La Gestion versus l'analyse dans l'approche électronique des textes

JEAN-GUY MEUNIER, SUZANNE BERTRAND-GASTALDY, LOUIS-CLAUDE PAQUIN

1 INTRODUCTION

1.1 La modification du rapport au texte

Un des résultats inattendus de l'introduction de l'ordinateur dans l'environnement de travail, tant des individus que des institutions, est la modification de leur rapport à l'information textuelle. En effet, la possession de plus en plus répandue de micro-ordinateurs a certes bouleversé le mode traditionnel de production et d'archivage de l'information textuelle, mais elle en a surtout profondément modifié la gestion et l'analyse. C'est, par exemple, une scène de plus en plus fréquente de voir des conseillers techniques d'une entreprise non seulement fouiller des textes archivés dans une base de données électronique pour récupérer de l'information contenant des politiques, des normes, etc., mais aussi pour analyser ces textes afin de prendre des décisions, orienter des politiques, cerner des tendances, etc.

Cependant, la nature électronique du support, surtout lorsque le corpus est grand, engendre des problèmes de traitement tout à fait nouveaux. En effet, contrairement à un classeur ou à une bibliothèque qui permet à une personne de fureter, lire, et analyser les fichiers ou les livres, un texte sur support électronique n'est pas accessible directement. Le document doit être atteint et parcouru via des intermédiaires, c'est-à-dire via des programmes spécialisés dans la gestion et l'analyse des textes. Or, ces programmes rencontrent de graves difficultés de traitement en raison de la nature originale d'un texte électronique. Celui-ci se trouve maintenant sur un support physique qui ne présente plus nécessairement un texte de manière linéaire comme cela était le cas sur du papier. De plus, cà ce qui se passe lors de la lecture humaine, le contenu n'est absolument pas transparent aux programmes qui le parcourent.

1.2 La double nature d'un texte

Un texte est un objet de nature originale. Il est certes un objet matériel que l'on doit *manipuler* physiquement. Un texte est quelque chose que l'on transcrit, copie, repère, transmet, imprime, archive, etc. Mais un texte est plus qu'un objet matériel. Il est aussi un objet *sémiotique* qui doit être *manipulé* cognitivement c'est-à-dire *lu* et *parcouru* pour en extraire les idées, les concepts, les propositions ou les thèses; pour en vérifier les démonstrations ou les preuves; en comprendre la narration, l'argumentation, etc. Bref, un texte présente toujours une double dimension, matérielle et sémiotique. Et c'est cette double dimension qui pose, pour son traitement informatique, des problèmes particuliers de design. Bien qu'en apparence seulement rien ne semble plus facile maintenant que cette question du traitement de texte par ordinateur, il faut constater que la problématique est des plus complexe. En effet, on ne comprend pas toujours la même chose sous ce concept de *traitement de l'information textuelle*. Aussi, pour clarifier le débat, nous distinguerons deux grands types de traitements que l'on peut effectuer sur l'objet textuel. L'un touche le support physique de l'information et l'autre aborde son contenu signifiant. Bien que les deux opérations portent sur l'information, elles n'ont pas les mêmes caractéristiques et ne relèvent pas non plus de la même modélisation formelle.

a) les opérations sur la matérialité de l'information

Le premier type d'opération sur les documents textuels consiste à les traiter en regard de leur dimension physique. Quel que soit le «texte», il doit se trouver sur un support physique quelconque. Celui-ci peut être de l'encre sur du papier ou encore des impulsions électriques gravées sur un disque magnétisé. Mais peu importe ce support, il n'affectera pas le contenu qui est exprimé. Que le texte *Roméo et Juliette* de Shakeaspeare soit inscrit sur un support papier ou sur un support électronique, on dira toujours, du point de vue du contenu, qu'il s'agit du même «texte».

Du point de vue physique cependant, il ne s'agit pas du même «texte». Le texte électronique présente des propriétés et des caractéristiques tout à fait différentes de celles du support papier. Et son traitement ne ressemble en rien à la manipulation d'une feuille de papier. De fait, un texte électronique n'est pas constitué d'une séquence de marques chimiques (encre) ou mécaniques (traces, gravures, etc.) mais d'une séquence de signaux électriques souvent distribués à travers tout le support magnétique de l'ordinateur. L'approche traditionnelle pour modéliser le traitement de ces signaux s'inspire de la théorie classique dite de *l'information* pour laquelle un texte est une séquence de signaux informationnels qui présentent une *structure linéaire, une fréquence, un degré de probabilité, d'enttopie, de taux de redondance, etc.* Ces concepts et leurs variantes se retrouveront à la base d'algorithmes pour la manipulation informatique des séquences de ces signaux. Ceux-ci se verront ainsi *encodés, décodés, compilés, transmis, filtrés, compressés, stockés, appariés, etc.*

Ces modèles ont permis de construire une technologie informatique importante. Pensons par exemple à la télécopie, à la télématique, à la saisie électronique des documents, à l'impression laser, etc. Plus spectaculaires encore sont les systèmes de traitement de textes dont les fonctions les plus appréciées sont la transcription, la copie, la mise en page, etc.

Bref, un texte peut être manipulé informatiquement en tant que séquence de signaux électriques. Cette technologie est essentiellement orientée vers le traitement du support physique du texte. Mais bien que cette technolgie modifie grandement notre rapport au texte, il faut voir qu'elle n'atteint jamais le contenu du texte même de la façon la plus superfielle. On ne peut pas dire sans anthropomorphisme qu'un système de traitement de texte non plus qu'une imprimante *lisent et analysent* un texte. Un FAX opère aussi bien sur une lettre française que sur une lettre chinoise! Il ne saisit aucunement la différence entre une fiche documentaire, une lettre de félicitation ou une facture de vente! La critique du modèle classique de la théorie de l'information a clairement montré qu'elle avait, dès le point de départ, éliminé l'aspect *contenu signifiant* de l'information (Bar Hillell, 1955). Les signaux y sont toujours manipulés sans égard à leur signification.

De plus en plus, les utilisateurs réclament que l'on manipule avec aisance non seulement le support physique d'un texte mais aussi qu'on accède à son contenu véritable.

b) les opérations sur le contenu du texte

Cette dimension matérielle d'un texte est certes importante, mais le véritable lieu de l'intérêt d'un texte - à moins d'être un collectionneur, un imprimeur ou un agent d'un service de télécommunication - réside dans son contenu informationnel signifiant c'est-à-dire dans son *propos* et les sciences littéraires et linguistiques apellent les *énoncés et le discours* dont il est le porteur. Mais parce que la question du contenu du texte apparaît dans un horizon informatique,la question paraît simple, du moins à première vue. La lecture d'un texte semble tellement transparente qu'on en vient à penser que l'ordinateur peut aisément faire la même chose. Il s'agit évidemment d'une illusion.

Un texte n'est pas quelque chose que l'on parcourt d'un seul trait. Pour le lire, il faut traverser différents niveaux et structures. Par exemple, il faut identifier, au-delà de la segmentation

en pages, sa structure éditoriale composée du titre et de sous-titres, de chapitres, de sections, etc. Le texte n'est compréhensible que si les normes de la langue sont respectées, si ses phrases sont signifiantes, si ses paragraphes sont bien campés, etc. Il faut être en mesure de suivre l'articulation de ses propositions, de son argumentation, de sa démonstration, de sa narration, etc. Bref, le contenu d'un texte touche à plusieurs niveaux d'organisation qui vont de la structure éditoriale à la langue et au discours qui y est exprimé.

