telles opérations mises en jeu par les locuteurs impliqués dans une pratique discursive. 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.[4].

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.).

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

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

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

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[5], l'analyse de texte consiste à utiliser ce métalangage pour isoler, par leur configuration et leur récurrence, des noyaux conceptuels. Suite à une classification des contextes, ces noyaux sont transformés en concepts multi-facettes et hiérarchisés. Suite à un examen des séquences où apparaîssent les concepts-clé, les concepts sont insérés des transitions d'états telles la modification, l'accroissement, l'intervention, etc. Ainsi, une analyse conforme aux principes de la logique naturelle s'intéresse aux entités nominales et aux entités verbales.

L'analyse des entités nominales d'un corpus de textes permet le dépistage des termes et leur structuration en concepts. L'effet de référence au réel dans un discours donné est tributaire de formes nominales qui consolident d'autres formes nominales en classes-objets. Ainsi, les marques référentielles[6] proviennent des configurations d'énoncés et des transformations linéaires engendrant la dynamique textuelle. Ces marques sont identifiables linguistiquement à partir des stratégies discursives[7] qui confèrent à certaines formes nominales une fonction de régie textuelle. Une fois que, parmi tous les substantifs, les concepts pertinents ont été retenus, les configurations nominales, appelés ingrédients, qui leur sont associés sont recherchées. Ainsi, par exemple pour le substantif "projet" on aura des configurations telles, l'assujettissement d'un projet, la pertinence d'un projet, etc. Les formes adjectivales présentes dans les contextes dépistés font apparaître les quantifications et les échelles

[5] 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.

[6] 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.

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

argumentatives qui positionnent virtuellement les autres valeurs qualitatives ou quantitatives possibles.

L'analyse des groupes verbaux permet le dépistage transitions d'état (opérations) définies sur les concepts. Les verbes jugés pertients au domaine servent à sélectionner des segments. Leurs flexions et leur contexte fournissent la modulation (actif, passif, nécessaire, facultatif, etc.), la localisation et la temporalité du processus en cours. La classification des segments est une opération qui consiste d'abord à examiner un large contexte des verbes jugés représentatifs, puis à délimiter le segment, selon un critère d'homogénéité, des bornes inférieure et supérieures sont assignées parfois arbitrairement et enfin à caractériser la structure du segment. Les énoncés peuvent être structurés en segments de textes significatifs de manière plus ou moins complexe (configurations à 1, 2, ... , n énoncés). Voici une liste partielle des connecteurs: conjonctions, concessions, restrictions, transitions, etc.

Au lieu d'une description arborescente de chacune des phrases qui s'avère lourde et difficile à valoriser l'analyse par la logique naturelle produit des inventaires, des classifications ou encore des partitions du texte. Cette approche à l'analyse de texte dans le cadre de la construction d'un hypertexte permettrait une indexation des itinéraires de lecture selon un classement méréologique s'approchant de la pensée naturelle. Ces itinéraires instancient en quelque sorte une forme de représentation socio-cognitive en offrant la dynamique d'un espace topologique.
<idée de géométrie variable du processus de la séméiosis (Veron)>.

Le passage du texte à l'hypertexte, dans le cadre d'une analyse des schématisations, est constitué d'une suite d'opérations: la description morpho-syntaxique, l'extraction des termes pertinents, la classification des contextes dépistés, délimitation du segment, étiquetage des liens. Plutôt que de dessiner et développer un logiciel offrant un ensemble de fonctions permettant d'appliquer un modèle théorique donné, nous avons privilégié une approche interactive "atelier logiciel" où la dimension heuristique prime. Nous croyons fermement que non seulement la validation, mais aussi la gouverne (contrôle) des opérations doit être laissée aux lecteurs/auteurs chargés de la construction de l'hypertexte. Notre contribution consiste à développer des logiciels utiles pour analyser les textes, à assister les lecteurs/auteurs dans leur démarche avec nos outils. et à adapter ces applications à leurs besoins spécifiques dans la mesure du possible. Ces

trois temps caractérisent une démarche de type recherche-action: les outils sont en évolution constante, la méthodologie est redéfinie par les nouveaux contextes d'application et enfin les buts de la recherches sont dictés en permanence par un besoin concret dans les organisations gouvernementales commenditaires.

