LE TRAITEMENT DES TEXTES PRIMAIRES ET SECONDAIRES POUR LA CONCEPTION ET LE FONCTIONNEMENT D'UN PROTOTYPE DE SYSTÈME EXPERT D'AIDE À L'ANALYSE DES JUGEMENTS

Par

Suzanne Bertrand-Gastaldy[a,b],Louis-Claude Paquin[b],Gracia Pagola[a], François Daoust[b]

a École de bibliothéconomie et des sciences de l'information, Université de Montréal
b Centre de recherche en information et cognition ATO.CI, Université du Québec à Montréal

RÉSUMÉ

Afin d'assister les conseillers juridiques de la Société Québécoise d'Information Juridique (SOQUIJ), un prototype de système expert pour l'aide à la sélection, à la classification, à la lecture et à l'indexation des jugements a été implanté sur ACTE (Atelier Cognitif et TExtuel) développé au Centre ATO.CI. À partir d'un corpus d'apprentissage de textes déjà analysés et grâce à des traitements statistico-linguistiques sur SATO (système d'analyse de textes par ordinateur) et SPSS, on a confronté les données issues de l'analyse humaine à celles des textes intégraux; comment des tendances et des anomalies ont pu être décelées qui ont servi à questionner les outils et les pratiques ainsi qu'à réorienter ou à corroborer l'enquête cognitive des savoir-faire. Une fois identifiés les types d'unités linguistiques et leurs propriétés généralement retenus par les spécialistes pour chacune des opérations d'analyse, on a mis au point une chaîne de traitements qui, pour chacun des modules du système expert et à partir du plus grand nombre possible de sources de connaissances, dépiste et décrit les indices pertinents, puis les transforme en faits. Ceux-ci s'avérant distincts les uns des autres, un diagnostic peut être porté, à chaque étape, selon un principe de convergence. Toutefois, quelques difficultés concernant le cumul des coefficients de certitude et l'intégration de statistiques nécessitent des études plus poussées.

INTRODUCTION

Au Québec, la cueillette, la sélection, le traitement et la diffusion de la jurisprudence sont sous la responsabilité principale d'un organisme parapublic**:** SOQUIJ. La loi constituant la Société québécoise d'information juridique, entrée en vigueur le 1er avril 1976, lui confie le mandat de "promouvoir la recherche, le traitement et le développement de l'information juridique en vue d'en améliorer la qualité et l'accessibilité au profit de la collectivité." À ce titre, SOQUIJ est le serveur des

bases de données du ministère de la Justice ainsi que le producteur-serveur de plusieurs autres bases de données, dont celles qui concernent la jurisprudence.

Or, la saisie électronique des jugements à la source mise en place progressivement par le ministère de la Justice du Québec multipliera presque par cinq le nombre de jugements acheminés à SOQUIJ: la quantité annuelle prévue est de 50 000. Pour maintenir son niveau de service sans accroître indûment son personnel, cet organisme a envisagé de recourir à des méthodes automatiques pour assister certaines des opérations intellectuelles d'analyse effectuées par des conseillers juridiques et a confié à notre équipe de recherche le mandat de concevoir un prototype de système expert. Celui-ci a été réalisé sur le logiciel ACTE (Atelier Cognitif et TExtuel) développé au Centre ATO.CI.

Après avoir expliqué en quoi consistent les différentes fonctions d'analyse qu'il nous a fallu modéliser, nous évoquerons les éléments théoriques sur lesquels nous avons appuyé notre démarche et nous montrerons comment celle-ci combine des approches complémentaires (statistiques, linguistiques et cognitives); nous l'illustrerons ensuite par quelques exemples d'indices extraits pour chaque type d'analyse. Puis nous mentionnerons les principaux enrichissements du thésaurus nécessités par les nouvelles fonctions qu'il est appelé à remplir dans un système automatique. Finalement, nous exposerons l'implantation des stratégies d'analyse dans un programme de chaîne de traitement et l'intégration d'informations de sources et de valeurs différentes.

LA MODÉLISATION DES TACHES D'ANALYSE

Les fonctions d'analyse à assister

Après entente avec les représentants de SOQUIJ, nous avons convenu de concevoir un système qui assisterait les tâches suivantes: 1) élimination à la source de certains jugements (étape de la sélection); 2) détermination du (ou des) domaine(s) du droit et, le cas échéant, du sous-domaine (selon un plan de classification préétabli) auquel chaque décision retenue appartient (étape du tri et de la classification); 3) prise de connaissance du contenu des textes en vue de la rédaction d'un résumé informatif; 4) sélection de termes d'indexation à partir du résumé rédigé par les conseillers juridiques.[1]

Théories sous-jacentes

La méthodologie que nous avons élaborée est sous-tendue par au moins trois orientations théoriques: le texte comme objet sémiotique, les analyses documentaires comme des applications particulières d'un processus de lecture et l'intertextualité.

Le texte comme objet sémiotique

Le texte est envisagé comme un objet sémiotique complexe dans lequel un lecteur humain ou informatique sélectionne, à des niveaux multiples, des indices pertinents en fonction de ses objectifs d'analyse (Meunier, 1992; Meunier *et al.*, 1994). Les chaînes de caractères ou «mots» sont autant de porteurs de traits signifiants. Les processus cognitifs d'interprétation humaine étant fonction de nombreux éléments dont la plupart ne sont guère formalisables (systèmes de croyance, intentions, connaissances du contexte textuel et extra-textuel, etc.), un système entièrement automatique est impossible à envisager: seul un mécanisme d'aide à l'interprétation est réalisable qui permet d'identifier et de manipuler certains des indices pertinents décelables par divers analyseurs.

Les analyses documentaires comme des applications particulières d'un processus de lecture

Les opérations d'analyses effectuées dans un service documentaire sont envisagées comme des lectures particulières dirigées par des tâches spécifiques à accomplir: attribution d'une rubrique de classification, assignation de mots-clés, condensation du texte, entre autres. Ces lectures mettent en jeu diverses opérations cognitives de sélection, rejet, généralisation (Van Dijk, 1977), stratégies de confirmation et contrôle, etc. (David, 1990) portant sur des indices ou configurations d'indices dont la pertinence varie en fonction de chaque type de lecture. À chaque tâche d'analyse correspond donc un parcours particulier du texte. La décision d'inclure un document dans une base de données - ou de le rejeter - n'exige pas la prise en compte du même nombre ni des mêmes types d'indices que l'opération d'indexation. La rédaction d'un résumé requiert une prise de connaissance plus approfondie du contenu textuel que l'attribution d'une rubrique de classification, mais exige un examen moins attentif cependant que la comparaison des thèses défendues par un texte avec celles d'un autre texte.

L'intertextualité

De par la nature même de leur condition de production, les textes secondaires sont en relation d'intertextualité avec les textes primaires dont ils sont issus (Beacco et Darot, 1984) ainsi qu'avec les outils documentaires - thésaurus et plan de classification - servant à effectuer l'analyse (Begthol, 1986). Comme nous l'avons exposé dans Bertrand-Gastaldy (1993), la comparaison des propriétés des éléments présents dans les textes de départ et retenus dans les différents textes d'arrivée

(rubriques de classification, termes d'indexation, résumés) avec celles des éléments qui ont été éliminés permet de découvrir des tendances et des anomalies qui servent à orienter ou approfondir l'enquête cognitive auprès des experts.

Notre approche

Nous n'avons donc pas cherché à mettre au point un outil d'analyse qui serait performant en dehors de tout contexte (par exemple un analyseur morphologique, un extracteur de lexies complexes), mais bien au contraire de comprendre en quoi le contexte de la tâche faisait varier les objets textuels et les propriétés des objets susceptibles de retenir l'attention des experts. Notre approche a dès lors consisté d'une part à modéliser les stratégies cognitives mises en oeuvre par les experts du domaine lors des différentes lectures effectuées en fonction des produits attendus (liste des documents à éliminer, tri et classification, résumé, indexation), d'autre part à faire évoluer les outils documentaires pour les rendre aptes à répondre à l'utilisation automatique que nous voulions en faire.

Les sources de données

Pour mener à bien notre travail, nous disposions de deux types de sources. Nous avions accès à une demi-douzaine de conseillers juridiques avec lesquels nous avons tenu plusieurs sessions de travail afin d'arriver à connaître les critères explicites ou implicites auxquels ils recourent pour prendre leurs décisions aux différentes étapes de leur analyse. D'autre part, les données de nature linguistique se trouvaient déjà presque toutes sur support informatique. Il s'agit des produits issus des différentes opérations d'analyse: textes intégraux rejetés ou retenus, notices bibliographiques accompagnées des résumés, index, ainsi que des outils utilisés pour l'analyse: plan de classification et thésaurus.

