La lecture experte[1]

Louis-Claude Paquin

Introduction

Malgré tout ce que l'on a pu dire de l'importance que prennent les médias alternatifs, il reste qu'une très grande partie de l'information utilisée, tant par les chercheurs universitaires, que par les agents oeuvrant dans les organisations[2] est toujours véhiculée par l'écrit. Alors que les objectifs de productivité sont de plus en plus élevés et que la masse des textes ne cesse de s'accroître, les ressources humaines hautement qualifiées qui peuvent être allouées à leur lecture sont limitées par des contraintes temporelles et financières. Tel est le contexte dans lequel s'inscrit le recours à l'ordinateur.

La présente contribution consiste à proposer un cadre computationnel pour l'analyse des textes qui puisse accommoder une grande diversité de points de vue, satisfaire les besoins véritables de ceux qui sont aux prises avec les textes et enfin couvrir l'ensemble du processus d'analyse. La réalisation de ces objectifs nécessite qu'un renversement de perspective sur l'analyse des textes soit opéré. Plutôt que chercher à mettre au point un analyseur général et exhaustif de l'ensemble des structures des textes, nous proposons d'implanter un modèle de la lecture particulière telle qu'effectuée par des experts sur un ensemble particulier de textes. Cette approche, désignée par l'appellation de «lecture experte», s'est imposée à l'occasion de projets de recherches[3], mais surtout de projets pilotes dans les

---

1 Paru dans *Technologie, idéologie et pratique*, numéro spécial consacré au colloque "Intelligence artificielle et siences sociales" Volume X no 2-4, 1992, pp.209-222.

2 Le terme «organisations» est utilisé ici dans une acception générique qui désigne le milieu de travail, autant celui de l'administration publique que celui des entreprises oeuvrant dans le domaine industriel ou des services.

3 D'abord une recherche pour déterminer les spécifications d'un Système d'Analyse de Contenu Assistée par Ordinateur (SACAO) [1988-1989] sous la direction de Jules Duchastel, département de sociologie UQAM, et financée

organisations[4]. Elle justifie pour une large part la commandite[5] du développement d'un Atelier cognitif et Textuel (ACTE).

---

dans le cadre des Actions spontanées (FCAR). Pour une description, voir : J. Duchastel, L. Dupuy et F. Daoust, "Système d'Analyse de Contenu Assistée par Ordinateur (SACAO)" *Actes du Colloque La description des langues naturelles en vue d'applications linguistiques*, Québec, Centre international de recherche sur le bilinguisme, 1988 : 197-210.

Ensuite une recherche ayant pour but l'*Élaboration d'un système expert (LOGE-EXPERT) expliquant la Loi de la Régie du Logement pour consultation par les citoyens dans une formule libre-service* [1988-] sous la direction de Claude Thomasset, Groupe recherche informatisation du droit (GIRD) UQAM; financée par la Fondation canadienne Donner, le Conseil de recherche en sciences humaines du Canada et le ministère québécois de l'Enseignement supérieur et science. Pour une description voir : C. Thomasset, L.-C. Paquin, et F. Blanchard, "Legal Knowledge Elicitation: from textual databases to expert systems", *DEXA '90 Internationnal Conference on Data Base and Expert Systems Applications*, Vienne, 1990 : 167-172.

Enfin une recherche qui vise la mise au point d'un *Système d'assistance à la classification, au tri et à l'indexation de textes de jurisprudence* [1991 -] sous la direction de Suzanne Bertrand-Gastaldy de l'École de bibliothéconomie et des sciences de l'information de l'Université de Montréal; financé conjointement par le Centre Francophone de Recherche pour l'Informatisation des Organisations (CEFRIO) et la Société québécoise d'information juridique (SOQUIJ).

4 Voici quelques projets pilotes qui ont eu lieu dans l'administration publique québécoise :

«Projet lexical et sémantiques de domaines» (PLSD) [1987-1988], sous la direction de R. Parent, Ministère des communications qui visait à sensibiliser différents intervenants de l'administration publique aux technologies de l'information.

«Système de gestion des évaluations environnementales» (SAGÉE) [1988-1990], sous la responsabilité de L. Valiquette, Ministère de l'environnement qui visait à assurer le suivi d'un avis de projet jusqu'à l'élaboration de la directive ministérielle. Pour un aperçu voir : L.-C. Paquin, L. Dupuy, et Y. Rochon "Analyse de texte et acquisition des connaissances : aspects méthodologiques", *Gestion de l'information textuelle ICO Québec* 2 (3), 1990 : 95-113.

«Système assisté pour la gestion de l'attribution des contrats» (SAGAC) [1988-1989], sous la responsabilité de M. Gingras, secrétariat du Conseil du Trésor du Québec qui visait à évaluer les systèmes experts comme moyen de diffusion de la Politique administrative gouvernementale.