ACTE

L'idée de développer un environnement computationnel intégré pour effectuer de l'analyse de textes par ordinateur est dans l'air depuis 1986 Le projet SACAO[8] visait à définir un Système d'Analyse de Contenu Assisté par Ordinateur avec les contraintes suivantes: convivialité de l'interface, précision, régularité, transparence et validité des procédures. Il s'en est suivi une réflexion sur les fonctionnalités de base de l'analyse de textes par ordinateur et leur combinaison en des opérations complexes. Ces fonctionnalités qui sont en nombre limité : le tri, l'étiquettage, le filtrage, le stockage doivent cependant être effectuées très rapidement et sur de très très grandes masses de textes. Pour qu'un outil interactif soit intéressant, il faut une grande rapidité d'exécution entre les interactions, sinon perte d'intérêt. Pour que les indications soient repérables, l'analyse doit être menée sur un vaste échantillonnage représentatif de textes: le vocabulaire doit être stable <note sur la loi de Zipft>.

Les opérations complexes doivent être signifiées à l'aide d'un formalisme. Ce formalisme doit être associé à une structure de contrôle sur le déroulement de ces opérations. Nous avons retenu l'approche système expert et non pas un (autre) langage ad hoc, pourquoi? D'une part parce que les règles d'inférences en tant que moyen de modélisation tendent à se répandre de façon constante. Comme aucun apprentissage de langage de programmation n'est requis, il facilite le transfert de technologie qui doit s'effectuer à l'usager-final afin que celui-ci soit en mesure d'être son propre développeur. D'autre

[8] Ce projet a été initié en 1986. Une subvention de type "Action spontanée" a été attribué pour deux ans (1987-1989) par l'organisme provincial FCAR. Le responsable était J. Duchastel, professeur de sociologie et directeur du Centre d'ATO.
Le projet SACAO a été présenté au congrès du SCACC tenu Trier (RFA) en 1987;
voir aussi Daoust F., Duchastel J., Dupuy L. "Système d'analyse de contenu assistée par ordinateur" Actes du Colloque La description des langues naturelles en vue d'applications linguistiques, Centre international de recherche sur le bilinguisme, Québec, 1989, 197-210.

part, en autant que le moteur d'inférences offre des performances satisfaisantes, l'approche par les règles permet un développement modulaire et par tentative (trial and error). L'assemblage de règles d'inférences construites en pointant avec la souris dans les options d'un menu ou dans le dictionnaire de connaissance permet la construction graduelle d'analyseurs plus sophistiqués et plus spécifiques basés sur des stratégies de contrôle sensibles au contexte. Ces analyseurs pourront par la suite être incorporés en exécutable (run-time).

Le devis de ACTE, un acronyme pour Atelier Cognitif et Textuel, a été conçu en 1989, il a été accepté par un consortium de ministères du Gouvernement du Québec; il est entré en phase développement depuis février 1990. L'originalité de ACTE[9], en tant qu'approche computationnelle à l'analyse de texte, repose sur l'intégration de deux logiciels: SATO[10] une base de donnée textuelle et un ensemble de fontionnalités pour l'analyse de textes et D_expert[11] un générateur de systèmes à base de connaissances. le troisième élément est une série de bases de données lexicales. La principale comportent les catégories morphologiques[12], une autre des locutions, une autre encore une grille de catégories sociologiques[13], etc. L'interaction avec le systéme est standardisé: menus déroulants emboîtés et souris, les messages et requêtes explicités on-line, etc. ACTE est aussi conçu comme un atelier ouvert. Ainsi, il pourra communiquer avec des serveurs d'information, intégrer l'information obtenue au traitement en cours d'une part et de

[9] This software is now under development within our research center. It should be deliver to a provincial governmental consortium by 1992. For an extensive description see: L. C. Paquin, L. Dupuy and F. Daoust "ACTE: a workbench for knowledge engineering and textual data analysis in the social sciences" Proceedings of the Fourth International Conference on Symbolic and Logical Computing (ICEBOL4), Madison 1989, Dakota State University, 122-136.

[10] F. Daoust, Système d'analyse de textes par ordinateur (SATO), version 3.5, 1989, Centre d'ATO, UQAM.

[11] L. C. Paquin, D_expert, (formerly Déredec-EXPERT) 1989 , Centre d'ATO, UQAM.