Les traitements sur les données linguistiques

Les caractéristiques attribuées aux données, en contexte ou hors contexte, ont consisté en l'ajout d'informations de nature diverse décrivant le statut sémiotique des constituants du texte et enrichissant les chaînes de caractères immédiatement accessibles à l'ordinateur. Ces caractéristiques proviennent de connaissances générales de la langue (type de langue, nature grammaticale des lexèmes), de connaissances générales sur la structure des textes (phrases, paragraphes), d'informations de nature éditique (conventions typographiques -- capitales, caractères gras ou italiques -- dans les enregistrements), de connaissances spécifiques au domaine (vocabulaire de spécialité, structure des jugements et de leurs résumés, mention de loi, de jurisprudence et de doctrine), de connaissances "documentaires" (champs d'une notice, appartenance ou non des lexèmes aux langages documentaires), de propriétés statistiques (fréquence absolue ou relative, indice de répartition, valeur discriminante, chi 2, etc.). Ces informations ont été obtenues par des

algorithmes développés avec le logiciel SATO (système d'analyse et textes par ordinateur) et ont fait l'objet d'un marquage approprié (propriété et valeur de propriétés dans SATO). On pourra consulter une publication du Centre ATO.CI pour plus de détails (Bertrand-Gastaldy *et al.* , 1993). Nous présentons ci-dessous un extrait de texte dans lequel apparaissent diverses propriétés et leurs valeurs:

- les caractères typographiques ***typo**, avec les valeurs *italique* et *nil*.
- les subdivisions ***par**, (avec les valeurs *manchette, litige, contexte, décision*), l'appartenance aux outils documentaires d'où sont tirés les mots-clés (***term**) avec les valeurs *Ta* pour descripteurs du thésaurus acceptés, *Tr* pour descripteurs rejetés du thésaurus, *Tl* pour termes libres du domaine tels qu'identifiés par les experts, *Clas* pour rubrique de classification (ces valeurs peuvent être spécifiées par les premières lettres du sous-domaine du droit auquel appartiennent les termes: par exemple, *TlAss* pour terme libre caractérisant le champ assurances);
- la numérotation des phrases ***phr** et leur ordre ***ord** (*pr* pour première, *deux* pour deuxième, *ad* pour avant-dernière, *de* pour dernière);
- la position (***marque**) des mots dans la macrostructure (manchette, litige, contexte, décision): un terme portant la valeur *mancondéc* se trouve donc à la fois dans la manchette, dans le contexte et dans la décision.

On remarque également, dans le texte qui suit, certains résultats du prétraitement automatique ou semi-automatique: le doublement des traits d'union séparant deux éléments grammaticaux différents (soit--elle), le doublement des points d'abréviation et l'ajout d'une barre oblique devant les majuscules de noms propres (\C..\C..). La détection de termes complexes dans les outils documentaires ou la liste de termes libres du domaine préparés en cours d'expérimentation a résulté en la substitution d'un trait d'union au caractère blanc figurant entre les composants des termes. Tous les ajouts figurent en caractères gras dans notre exemple:

***par=ident*typo=nil**<ND>91-3 ***par=prov**<HD>COUR_D_'APPEL

par=manchette** ASSURANCEterm=(TaAss,ClasAss) *marque=mandéc** -- assurance_de_responsabilité***term=ClasAss *marque=man** -- recours***term= Clas *marque=mancondéc** contre le tiers responsable -- option***term=Ta *marque=mancondéc** -- article 2603 C.C. -- interdiction_de_cumul***temr-Tl** -- amendement***term=(Ta,Clas) *marque=mancondéc**.

par=litige *phr=1 *ord=(ad,pr)** Appelterm=(Ta,Clas) *marque=li** d'un jugement***term=Ta *marque=li** de la \Cour supérieure ayant accueilli une requête_en_irrecevabilité. ***term=Tl *phr=2 *ord=(de,deux)** Rejeté, avec dissidence.

par=contexte *phr=1 *ord=pr** Le 18 février 1988, l'appelante a intenté une actionterm=Tr *marque=condéc** contre la mise_en_cause ***term=Tr *marque=con** \Fontaine, lui réclamant 23 688$ à titre de

dommages***term=Class *marque=con** à la suite d'un incendie ***term=TaAss *marque=condéc** provoqué par sa négligence***term-Tr *marque=con *phr=2 *ord=deux**. Quelques mois plus tard, l'appelante a fait signifier une déclaration***term=Ta *marque=con** amendée qui ajoutait la compagnie_d'_assurances***term=TlAss** intimée à titre de défenderesse et qui concluait à la condamnation conjointe et solidaire des codéfenderesses. ***phr=3 *or=au** L'intimée a alors présenté une requête_en_irrecevabilité***term=Tl** fondée sur le fait que l'appelante n'avait aucun recours***term=Clas *marque=mancondéc** contre elle puisque, en poursuivant \Fontaine, elle avait exercé l'option***term=Ta *marque=mancondéc** prévue à l'article 2603 **\C..\C..** . ***phr=4 *ord=au** La requête_en_irrecevabilité***term=Tl** a été accueillie malgré la demande verbale d'amendement*term=(Ta,Clas) *marque=mancondéc présentée par l'appelante visant à modifier la désignation des parties et à ne maintenir que l'intimée à titre de défenderesse, reléguant \Fontaine au rang de mise_en_cause***term=Tr*marque=con**. [...]

par=décision *typo=italique *phr=1*ord=pr** \Mme la jugeterm=Ta *marque=condéc** \Tourigny et **\M..** le juge*Term=Ta ***marque=condéc** \Proulx: ***typo=nil** Les dispositions du ***typo=italique** Code de procédure_civile***term=(Ta,Clas) *marque=déc *typo=nil** relatives à l'amendement doivent recevoir une interprétation aussi large que possible. ***phr=2 *ord=deux** Cependant, une interprétation, aussi large soit**--**elle, ne peut écarter une disposition de droit substantif incluse dans le ***typo=italique** \Code civil. ***typo=nil *phr=3 *ord=au** Le législateur a voulu que, en intentant un recours***term=Clas *marque=mancondéc**, la partie demanderesse fasse un choix, ainsi que l'a confirmé **\M..** le juge***term=Ta *marque=condéc** \Mayrand dans l'arrêt ***typo=italique** \L'\Union québécoise, mutuelle [...]

En se positionnant sur un mot, on peut visionner, à l'aide d'une commande du logiciel SATO, toutes les valeurs de propriétés (ou traits) qui lui ont été attribuées, y compris celles qui résultent de calculs statistiques effectués par le logiciel:

assurance_de_responsabilité

*alphabet	=	fr
*fréqtot	=	3
*longueur	=	27
*term	=	ClasAss
*marque	=	man
*par	=	manchette
*poids	=	27
*discri	=	92
*chi2	=	2958
*gramr	=	tcomposé

mise_en_cause

*alphabet	=	fr
*fréqtot	=	2
*longueur	=	13
*term	=	Tr
*marque	=	con
*par	=	contexte
*phr	=	(1,4)
*ord	=	(pr,au)
*poids	=	26
*discri	=	133
*chi2	=	4890
*gramr	=	tcomposé

Une fois caractérisées, les données ont été filtrées en fonction des différents indices et soumises à une analyse de discrimination sur SPSS qui a fait ressortir les meilleurs prédicteurs pour expliquer les résultats des diverses opérations d'analyse.

L'enquête cognitive

Les résultats des analyses ont ensuite été confrontés aux données recueillies dans une première phase de l'enquête cognitive, puis soumis aux experts du domaine qui avaient pour tâche de confirmer les tendances observées et d'expliquer les anomalies, et surtout de décider d'une éventuelle réingénierie des processus ainsi que d'une éventuelle modification des outils documentaires.

L'enquête cognitive (comportant entrevues, observation et recueil de commentaires sur les résultats de nos traitements) a donc permis à la fois de compléter les analyses de données exposées précédemment et de les orienter. Nous cherchions les techniques et les stratégies employées pour parcourir un texte, les différentes parties du texte examinées pour prendre une décision de sélection, de tri-classification, de résumé et d'indexation, les connaissances utilisées (importance de tel ou tel tribunal, poids à accorder à la nature des parties en cause, valeur discriminante de telle ou telle mention de loi, de tel ou tel lexème, marqueurs du raisonnement du juge; références au contenu de la base de données, aux besoins des utilisateurs, à l'actualité, etc.), les catégorisations effectuées, les inférences faites pour passer des expressions en langue naturelle à leurs équivalents dans le thésaurus. Nous avons donc procédé selon une boucle: textes --> conseillers juridiques --> textes.