Un système de traitement électronique de l'information textuelle devra toujours distinguer entre les opérations qui s'appliquent au texte comme objet physique et les opérations qui s'appliquent à sa nature sémio-cognitive. Les premières permettront la manipulation du texte en tant qu'objet physique, les secondes permettront sa gestion et son analyse. Si on ne distingue pas ces deux dimensions, on confondra les questions du traitement matériel du texte et celles de l'analyse du texte. Or, il faut maintenir une différence radicale entre ces deux dimensions. En effet, si toute analyse de texte nécessite une manipulation matérielle, l'inverse n'est pas vrai. Les deux types d'opérations se complètent mais ne sont pas identiques.

Dans l'étude qui suivra nous tenterons de préciser davantage la nature des opérations sémio-cognitives impliquées dans le traitement électronique de la documentation textuelle. Nous étudierons la complexité de la tâche en jeu - la gestion et l'analyse des textes par ordinateur - et proposerons un modèle fonctionnel du flux de traitement. Enfin, nous étudierons les solutions logicielles existantes pour réaliser ces tâches.

2 NATURE DE LA PROBLÉMATIQUE

Il arrive souvent que la problématique du traitement électronique de la documentation textuelle soit présentée uniquement sous l'angle de la grande masse de documents à stocker et à repérer. Or, il faut bien voir qu'il s'agit là d'une question parmi plusieurs autres. Une étude approfondie des tâches reliées au traitement de la documentation textuelle révèle que la gestion et l'analyse de texte sont confrontées à deux grands types de problèmes. Un premier type vient de la complexité de la documentation à traiter, alors que le second est inhérent au processus cognitif de l'accès au contenu de l'information textuelle.

2.1 La complexité de la documentation textuelle

Le premier problème auquel est confronté tout système de gestion et d'analyse de texte par ordinateur est celui de la complexité de la documentation textuelle. D'ailleurs, la littérature a régulièrement mis en évidence plusieurs dimensions de cette complexité. S. Bertrand-Gastaldy (1990) pour sa part, retient les six dimensions suivantes :

La première dimension, et la plus évidente, est son *volume.* On mesure maintenant au kilomètre et même au poids le volume de la documentation textuelle à traiter. Des statistiques crédibles montrent que les entreprises, comme l'administration française, génèrent annuellement quelque trois à quatre cents milliards de pages! On dit que la production d'un Boeing génère son poids en documents afférents. Un utilisateur, individuel ou collectif, n'est souvent confronté qu'à une fraction de ce volume, mais celui-ci demeure imposant. Un employé, dit-on, génère annuellement l'équivalent de son poids en documentation textuelle!

La seconde dimension relève de la *diversité* des documents eux-mêmes. Dans une organisation, un corpus textuel n'est que rarement homogène dans ses types de documents. Se

côtoient souvent pour un même domaine des textes légaux, administratifs, des décrets, des procès-verbaux, de la correspondance, des manuels techniques, de la documentation afférente, etc.

La troisième dimension touche la temporalité du texte. En effet , un corpus textuel est souvent une réalité *dynamique.* Tous les documents n'ont pas le même cycle de vie. Certains sont stables, presque éternels (une constitution) alors que d'autres ne durent que le temps de la communication (les mémos). Enfin, certains documents sont en constante évolution; ils peuvent être modifiés quotidiennement.

Une quatrième dimension, corrollaire aux précédentes, est l'*interdépendance* des types de documents. En effet, non seulement les documents sont-ils divers, mais un grand nombre d'entre eux sont interreliés. Ainsi, par exemple, les griefs découlent d'un texte de convention collective.

Une cinquième dimension résulte du processus de la production textuelle elle-même. S'il arrive quelque fois qu'un texte soit produit d'un bout à l'autre par une seule personne, il est très souvent le résultat d'un *travail collectif*. Par exemple, le spécialiste, le technicien, le gestionnaire, l'utilisateur éventuel, etc. participent à l'écriture d'un manuel de référence pour un nouveau système. Les uns le rédigent alors que les autres l'annotent, le commentent, le corrigent, le révisent, le traduisent, etc. Une telle dynamique de production entraîne évidemment de nouveaux besoins en termes d'homogénéité de vocabulaire et de style, de validation du contenu, etc.

Enfin, la dernière dimension de la complexité des documents textuels concerne leur nature potentiellement composite. Non seulement le corpus peut-il contenir des documents textuels, mais, de plus en plus, il en vient à marier le langage naturel à des documents présentant d'autres formes sémiotiques. On trouvera ainsi à côté de données linguistiques, des données numériques, des images fixes, des graphiques et des plans, et même des images dynamiques (vidéo) ou encore du son. Le corpus devient alors *multi-sémiotique* ou multi-modal.

Bref, les données à traiter sont complexes. *Elles sont souvent volumineuses, hétérogènes, dynamiques, interdépendantes, collectivement produites et multimodales.* Les systèmes informatiques qui sont confrontés à ce type de corpus ne peuvent laisser de côté cette complexité car celle-ci est inhérente à la nature même des documents. Aussi, les systèmes de gestion et d'analyse des documents devraient-ils, entre autres, être en mesure :

- *de traiter de grandes quantités de documents, tant textuels que multisémiotiques,*
- *d'offrir des stratégies différenciées d'analyse adaptées à la diversité des documents,*
- *de permettre une mise à jour en temps réel et une gestion du cycle de production,*
- *de supporter des liens intertextuels, hypermédia et multimodaux,*
- *d'assister la rédaction et la validation du contenu,*
- *de contrôler la diffusion et la confidentialité,*
- *d'offrir un interface convivial permettant un accès ergonomique.*

Bref, les systèmes doivent être polyvalents et souples; ils doivent mettre les utilisateurs en relation directe avec la complexité même du corpus textuel, au moyen d'un interface commun et convivial (Belkin *et al.*, 1991). Un système trop spécialisé dont l'application serait restreinte à certains types de documents et qui n'effectuerait qu'un nombre restreint de fonctions sera peut-être informatiquement efficace, mais il deviendra vite inadéquat et source de frustation pour un

utilisateur. Il sera vite incapable de soutenir un cheminement de *gestion et d'analyse* de la documentation telle qu'elle se présente en réalité.

2.2 la complexité du processus cognitif d'accès à l'information textuelle

Non seulement un système de traitement électronique des textes doit-il s'ajuster à la nature complexe du corpus textuel, mais il est aussi confronté à la complexité du processus d'accès à l'information textuelle. En effet, le but ultime d'un traitement électronique de textes est de permettre à un utilisateur d'extraire du corpus de l'information pertinente. Pour ce faire, les textes doivent être lus et interprétés. Il ne faut jamais oublier que ce n'est que par métonymie que l'on dit que l'ordinateur traite l'information textuelle. Seul l'humain, en dernière instance, peut lire et interpréter un texte. Ce processus est d'ordre cognitif et présente des caractéristiques propres suivantes.

La première caractéristique de ce processus est que de la part du lecteur il s'agit toujours d'un acte privé. Les théories classiques sur le processus cognitif de l'accès au contenu d'un texte ont toujours soutenu que cette activité dépend toujours des objectifs et des projets que se donne le lecteur:

> «Quiconque veut comprendre un texte a toujours un projet. Dès qu'il se dessine un premier sens dans le texte, l'interprète anticipe un sens pour le tout. À son tour, ce premier sens ne se dessine que parce qu'on lit déjà le texte, guidé par l'attente d'un sens déterminé. C'est dans l'élaboration d'un tel projet anticipant, constamment revisé, il est vrai, sur la base de ce qui ressort de la pénétration ultérieure dans le sens du texte, que consiste la compréhension de ce qui s'offre à lire [...]. Ce processus est donc le renouvellement incessant du projet qui entretient le mouvement de la compréhension et de l'interprétation.»
> (Gadamer; 1976: 196)

Autrement dit, il faut toujours situer l'accès au contenu d'un texte ou ce que traditionnellement on appelle l'*interprétation* dans un horizon d'action. C'est le projet qui permet l'avènement ultime du sens du texte.