«Système à base de connaissances pour assister l'entraînement des vérificateurs fiscaux» [1989-1990], sous la responsabilité de C. Boivin, Ministère du Revenu

Un examen de la problématique reliée à l'analyse des textes par ordinateur (ATO)[6] nous a amené à abandonner le découpage de la tâche comme moyen de réduire sa complexité au profit d'une focalisation sur la lecture en tant qu'expertise à modéliser dont nous dégageons les opérations fondamentales. Nous proposons ensuite un modèle de traitement, un modèle computationnel pour en faire l'implantation et une méthodologie de développement. Enfin, les fonctionnalités dont l'ACTE sera à terme doté sont décrites en regard de l'ATO.

Problématique

Le processus analytique est habituellement conçu en quatre étapes successives : le découpage des unités significatives du texte, la description de ces unités, l'extraction d'informations à partir des descriptions et l'interprétation des informations. On arrive généralement à effectuer au moyen de l'ordinateur, à des degrés divers d'automatisme et avec plus ou moins de succès, les trois premières étapes. Cependant, à notre connaissance, aucun système basé sur l'un et/ou l'autre des trois modèles de traitement suivants : statistique[7], linguistique[8] ou

---

Québec qui visait à déterminer la technologie la mieux adaptée au contexte de la formation en situation de travail. Pour un aperçu voir : L.-C. Paquin et C. Boivin, "Une approche technologique pour l'entraînement à la vérification fiscale" *La formation Intelligemment assistée ICO Québec* 3 (1), 1991 : 65 - 72.

5 La réalisation de l'ACTE est financée depuis 1990 par DELTA, un consortium de ministères et organismes du Québec sous la coordination de la Direction générale des technologies de l'information, Ministère des communications.

6 Pour un exposé systématique voir : J.-G. Meunier, "Le traitement et l'analyse informatiques des textes" *Gestion de l'information textuelle ICO Québec* 2 (3), 1990 : 9-18.

7 Voir L. Lebart, A. Salem, *Analyse statistique des données textuelles*, Paris, 1988.

8 Voir les analyses détaillées de D. Coulon et D. Kayser, "Informatique et langage naturel: présentation générale des méthodes d'interprétation des textes écrits", *Technique et science informatique* 5 (2), 1986, ainsi que de B.J. Grosz et al. *Readings in Natural Language Processing*, California, 1986.

associationniste[9], ne s'est attaqué avec succès à l'interprétation qui pourtant constitue l'étape cruciale de l'analyse.

Quelque soit le modèle de traitement privilégié, c'est la stratégie «étapiste» mise en oeuvre dans le développement des systèmes d'ATO qui, à notre avis, est déficiente. Étant donné les difficultés rencontrées à chacune des étapes, celles-ci deviennent des finalités. Le découpage du texte en mots[10] ne donne pas pour autant des unités significatives; celles-ci sont souvent composées[11] de plusieurs mots et leur délimitation peut donner lieu à des traitements élaborés[12]. La description linguistique des unités et de leurs relations à l'intérieur de la phrase est elle-même

---

9 Pour une introduction générale, voir entre autres: S. Grossberg, *Neural Networks and Natural Intelligence*, Cambridge Mass.,1988. Pour l'application aux informations de type linguistique voir : G. W. Cotrell, et S. L. Small, "A connextionnist scheme for modelling word sens disambiguation". *Cognition and Brain Theory* 6, 1983 : 89-120; D. Waltz, J. B. Pollack, "Massively Parrallel Parsing: A Stongly Interactive Model of Natural Language Intepretation". *Cognitive Science* 9, 1985 :.51-74.

10 Le terme «mot» est utilisé ici dans une acception informatique en tant que chaîne de caractères délimitée par des blancs.

11 Les expressions «traitement de textes» et «assemblée générale» sont des termes composés, aussi appelées dans d'autres contextes «segments répétés», «synapsies» ou encore «unités polylexicales».

12 Des traitements statistiques et linguistiques plus ou moins sophistiqués ont été mis à contribution pour leur dépistage. Pour un aperçu voir L.-C. Paquin, "Du terme au concept", *Actes du Colloque international 'Les industries de la langue: Perspectives des années 1990'*., Montréal, 1991 : 313-333. Plusieurs logiciels ont même été développés spécifiquement à cette fin : S. David et P. Plante, "Le progiciel TERMINO : de la nécessité d'une analyse morphosyntaxique pour le dépouillement terminologique", *Actes du Colloque international 'Les industries de la langue: Perspectives des années 1990'*., Montréal, 1991 : 71-88; D. Bourigault, *Système d'aide à la détection des groupes nominaux terminologiques apparaissant dans un corpus*, Direction des Études et Recherches, Électricité de France. Toutefois ces logiciels destinés à la terminologie, à l'ingénierie cognitive ou aux sciences documentaire offrent un faible à l'ATO comme telle. En dernière analyse, le problème s'avère parfois insoluble sans le recours à une connaissance étendue du domaine de référence. Voir S. Bertrand-Gastaldy, "L'indexation assistée par ordinateur : un moyen de satisfaire les besoins collectifs et individuels des utilisateurs de bases de données textuelles dans les organisation", *Gestion de l'information textuelle ICO Québec* 2 (3), 1990 : 79.

l'objet d'une stratégie «étapiste»[13] et chacune des étapes constitue en elle-même un champ de recherches.