Complémentarité des approches

La complémentarité des approches utilisées pour la modélisation, recommandée à plusieurs reprises (Chaumier et Dejean, 1992; Doszkocs, 1986; Blosseville *et al*., 1992 ; Meunier *et al*., 1987), permet de tenir compte de la multiplicité des connaissances mises en oeuvre pour l'analyse du matériau textuel orientée vers des fins documentaires. Elle constitue, nous semble-t-il, un heureux compromis qui tient compte de caractéristiques exigeant parfois des solutions contradictoires, dans l'état de développement actuel des technologies: matériau textuel très complexe à analyser, mais nécessitant néanmoins des approches de nature linguistique et cognitive, volume important des données prohibant des analyses très fines et pouvant bénéficier des effets de nombre, savoir-faire de plusieurs experts à expliciter, selon des méthodes appropriées à leur mode d'inscription, de façon à respecter la culture de l'organisation.

Exemples d'indices pertinents retenus pour modéliser Les différentes opérations d'analyse

Nous donnons ci-dessous quelques exemples d'indices utilisés par les conseillers juridiques pour chacune des opérations d'analyse qu'il nous a fallu modéliser et nous fournissons des indications sur la place qui leur a été réservée dans le prototype.

L a s é l e c t i o n

L'annexe 2 au Règlement sur la cueillette et la sélection des décisions judiciaires (*Loi sur la Société québécoise d'information juridique* (L.R.Q., chap. S-20, art. 21) indique qu'une décision peut être sélectionnée si elle contient un des éléments suivants**:** 1) un point de droit nouveau; 2) une orientation jurisprudentielle nouvelle; 3) des faits inusités; 4) une information documentaire substantielle; 5) une problématique sociale particulière.

Notons tout de suite que tous les jugements de la Cour suprême sont gardés ainsi que tous les jugements de la Cour d'appel à moins que ces derniers ne soient pas motivés. Pour les autres cas, les conseillers juridiques nous ont fourni, pour chacun des critères mentionnés dans le règlement, au moins un exemple d'indice textuel, mais, à part le nombre de citations aux lois et à la jurisprudence, ce sont des indices qui se détectent difficilement par une analyse automatique. Nous avons été à même de constater que l'étape de la sélection repose sur des opérations cognitives complexes mettant en jeu de nombreuses connaissances spécialisées.

La détection de la plupart des indices pertinents nécessite une compréhension du sens des phrases ou de plus larges portions du texte (par exemple, lorsque le juge exprime son désaccord - critère no2) et des connaissances sur le monde, en particulier sur l'actualité (en Responsabilité civile, il faut détecter le fait inusité - critère no3 - comme un traitement médical nouveau ou la chute d'une personne aveugle sur un trottoir). Ou bien il faut identifier, à l'intérieur des textes, certaines catégories d'information et apprécier leur importance relative, par exemple la nouveauté du jugement par rapport à ceux qui ont été publiés dans des numéros antérieurs, etc. C'est pourquoi la prise de décision restera toujours la prérogative des conseillers juridiques.

En plus de consulter les experts, nous avons procédé par apprentissage sur corpus. Après examen d'un certain nombre de jugements rejetés (disponibles sur support papier seulement) et d'une série de jugements retenus, nous sommes arrivés à la conclusion que quelques critères formels simples permettent néanmoins de déclarer candidats au rejet un certain nombre de textes: 1) les jugements sont courts; 2) les jugements sont de type formulaire; 3) ils proviennent de la Cour des petites créances; 4) ils entérinent une convention. Une liste de types de requêtes ne faisant généralement pas l'objet de sélection a été constituée, mais n'est pas encore validée définitivement. Le prototype inclut seulement le premier et le dernier critères et nous envisageons d'y rajouter celui qui s'appuie sur la structure physique des jugements. La tâche sera d'autant plus facile que les textes seront saisis selon la norme SGML (Standard General Markup Language), ce que malheureusement n'a pas prévu pour le moment le ministère de la Justice.

L e t r i - c l a s s i f i c a t i o n

Un document *Savoir-faire des conseillers juridiques pour le tri* a été constitué à partir des entrevues effectuées auprès des conseillers juridiques. Il explicite les critères de tri utilisés pour chacune des 57 grandes classes du plan de classification.

Il s'avère que l'appartenance d'un jugement à un domaine du droit peut être décelée, dans plusieurs cas (par exemple: DROIT PÉNAL, FAMILLE, TRAVAIL), d'après quatre types de renseignements contenus dans la première page: le tribunal, le nom des parties ou la procédure entreprise, le numéro de greffe, l'intitulé du jugement le cas échéant.

- Ainsi, un jugement provenant de la Chambre d'expropriation de la Cour du Québec traitera assurément du domaine de l'expropriation. Par ailleurs, un jugement dont l'une des parties est un syndicat pourrait vraisemblablement aborder le droit du travail. Enfin, un jugement qui mentionne qu'il s'agit d'une requête en irrecevabilité à l'encontre d'une action en dommages-intérêts pourrait être classé en procédure civile.
- Dans certains cas, le numéro de greffe permet de classer immédiatement le jugement sous la bonne rubrique: par exemple, lorsque le chiffre qui suit le premier tiret est 11 (500-11-222222), il s'agit de FAILLITE, 41 pointe vers FAMILLE-PROTECTION DE LA JEUNESSE, 12 ou 04 vers FAMILLE, 43 vers FAMILLE-ADOPTION.

Mais comme il existe des chevauchements entre plusieurs rubriques de classification (par exemple, le DROIT CIVIL recoupe OBLIGATIONS, VENTE, CONTRATS, entre autres) et comme plusieurs rubriques (quatre au maximum) peuvent être attribuées à un même jugement en vertu des politiques implicites de classification, il est parfois nécessaire de consulter le texte du jugement, pour prendre connaissance soit des lois ou articles du code civil cités, soit du vocabulaire employé par le juge (on retrouve dans ce vocabulaire beaucoup des termes répertoriés dans le thésaurus ou le plan de classification). Pour le domaine ASSURANCE, par exemple, sur la première page, le tribunal qui rend la décision n'est pas un bon indice. Si le nom d'une des parties désigne une compagnie d'assurances, il est possible mais pas certain qu'il faille classer le jugement dans ASSURANCE; une compagnie d'assurances qui a indemnisé son assuré peut, en effet, poursuivre la personne qui lui a causé des dommages et il faudrait alors classer le jugement dans RESPONSABILITÉ. Le fait que, dans le texte du jugement, les articles 2468 à 2676 du *Code civil* ou bien la *Loi sur les assurances* soient cités, vient renforcer le second indice. Si, de surcroît, le jugement comporte les termes comme: «assurance-automobile, assurance collective, assurance de choses» ou ses spécifiques: «assurance-incendie, assurance-vol, assurances de personnes» ou à nouveau les spécifiques de ce dernier terme: «assurance-vie, assurance-invalidité, assurance-accident», ou encore «assurance (de) responsabilité, assurance maritime» , alors on peut prendre la décision de le classer dans ASSURANCE avec une quasi-certitude de ne pas se tromper.

Notre enquête cognitive a révélé l'utilité d'autres types de combinaison d'indices comme la présence d'un terme associée à sa position (par exemple, «requête en liquidation d'une compagnie» qui, se trouvant dans les premières pages du jugement, entraîne la décision de le classer dans COMPAGNIES, de même que «demande en divorce» qui permet de classer dans FAMILLE) ou la co-présence et la proximité de deux termes surtout dans le cas où l'un des termes est vague et peut

pointer vers plusieurs domaines du droit («délégation» près du terme «obligation» est un bon indice pour le classement sous la rubrique OBLIGATIONS, de même que «divorce» et «pension alimentaire» ou «prestation compensatoire» pour FAMILLE). Mais ceci n'a pas été implanté dans le prototype actuel.

Enfin, notons que certains indices permettent de classifier immédiatement le jugement dans une sous-rubrique sans attendre une analyse plus approfondie de la part du conseiller juridique responsable du domaine: ainsi, le nom du tribunal «Cour du Québec - Chambre de la jeunesse» et la mention de la «Loi sur la protection de la jeunesse» pointent sans ambiguïté vers la sous-rubrique PROTECTION DE LA JEUNESSE dans la rubrique FAMILLE.

On constate que l'analyseur textuel doit repérer plusieurs indices différents. Il faut pour cela que ces éléments fassent l'objet d'une fouille appropriée, ce qui est réalisé grâce au système de marquage de propriétés dans SATO et pourrait l'être au préalable avec SGML, dans les cas comme des citations de lois et de jurisprudence, par exemple.