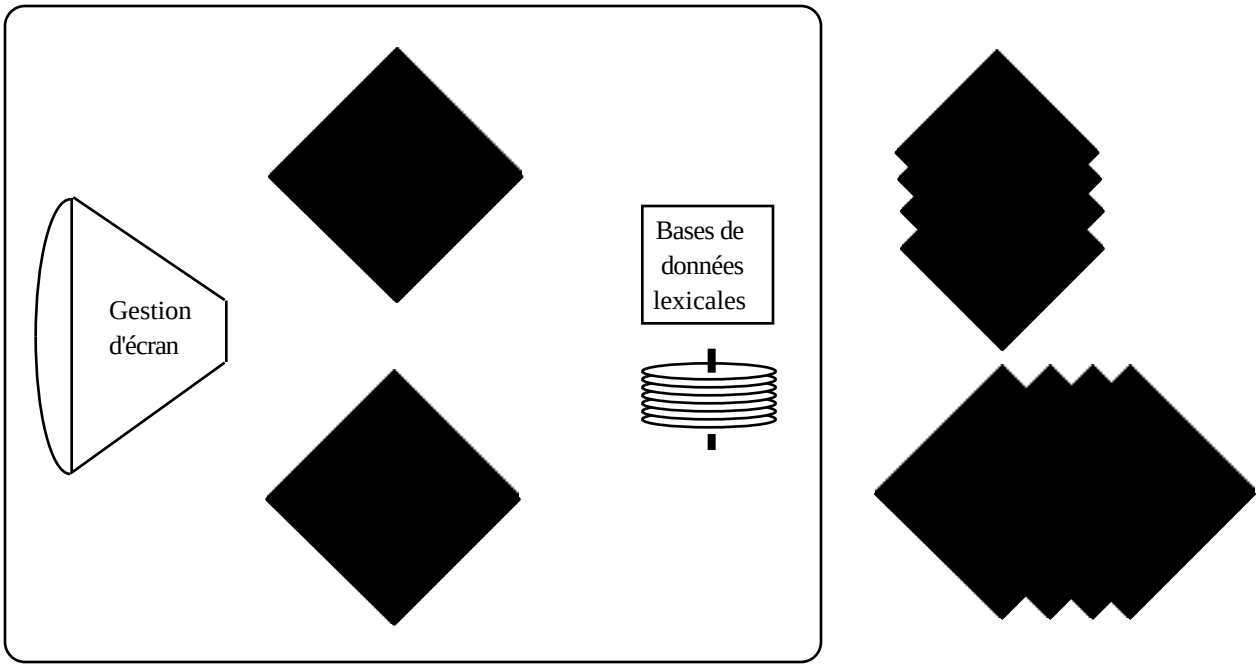
[12] L. Dupuy, <donner le nombre de formes> 1989, Centre d'ATO, UQAM.

[13] G. Bourque, J. Duchastel Projet Duplessis une hiérarchie de 100 catégories projetés sur 8,500 formes:

économie
question de l'État
institutions sociales
dimensions anthropologiques fondamentales
valeurs traditionnelles, typiques, existentielles, discipline personnelle
contrôle social
catégories évaluatives dans l'analyse: quantités, qualités, états d'esprit

G. Bourque and J. Duchastel, Restons traditionnels et progressifs, Montréal, 1988, Boréal: 78-80.

l'autre accueillir des descriptions linguistiques complexes. Voici un schéma des modules décrits:



L'intégration des fonctionnalités permet l'extension de l'espace de recherche (base de faits) usuel au texte dans son plein format. La prémisse des règles d'inférences peuvent filtrer des configurations complexes d'énoncés. Les registre des actions est étendu à la manipulation des textes: la catégorisation, le filtrage, la comparaison de lexiques ou sous-textes. Les résultats de ces actions des valeurs numériques ou symboliques, des sous-lexiques, des segments de texte (concordance), peuvent devenir des faits. C'est ainsi que s'effectue une chaîne interprétative (cf. supra, p.5). Plus profondément, le recours à la technique des systèmes experts change la façon dont les analyses sont menées. Dans le domaine des analyseurs syntaxiques (parseurs), l'utilisation de PROLOG, un moteur d'inférence avec variables dont le principe est l'unification, a rétro-agit sur la théorie du parsage voire la théorie syntaxique; on parle maintenant de «grammaire d'unification». L'analyse n'est plus conçue comme un algorithme avec des stratégies de contrôle peut-être souples (sensibilité aux contexte), mais toujours déterministes, modulaire certes, mais monolithique; le contrôle est assuré par les faits dont la présence fait déclencher des règles d'inférences dont les actions consistent en des interventions textuelles ou l'ajout de faits qui, à leur tour feront déclencher d'autres règles.