Par ailleurs, la description syntaxique d'un texte, même si elle permet à terme d'accéder à une certaine sémantique relationnelle des unités, ne contribue qu'en partie à la description des niveaux supérieurs d'organisation du texte[14], tels la structure argumentative ou encore les «figures»[15], dont la prise en compte est pourtant essentielle à l'interprétation du texte. C'est qu'un texte n'est pas seulement un fait de langue mais aussi de discours. Ce dernier aspect fait appel à des connaissances externes au texte analysé, celle des autres textes et celle des conventions sociales qui en régissent la production[16]. Voilà pourquoi l'étiquetage contextuel de catégories autres que linguistiques aux unités significatives est si difficile à réaliser à l'aide de l'ordinateur.

Plutôt que de réduire la complexité du processus analytique en le découpant en étapes et de consacrer des efforts théoriques et computationnels à la résolution automatique de tous les problèmes qui se posent pour chacune des étapes, nous

---

13 Une série de descriptions linguistiques sont réalisées successivement à partir du résultat des précédentes. D'abord on procède à une description morphologique hors contexte de chacun des mots à l'aide d'un dictionnaire. Dans le cas où plusieurs morphologies sont possibles, une description en contexte est nécessaire pour déterminer, à l'aide des mots voisins, laquelle des formes possibles est effectivement réalisée. Puis, une description syntaxique vient préciser les relations qu'entretiennent les mots à l'intérieur des propositions et les relations entre les propositions à l'intérieur de la phrase. Voir D. Coulon et D. Kayser, *art. cit.* : 107. Cette stratégie «étapiste» repose sur les postulats qu'un texte est composé d'un ensemble de structures dont les principes d'organisation peuvent être explicités que, quoique intereliées, ces structures sont suffisamment distinctes pour être décrites les unes après les autres. Or rien n'est moins sûr, la désambiguation d'une catégorie morphologique exige parfois une analyse syntaxique complète ou encore le recours à un contexte étendu.

14 Certains linguistes font l'hypothèse que les niveaux supérieurs de l'organisation textuelle, tels les paragraphes, qu'ils répondent à des lois similaires à celles de la phrase. Voir entre autres M. A. K. Halliday, *An Introduction to Functional Grammar*, London, 1987.

15 Voir, J. Dubois *et al*., *Rhétorique générale*, Paris, 1970.

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

proposons de focaliser sur les besoins et les pratiques de ceux qui effectuent l'analyse des textes, les lecteurs. Au lieu de chercher à mettre au point un algorithme général d'ATO, nous mettons l'accent sur la modélisation d'une lecture particulière des textes et le passage d'une lecture humaine vers une lecture machine.

La lecture en tant qu'expertise

La lecture nous apparaît relever plus d'un savoir-faire plus ou moins implicite[17], que d'un savoir exact formalisé dans une théorie. Il s'agit d'une «expertise», acquise non pas tant par apprentissage mais au fil d'une pratique[18]. De plus, cette pratique constitue rarement une fin en elle-même, est partie intégrante d'une activité professionnelle, comme par exemple, déterminer l'admissibilité d'un dossier ou encore à analyser des récits de vie, etc. Nous explorons un modèle empirique de la lecture pour en dégager les composantes opérationnelles mises en oeuvre par les experts de domaine. Dans cette perspective, le lecteur expert effectue sur les textes quatre opérations fondamentales dont la complexité est croissante : segmenter, filtrer, déchiffrer et interpréter.

La segmentation consiste à découper la suite des caractères du texte en mots significatifs. Tous les mots d'un texte ne sont que rarement pris en compte par le lecteur. La plupart du temps, un filtrage plus ou moins sévère est effectué en vertu de ses objectifs, de sa connaissance du domaine de référence, etc. Sa lecture se trouvera d'autant accélérée que le filtrage sera sévère. Toute l'information nécessaire à sa compréhension n'étant pas

---

17 «De manière générale, toute personne pratiquant intensément une activité acquiert des automatismes. Cette automatisation apporte une plus grande efficacité à l'expert, qui agit alors plus au niveau réflexe qu'au niveau conscient(...). Mais l'automatisation limite aussi l'expert lors d'un accès conscient à l'activité. Ainsi, l'expert a du mal à expliquer ce qu'il fait (il possède un «savoir-faire» mais n'a pas de disposition pour le «faire savoir»). J.-F. Gallouin, "Systèmes experts et psychologie cognitive", *Micro-Systèmes*, décembre 1988 : 159.