Nous avons ajouté à cette approche linguistico-cognitive, une approche purement statistique qui a consisté en une analyse discriminante (effectuée sous SPSS) des mots par rapport à un corpus d'apprentissage: l'algorithme utilisé est capable de produire un indice de confiance dans le résultat obtenu. [2]

La prise de connaissance du contenu du jugement en vue de la rédaction du résumé

En observant les conseillers juridiques en train de parcourir et d'annoter les textes de jugements et en recueillant les commentaires qu'ils ont bien voulu faire pendant ou après l'exécution de leur tâche, nous avons pu brosser un portrait de la façon dont ils prennent connaissance du contenu. Nous avons constaté que certains éléments utiles pour le tri-classification peuvent ensuite être réutilisés avec d'autres indices pour la rédaction du résumé et l'indexation. Nous avons ensuite pu établir une liste des éléments textuels importants pour chacun des experts selon les domaines de droit dans lesquels il oeuvre.

Chaque spécialiste possède un schéma de la structure d'exposition des jugements dans tel ou tel domaine et recherche les énoncés-clés dans les parties réputées les contenir: questions de droit au début du jugement, motifs d'accusation («motifs suivants», «chefs d'accusation») et peine dans les premières lignes, énoncés des faits («les faits se résument comme suit») au début du jugement, moyens de procédure.

Les unités lexicales, particulièrement celles qui figurent dans le thésaurus et, le cas échéant, dans les listes de termes supplémentaires élaborées par quelques conseillers ainsi que les expressions pouvant indiquer qu'il y a discussion, lien de causalité, interprétation, etc. semblent constituer de

bons déclencheurs dans certains domaines, mais aussi la citation d'un article de loi, la mention du *Code civil* ou du *Code de procédure civile*, de la *Charte québécoise des droits et libertés*, etc.

Certaines divergences de lecture tiennent tout simplement au style cognitif des conseillers juridiques, mais peuvent en même temps être déterminées par la plus ou moins grande complexité du domaine ou les possibles recoupements entre domaines dans lesquels les jugements peuvent être classés: là où une personne lit intégralement le texte pour en prendre connaissance, plusieurs autres se contentent d'une lecture rapide favorisant qui le début et la fin du texte, qui le début et la fin de chaque paragraphe.

Dans le prototype de système expert, nous avons, pour le moment, retenu trois profils de lecture qui conviennent à tous: 1) les termes appartenant aux outils documentaires (thésaurus et plan de classification); 2) les intervenants (Juge, cour, parties, etc.); 3) les sources du droit (lois, articles, jurisprudence, etc.). Des couleurs différentes les mettent en relief et l'on peut les visualiser, au choix, dans le texte intégral, dans leur contexte immédiat, dans les phrases dans lesquelles ils sont insérés ou encore dans les paragraphes.

Pour une étape ultérieure de notre recherche, nous envisageons, en outre, une aide à la lecture personnalisée en fonction du domaine de droit dans lequel le jugement aura été préalablement classé, cette aide consistant tout simplement à mettre en relief par des couleurs les indicateurs particuliers à ce domaine. Par exemple, en RESPONSABILITÉ, le système surlignerait: «liens de causalité», «faute», «dommage exemplaire», «Charte des droits et libertés», etc.

Finalement des études exploratoires nous ont montré qu'il serait possible de mettre en lumière de façon différenciée, les parties du jugement qui traitent du litige, du contexte et de la décision, d'après des constantes de vocabulaire observées dans les résumés où ces trois parties sont très nettement distinguées (occurrences de lexèmes très différents, temps des verbes, etc.).

L'indexation

La tâche d'indexation, plus complexe que les tâches précédentes, n'a pas été aussi bien explicitée par les experts et nous avons dû nous appuyer sur la littérature - très peu diserte cependant - pour formuler des hypothèses en vue des traitements. En effet, l'étude cognitive des opérations d'analyse documentaire ne bénéficie pas d'une longue tradition en sciences de l'information (Bertrand, 1993; Bertrand-Gastaldy *et al.*, 1994; David, 1990; Endres-Niggemeyer, 1990; Farrow, 1991).

Mais il est clairement apparu que l'assignation des termes à insérer dans la manchette puis dans l'index est effectuée d'après le résumé. D'ailleurs, pour expliquer ses choix, une personne nous a précisé: «En lisant le résumé, il y a des mots qui clignotent. Question d'expérience, de flair.» Le dispositif auquel nous recourons pour mettre les termes importants en valeur permet justement de faire clignoter les termes marqués.

Le type de termes retenus, leur localisation dans le résumé, leur forme, leur ordre d'inscription semblent répondre à de très nombreuses règles mises au point par chacun au fil de l'expérience, selon les domaines. Si le système expert doit reproduire ces règles, la tâche va être longue et surtout va nécessiter de entrevues supplémentaires: chaque cas est un cas particulier ou presque. Par contre, c'est à ce prix que le système pourra faciliter la cohérence, -- du moins la cohérence intra-indexeur --, alléger le fardeau des conseillers juridiques et les libérer pour les prises de décision les plus délicates, notamment pour les cas-frontières.

En attendant de pouvoir approfondir cette enquête cognitive, nous nous sommes livrés à une étude comparée des propriétés des termes présents ou pas dans les résumés et retenus ou pas dans les manchettes. Toutes nos études ont pris appui sur les phénomènes d'intertextualité entre les résumés, les manchettes et les outils documentaires. Nous avons, entre autres, examiné l'importance de critères comme la position des termes dans la macro et la meso-structure des résumés, leur fréquence, leur valeur discriminante. Pour concevoir une aide à l'indexation directement à partir des textes intégraux, notamment pour ceux qui ne feront pas l'objet de résumé (c'est une perspective envisagée par SOQUIJ à plus ou moins long terme), il faudrait inclure ceux-ci dans l'étude des phénomènes d'intertextualité.

L'enquête cognitive a révélé, en outre, que, dans plusieurs domaines, l'indexation obéit à une sorte de grille implicite: le premier descripteur est chargé d'apporter tel type d'information, le second tel type de précision, etc. Par exemple, en droit pénal, on respecte l'ordre suivant: la rubrique, la sous-rubrique, le type d'infraction commise, les principes de droit étudiés dans la décision, les mentions sur l'appelant, le contexte de l'infraction, la peine imposée, alors qu'en procédure civile, on retient successivement: l'identification de la procédure, le moyen de procédure, le type de défense.

Sachant que, pour les experts de SOQUIJ, l'ordre d'inscription des termes a une signification, il nous sera possible de mettre au point, dans une phase ultérieure, des traitements plus complexes permettant de comparer, dans chaque domaine du droit, les listes de termes assignés en première, deuxième, troisième positions, etc. pour faire surgir des grilles utilisées de façon peut-être inconsciente. Pour le moment, le prototype de système expert d'aide à l'indexation ne fait que surligner de façon différenciée les différents mots-clés potentiels et produire une liste de ces mots-clés triés selon le domaine de classification et classés par ordre de fréquence décroissante.

Avant d'implanter toutes les fonctionnalités envisagées (prise en compte de la valeur discriminante, de la position dans la macro-structure et la micro-structure), il faut que les experts prennent plusieurs décisions sur leurs politiques d'indexation en fonction de nos observations et recommandations et se prononcent également sur les modifications des outils documentaires. Nous pensons qu'en les confrontant aux résultats produits par un système expert encore rudimentaire, nous les aiderons à expliciter davantage les choix qu'ils feront à partir des suggestions de la machine.

Les propositions d'enrichissement des outils documentaires

Tout au cours de notre projet, nous avons été amenés à étudier l'utilisation des outils documentaires et à introduire des modifications qui facilitaient le travail de marquage automatique, modifications qui pourraient même être utiles dans le contexte d'une analyse humaine.

Les modifications à apporter au plan de classification

D'après le taux d'utilisation des différentes rubriques et sous-rubriques, le plan de classification nous a semblé répondre au volume et au rythme de publication des analyses de jugements dans *Jurisprudence Express*. Nous avons simplement recommandé d'examiner la possibilité de subdiviser deux classes fortement représentées et de recourir davantage aux subdivisions pour le repérage des notices dans une base de données automatisée, de façon à permettre une sélection relativement fine aux utilisateurs ayant un domaine particulier en tête. Le besoin de sélectivité n'est pas le même dans les publications imprimées, surtout celles qui paraissent à un rythme hebdomadaire comme *Jurisprudence Express.*

Les études effectuées sur l'utilisation du thésaurus

Le marquage des termes nécessaire à plusieurs de nos traitements, de même que le désir de mieux évaluer dans quelle mesure le thésaurus répondait aux besoins d'indexation tels qu'implicitement fixés par les conseillers juridiques, nous ont conduits à effectuer une série d'études complémentaires. Nous avons, par exemple, recherché les variantes morphologiques et les variantes syntaxiques, la présence de descripteurs et non-descripteurs à l'intérieur des mots-clés libres dans les manchettes (qui constituent environ 60% des mots-clés), étudié les structures les plus fréquentes pour la formation de ces mots-clés libres, fait la liste des descripteurs jamais employés ou jamais employés seuls, etc. Nous avons aussi tenu compte des cooccurrences des différents termes (descripteurs et non-descripteurs, mots clés libres) entre eux et avec les rubriques de classification. En outre, en calculant la force d'association des termes avec les rubriques (selon une méthode qui tient compte de la fréquence), nous sommes désormais en mesure d'amorcer une structuration du vocabulaire par domaine de droit et donc de concevoir une réconciliation de deux outils documentaires (qui se recoupent et se contredisent parfois).