SATO produit des lexiques et des concordances à partir d'une chaîne comportant jusqu'à 5 patrons de fouilles complexes sur de larges corpus de textes <formuler un patron de fouille et donner le temps de réponses sur 4 megs de texte avec un 80386)>. Il est suffisamment performant pour permettre une interrogation interactive. SATO est surtout doté de fonctionnalités permettant non seulement l'annotation de mots ou de segments de textes, mais aussi l'annotation du dictionnaire des mots. L'inscription de propriétés (symboliques ou numériques) autant au lexique qu'au texte qui peuvent ensuite être questionnées, permet le dépassement du mot à mot du texte, des formes différentes d'un même mot. Les propriétés ont un type: numérique (divers types de fréquences ou pondérations) ou symbolique (étiquettage en langue naturelle) peut être structuré. Les propriétés couvrent les aspects paradigmatiques <expliciter le terme> et les aspects syntagmatiques <expliciter le terme>.des mots. L'usager peut très facilement projeter sur le texte ses propres systèmes de catégories issus dont les hypothèses quant à l'interprétation du texte sont préférablement explicites. Ainsi, les dénombrements pourront être effectués sur les catégories tout autant que sur les mots.

Cette façon de faire amène le lecteur à expliciter les éléments textuels susceptibles d'être porteurs de sens et à arrêter les critères à partir desquels ceux-ci seront retenus et comptabilisés. La neutralité de l'instrument, qui permet la coexistence de plusieurs niveaux d'analyse potentiellement contradictoires, favorise une démarche d'aller-retour entre la constitution de modèles sur les textes et leur validation empirique. Il faut voir en effet, qu'il n'y a pas dans ACTE de projection déterministe d'un modèle pré-construit sur le texte. Le savoir sémantique et procédural, qui sera inscrit dans le format de règles d'inférences appartient à l'usager. L'approche privilégiée par l'atelier est donc la mise à jour de l'organisation du texte par l'ajout de descriptions successives du texte en alternance avec l'exploration de résultats provisoires.

La référence au réel.est le fait des nominaux (cf. supra). Le dépistage et la structuration des concepts en vue d'une indexation hypertextuelle, passe donc par une catégorisation morphologique des mots du texte. SATO permet la projection de dictionnaires ou bases de données lexicales; dont celle des parties du discours {nom, pronom, adjectif, verbe, adverbe, préposition, conjonction, etc.}. Sur la base de

configurations morphologiques, le dépistage des concepts et leurs relations sera exhaustif, objectif, reproductible parce qu'indépendant des problématiques définies dans les textes. Ces configurations sont formulées sous forme séquence de patron de fouille.filtrant la co-occurrence de catégories particulières (par ex: traitement de textes {[nom] de [nom]}). Lorsque les patrons de fouilles sont réalisés, SATO permet l'assigantion de valeur à des propriétés. Il en va ainsi pour le blocage des locutions: littéralement ajoute la valeur «lié» à la propriété édition[14].

SATO permet de re-catégoriser un mot ou un segment de texte n'importe où dans le texte ou encore dans le dictionnaire des mots à partir d'un certain point et selon certaines conditions dont les seuil peuvent être exprimés sous forme de règle d'inférences:

> Si <condition> et <condition> ... alors <action> ...
> condition ::= test sur les faits / patron SATO
> action ::= inférence / question / action SATO (fouille, catégorisation)

Le moteur d'inférence en chaînage avant offre une structure de contrôle qui permet le dépassement d'une lecture linéaire, du début à la fin du texte. Une hypothèse est d'abord formulée sous forme de patron de fouille; puis, de multiples analyses approfondies sont menées séquentiellement sur chacun des contextes proches. Chacune des analyse peut mener à une re-catégorisation et est susceptible d'aboutir dans la formulation d'un autre patron de fouille, etc.

La tâches finale du moteur d'inférences sera génération des codes nécessaires pour la conversion en hypertexte. Les principaux standards d'importation des fichiers textes déjà formattés seront appris, c'est-à-dire convertis en forme de règles d'inférences. Ultérieurement une révision pourra se faire par l'éditeur hypertextuel du système utilisé.