Cela signifie, en conséquence, que la lecture variera presque nécessairement d'une personne à une autre, d'un moment à un autre. C'est en ce sens qu'on dira qu'elle est un acte privé. Un avocat ne «lira» pas un texte comme le ferait un administrateur parce que chacun possède son propre projet de lecture. Il s'ensuit que même si un texte se présente dans une langue spécifique, par exemple, le français, et même si l'ensemble des expressions linguistiques qui le constituent sont relativement stables et susceptibles d'être partagées socialement, son contenu présentera toujours une part d'indétermination.

> «Le texte contient une composante d'indétermination. Ce n'est pas un défaut, mais bien une condition fondamentale de la communication du texte; elle permet la participation du lecteur à l'intention du texte.» (Iser, 1985: 15).

La deuxième caractéristique de ce processus est qu'il est social. L'accès au contenu d'un texte est non seulement marqué par la subjectivité d'un lecteur, il l'est aussi par le milieu social dans lequel cette lecture s'effectue et pour lequel un texte sert de médium de communication. La recherche contemporaine dans le domaine de l'analyse du discours et des textes a amplement montré que le texte se construit en regard du tissu social dans lequel il s'insère. Un texte n'est jamais isolé des autres discours auquel il renvoie (Foulcault, 1969). Il sert à consolider les normes d'action (Brenner, 1990), à étayer le savoir et à consolider la mémoire (Kinstch, 1977). Dans une organisation sociale, un texte remplit donc plusieurs fonctions importantes, non seulement pour assurer la transmission de l'information mais aussi pour consolider l'existence même de l'organisation.

Il s'ensuit alors que le texte subira un ensemble extrêmement diversifié de parcours interprétatifs. La «lecture» des textes changera en fonction de l'évolution des besoins et de l'état des connaissances de l'organisation sociale:

> «[...] different persons, in different occupations may possess different world views and make different demands upon sources of knowledge as a consequence. For example, some occupations may require no more than 'recipe knowlege' for their effective performance; others, falling short of a need for 'expert' knowledge, may demand more in the nature of 'reasoned opinion' and, hence, a greater need for access to sources of information.» (Wilson, 1984: 200)

Ces deux dimensions de l'accès au contenu du texte, à savoir son caractère privé et en même temps social, nous amènent à une conclusion théorique importante: il est impossible de construire un système de lecture et d'analyse automatiques des textes. Certaines thèses issues des recherches en intelligence artificielle (Schank et Abelson, 1977) de la linguistique computationnelle (Pêcheux, 1972) et même du repérage de l'information (Salton et Mc Gill, 1983) ont donné à croire qu'il était possible de construire des systèmes d'accès au contenu d'un texte qui soient automatiques[1]. Selon nous, «la lecture automatique du discours ou du texte» sont, dans cette perspective, des contradictions dans les termes. Lire et comprendre un texte est idiéosyncratique à l'activité d'intégration et d'adaptation des *humains* à leur environnement. Autrement dit, la lecture et l'interprétation de textes sont des activités cognitives humaines qui, à ce titre, ne peuvent être automatisées.

Cependant, si on ne peut construire des systèmes qui simulent la lecture et la compréhension humaines, il ne s'ensuit pas que l'ordinateur ne peut en rien être utile dans ce processus. Au contraire, on peut penser outiller les utilisateurs pour faciliter ce processus, plutôt que de les en déposséder au profit d'un automate qui ne pourra de toute façon que construire des représentations rudimentaires, stéréotypées et insensibles au projet de l'utilisateur. Autrement dit, s'il est impossible de doter un ordinateur de toutes les connaissances et habiletés nécessaires pour «comprendre» un texte, il est cependant réaliste de concevoir des outils informatiques capables d'assister l'utilisateur dans la transformation des données textuelles en éléments structurés, significatifs et porteurs de connaissances. Il sera alors plus aisé pour ce dernier de manipuler, de classifier, de relier et d'interpréter de tels éléments.

Pour ce faire, il faut donc abandonner la voie des systèmes informatiques qui se donnent comme des robots-lecteurs au profit de celle de systèmes «adjuvants» dans l'activité cognitive de la lecture humaine (Meunier, 1992) qui laissent la maîtrise ultime du traitement de l'information entre les mains de l'expert qu'est l'humain. Ce n'est que dans cette perspective qu'on peut garantir une amélioration de la qualité du travail de gestion et d'analyse, tant en termes de volume, de rigueur et de systématicité (Paquin et Beauchemin, 1988).

Par conséquent, un système informatique de gestion et d'analyse des textes ne peut être clos et autonome; il doit au contraire être ouvert et offrir une grande polyvalence de fonctions. Un peu à la façon d'une boîte à outil, le système doit mettre à la disposition de l'utilisateur - lecteur et interprète - un évantail de modules et de fonctions avec lesquelles il pourra gérer et analyser son corpus. Seul l'humain doit ultimement contrôler ce processus de gestion et d'analyse. Sur le plan de la gestion, des modules devront être en mesure d'assister l'utilisateur dans tout le processus de production, de transmission et de classification de l'information, c'est-à-dire sa chaîne de traitement. Et sur le plan de l'analyse, des modules devront lui permettre de pénétrer le contenu du texte c'est-à-dire de participer à la dynamique de sa construction, de sa description et d'en extraire des connaissances.

---

1 Des arguments semblables ont été invoqués dans la critique des projets de traduction automatique. On parle plutôt maintenant de projets en traduction assistée par ordinateur (TAO).

3 UNE APPROCHE SENSIBLE AU CONTENU

La conception de la problématique du traitement électronique des textes que nous avons brièvement esquissée ci-haut nous oblige donc à mieux préciser ce que nous entendons par la gestion et l'analyse des textes. Pour expliciter ces concepts, nous proposerons un modèle comportant les principales tâches ou fonctions opératoires qui sont en jeu dans une chaîne de traitement, de gestion et d'analyse de texte par ordinateur. Il nous faut cependant faire quelques remarques préliminaires.

Premièrement, il faut constater que les systèmes traditionnels de traitement électronique des documents mettent habituellement à la dispositon d'un utilisateur un vaste échantillon d'opérations à effectuer sur le corpus. Voici, à titre d'exemple, une liste non exhaustive de celles-ci: *création, sélection, acquisition, stockage, organisation, description, indexation, évaluation, synthèse, conservation, repérage, diffusion, mise à jour, etc.*

Les modèles provenant des sciences documentaires et de l'information préciseraient assurément davantage cette liste. On pourrait même y ajouter des opérations plus simples comme *paginer, souligner, ordonner, segmenter,* etc. Notre souci n'est pas ici d'en faire la liste exhautive mais d'en comprendre la nature plus formelle.