Bref, la richesse des analyses effectuées permet d'offrir une multitude de points de vue sur l'utilisation effective (plutôt que souhaitée lors de la conception) de ces outils, au fil des ans, par plusieurs personnes. Nous avons donc soumis à SOQUIJ non seulement un portrait des outils et des pratiques actuelles, mais des suggestions très détaillées pour l'enrichissement et la modification de ces outils et de ces pratiques.

Les résultats de nos différentes études, nous ont amenés à conclure que le thésaurus devait être enrichi; d'abord pour contrôler une indexation qui s'avère, dans les faits, plus spécifique que ce que permet l'outil actuel, ensuite parce que le système expert doit pouvoir repérer toutes les formes possibles d'un descripteur dans les résumés et éventuellement, dans les textes intégraux pour les ramener aux formes souhaitées pour l'indexation, enfin parce que, une fois les descripteurs organisés selon une hiérarchie stricte, il devient possible d'opter pour différents niveaux de généricité selon les produits documentaires (en fonction notamment de leur périodicité, de leur support et de leur couverture du domaine).

LA RÉALISATION DU PROTOTYPE DE SYSTÈME EXPERT

La section précédente exposait la modélisation, méthode et résultats, effectuée pour l'aide à la sélection, à la classification, à la lecture et à l'indexation des jugements; la présente section traite de l'implantation réalisée du modèle. Les aspects suivants sont touchés : la tâche à informatiser; la motivation du choix de la technologie des systèmes experts; l'incertitude reliée à cette entreprise; l'arrimage du système expert avec l'analyse de textes par ordinateur; le design de la chaîne de traitement des documents; les aménagements apportés au traitement standard de l'incertitude; la réalisation de la base de règles par apprentissage et les problèmes laissés en suspens.

La tâche à informatiser

Comme on a pu le constater dans la section précédente, les tâches à informatiser présentent un haut niveau de complexité et sont accomplies dans un contexte de production. Rappelons que les publications de SOQUIJ connaissent des échéances et sont assujetties aux lois du marché. Ces tâches sont dites cognitives en ce que leur accomplissement requiert la mise en oeuvre particulière et discrétionnaire de connaissances et de stratégies générales accumulées durant l'exercice répété et supervisé des tâches mêmes. Pour leur réalisation, de nombreuses informations de source et de valeur diverses et parfois contradictoires doivent être recueillies et synthétisées. Conséquemment, une méthode mixte de modélisation a été déployée; il s'agit de faire converger les résultats d'une enquête cognitive auprès des conseillers juridiques, d'un traitement statistique de la distribution des indices et d'une analyse de texte plus qualitative.

La stratégie retenue est de recourir au plus grand nombre de sources de connaissances, identifiées lors de la modélisation, pour lesquelles des indices sont repérables dans les jugements. Par source de connaissance nous entendons, par exemple, la longueur du jugement, les lois qui y sont mentionnées, le tribunal qui a rendu le jugement, etc. Dans la mesure où ces sources de connaissances s'avèrent distinctes les unes des autres, il est possible de fonctionner avec un principe de convergence. Ainsi, on est d'autant plus certain qu'un jugement pointe vers le domaine «pénal» que son numéro de greffe comporte en deuxième section, l'une des combinaisons suivantes {01, 03, 10, 27 ou 36}, que ce jugement a été rendu dans la «Chambre criminelle et pénale»; que «La Reine»

est une des parties impliquées et que le *Code criminel* y est mentionné, etc. Cette stratégie présente l'avantage de fonctionner, la plupart du temps de façon satisfaisante, dans des conditions de bruit. Le bruit étant essentiellement causé ici par les indices qui pointent vers plus d'un domaine.

Motivation du choix de la technologie des systèmes experts

Pour réaliser une implantation informatique du modèle cognitif obtenu, nous avons retenu la technologie des systèmes experts (SE) pour plusieurs raisons. L'implantation d'algorithmes incomplets et/ou sujets à de fréquentes révisions est possible car les règles d'inférences qui tiennent lieu des instructions d'un programme conventionnel sont indépendantes les unes des autres et leur enchaînement est assuré par un mécanisme général appelé moteur d'inférences. Il n'est pas donc plus nécessaire de prévoir à l'avance le déroulement complet de la solution définitive du problème: l'implantation peut être modulaire et évolutive. La réalisation d'un prototype se trouve à jouer un rôle heuristique en permettant d'achever la conception par des boucles de tests/ajustements en situation. La structure des règles d'inférences permet une implantation quasi directe du modèle cognitif qui a été développé : un ou plusieurs indices détectés dans le texte du jugement (la prémisse) sont mis en relation avec une rubrique du plan de classification (la conclusion). De plus, la certitude de ces relations peut être qualifiée au moyen d'un coefficient numérique qui sera cumulé tout au long de la consultation. Ce «cumul» d'une part atténue la valeur des validations subséquentes lorsqu'une validation est affectée d'un coefficient incertain et, d'autre part, renforce la valeur d'une validation qui a déjà été réalisée. Le chaînage avant des règles permet enfin d'obtenir toutes les «réponses» valides et non une seule; un jugement peut donc être classifié dans plus d'un domaine avec une certitude différente pour chacun. Le découpage en règles d'inférence facilite la génération en contexte d'un rapport qui permet de valider les associations indices/rubrique du plan de classification, de localiser précisément les dysfonctionnements et finalement d'entraîner des conseillers juridiques novices.

Les incertitudes reliées à cette entreprise

Une fois la technologie des SE retenue en raison de caractéristiques qui apparaissaient souhaitables étant donné le projet, plusieurs incertitudes demeuraient; certaines ont été résolues lors de la réalisation du prototype et les solutions retenues feront l'objet des prochaines sections. Une incertitude provenait de ce que les indices nécessaires pour la classification des jugements sont essentiellement de nature textuelle. Ainsi, contrairement aux situations habituelles de développement des SE, les indices ne sont pas fournis directement au système par l'utilisateur ou des senseurs. Cet état de fait implique le partage du traitement entre le SE pour interpréter les indices et un logiciel d'analyse de texte par ordinateur (ATO) pour les dépister dans le texte des jugements. De plus, ces indices dépistés par le logiciel d'ATO doivent être transformés pour être admissibles au SE. Une autre incertitude consistait à développer et implanter une chaîne de traitements qui soit conforme au traitement accompli par les conseillers juridiques, notamment en dépistant les mêmes types

d'indices. La difficulté est double, le dépistage en lui-même et le regroupement des indices de nature différente. Une autre incertitude enfin était liée à l'utilisation des coefficients de certitude pour rendre compte du fait que les indices sont rarement totalement fiables. En effet, un indice peut pointer vers plus d'un domaine du droit ou encore la présence d'un indice peut être considérée comme accidentelle et constituer en quelque sorte du «bruit». De plus, le mode de cumul des coefficients présente certains problèmes qui sont documentés. Certaines autres incertitudes sont toutefois demeurées, principalement parce que des recherches plus fondamentales dont l'envergure dépassaient le mandat s'avèrent nécessaires; celles-ci sont présentées dans la dernière section.