18 L'apprentissage d'une langue étrangère est la meilleure façon de vérifier empiriquement cette assertion.

repérable à la surface du texte[19], le lecteur procède au déchiffrement d'indices lui donnant accès à l'information ou encore projetant un éclairage particulier sur celle-ci.

Par ailleurs, déchiffrer c'est aussi prédire dans la mesure où, même sans la co-présence de tous les indices, le lecteur peut quand même accéder à l'information du texte. Son expertise lui permet de «lire entre les lignes», de combler l'absence d'indices ou encore de discriminer la relation probable entre les indices de celle qui ne l'est pas dans des configurations ambiguës. Il utilise alors les indices déjà relevés au cours de la lecture ou fait appel à des lectures précédentes. Dans le cadre des théories cognitives, on dirait que le lecteur a en tête des «prototypes» qu'il retrouve dans les textes dans un état plus ou moins complet[20].

Quant à l'interprétation, dans la perspective opérationnelle qui est la nôtre, c'est une relation que fait le lecteur entre l'information obtenue par le dépistage des indices et la connaissance qu'il a du monde de référence.

Le modèle de traitement

Le modèle de traitement qui a été élaboré pour prendre en compte les quatre opérations de la lecture dégagées précédemment tente de répondre aux contraintes suivantes : demeurer assez simple pour permettre une implantation sur micro-ordinateur dotée d'une couverture satisfaisante et être assez général pour demeurer unique malgré des expertises de lecture différentes[21]. Chacune de ces opérations est en quelque sorte une transformation effectuée sur un texte d'origine et dont le résultat est un autre texte[22].

---

19 D'où l'échec de l'ATO d'allégeance strictement statistique.

20 Pour une vue d'ensemble de l'application de la notion de «prototype» aux sciences du langage, voir : J. R. Taylor, Linguistic Categorization, *Prototypes in Linguistic Theory*, Oxford, 1989.

21 Comme par exemple l'expertise des conseillers juridiques, celle des aviseurs agricoles, celle des psychanalystes, des sociologues, des herméneutes, etc.

22 Cette façon d'exprimer la compositionnalité des opérations est empruntée à J.-G. Meunier, *art. cit.* : 15-16.

Avant même d'effectuer la segmentation du texte en unités significatives, tout comme le lecteur devant un mot nouveau consulterait le dictionnaire, le premier traitement consiste à projeter sur les mots du texte des descriptions pertinentes. Cette information catégorielle, appelée «indices» peut être assignée, soit hors contexte à toutes les occurrences d'un mot donné, soit à une occurrence donnée dans un contexte particulier. Selon les besoins du modèle de lecture, l'information pourra être de nature linguistique, relative à la morphologie, au lemme; à la syntaxe, etc.[23] L'information pourra aussi être la distribution du mot au travers du texte, du corpus et/ou d'une partition de celui-ci; l'information sera enfin relative au domaine d'expertise. Cette information a assurément nécessité la constitution d'immenses dictionnaires électroniques et/ou la réalisation d'analyses très complexes, et ce dans des disciplines et des cadres théoriques variés afin de réaliser des objectifs particuliers, souvent différents de celui qui leur est réservé ici. C'est pourquoi le présent modèle de traitement peut être qualifié d'intégrateur.

La segmentation en unités pertinentes sera réalisée par des regroupements de mots faits à partir des indices déposés sur les mots[24] et selon des critères explicités lors de la constitution du

---

[23] À titre d'illustration, les dictionnaires et analyseurs développés au centre d'ATO sont mentionnés ici :

Une base de données lexicales (BDL) a été développée par Luc Dupuy pour fournir hors-contexte des informations morpho-syntaxiques; elle regroupe présentement 358,820 mots du français écrit en une quinzaine de collections d'unités lexicales, telles substantifs, verbes à l'infinitif, verbes conjugués, adjectifs qualificatifs, pronoms, conjonctions, prépositions, adverbes, déterminants.

Un progiciel pour la lemmatisation et la caractérisation morphologique du français (LCMF) a été développé par Lucie Dumas en collaboration avec P. Plante, D. Perras et A. Plante. Pour une description voir : S. David et P. Plante, *art. cit.* : 82-83.

Un analyseur lexico-syntaxique du français (ALSF) est en développement sous la responsabilité de J.-M. Marandin, S. David et P. Plante.

Toutefois, l'implantation du modèle de traitement permettra l'intégration de descriptions linguistiques provenant d'autres horizons via un protocole.

[24] Une procédure appelée MARQUELO mise au point au Centre d'ATO permet rapidement et sur une grande échelle de dépister des termes qui sont composés

modèle de lecture experte. Il sera aussi possible de bénéficier de l'apport de logiciels développés à cette fin[25] via un protocole. Tout comme les mots, les segments pourront être l'objet d'une description.