Sauf pour certaines, comme le *stockage, la diffusion, la conservation,* il faut voir que ces opérations ne portent pas sur la matérialité de l'information, mais sur son contenu. En effet, malgré les apparences, ces opérations sont d'ordre sémio-cognitives c'est-à-dire qu'elles participent à l'interprétation par un humain du donné informationnel. Même une opération aussi simple que de *paginer un document* est une opération cognitive complexe. Que ce soit au moment de la création de l'édition ou de l'utilisation, il s'agit d'une interprétation appliquée à une réalité physique (un segment de signaux).L'association d'un nombre et d'un ordre aux segments est toujours assujettie à un projet de manipulation humaine. Paginer est une opération descriptive d'ordonnancement qui est certes plus simple que celle de reconnaître une catégorie syntaxique comme le syntagme nominal, mais qui n'en demeure pas moins une opération cognitive de haut niveau.

Deuxièmement, il faut constater que cette liste d'opérations a été établie surtout en fonction de systèmes de gestion de l'information pour des documents de type fiches plutôt que pour des documents de type plein texte. En effet, ces opérations ont été surtout utilisées dans des approches traditionnelles de gestion documentaire pour des documents bibliographiques ou des dossiers d'archives et ce à des fins de repérage d'information (information retrieval) dont les performances sont habituellement évaluées en termes de *taux de rappel* et de *taux de précision* (Salton et McGill, 1983).

Si, dans une perspective de *repérage d'information*, des stratégies numériques non sensibles aux dimensions linguistiques ou discursives des documents se sont avérées efficaces, il faut voir que l*'accès au contenu* d'un texte ne peut aucunement se réduire à ce type de stratégies. Repérer un document n'est qu'une activité - importante dans un contexte de recherche documentaire - parmi plusieurs autres qu'un humain peut vouloir effectuer sur un texte. On imagine l'absurde d'une situation où l'apprentissage de l'usage d'un texte n'aurait pour seul but que de montrer comment repérer de l'information. Lire Shakespeare consisterait alors à trouver où la phase «to be or not to be» est imprimée!

Ces constatations entraînent un ajustement de la précédente liste d'opérations. En effet, des difficultés surgissent lorsqu'on veut étendre leur domaine d'application au plein texte et ce à des fins d'analyse. Rappelons que l'analyse des documents textuels vise leur contenu et, pour ce faire, doit

tenir compte de leurs aspects linguistiques, discursifs, conceptuels, argumentatifs, etc. C'est ainsi que certaines opérations ne s'appliqueront pas aussi facilement au plein texte, alors que dans certains cas d'autres opérations se transformeront en de véritables stratégies d'analyse. Voici, à titre d'exemple quelques-unes de ces opérations:

Ordonnancement :

Par ex.: établir une pagination.

Segmentation :

Par ex.: diviser le texte en chapitres à des fins de catalogage et d'indexation.

Sélection :

Par ex.: le choix de textes selon divers critères divers (coût, durée de vie, pertinence technique).

Catégorisation éditique :

Par ex.: les parties d'un texte comme le titre et les sous-titres selon des normes internationales (SGML).

Classification :

Par ex.: regroupement de textes ou de segments comme unités documentaires dans une base de données.

Liaison :

Par ex.: relier des parties de textes entre eux et à d'autres textes (liens hypertextuels).

Indexation

Par ex.: fabriquer un index de mots-clés pour accéder aux textes.

Contrôle du vocabulaire

Par ex.: identifier, contrôler et structurer le vocabulaire technique d'un domaine, (aréonautique, biologie, éducation).

Indexation pour la diffusion

Par ex.: faire le résumé d'un article de périodique.

Identification des synomymes ou paraphrases:

Par ex.: recherche d'expressions similaires dans des documents, tels *pollution sonore* et *problème de bruit.*

Lisibilité :

Par ex.: indice de la complexité lexicale, syntaxique, etc. des textes produits par une entreprise.

Étude de la détermination d'un terme générique :

Par ex.: congé *sans solde*, congé *sans salaire*, congé *de maternité.*

Description des relations conceptuelles d'un terme avec d'autres

Par ex.: *ascenseur* vs *monte charge* vs *élévateur*.

Construction d'une thématique conceptuelle

Par ex.: interprétation légale de concepts aux frontières floues, tels : *meurtre au premier degré* et *contrat de bonne foi.*

Comparaison de l thématique des propositions

Par ex.: les normes dans les conventions collectives.

Repérage des arguments pour ou contre une décision

Par ex.: dans les rapports sur l'établissement d'un site d'enfouissement de déchets dangereux

Regroupement des arguments pour une décision stratégique

Par ex.: investir ou ne pas investir dans l'amiante.

Regroupement de réponses dans une enquête à questions ouvertes

Par ex.: arguments pour ou contre l'euthanasie.

Identification des défendeurs d'une idée, d'un mouvement, etc.

Par ex.: qui soutient le développement d'un barrage dans un territoire amérindien?

Description de l'évolution d'une argumentation pour ou contre une politique

Par ex.: les positions du gouvernement relativement à l'avortement depuis 1940

Mise à jour et comparaison des politiques et des règlements d'une institution

Par ex.: une politique d'hypothèque dans une banque

Comme on le voit, les opérations que l'on peut effectuer sur un texte sont nombreuses et diversifiées. Certaines recoupent les opérations que l'on peut faire sur une description bibliographique, mais la majorité cherchent véritablement à atteindre le contenu informationnel. En ce sens, elles sont toutes d'ordre sémio-cognitives c'est-à-dire qu'elles participent aux activités de lecture et d'interprétation du matériau textuel. Certaines sont plus orientées vers la gestion alors que d'autres sont orientées vers de l'analyse.

3.1 La chaîne de traitement dans la gestion et l'analyse de texte par ordinateur

Nous avons dit que les opérations énumérées ci-haut portent sur diverses dimensions de la gestion et de l'analyse textuelle. Cependant, présentées sous la forme d'une liste, on ne voit pas

clairement le principe de leur distinction. Pour rendre ce point plus clair, on peut les regrouper dans une perpective de chaîne de traitement (work flow). Le terme chaîne de traitement recouvre ici l'ensemble des opérations de base, séquentielles et récursives, que l'on peut appliquer dans la gestion et l'analyse électroniques des textes. On distiguera alors les opérations qui sont liées à a) la *production* du texte, b) sa *description*, c) *l'extraction* d'information et d) à la *représentation* des connaissances.

Chacune des étapes ou moments du processus se décompose lui-même en plusieurs autres opérations. Nous ne pouvons ici les décrire de manière exhaustive, mais nous pouvons en faire une présentation schématique.

3.2 Les opérations liées à la production du texte

L'entrée des textes dans l'ordinateur ne s'effectue jamais directement; plusieurs opérations parallèles et séquentielles sont toujours requises. Un système informatique intégré devra offrir de l'assistance pour certaines d'entre elles.

Le cas le plus simple de la production d'un texte est celui où l'on a directement recours à un système classique de traitement de texte, comme MSWord, Word Perfect, etc. L'écrivain, unique ou collectif, pourra tirer un grand bénéfice d'une interaction avec des dictionnaires électroniques, etc., avec des correcteurs, avec un gestionnaire de références bibliographiques ou encore avec des versions antérieures du texte, des traductions ou d'autres textes. Par ailleurs, dans certaines organisations, l'écriture sera collective et passera par plusieurs instances pour la révision, la validation, l'approbation, la mise en page, etc. La progression des documents dans un tel circuit pourrait avec avantage être supportée par une base de données. De plus, la génération même du texte peut être assistée (Rada, 1991).