L'arrimage du système expert avec l'analyse de textes par ordinateur

Le générateur de système expert (GSE) utilisé est l'Atelier Cognitif et TExtuel (ACTE) développé au Centre ATO.CI. Le développement de ACTE a démarré en février 1988, sur la commande d'un consortium de ministères et organismes québécois appelé DELTA; il s'agit d'une intégration logicielle de SATO et d'une version optimisée du D_expert (GSE en LISP)[3]. Cette intégration permet de faire du diagnostic textuel, c'est-à-dire de ne plus modéliser comme tel le contenu des textes, mais bien les opérations cognitives de lecture et de compréhension, opérations qui sont en jeu pour la classification des jugements[4]. La séquence qui a été retenue consiste à effectuer en lot une série de fichiers de commande SATO qui dépistent et identifient, principalement à l'aide de concordances, les différents types d'indices. Voici, par exemple, un extrait d'un tel fichier de concordances identifiant certaines requêtes qui, lorsque le jugement n'est pas motivé, entraînent le rejet :

```
Concordance stricte  pension alimentaire
Concordance ordonnée rectification registre état civil
```

L'interface entre les traitements effectués par SATO et le SE se fait par la consignation dans un fichier du résultat - succès ou échec - de chacune des concordances. Ce fichier est alors traité pour ne conserver que les résultats positifs qui sont identifiés à l'aide d'une table, ce qui permet de normaliser le segment dépisté. Si l'on poursuit l'exemple précédent et que la deuxième concordance est réussie, peu importe la formulation exacte du segment dépisté, l'appellation normalisée sera transmise au SE.

Le design de la chaîne de traitement des documents

La modélisation cognitive de la tâche a été transformée en une suite séquentielle de traitement dont une schématisation est adjointe en appendice.

La première étape consiste en un prétraitement qui est requis pour rendre les jugements admissibles à SATO. Reçus en format «WordPerfect», ils sont d'abord convertis en ASCII sans perdre les codes indiquant les attributs graphiques (gras, souligné, etc.) à l'aide d'un fichier de configuration d'imprimante élaboré à cet effet. Les codes de début et de fin deviennent des valeurs de la propriété **typo**:

(...) l'arrêt Laurentide Motels Ltd. c. Ville de Beauport, (...)
(...) l'arrêt *typo=+soul Laurentide Motels Ltd*typo=-soul. c. *typo=+soul Ville de Beauport *typo=-soul (...)

Puis, un programme en ICON[5]. procède à la désambiguisation des marques de phrase et de paragraphes. Le point marque habituellement la fin des phrases, mais il est aussi utilisé dans la notation de nombres décimaux, dans des sigles et il marque l'abréviation. Surtout dans les domaines législatifs et administratifs, la mise en page d'une énumération ne se distingue que difficilement d'une suite de paragraphes. Enfin, les commandes nécessaires pour que le programme SATOGEN transforme le texte en matrice admissible à SATOINT sont ajoutées : l'alphabet, les séparateurs, les valeurs de la propriété **typo**.

Dans une deuxième étape, les indices textuels relatifs à chacune des sources de connaissances, sont dépistés, principalement par l'exécution de fichiers de commande SATO renfermant des concordances, à l'exception du numéro de greffe qui est dépisté par un programme *ad hoc* en ICON.

La troisième étape est celle de la mise en relation des indices dépistés avec les domaines du droit pertinents à l'aide du SE. Ce faisant, un rapport de la consultation est produit où sont consignés, pour chacune des sources de connaissance, les indices dépistés ainsi que les associations qui sont faites avec des domaines; une justification en contexte est, à l'occasion, fournie; un exemple de rapport est joint en annexe.

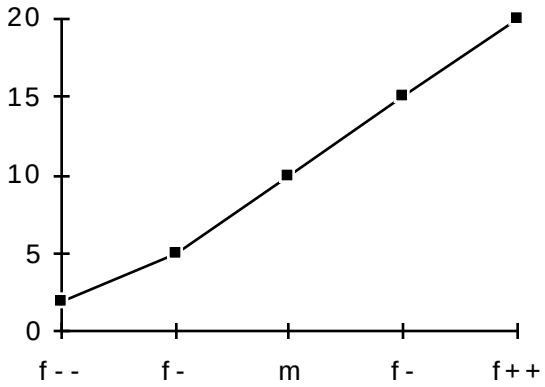
La quatrième étape, appelée assistance à la lecture, est optionnelle. Elle consiste en l'affichage des indices dépistés pour effectuer la tâche de classification ou encore d'autres indices. La distinction entre les types d'indices est produite par l'utilisation de couleurs différentes. Cette visualisation peut être effectuée selon un ou plusieurs profils. Les profils offerts actuellement ont été mentionnés plus haut, il s'agit des intervenants [Juge, cour, parties, etc.]; des sources du droit [lois, articles, jurisprudence, etc.] et des outils documentaires [thésaurus et plan de classification]. Voici, à titre d'illustration un extrait de jugement :

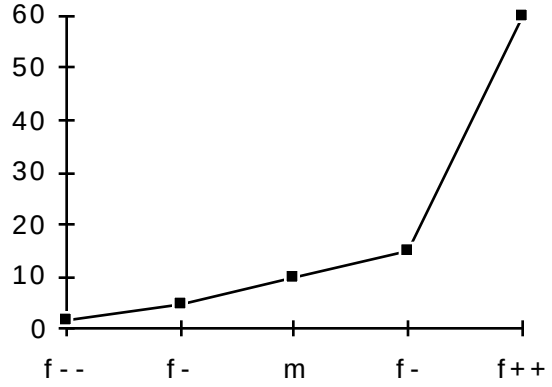
> Il s'agit d'une procédure assez exceptionnelle puisque, le requérant allègue certaines erreurs de droit du juge de paix. (...) Il y a évidemment ici la gravité objective de l'accusation. ° C'est une des plus sérieuses, une des plus graves que le Code criminel contient -- plutôt que la Loi des stupéfiants contient (...)

Les aménagements apportés au traitement standard de l'incertitude

Chacun des indices peuvent pointer vers plusieurs domaines, avons-nous dit précédemment. De plus, une même confiance n'est pas accordée à toutes les relations établies entre les indices et les domaines. Les SE offrent la possibilité d'implanter des structures conditionnelles pondérées par des coefficients numériques, de même qu'une fonctionnalité pour leur cumul. Le cadre théorique le plus souvent utilisé est celui des coefficients de certitude développé pour le système Mycin[6]. Rappelons que le principe du cumul des coefficients est le renforcement; en voici un exposé simplifié : si on arrive à une même conclusion à partir de deux sources de connaissances distinctes, on attribue à

cette conclusion un coefficient supérieur au coefficient le plus élevé. Ce principe permet donc de discriminer les différents domaines du droit vers lesquels l'ensemble des indices repérés pointent.
Lorsqu'appliqué à notre système, ce cadre théorique pose toutefois deux ordres de problèmes[7]. Il a été démontré d'une part qu'il était très difficile pour des experts d'exprimer leur confiance dans les relations qu'ils établissent entre des faits vérifiés ou tenus pour vrais et des conclusions sous la forme d'un coefficient numérique. D'autre part, un calibrage des coefficients doit être effectué en fonction du nombre de renforcements qui sont susceptibles de se produire pour que l'effet discriminant soit optimal. En effet, si beaucoup de renforcements ont lieu alors que les coefficients sont élevés, plus le résultat tend vers 100, plus il perd de la valeur discriminante; la valeur des coefficients ne doit pas être élevée. Par ailleurs, si les renforcements ne sont pas nombreux et que les coefficients sont très bas, les résultats ne seront pas convaincants. Pour remédier à ces deux problèmes, une approche modulaire a été développée. L'expression de la confiance quant à la relation entre les indices et les domaines du droit est séparée de l'algorithme de cumul par renforcement qui intervient lors d'une consultation.
Cette confiance est exprimée par des coefficients symboliques distribués sur une échelle bi-polarisée qui comporte cinq valeurs : forte [f++], moyenne-forte [f+], moyenne [m], moyenne-faible [f-] et faible [f--]. La conversion de ces coefficients «symboliques» en des coefficients numériques admissibles à l'algorithme de cumul. L'échelle numérique des coefficients est de 1 à 100. Cette fonction a deux rôles : exprimer l'écart entre les coefficients symboliques et ajuster leur valeur en fonction du nombre potentiel de renforcements. L'écart entre les coefficients détermine leur aspect discriminant. La figure de gauche montre une discrimination plutôt constante, celle qui est présentement implantée, la figure de droite montre une forte discrimination; l'utilisation du coefficient le plus élevé indique une relation prépondérante :





Le calibrage de la valeur numérique attribuée à chacun des coefficients symboliques en fonction du nombre potentiel de renforcements, se fait par essai-erreur. Des recherches supplémentaires sont requises pour déterminer une méthode exacte.