L'opération de déchiffrement consiste à mettre en relation un ou plusieurs indices provenant des descriptions effectuées préalablement avec une hypothèse de représentation (HR)[26]. L'HR constitue en quelque sorte une variable intermédiaire entre les indices et leur interprétation. La catégorisation des unités textuelles permet de réduire sensiblement la diversité des formes en présence et, par conséquent, le nombre de configurations différentes mises en relation avec une HR donnée afin de couvrir l'ensemble des cas d'espèce.

L'opération d'interprétation c'est la mise en relation d'une ou plusieurs HR avec un élément extérieur à l'univers textuel, appelé «interprétation». Cette relation n'est pas de nature formelle comme la précédente mais subjective. Ainsi deux interprétations différentes peuvent être données à une même HR, alors qu'une seule HR devrait être assignée à une configuration donnée d'indices.

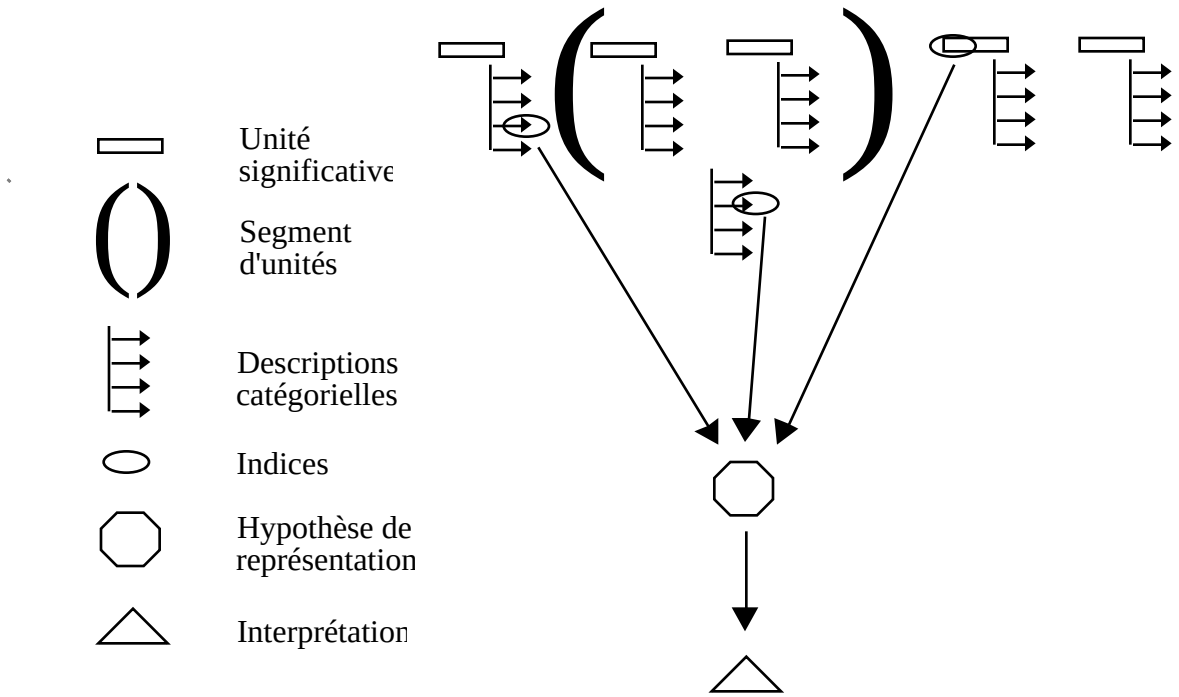
Voici une schématisation du modèle de traitement proposé :

---

de plusieurs mots et dont la formation répond à une régularité. Elle consiste en la projection de filtres sur l'information morphologique fournie par les BDL, comme par exemple : non + de + nom. Pour une description voir : L.-C. Paquin, L. Dupuy, et Y. Rochon, *art. cit.* : 102-105.

25 Tel le progiciel TERMINO développé au centre d'ATO. Pour une description, voir J. Perron, "Présentation du progiciel de dépouillement terminologique assisté par ordinateur : TERMINO" *Actes du Colloque international 'Les industries de la langue: Perspectives des années 1990'.*, Montréal, 1991 : 715-755.

26 Ainsi, par exemple, la phrase suivante : «Ce barrage provoquera une augmentation du débit d'eau de la rivière Rouge». Une fois le segment «débit d'eau» bloqué puisqu'il désigne un seul concept dans le domaine des sciences environnementales, l'hypothèse de représentation (HR) «impact» en vertu du déchiffrement des indices suivants obtenus à partir d'un dictionnaire : barrage est une «activité»; l'augmentation est un «changement d'état»; le débit d'eau est une «composante» et enfin la rivière Rouge est un «milieu naturel».



Le modèle de traitement permet la récursion : les interprétations devenant à leur tour les indices d'une méta-lecture, etc.

Le modèle computationnel

L'implantation du modèle de traitement exposé précédemment fait appel à deux modèles computationnels différents. Un premier modèle pour la description des mots et des segments de mots et un second modèle pour la mise en relation successive des indices avec les HR et des HR avec une interprétation.