D'autres textes proviennent d'archives. Si les textes sont sur support papier, il faudra effectuer une tranformation du texte original en une représentation informatique. Si leur état le permet, ils peuvent être saisis par balayage optique et les images obtenues transformées en caractères ASCII. Si les textes sont sur support magnétique, il faut s'assurer de la conformité de leur représentation informatique avec les formats électoniques contemporains. Le cas échéant, les formats devront être interprétés et convertis de façon à n'encourir aucune perte d'information. Par la suite, dans les deux cas, le résultat obtenu devra être révisé pour s'assurer de la conformité avec l'original. Cette révision pourra être effectuée avec profit à l'aide d'un correcteur.

Toutes ces opérations ne sont que des portes d'entrée pour accéder au contenu des textes mais il ne l'atteignent pas encore.

3.2 Les opérations descriptives sur un texte

Le contenu d'un texte n'est pas transparent pour un ordinateur. Et les multiples travaux de l'intelligence artificielle nous ont démontré que tout système intelligent de traitement de l'information devait disposer au préalable des «connaissances» qui servent de point de départ pour les analyses. Cette hypothèse rejoint la longue tradition philologique qui dit que la lecture analytique d'un texte est une opération de commentaire, c'est-à-dire d'ajout descriptif et explicatif. Dans le cas d'un traitement informatique, ceci signifiera qu'il faut lui ajouter des informations qui rendent explicites les multiples niveaux de sa structure signifiante. Or ces niveaux sont multiples (Mennik et al, 1987; Duchastel 1991: 601). Cet ajout d'information touchera les structures éditique, linguistique, discursive, argumentative, etc. Autrement dit, pour qu'une suite de signaux électroniques soit considérée comme un phénomène textuel, il faut que des analyseurs aient ajouté au texte des informations décrivant les différents niveaux de sa structure sémiotique.

Par exemple pour qu'une suite de caractères comme *Je pense donc je suis* soit considérée comme une unité textuelle éventuellement analysable, il faut préciser plusieurs choses. On doit par exemple indiquer son statut éditique c'est-à-dire décider si elle appartient à un titre ou au corps d'un paragraphe, etc. : *Je pense donc je suis* **(titre)**. On peut aussi identifier ses constituants syntaxiques : *je* **(pronom)** *pense* **(verbe)** *donc* **(composition)** *je* **(pronom)** *suis* **(verbe)**. On peut vouloir préciser la signification de certaines expressions : *Je* **(personne)** *pense* **(action)** *donc je* **(personne)** *suis* **(état)**. Dans certains cas on pourra ajouter des informations sur le locuteur, l'allocutaire, la situation, le contexte, le temps, le statut illocutoire, etc.

Bref, le texte doit être soumis à des analyseurs qui en décrivent les constituants sémiotiques. Il est important de noter que le texte qui est alors l'objet de l'analyse n'est plus le texte de départ mais le texte amplifié de ces multiples niveaux de description. Ces descriptions qualifiant les unités d'information d'un document textuel, obtenues manuellement ou par des analyseurs spécialisés, sont ajoutées au texte lui-même par le biais d'une catégorisation. Celle-ci consiste en l'ajout d'étiquettes qui décrivent le statut sémiotique (éditique, linguistique, logique, etc.) des constituants du texte auxquels ils sont adjoints. En ce sens, il s'agit d'un mode d'opérationnalisation formel de l'une des étapes du processus d'interprétation du texte.

Aucun système informatique n'est actuellement en mesure d'interpeler des modules qui permettent de réaliser une catégorisation automatique, complète et fiable sur tous les niveaux en jeu. Aussi, devant l'étroitesse de la couverture des analyseurs disponibles et surtout devant la complexité des descriptions à effectuer, l'opération de catégorisation est souvent effectuée partiellement ou entièrement à la main. Mais comme la qualité d'une analyse dépend de cette catégorisation, il faut malgré tout tenter d'y recourir le plus systématiquement possible.

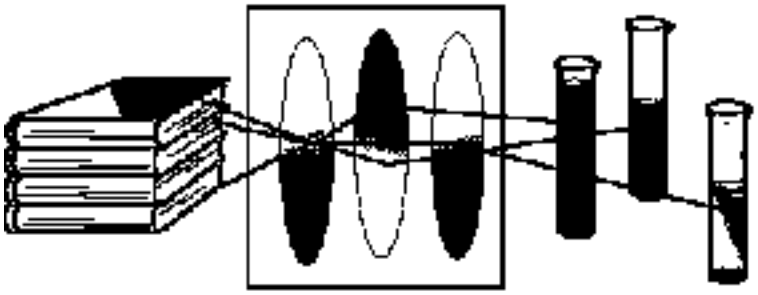
Un bon système informatique devra alors être un adjuvant souple pour la réalisation de ces opérations descriptives. La transparence et l'interactivité seront plus importantes que le niveau d'automatisation.

3.3 Les opérations liées à l'extraction de l'information

Un texte catégorisé n'est intéressant que dans la mesure où il permet l'extraction d'information. Tout ce qui est dit ou énoncé dans un texte n'est pas pertinent en soi. Il ne le devient, avons-nous dit, que relativement à un projet de lecture. Un même texte peut être utilisé par de nombreuses personnes et toujours être vu sous un angle nouveau. Dans une entreprise, les conventions collectives seront lues différemment par un avocat, un représentant syndical, un directeur du personnel, un arbitre, etc.

Pour certains analystes, la dimension lexicale d'un texte peut être le lieu de son intérêt. Pour un autre, il faut rejoindre les concepts alors que pour un troisième, le repérage d'information peut être l'aspect important. Bref, l'accès au contenu des textes est toujours orienté par un projet. L'utilisateur focalise toujours son attention selon une perspective particulière. Le même texte permettra donc d'extraire autant d'informations différentes que de projets de lecture différents pourront être formulés.

En termes métaphoriques, nous pourrions comparer le processus d'extraction de l'information à une opération de production de précipités qui permet sur des composés chimiques d'isoler certains constituants :

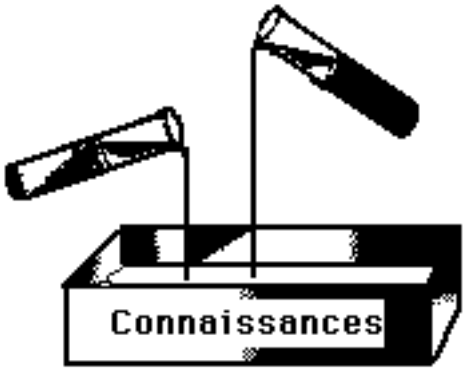


Dans le cas du texte, cette extraction peut être effectuée par des procédures complexes appelées «patrons de fouille» qui sont formulés tant à partir des mots du texte original qu'à partir des catégories ajoutées lors de la description. Les résultats obtenus peuvent être présentés sous différentes formes. Certains peuvent prendre une forme structurale, telles une liste, un lexique, une concordance, un index, etc. D'autres peuvent prendre une allure numérique ou statistique, tels un tableau, un histogramme, une courbe, etc.

3.4 Les opérations liées à la configuration des connaissances

L'extraction de l'information d'un texte est souvent complexe et livre des résultats dont la manipulation n'est pas toujours facile. Un bon analyste de texte tentera de reconfigurer les informations obtenues de manière à les rendre le plus lisibles et compréhensibles possible. Il procèdera habituellement à une configuration quelconque de ces résutats. Il constituera ainsi des réseaux lexicaux, des réseaux sémantiques, des hiérarchies de concepts, des thésaurus, des suites d'arguments, etc. Si une exploitation informatique est prévue, les résultats seront configurés puis déposés dans une plate-forme particulière. Ainsi, par exemple, si le résultat de l'extraction est un thésaurus a posteriori, il peut être versé dans une base de données plein texte pour l'interrogation du corpus. L'information extraite, lorsque configurée par l'analyse, prend le statut de «connaissance».