La réalisation de la base de règles par apprentissage

La technologie des SE ne présente comme tel aucun mode d'apprentissage, les règles d'inférences doivent être écrites et modifiées de la même manière : une à une à l'aide d'un éditeur spécialisé. Afin de pallier cette carence, deux solutions ont été combinées : une approche tabulaire et une étude de corpus. Comme elles expriment des relations simples, les règles d'inférences peuvent s'exprimer sous forme de tableau, en trois colonnes: l'indice, le domaine et le coefficient de confiance, géré par une base de données ou un tableur; ainsi par exemple un extrait du tableau des numéros de greffe :

```
03    pena    f++        36    pena    f++        12    fami    f++
27    pena    f++        53    drli    m          04    fami    f++
01    pena    f++        06    proc    f+         43    fami    f++
10    pena    f++        41    fami    f++
```

Le passage aux règles d'inférences est le fait d'un programme en ICON qui constitue la prémisse à partir de la première colonne, la conclusion à partir de la deuxième et opère le passage du coefficient symbolique en un coefficient numérique :

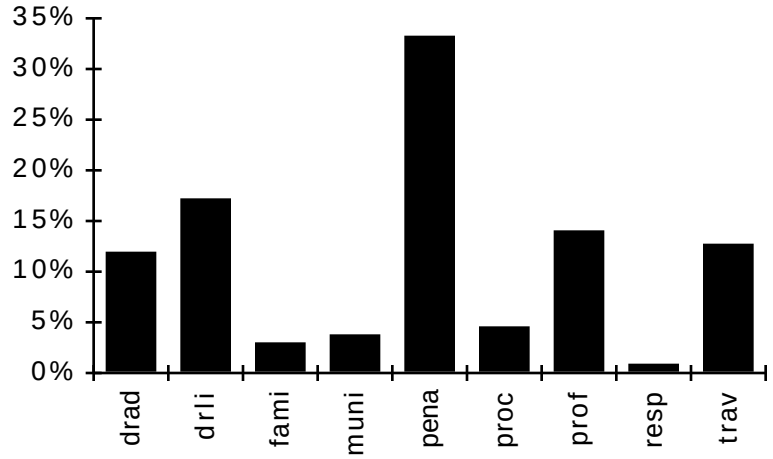
```
Connaissance Règle Définir 201 **
 Note "TRI -> 1ère page -> no de greffe : 03" **
 Auteur automatik **
 Création 1904-01-01 00-00-00 **
Si  **
 Base TRI **
  Granule "Indices de première page" **
   ( Trait "section du no de greffe" =   Chaîne "03" ) **
Alors **
 Base TRI **
  Granule Document **
  ( Trait domaine =   Chaîne "Pénal" Coef 20 ) **
CanalEcrire ( Canal rapport **
  Message " Ce no. de greffe pointe vers le domaine " **
    Base TRI **
     Granule Document **
      Trait domaine **
  Message " avec une confiance forte" **
  Message " ; cumulatif de : " Coefficient **
  Message "%"  Retour **
)
```

L'étude d'un corpus constitué de jugements déjà classifiés a permis de littéralement découvrir des indices pour la plupart des sources de connaissances. À titre d'illustration, le cas des lois citées sera décrit. Pour chacun des domaines du droit, un sous-texte a été constitué de tous les passages en italique à l'aide de SATO. Ces sous-textes ont été épurés de façon à ne contenir que les lois citées dans les jugements et ont été constitués en tableaux avec le domaine attribué :

Loi de l' aménagement	municipal
Loi de l' assistance publique	municipal
Loi de l' évaluation foncière	municipal
Loi de la qualité de l' environnement	municipal
Loi de police	municipal

Ensuite, ces tableaux ont été fusionnés et triés, de façon à regrouper pour chacune des lois tous les domaines pointés. Cette distribution, suite à une validation par les conseillers juridiques pour éliminer les aberrations, a guidé l'attribution des coefficients de confiance. La règle suivie est que si

la distribution est égale, un coefficient faible est attribué, par contre, sinon la force du coefficient est proportionnel à la distribution. Ainsi, par exemple, la *Charte canadienne des droits et libertés* apparaît dans les domaines suivants avec cette distribution :



À partir de cette distribution, les coefficients suivants ont été attribués :

```
Pénal (pena)                    f++
Droits et libertés (drli)       f+
Droit administratif (drad)      m
Professions (prof)              m
Travail (trav)                  m
Famille (fami)                  f--
Municipal (muni)                f--
Procédure civile (proc)         f--
Responsabilité (resp)           f--
```

À la suite de ces opérations, on obtient un tableau en trois colonnes qui permet de générer les règles d'inférences. Les indices qui pointent vers plusieurs domaines sont regroupés dans une même règle.

Les problèmes laissés en suspens

Malgré tous les efforts déployés, des difficultés ont été laissées en suspens, parce qu'elles demandent des recherches dont l'envergure dépassait le mandat, mais dont l'intérêt apparaît évident étant donné le succès du prototype, par exemple :

- l'intégration du coefficient de confiance dans le résultat obtenu par l'algorithme d'analyse discriminante utilisé au cumul des coefficients de certitude des règles d'inférences;
- l'intégration des occurrences différentes des termes lemmatisés du plan de classification et du thésaurus dans le cumul des coefficients de certitude et la prise en compte de leur fréquence. Est-ce que plusieurs termes différents pointant vers un même domaine valent plus cher que des fréquences élevées de quelques termes ?
- la modification du modèle de cumul des coefficients par renforcement qui ne permet que l'accroissement linéaire, pour prendre en compte des indices qui invalident un ou plusieurs domaines, ce qui serait plus conforme au fonctionnement cognitif des conseillers juridiques.

CONCLUSION

L'expérience que nous venons d'exposer a permis à l'équipe de recherche de vérifier: 1) qu'il est possible de modéliser les opérations cognitives des experts dans diverses situations de lecture à partir d'une enquête cognitive et d'une analyse sémio-statistique des textes analysés et des résultats de plusieurs types d'analyses; 2) qu'il est possible d'implanter un système expert s'appuyant sur des stratégies dépistant dans les textes certains des indices détectés par les humains, à différents niveaux d'organisation des textes (éditorial, morpho-syntaxique, intra-phrastique, intra- et inter-textuel, sémantique, pragmatique, etc.). Il a également été constaté que plusieurs de ces indices sont utilisables pour faciliter la lecture selon les objectifs poursuivis par chacune des opérations de sélection, de tri-classification et d'indexation. Pour cela, il faut disposer d'un logiciel qui, non seulement autorise le marquage des unités textuelles et lexicales selon autant de caractéristiques que les hypothèses le suggèrent, mais aussi facilite auparavant la découverte de ces propriétés puis leur manipulation au même titre que la manipulation des chaînes de caractères. Nous avons également appris que la performance de nos méthodes de modélisation/formalisation était optimale lorsque l'information requise par les tâches cognitives se trouvait dans les textes sous la forme d'indices repérables. Ainsi, même lorsque la prise de connaissance du contenu semble superficielle (par exemple pour la sélection), si la prise de décision fait appel à des connaissances autres que celles de la langue et du cadre textuel, la tâche échappe en grande partie à nos méthodes. Par conséquent, comme les tâches requièrent pour la plupart de telles connaissances, les experts doivent toujours garder le contrôle des systèmes. En ce qui concerne l'implantation de la chaîne de traitement, nous avons constaté que le succès reposait moins sur la complexité de la technologie ou des formules mathématiques que sur la maîtrise une à une de chacune des sources de connaissances requises et des indices qui les désignent dans les textes.

Ajoutons qu'un des bénéfices importants de la recherche a consisté dans le portrait des politiques et procédures d'analyse suivies par la dizaine de conseillers juridiques telles que révélées par l'analyse des données et l'enquête cognitive, dans la constatation de quelques divergences dont certaines devaient être corrigées pour accroître la prédictibilité des index, et enfin, dans la production d'un thésaurus considérablement enrichi et l'amorce d'un meilleur arrimage entre thésaurus et plan de classification.

REMERCIEMENTS

Le projet a été soutenu financièrement par les institutions suivantes: Centre francophone de recherche en informatisation des organisations (CEFRIO), Société québécoise d'information juridique (SOQUIJ), Ministère des Communications du Québec, École de bibliothéconomie et des sciences de l'information, Université de Montréal, Centre de recherche en cognition et information ATO.CI, Université du Québec à Montréal.

Plusieurs personnes ont été impliquées à diverses étapes du projet: Jean-Guy Meunier, directeur du Centre ATO.CI; Sylvie Michaud, bibliothécaire professionnelle; Myriam Desclos-Lalaude, stagiaire de l'I.E.P. (Cycle supérieur de spécialisation en information et documentation, Institut d'Études Politiques), Paris; Luc Dupuy, agent de recherche, centre ATO.CI, Yves Khawam, professeur adjoint, ÉBSI et plusieurs étudiants en bibliothéconomie et sciences de l'information ainsi qu'en linguistique.

BIBLIOGRAPHIE DES SOURCES CITÉES

Beacco, J.-C.; Darot, M., 1984. *Analyse de discours; lecture et expression.* Paris: Hachette / Larousse; 1984.