La description des mots et des segments est effectuée à l'aide du système d'analyse de textes par ordinateur SATO[27]. Ce système offre un cadre interactif pour l'annotation, le filtrage et le décompte des mots de très grands corpus de textes. On peut adjoindre des propriétés aux mots dont les valeurs seront numériques ou symboliques. «Lexicales», les propriétés affectent toutes les occurrences d'un mot; «textuelles» elles en affectent une occurrence particulière. Les annotations peuvent être faites manuellement ou systématiquement, soit par la projection d'un dictionnaire ou lors de la réalisation d'un filtre arbitrairement complexe. Le filtrage s'effectue à l'aide d'un langage

[27] Développé par François Daoust du Centre d'ATO.

d'interrogation complet qui s'applique sur le lexique ou sur le texte lui-même. Une propriété peut hériter des valeurs d'une propriété mère. Un mécanisme de répartition peut, pour les segments de texte désignés, fusionner les valeurs d'une propriété symbolique de tous les mots de ce segment et les rapporter sur chacun d'eux. Enfin, SATO permet de produire des statistiques sur la distribution des éléments lexicaux à travers l'ensemble du corpus ou entre diverses partitions de celui-ci.

Le modèle computationnel qui nous apparaît convenir le mieux pour implanter les mises en relation est celui des règles de production[28]. Parce qu'elle repose sur les propriétés de la logique des propositions, cette représentation des instructions permet la compositionnalité et un accès systématique à de nouvelles informations sans compromis quant à leur modularité. Les systèmes à base de connaissances (SBC), de par leur architecture favorisent le développement et l'exploitation de grands ensembles de règles. Les règles, tout comme les données pour les systèmes conventionnels, sont tenues séparées du dispositif computationnel appelé «moteur d'inférences» qui a pour seule fonction la validation de leur «prémisse» en regard de «faits» et, le cas échéant, l'exécution de leurs «actions» consistant en la production de nouveaux «faits» qui permettront la validation d'autres règles, etc. Pour implanter le modèle de traitement, il suffit de traduire en «règles d'inférences» les relations établies lors de la modélisation. Les indices réunis par une conjonction ou une disjonction sont mis en relation avec une hypothèse de représentation (HR)[29]. Il en va de même pour HR et l'interprétation correspondante.

Une lecture experte s'accomplirait de la façon suivante. D'abord les descriptions qui fourniront les indices sont effectuées et projetées sur les mots du texte. Ensuite, une à une, des entités logiques du texte, telles le paragraphe, l'article, le chapitre, etc.

---

[28] Voir entre autres A. Newell, "Physical symbols systems". *Cognitive Science* 4(2) (1983) : 135-183.

[29] Dans l'exemple précédent la règle d'inférences serait formulée ainsi :

Si «activité» **et** «changement d'état» **et** «composante» **et** «milieu naturel»

Alors «impact»

sont constituées en faits de départ et soumis au «moteur d'inférences». Une chaîne inférentielle est alors initiée. Les règles qui testent les indices en présence sont une à une validées et, le cas échéant, les HR sont «inférées», c'est-à-dire qu'elles acquièrent le statut de «faits». Au cycle d'inférences suivant, les règles qui testent les HR affirmées sont validées et, le cas échéant, des interprétations sont à leur tour «inférées» et sont consignées dans un «rapport». Plus tard, les éléments du rapport peuvent, à leur tour, constituer les indices d'une méta-analyse et ainsi de suite.

De plus, en ayant recours à un atelier logiciel pour générer un SBC, un non-informaticien peut lui-même jouer le rôle de développeur en modélisant sa propre expertise de lecture. De tels ateliers logiciels, aussi appelés «générateurs» offrent en plus d'un moteur d'inférences, un éditeur des règles d'inférences et des utilitaires de validation pour s'assurer de la cohérence des déclarations.

Contrairement à la programmation conventionnelle où tout changement en cours de route dans le modèle est pénalisant, la phase d'implantation dans un SBC, participe de la démarche exploratoire. Ainsi, les SBC favorisent le prototypage, c'est-à-dire une implantation schématique qui, par essai et erreur sera raffinée. Il est alors possible de construire graduellement et de façon modulaire des systèmes de plus en plus complexes. Les SBC présentent un autre avantage au niveau de l'efficacité : contrairement aux implantations sous la forme procédurale d'automates où pour toutes les analyses possibles, toutes les conditions sont testées, ce sont les indices qui déclenchent les analyses pertinentes.

Les coefficients de certitude

La capacité de prédiction du lecteur qui lui permet de déchiffrer l'information même en l'absence d'indices pourtant nécessaires est implantée dans notre modèle de la lecture experte au moyen

de coefficients numériques. Les générateurs de SBC offrent généralement la possibilité d'adjoindre aux conclusions des règles d'inférences un coefficient afin que le développeur du système puisse exprimer la confiance qu'il prête à ses associations, ici d'indices avec l'HR correspondante. Le moteur d'inférences incorpore un mode de cumul de ces coefficients pour propager l'incertitude ou encore la confiance tout au long de la chaîne inférentielle. Les formules utilisées[30] pour cumuler ces coefficients selon la conjoncture provoquent, soit leur atténuation, soit leur renforcement; dans le cadre de la lecture experte ces opérations présentent les avantages suivants.