[14] Voici la commande SATO:

```
concordance stricte                    **
    $*discours=nom$*édition:lié        **
    $*discours=artprep*édition:lié     **
    $*discours=nom$
```

Conclusion

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

Depuis plus de deux ans nous intervenons au sein de la structure administrative du Gouvernement du Québec[15] en analyse-conseil en ingénierie cognitive. Dans tous les cas, le mandat de superviser la construction d'un système expert, qui constituait le point d'entrée de notre intervention, _ objet d'une entente contractuelle avec livraisons spécifiques _ a été élargie en une intervention cognitive en profondeur: analyse documentaire[16] constitution d'une base de données textuelles, consolidation terminologique: blocage des multi-termes, constitution d'un dictionnaire de concept (index a posteriori), etc. Face à l'intérêt suscité par la démonstration de l'analyse de texte, la demande se fait forte pour des plateforme de livraison du corpus de textes alternatives aux bases de données. Au cours de la prochaîne année, une ou plusieurs implantations hypertextuelles expérimentales seront faites sur des corpus tels: la politque administrative, les lois et règlements relatifs à la loi sur l'impôt, les avertissements agricoles, etc.

---

[15]- Commission des normes du travail du Québec: Système à base de connaissance pour supporter le travail de l'inspecteur-enquêteur. Validation de la normalisation des règlements.

- Ministère du Revenu du Québec: Système expert pour l'assistance à la formation en vérification fiscale. Ce projet intègre à l'ingénierie cognitive du domaine une approche de développement orientée vers l'enseignement intelligemment assistée par ordinateur.

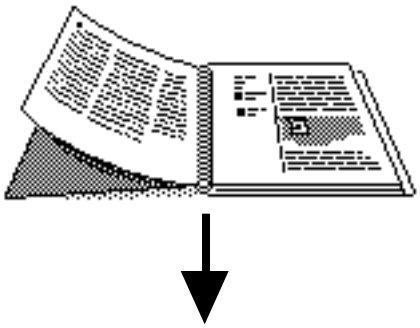
- Secrétariat du Conseil du Trésor du Québec: Système expert d'aide à l'attribution des contrats de services du gouvernement. Ce système permet gestionnaires non-spécialistes de remplir un formulaire de demande respectant les prescriptions de cet aspect de la politique gouvernementale.

- Ministère de l'environnement du Québec: Système Assisté de Gestion des Évaluations Environnementales.

- Ministère de l'agriculture des pêcheries et de l'alimentation du Québec Système d'Aide au Diagnostic Appliqué à la Pathologie en santé animale: systèmes digestif et respiratoire des volailles.

[16]- Les auteurs sont consultants au projet intitulé: De l'indexation et du contrôle du vocabulaire assistés par ordinateur à l'extraction et à la représentation des connaissances: applications à trois corpus ministériels avec les logiciels SATO, D_expert et SECONDE. Rapport de synthèse et guide de procédures. Financé par la Direction des méthodes de gestion des technologies du Ministère des Communications (Gouvernement du Québec), sous la responsabilité de Suzanne Bertrand-Gastaldy.

En dernière analyse, au même titre que les systèmes experts et les bases de données "plein texte" mais avec des modalités différentes que le systéme nous considérons l'hypertexte comme une plateforme d'accès à l'information qui répond à une gamme de besoin dans les organisations où les objets et concepts manipulés sont si complexes que leur documentation fait problème sur les supports séquentiels conventionnels comme le livre imprimé. Voici en terminant un schéma de l'ingénierie textuelle

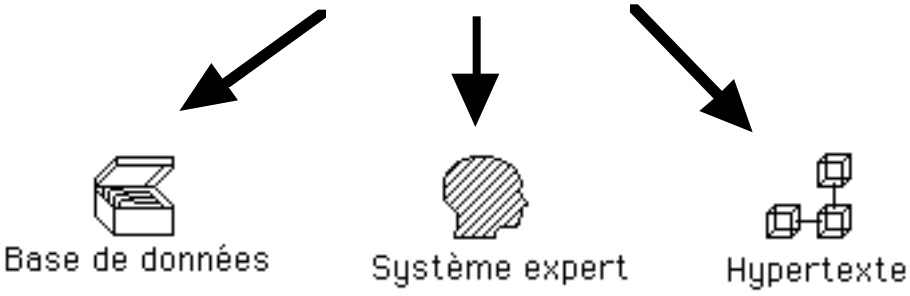


analyse de textes par ordinateur

des constituants
blocage des locutions
distribution lexicale

des relations
séquences nominales
détermination
séquences verbales
cas

topographies
inventaires, classifications, partitions



<ajoute une courte bibliographie>