Beghtol, C.; Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation*; June 1986; 42(2): 84-113.

Bertrand, A., 1993. *Compréhension et catégorisation dans une activité complexe: l'indexation de documents scientifiques.* Thèse de doctorat, Équipe de psychologie du travail ER 15- CNRS, Université de Toulouse-Le Mirail, France, 1993.

Bertrand-Gastaldy, S. , 1993. Analyse documentaire et intertextualité. *Les Sciences du texte juridique: Le droit saisi par l'ordinateur* . Sous la direction de Claude Thomasset, René Côté et Danièle Bourcier. Textes présentés à un séminaire tenu à Val-Morin, Québec, du 5 au 7 oct. 1992 sous l'égide du Laboratoire Informatique, droit et linguistique du CNRS et du Groupe de recherche Informatique et droit de l'Université du Québec à Montréal. Cowansville: Les Éditions Yvon Blais; 1993: 139-173.

Bertrand-Gastaldy, S.; Daoust, F.; Pagola, G.; Paquin, L.-C., 1993a. *Conception d'un prototype de système expert d'aide à l'analyse des jugements : rapport final présenté à SOQUIJ. Vol. 1 : synthèse des travaux.* [Montréal]: Université de Montréal. École de bibliothéconomie et des sciences de l'information / Université du Québec à Montréal. Centre de recherche en information et cognition ATO.CI; juillet 1993: 88 p. + annexes.

Bertrand-Gastaldy, S.; Giroux, L.; Lanteigne, D.; David, C., 1994. Les produits et processus cognitifs de l'indexation humaine. *ICO Québec*; 6(1-2); printemps 1994: 29-40.

Blosseville, M.J.; Hébrail, G.; Monteil, M.G.; Pénot, N. Automatic document classification: Natural language processing, statistical analysis and expert system techniques used together. *SIGIR 92, Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, June 21-24, 1992*: 51-57.

Chaumier, Jacques; Dejean, Martine. L'indexation documentaire: de l'analyse conceptuelle humaine à l'analyse automatique morpho-syntaxique. *Documentaliste* ; 27(6); novembre-décembre 1990: 275-279.

David, C., 1990. *Élaboration d'une méthodologie d'analyse des processus cognitifs dans l'indexation documentaire.* Montréal: Université de Montréal, Département de communication; 1990. Mémoire de maîtrise.

Doszkocs, Tamas. Natural language processing in information retrieval. *Journal of the American Society for Information Science*; 1986; 37(4): 191-196.

Endres-Niggemeyer, B., 1990. A procedural model of abstracting, and some ideas for its implementation. *TKE'90*; *Terminology and Knowledge Engineering*. Frankfurt: Indeks Verlag; 1990: 230-243.

Farrow, J., 1991. A cognitive process model of indexing document. *Journal of documentation;* 47 (2); June 1991: 149-166.

Meunier, J.-G., Bertrand-Gastaldy, S.; Lebel, H.. 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, Jean-Guy. SATO: un philologue électronique. *Documentation et bibliothèques*; 38(2); avril-juin 1992: 65-69.

Van Dijk, Teun A. Perspective paper: complex semantic information processing. In: Walker, D.E.; Karlgren, H.; Kay, M. *Natural Language in Information Science; Perspectives and Directions for Research.* Stockholm: Skriptor, 1977:127-163.

Chaîne de traitement combinant analyse de texte et système expert

SATO

Prétraitement

- conversion WordPerfect -> ASCII
- marquage
  . des paragraphes
  . des phrases
- SATOGEN

Description

- longueur
- le type de requête
- le no. de greffe
- le tribunal
- les parties
- l'intitulé
- lois mentionnées
- art. du C.c. mentionnés
- par les descripteurs

système expert

SÉLECTION

- longueur > 40 lignes ou > 400 mots
- type de requête

CLASSIFICATION Indices de 1ère page

- le no. de greffe
- le tribunal
- les parties
- l'intitulé

CLASSIFICATION Sources mentionnées

- lois
- articles du Code civil

CLASSIFICATION Distribution

- statistique lexicale
- documentaire

Assistance à la lecture

- dépistage d'information selon des profils
- surlignage

Projet CEFRIO-SOQUIJ
RAPPORT

SYSTÈME EXPERT POUR SÉLECTIONNER ET CLASSIFIER LES JUGEMENTS

prototype pour SOQUIJ, le CEFRIO et le MCQ
Suzanne Bertrand-Gastaldy resp., Gracia Pagola
EBSI : École de bibliothéconomie et des sciences de l'information, UdeM
François Daoust et Louis-Claude Paquin
Centre ATO-CI: Centre de recherche en cognition et information à l'UQAM

Le système expert a pour tâche principale de sélectionner les jugements
et de désigner les rubriques de classification les plus probables.

ÉTAPE DE LA SÉLECTION
Les jugements sont retenus pour traitement selon les critères suivants : la longueur dont le seuil est de 40 lignes ou 400 mots le type de requête.

Ce jugement compte 179 lignes et 2002 mots.
Le jugement est sélectionné. il a la longueur suffisante; il n'est pas répertorié parmi les requêtes rejetées.

ÉTAPE DU TRI-CLASSIFICATION
Cette étape comporte quatre analyses :
- les indices de la première page;
- les lois et articles du code civil mentionnés;
- la discrimination lexicale;
- la discrimination par les outils documentaires.

A. L'analyse des indices de la première page du jugement touche les indices suivants :
1) le numéro de greffe : 500-05-001562-899
- ne pointe vers aucun domaine

2) le tribunal : COUR SUPÉRIEURE
- ne pointe vers aucun domaine

3) Le nom des parties
- «La Reine» pointe vers le domaine Pénal avec une confiance forte ; cumulatif de : 20%

4) L'intitulé du jugement
- ne pointe vers aucun domaine

B Lois et articles du Code civil mentionnés
«code criminel»
- pointe vers le domaine Municipal avec une confiance faible ; cumulatif de : 2%
- pointe vers le domaine Pénal avec une confiance forte ; cumulatif de : 36%
- pointe vers le domaine Professions avec une confiance faible ; cumulatif de : 2%

«L'article 614.3 C.C.»
- pointe vers le domaine Famille avec une confiance forte ; cumulatif de : 20%

C Analyse de la discrimination documentaire

Cette analyse est effectuée par la projection des termes appartenant au thésaurus et au plan de classification.

1 descripteur(s) pointe(nt) vers le domaine Droits et libertés
1 descripteur(s) pointe(nt) vers le domaine Sûretés
1 descripteur(s) pointe(nt) vers le domaine Travail
2 descripteur(s) pointe(nt) vers le domaine Responsabilité
3 descripteur(s) pointe(nt) vers le domaine Procédure civile
24 descripteur(s) pointe(nt) vers le domaine Pénal

_____ F i n _ d u _ t r a i t e m e n t _____

[1] Le mandat qui nous avait été confié incluait plusieurs contraintes dont celle de respecter à la fois les outils documentaires actuels (thésaurus et plan de classification) et les «habitudes» des experts - habitudes cependant très faiblement documentées.

[2] L'analyse discriminante établit le «portrait-robot» de chaque classe à partir du profil statistique de chacun des individus de la classe. Le meilleur résultat que nous ayons obtenu dans nos expérimentations avec un grand nombre de variables sélectionnées en fonction de leur chi2 et de certaines autres caractéristiques comme la langue a donné un taux de réussite de 92% en apprentissage et de 68% pour le groupe-test. Nous prévoyons plusieurs améliorations avec le dépistage des multitermes et des lois citées pour augmenter la précision du vocabulaire (nous envisageons aussi de procéder à la sélection de termes identifiés par les experts pour chaque domaine, car l'expérience de classement à l'aide de SATO a montré la pertinence de cette approche).

[3] Pour une description de ACTE, voir L.-C. Paquin, L. Dupuy et F. Daoust (1989) "ACTE: a workbench for knowledge engineering and textual data analysis in the social sciences" *Proceedings of the Fourth International Conference on Symbolic and Logical Computing (ICEBOL4)*, Dakota State University, Madison: 122-136.

[4] Pour un exposé formel de cette approche voir L.-C. Paquin, (1992) "La Lecture experte", *Technologie, idéologie et pratique* (10) 2-4, pp.209-222.

[5] ICON est un langage du domaine public principalement pour le traitement de chaînes de caractères qui est supporté à l'Université d'Arizona. Pour une documentation voir Griswold, R. et Griswold, M., (1990) *The Icon Programming Language*, Prentice Hall, New York.

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

[7] Nous tenons à remercier M. Claude Boivin, Ministère du revenu (Québec) pour le support théorique fourni dans ce développement.