En termes métaphoriques, on déversera le résultat des multiples analyses dans des moules adaptés à l'assimilation et la compréhension efficace :



Ces quatre grands types d'opérations se retrouvent dans tout processus de traitement de textes par ordinateur qui veut atteindre le contenu. Qui plus est, on peut aussi imaginer une méta-fonction qui gère l'ensemble de ces opérations et soutient l'utilisateur dans les dédales de sa chaîne de traitement. On créera ainsi un genre de station de travail ou d'atelier cognitif de traitement

analytique des textes. C'est en cela que nous dirons qu'il s'agit de systèmes informatiques de gestion et d'analyse intelligement assistée de textes.

4 LES SOLUTIONS INFORMATIQUES ACTUELLES[2]

On observe actuellement sur le marché, un foisonnement de systèmes et programmes consacrés au traitement électronique des documents. Ceux-ci peuvent être regroupés autour de quatre grands types : des logiciels documentaires, des logiciels de gestion de documents saisis en mode image, des logiciels de repérage de plein texte et enfin des logiciels d'analyse de texte. Nous allons brièvement passer chacun de ces types en revue en nous attardant à la façon dont ils attaquent la problématique que nous avons esquissée plus haut.

4.1 Les logiciels documentaires

Un premier type de logiciels a été conçu pour le repérage de documents d'archives ou de bibliothèques via l'interrogation de bases de données bibliographiques. Ces systèmes assurent généralement des fonctions de gestion des collections: l'acquisition, le classement, le prêt, la conservation, la préservation de la confidentialité, etc. La plupart de ces logiciels ont été développés il y a plusieurs années pour remplir des besoins bien spécifiques sans souci d'intégration des différents services d'information. À cette époque, pas si lointaine, l'accès au plein texte relevait encore de l'utopie (Karivalo, 1989).

Sur le plan logiciel, la structure interne des données de chacun des systèmes empêche le partage de l'information; les stratégies d'interfaces sont variées et, la plupart du temps, cryptiques de sorte que la manipulation des systèmes requiert souvent une période d'entraînement longue et intensive. Sur le plan conceptuel, chacun de ces systèmes a déterminé sa propre grille d'analyse et ses propres catégories d'accès (Bertrand-Gastaldy, 1990b: 74). Tout ceci constitue un obstacle à la collecte ponctuelle de renseignements. Or, la gestion d'information ne constitue pas la tâche accomplie par les professionnels des organisations. Ces derniers ont un problème à résoudre, une décision à prendre, un dossier à évaluer, etc. Toutefois, les enjeux qui sont reliés à l'accomplissement de ces tâches exigent un accès rapide et précis à plusieurs types d'informations situées sur des systèmes différents et ce par la personne sans que celle-ci n'ait préalablement subi un entraînement intensif.

4.2 Les logiciels de gestion des documents saisis en mode image

Le succès des systèmes de gestion électronique des documents (GED)[3] s'explique par le fait qu'ils résolvent la plupart des problèmes liés à la manipulation du support papier. Ils réunissent sur un même support les documents composites autrefois dispersés dans plusieurs systèmes de stockage. Tout en préservant la présentation visuelle des documents originaux, ces systèmes réduisent considérablement les coûts de stockage. Les coûts et les délais de manipulation connaissent également une diminution importante. De plus, l'ergonomie de la consultation ne

---

2 Cette section reprend en partie l'argumentation exposée dans le texte suivant: S. Bertrand-Gastaldy, J.-G. Meunier et L.-C. Paquin, "De la nécessité de repenser la gestion et l'analyse de l'information textuelle dans les organisations", *Actes du colloque ICO 93* en collaboration avec ; (à paraître).

3 Des logiciels comme Desktop Document Manager, Inspire VisionQuest, Optix, etc. sur le marché nord-américain et Taurus en France. Ces logiciels sont aussi appelés DIP (Document Image Processing)

change pas trop les habitudes par rapport au support papier. On peut agrandir ou rétrécir les documents sur l'écran, les faire pivoter, les envoyer à un télécopieur, etc. Enfin, il est souvent possible de leur ajouter des annotations ou même des messages vocaux (Benmergui-Perez, 1989; Chevreau et Kelly, 1989).

Cependant, ces logiciels automatisent surtout les tâches effectuées sur le document textuel en tant que porteur physique d'information. L'accès au contenu qui mobilise une grande partie du temps des professionnels dont il a été question plus haut pose les mêmes problèmes que lorsque l'information est sur support papier. Il faut soi-même fournir des mots-clés pour décrire le ou les thèmes principaux traités dans les documents et renoncer à repérer directement l'information spécifique selon de multiples points de vue. On voit cependant apparaître des logiciels de GED interfacés avec des systèmes de repérage en plein texte[4] qui eux travaillent sur les textes codés en ASCII après reconnaissance optique des caractères (ROC).

4.3 Les logiciels de repérage en plein texte

Aux États-Unis, le marché du repérage de l'information textuelle a presque atteint un stade de maturité, d'après Delphi Consulting Group (1992) qui a dénombré 107 000 sites où sont installés des logiciels. La croissance de ce marché est considérable si l'on en juge par l'analyse que ce groupe en a faite[5]. Conçus à l'origine d'après les logiciels de repérage des données bibliographiques, il ont évolué vers un niveau plus élevé d'interactivité, des capacités de sélectivité plus étendues et vers une convivialité plus grande. Dans certains cas, le repérage s'appuie sur des analyses statistiques et permet de réinjecter une réponse pertinente à titre de nouvelle question. Différents opérateurs sont fournis pour travailler sur les chaînes de caractères (masque, troncature, etc.) et sur leur position dans la phrase. Quelques logiciels offrent de plus des possibilités de navigation hypertextuelle : l'utilisateur peut alors, comme avec le support papier, s'appuyer sur l'organisation éditique des documents en sections, chapitres, paragraphes, illustrations, tableaux, etc. Mais, pour être exploitable électroniquement, cette organisation éditique doit avoir été préalablement décrite et cette description nécessite l'accès au contenu des documents.

Cependant, la plupart de ces logiciels de repérage en plein texte[6] n'offrent pas d'autres possibilités d'accès à l'information que les chaînes de caractères qui forment les mots du texte. Comme aucune catégorisation n'est supportée, la mise à disposition brute de très nombreux textes, à la limite, accroît les problèmes d'accès à l'information plus qu'elle ne les résoud. En effet, l'ambiguïté inhérente au langage naturel empêche la formulation de requêtes précises et un repérage vraiment efficace avec pour conséquence que les utilisateurs sont innondés de textes non pertinents. De plus, des phénomènes courants comme l'anaphore, l'ellipse, la paraphrase, etc. nuisent au repérage de tous les textes pertinents. La segmentation des textes en paragraphes et en phrases réduit l'abondance, mais ne constitue pas une solution suffisante aux problèmes de bruit et de silence.