D'une part, l'association d'une configuration particulière d'indices peut être associée à une HR donnée malgré l'absence de certains indices, pourtant jugés nécessaires lors de l'élaboration du modèle. Le système fera quand même l'association mais avec un coefficient de certitude moindre qui, à son tour, provoquera une atténuation de la certitude lors de l'association subséquente de l'HR avec l'interprétation correspondante. L'utilisation des coefficients permet un dépassement du cadre strict de la logique booléenne. En effet comme le modèle prévoyait associer une HR à une conjonction d'indices, si l'un d'eux venait à manquer, sans le recours aux coefficients, l'association n'aurait pu être établie.

D'autre part, il n'est plus besoin d'inclure dans une seule règle d'inférences tous les indices dont la conjonction est associée à une HR donnée. Les indices peuvent être associés individuellement à l'HR avec un faible coefficient qui sera renforcé au fur et à mesure du déclenchement des règles filtrant les autres indices. L'utilisation des coefficients permet une réduction de la complexité des règles nécessaires pour implanter un modèle donné et permet le dépassement de l'unicité pour déboucher sur des associations plurielles mais différenciées. Ainsi, par exemple, il est possible dans une même règle d'exprimer qu'un même indice dénote très faiblement une HR donnée, car la co-présence de plusieurs indices est nécessaire, et que ce même indice dénote moyennement une autre HR.

---

30 Voir B. G. Buchanan et E.H. Shortliffe, *Rule-Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project.* Reading, MA.,1984.

Le développement des systèmes de lecture experte

Cette nouvelle façon de concevoir l'ATO entraîne une redistribution des rôles dans le développement des systèmes. Plutôt que la mise au point en vase clos par des spécialistes d'un programme dont seuls les résultats sont présentés aux utilisateurs, ceux-ci sont appelés à intervenir plusieurs fois dans le cycle de développement. D'abord pour effectuer un transfert de leur expertise de lecture dans le modèle de traitement et ensuite pour valider les résultats générés par le programme en fonction de leurs besoins. Cette validation entraînera la révision ou l'adjonction de nouvelles descriptions et/ou règles d'inférences au système qui sera validé à nouveau et ainsi de suite jusqu'à ce qu'il ne soit plus rentable d'effectuer de nouveaux cycles de développement. La rentabilité est décroissante lorsque les ajouts ne concernent que des exceptions dont la fréquence d'occurrence est marginale par rapport à la majorité des cas traités.

Il ne s'agit plus de modéliser les multiples aspects ou dimensions des textes, mais de modéliser une lecture particulière qui en est faite. Les seuls aspects retenus seront ceux qui sont pertinents pour une pratique de lecture donnée, ce qui facilite la tâche de la modélisation. De plus, des facteurs extra-textuels devront être pris en considération, tels les objectifs du lecteur, ses particularités «culturelles», ce qu'il entend faire des résultats obtenus, etc. Toutefois, le lecteur, expert d'un domaine, éprouve le plus souvent de la difficulté à structurer de lui-même un schéma pédagogique des indices qui lui permettent d'interpréter le texte. C'est ainsi qu'il nous apparaît opportun d'avoir recours à des techniques de transfert de connaissances et de modélisation empruntées à l'ingénierie de la connaissance[31]. Cette discipline vise à produire des spécifications complètes et claires permettant de développer un système expert lorsqu'un domaine d'expertise n'a pas encore été modélisé informatiquement. Pour ce faire, un modèle de l'expertise est élaboré à partir de l'analyse des verbalisations effectuées par les experts qui doivent subir un certain entraînement pour arriver à décrire leur expertise. Ce

[31] Pour une revue voir J. Boose et B. Gaines, *The Foundations of Knowledge Acquisition*, New York, 1990.

modèle sera par la suite traduit dans les termes du modèle de traitement puis implémenté et soumis aux experts pour validation.

Les protocoles[32] sont des verbalisations effectuées par l'expert en situation de résolution de problème; elles servent à capter leur performance effective. On demandera aux lecteurs de dire à haute voix et sans censure tout ce qui leur passe par la tête au moment où il exercent leur expertise de lecture sur les textes. Chacune des étapes ou opérations, même si elles apparaissent insignifiantes doivent être mentionnées. Les entrevues sont des échanges où les experts, en dehors du contexte de leur activité professionnelle, fournissent une description générale du domaine. L'entrevue n'offrant pas de garantie d'objectivité; l'information recueillie doit être recoupée avec le contenu des protocoles.

Parallèlement à l'analyse des verbalisation, les concepteurs du système de lecture experte tireront profit d'une analyse de contenu[33] d'un sous-corpus de textes représentatif du domaine de référence. En plus de leur fournir des questions pertinentes à poser aux experts lors des entrevues, cette démarche leur permettra d'évaluer la faisabilité d'une modélisation et de valider les différents aspects du modèle envisagé.

Le cas échéant, une analyse des textes qui résultent de la lecture par les experts doit aussi être effectuée. La comparaison des textes produits par les lecteurs avec les textes d'origine fournit habituellement de l'information très utile sur la lecture qui a été faite. Ainsi, par exemple, dans le cas d'un résumé, la comparaison permettra d'identifier ce qui a été sélectionné. Plus tard l'entrevue fournira les critères et la justification de la sélection.

Un atelier cognitif et textuel (ACTE)

Un atelier cognitif et textuel (ACTE) qui intègre dans un même environnement informatique SATO et un générateur de SBC est

---

32 Voir A.K. Ericsson et H.A. Simon, *Protocol Analysis, Verbal Reports as Data*. Cambridge MA., 1984.

33 Voir entre autres les ouvrages méthodologiques suivants : L. Bardin, *L'analyse de contenu*, Paris, 1977; R. Ghiglione *Manuel d'analyse de contenu*, Paris, 1980.

présentement en cours de développement au Centre d'ATO. La lecture experte sera implantée et expérimentée dans l'ACTE, dans son état actuel de développement. Présentement le moteur d'inférences du générateur de SBC est complété, de même qu'une version partielle d'un éditeur de définitions des données cognitives et des règles d'inférences. Les données cognitives qui servent à formuler les règles d'inférences prennent la forme d'objets valués, ce qui permet de traiter directement des unités textuelles préalablement décrites en SATO.

À terme, les commandes de SATO pourront être exécutées en conclusion des règles d'inférences. Ainsi, le filtrage et l'annotation des textes pourra se faire à l'intérieur même d'une chaîne inférentielle ce qui autorise des structures de contrôle de description et d'analyse des textes complexes et fines tout en demeurant lisibles. De plus, le résultat des fouilles et des analyses de SATO pourra être admissible au filtrage par les prémisses des règles d'inférences.

Du point de vue de l'analyse des textes, l'ACTE permettra le dépassement de l'approche séquentielle habituelle pour expérimenter, en temps réel et sur des corpus d'envergure, une approche topologique[34]. Le texte pourra être conçu, non plus seulement comme une suite linéaire de séquences mais comme un espace multidimensionnel. Les régularités et les ruptures, tant formelles que sémantiques pourront être traitées en réseaux multiples qui s'enchevêtrent à l'intérieur de l'espace textuel. Ce dispositif computationnel, par ses capacités de filtrage et d'étiquetage, permet, à partir de n'importe quelle unité lexicale ou segmentale, d'examiner, de catégoriser ou re-catégoriser n'importe quelle autre unité, en aval ou en amont de l'unité qui est analysée.

Du point de vue de la théories des SBC, cette intégration permet un élargissement de l'espace de filtrage des règles d'inférences dont la structure des unités est présentement contrainte par un

[34] Ce dispositif permettrait l'implantation d'analyseurs topologiques tels que proposés dans la foulée des travaux de René Thom par J. Petitot-Cocorda, *Morphogénèse du Sens*, Paris, 1985.

modèle formel, les *frames*[35], les prédicats du premier ordre, etc. aux unités textuelles en tant que tel.

Conclusion

Nous croyons que la lecture experte a le potentiel de pallier au manque de dispositifs computationnels pour assister, autant les agents des organisations que les chercheurs en sciences humaines dans les phases de l'analyse des textes consécutives à leur catégorisation, soit le déchiffrement d'indices et leur interprétation. Dans la mesure où le déploiement du modèle de la lecture humaine sur le modèle de traitement sera possible, la lecture experte permettra le passage du cas par cas à la formulation de règles, ce qui garantirait l'uniformité et l'exhaustivité du processus interprétatif.

Des expérimentations avec l'ACTE seront menées afin de valider le cadre méthodologique et le modèle de traitement proposés pour la modélisation et l'implantation de la lecture experte. Les domaines d'application, de même que les types de texte et les perspectives de lecture seront aussi variés que possible[36]. Nous serons alors en mesure de délimiter le champ optimal d'application de la lecture experte, ses limites, de même que l'ampleur des efforts requis pour son implantation.

---

35 Voir M. Minsky. "A Framework for Representing Knowledge". dans Winston, P.H. *The Psychology of Computer Vision,* 1975 : 211-277.

36 Dans une perspective administrative : des mémoires déposés lors d'audiences publiques, les articles de la politique administrative; dans une perspective d'analyse du discours : des articles de quotidiens, des conférences constitutionnelles : dans une perspective documentaire : les décisions des tribunaux; dans une perspective littéraire : des nouvelles, etc.