Certes, il existe des logiciels qui tiennent davantage compte de la nature linguistique du matériau à traiter. Ils tentent de retrouver par delà les chaînes de caractères de véritables unités

---

4 Des logiciels comme BRS, BasisPlus, MicroQuestel, etc.

5 "Text retrieval was a $118+ million market in 1990. Both the PC and mini/mainframe markets are growing at an impressive rate. The PC market revenue is growing at a 45% CAGR. The mini/mainframe market revenue is growing at a 35% CAGR.
The market is expected to reach the critical 300+ million mark in approximately 2-3 years". (Delphi Consulting Group, 1992 :TR-12)

6 Des logiciels comme Book Manager, Basis +, Open TEXT, TOPICS, ConQuest, Elexir, Isys, Zyindex, etc., et plus près de nous : CEDROM, Édibase, Seconde, etc.

conceptuelles. Les meilleurs résultats s'arrêtent cependant à la reconnaissance de termes complexes susceptibles de dénoter des notions importantes dans le domaine de référence à condition que celles-ci soient «bien formées»[7]. Mais la possibilité d'explorer les textes dans une perspective autre que terminologique et de les analyser en fonction d'objectifs divers est quasiment absente. En dernière analyse, la plupart des logiciels constituent une "boîte noire" qui a pour fonction unique de mettre les utilisateurs en relation avec les textes ou passages de textes contenant telle ou telle expression ou traitant de tel ou tel sujet.

4.4 Les logiciels d'analyse de texte

Un dernier type de logiciel ou plutôt de plate-forme informatique commence à voir le jour, de sorte que le traitement électronique des documents textuels se modifie lentement. D'une part, les modalités classiques de l'organisation et du repérage de l'information sont de plus en plus ajustées en fonction de la diversité des données à consulter. D'autre part, des stratégies cherchent de plus en plus à atteindre le contenu même de l'information afin d'adapter les opérations aux tâches que les utilisateurs effectuent sur leur documentation textuelle. On voit ainsi apparaître des logiciels qui contribuent de plus en plus aux quatre phases du flux de traitement que nous avons présentées plus haut.

Par exemple, certains logiciels s'insèrent dans la *phase de production* des textes. Ils permettent l'interaction simultanée ou parallèle de plusieurs auteurs. Mais leurs fonctionnalités semblent relativement limitées; les cycles de consultation, de traduction, de révision, d'approbation, etc. ne sont pas couvertes. De même, la validation des éléments autres que le contenu est laissée pour compte : aucun automatisme ne permet de vérifier l'uniformité de la terminologie employée, la lisibilité en fonction du public visé, la conformité à une politique éditoriale. L'insertion de modules n'est pas prévue pour assister la création de thésaurus ou de bases de connaisances pourtant de plus en plus nécessaires dans les systèmes dits "intelligents", malgré l'intérêt qui émerge pour ce genre d'applications dans des publications récentes (Schmitz-Esser, 1990; RIAO Conference Proceedings, 1991). Les documentalistes, terminologues et cogniticiens sont contraints d'attendre "que les outils informatiques d'analyse de contenu des textes soient à la portée de tous" (Ranjard, 1991).

Quant à la phase de description, des analyseurs linguistiques robustes et dotés d'une grande couverture sophistiqués sont en cours de construction. Les uns touchent la catégorisation morphologique et grammaticale alors que d'autres s'attaquent à la sémantique lexicale. Des dictionnaires riches et complexes sont produits et des normes pour leur rédaction sont en cours d'élaboration, citons à cet effet le projet GENELEX. Malheureusement, la plupart des ces produits n'ont pas été élaborés dans la perspective d'une tâche d'analyse textuelle assistée par ordinateur. Un bon nombre de ces analyseurs ont vu le jour dans une perspective de traduction automatique. Ils sont souvent réalisés sur des plates-formes incompatibles et dotés de structure de données particulières. Dans leur forme actuelle, il est donc difficile de les interpeler au sein d'un flux intégré de traitements. On trouve encore moins de logiciels d'analyse de texte qui entrent en interaction avec des modules de description. Et même lorsqu'ils le font, on ne retrouve pas la description organisée de manière à ce qu'elle soit utilisable dans un processus d'analyse textuelle tel que nous l'avons décrit dans le flux de traitement.

La phase *d'extraction* a reçu cependant une plus grande attention. Elle s'est développée notamment dans les contextes de repérage d'information et dans le contexte des travaux d'analyse de texte effectués par les chercheurs en sciences humaines ou en linguistique. Les trois premières approches, exposées précédemment, n'ont malheureusement produit qu'une vision assez réductrice

---

7 Des logiciels comme ALETH de la firme GSI-ERLI et SPIRIT de la compagnie SYSTEX.

de l'analyse de texte: le repérage de passages en fonction d'une question thématique. Qui plus est, cette tradition théorique n'a pas cru toujours nécessaire de passer par les strates de l'organisation de la signification des textes soit les niveaux éditique, syntaxique, sémantique et encore moins pragmatique. Les stratégies numériques de types statistiques (indice de dicrimination, pondération différentielle, etc.) se sont avérées amplement satisfaisantes (Salton 1983). Ces stratégies ont ainsi permis de construire des modèles d'indexation et de classification compétitifs. Mais elles ne donnent que des résultats limités si l'on veut atteindre le contenu discursif du texte.

La tradition linguistique et celle d'analyse du discours a elle aussi offert des stratégies intéressantes d'extraction d'information. Du lexique à la concordance, de l'analyse stylistique à l'analyse thématique, elle a, à travers les années, produit un évantail important de stratégies d'extraction. Malheureusement, ces diverses stratégies sont demeurées relativement isolées dans le milieu de la recherche universitaire et n'ont que très peu été intégrées dans des plates-formes accessibles au grand public. Il y a cependant quelques exceptions, au Centre ATO un système d'analyse de textes par ordinateur (SATO) [Meunier, Daoust et Rolland, 1976] a été développé dès les années 1970. Ce système produit une représentation matricielle du texte qui supporte les annotations [Daoust, 1992]. Cette représentation permet une fouille efficace, autant à partir des unités du texte que des descriptions qui ont pu leur être adjointes. Les résultats obtenus peuvent être soumis à des analyseurs statistiques pour déterminer la co-occurrence lexicale, la distance entre des segments, etc. De plus, depuis peu, un générateur de systèmes à base de connaissances a été intégré à SATO pour constituer un atelier cognitif et textuel (ACTE) [Paquin et Daoust, 1993]. ACTE permet à des non-informaticiens de mettre au point des analyseurs spécifiques à leurs besoins incorporant des stratégies de contrôle sensibles au contexte. De plus, la prise en compte d'informations incertaines permet de dépasser le cadre strict de la logique booléenne pour déboucher sur la modélisation de l'interpétation de descriptions plurielles, différenciées par leur plausibilité. D'autres recherches sont en cours afin «d'enrichir et de faire évoluer les méthodes d'analyse et de traitement d'informations composites associant données quantitatives et qualitatives» en associant l'analyse statistique à l'analyse de contenu [Moscarola 1992].

Enfin, la phase de *configuration des connaissances* demeure à l'horizon de la recherche. Certes il existe de nombreuses plates-formes qui peuvent représenter des connaissances : les systèmes experts, les bases de données relationnelles, les bases de données orientées objet, les réseaux sémantiques, etc. Toutefois, ces dernières n'ont pas été conçues comme dépositaires des informations issues d'un texte, mais comme des matrices pour gérer des opérations ou modéliser un savoir. Ils peuvent être utilisés avec profit en relation avec l'analyse et la gestion textuelle mais ils doivent être réunis par des passerelles. Et dans ce secteur rien n'est automatique.

5 CONCLUSION

Ainsi donc la question du traitement électronique des documents textuels est complexe. Elle ne peut être ramenée au simple traitement du support matériel du texte. Par ailleurs, lorsqu'on touche au coeur du problème qui est l'accès cognitif au contenu des textes, on découvre que l'analyse et la gestion ne peuvent être encapsulées dans des processus automatiques. Ces opérations doivent plutôt être supportées par des systèmes informatiques sophistiqués qui assurent flexibilité et polyvalence pour respecter les projets de lecture. Malheureusement, il existe encore une barrière entre les divers systèmes de gestion et d'analyse électronique des documents. Cette étanchéité regrettable des logiciels les uns par rapport aux autres a d'ailleurs été soulignée à propos des tâches complexes de lecture et d'écriture qui "nécessitent la mise en oeuvre d'un grand nombre de nos facultés":

> Cette multiplicité se reflète dans le profusion des solutions informatiques proposées (traitements de textes, correcteurs, dictionnaires, analyseurs). Cependant, ces progiciels sont rarement pensés dans

> un cadre d'intégration. Tant que l'usager ne cherche qu'une aide ponctuelle pour effectuer une tâche spécialisée, il trouve généralement des systèmes adaptés à cette demande. C'est dans la mesure où un même usager requiert une aide globale pour effectuer un ensemble de tâches complexes de lecture et d'écriture que devient urgente leur intégration dans un cadre méthodologique complet. (Duchastel, 1991: 601)

La solution à ces questions réside essentiellement selon nous dans une vision intégrée de la chaîne de traitement et non pas uniquement dans l'intégration modulaire des logiciels. Seule une telle vision permettra de construire une plate-forme qui apporte assistance aux véritables opérations cognitives que les humains effectuent sur les textes.

RÉFÉRENCES ET BIBLIOGRAPHIE

Bar Hillel, Y. (1955). *An examination of Information Theory.* Philosophy of Science, 22, 86-105.

Barrett, E. (1985). *The Society of Text. Hypertext, Hypermedia, and the Social Construction of Information.* Cambridge, Mass.: MIT Press.

Belkin, N.J.; Marchetti, P.G.; Albrecht, M.; Fusco, L.; Skogvold, S.; Stokke, H.; Troina, G. (1991). *User interfaces for information systems.* Journal of Information Science; 17; 1991: 327-344.

Benmergui-Perez, M. (1988). *Charting the uncharted.* Office Equipments & Methods; November 1988: 26-29.

Bertrand-Gastaldy, Suzanne (1990). *L'indexation assistée par ordinateur: un moyen de satisfaire les besoins collectifs et individuels des utilisateurs de bases de données textuelles dans les organisations.* ICO Québec; intelligence artificielle et sciences cognitives au Québec; vol. 2, no. 3; septembre 1990: 71-91.

Brenner, J. S. (1990) *Acts of Meaning.* Cambridge, Mass. Harvard University Press.

Chevreau, J.; Kelly, T. (1989). *Paperless report.* Office Equipments & Methods; January-February 1989: 42-46.

Daoust, F. (1992). *Système d'Analyse de Texte par Ordinateur version 3.6, manuel de référence*, Centre ATO•CI, Université du Québec à Montréal.

Delphi Consulting Group.(1992) *Information Management: The Next Generation; Conferences and Seminars on Electronic Management Systems* ; 1992.

Duchastel, Jules (1991). *Pour une méthodologie d'aide à la lecture et à l'écriture.* Actes du coloque "Les industries de la langue: perspectives des années 1990, Montréal, 21-24 novembre 1990. [s.l.]: Office de la langue française / Société des traducteurs du Québec, 1991: 583- 601.

Eco, Umberto. *Lector in fabula ou la Coopération interprétative dans les textes narratifs*. Paris: Grasset; 1985.

Foucault, M. 1969 *L'Archéologie du savoir*, Paris, Gallimard.

Gadamer, H. G. (1976) *Vente et méthode*. Paris: Seuil; 1976.

Iser, W. (1976). *The Art of Reading A Theory of Esthetic response*, Baltimore 1976: John Hopkins University.

Karivalo, M.(1989). *Training for information management in a company.* Information Services & Use; 9; 1989: 341-346.

Kinstch, Walter (1977). *Memory and Technicians;* New-York, Wily Edition.

Meunier, J. G., Bertrand Gastaldy, S., & Lebel, H. (1987).*A Call for Enhanced representation of Content as a Means of Improving On line Full-Text Retrieval.* International Classicfication *14* (1), 2-10.

Meunier, J.-G. (1992). *SATO: un philologue électronique.* Documentation et bibliothèques; 38(2); avril-juin 1992: 65-69.

Meunier, J.-G., Bertrand-Gastaldy, S. et Lebel, H., (1987). *A call for enhanced representation of content as a means of improving on-line full-text retrieval.* International Classification, 14(1), 1987: 2-10.

Meunier, J.-G., Daoust, F., Rolland, S., (1976). "SATO: A System for Automatic Content Analysis of Text." Computer and the Humanities 10 (5): 281-287.

Moscarola, J. (1992). *L'Analyse de contenu et analyse de données assistés par ordinateur. Nouveaux outils et nouvelles pistes. Le projet ISIS*. Papier fourni par l'auteur.

Paquin, L.-C. et Daoust, F. (1993). *ACTE Atelier cognitif et textuel, version 1.0, manuel de référence*, Centre ATO•CI, Université du Québec à Montréal.

Paquin, L.-C. (1992). La lecture experte *Technologie, idéologie et pratique*, numéro spécial consacré au colloque "Intelligence artificielle et sciences sociales"; 10 (2-4): 209-222.

Paquin, L.-C.; Beauchemin, J. (1988). Apport de l'ordinateur à l'analyse des données textuelles. In: *RELAI: Recherche en linguistique aappliquée à l'informatique. Actes du colloque "La description des langues naturelles en vue d'applications informatiques"*. Université Laval, 7-9 décembre 1988. Québec: Centre international de recherche sur le bilinguisme; 1989: 197-210.

Pêcheux, M. (1972), *L'analyse autamatique du discours*, Paris, Maspéro.

Rada, R. (1991). *From Text to Expert text.* Mc Graw Hill.

Ranjard, S. (1991). L'indexation manuelle: une valeur ajoutée. *Archimag.* Hors série; novembre 1991.

RIAO 91 Conference Proceedings (1991). Intelligent Text and Image Handling, Universitat Autonòma de Barcelona, Barcelona, Spain, April 2-5, 1991. 2 vol.

Salton G., & Mc Gill, M. (1983). *Introduction to models of Information Retrieval,*. New York: Mc Graw Hill.

Schank R., &. Abelson A., R. (1977). *Scripts, Plans Goals and Understanding.* Hillsdale N.J: Laurence Erlbaum Associates.

Schmitz-Esser, Winfried (1990). Thesauri facing new challenges. *International Classification* ; 17 (3/4); 1990: 129-132.

Wilson, T.D (1984). *The cognitive approach to information-seeking behaviour and information use.* Social Science Information Studies; 4; 1984: 197-204.

AUTEURS:

Suzanne BERTRAND-GASTALDY, professeure
École de bibliothéconomie et des sciences de l'information
Université de Montréal
Case Postale 6128, Succ. A
Montréal, Québec
CANADA H3C 3J7
tél.: (514) 343-6048
fax : (514) 343-5753
GASTALDY@ERE.UMONTREAL.CA

Jean-Guy MEUNIER, directeur
Louis-Claude PAQUIN, chercheur
Centre ATO•CI
Université du Québec à Montréal
Case Postale 8888, Succ. A
Montreal, Québec
CANADA H3C 3P8
tél.: (514) 987-8256
fax : (514) 987-4567
R23325@UQAM.BITNET