Conférence TIA-2001, Nancy, 3 et 4 mai 2001

Classement automatique de documents et analyse terminologique de corpus

Guiraude Lame

Centre de recherche en informatique, Ecole nationale supérieure des mines de Paris,
35 rue St-Honoré, 77305 Fontainebleau Cedex
lame@cri.ensmp.fr
http://cri.ensmp.fr/~lame

Résumé

Nous traitons dans cet article du problème de la constitution d'un corpus homogène permettant de mener les analyses terminologiques nécessaires à la réalisation d'une ontologie destinée à un outil de recherche des documents juridiques français. Une méthode supervisée de classement automatique de documents juridiques est présentée. Elle repose sur la définition terminologique de catégories à partir d'un corpus constituant une taxonomie prédéfinie du domaine. Un langage d'indexation est élaboré qui permet de définir de façon homogène donc comparable les attributs de chaque catégorie et document. La construction de ce langage combine une analyse des connaissances terminologiques du domaine avec une analyse statistique, ce qui conduit à la représentation des documents et catégories par deux types de termes, les *descripteurs juridiques* et les *descripteurs contextuels*. Des mesures de similarités sont ensuite calculées entre les représentations des catégories et du document donné. Il résulte de ce processus un corpus *catégorisé* sur lequel sont menées les analyses terminologiques propres à la détection des termes du domaines et leurs relations, les éléments de l'ontologie.

1. Introduction

Le contexte de cette étude est le site Internet http://droit.org qui diffuse des documents juridiques français et communautaires. Le but de nos travaux est de faciliter la recherche et la navigation au travers des documents normatifs français que constituent les lois, décrets, arrêtés, avis de diverses autorités publiés tous les jours dans l'édition Lois et décrets du Journal Officiel de la République française (ci-après J.O.). Nous envisageons dès à présent une interaction avec l'utilisateur lui permettant d'affiner sa requête grâce à la suggestion de termes, ainsi que la possibilité d'un accès thématique aux documents. Pour atteindre cet objectif, nous envisageons d'élaborer une ontologie rassemblant les concepts du domaine et leurs relations.

Dans notre objectif final qui est celui d'élaborer une ontologie pour un outil de navigation et de recherche des documents juridiques français, le problème auquel nous répondons dans cet article est celui de la constitution d'un corpus homogène permettant de mener les analyses terminologiques nécessaires à la réalisation de cette ontologie. Nous choisissons en effet, pour la construction de notre ontologie, d'adopter une démarche principalement *bottom-up* (ou sémasiologique) [Frath, 2000] consistant à extraire les entités conceptuelles de l'ontologie des données textuelles d'un corpus et à identifier et qualifier les liens à établir entre ces entités sur la base de ce même corpus. Il faut pour ce faire disposer d'un corpus adapté. L'utilisation du

corpus même du J.O. comme base à l'élaboration de l'ontologie a été envisagée [Lame, 2000]. Mais l'expérience a montré les insuffisances d'un tel corpus, notamment le fait que les connaissances terminologiques du domaine portées par ces documents (une loi, un décret, un arrêté ... étant considérés chacun comme un document) sont relativement faible, la concentration des documents du J.O. en termes juridique étant plutôt basse. Nous avons dès lors décidé d'utiliser d'autres types de documents pour en extraire les connaissance terminologiques nécessaires à la construction d'une ontologie. Nous utilisons les codes qui sont des documents compilant les règles applicables à différents domaines appréhendés par le droit (Code civil, Code pénal, Code la voirie routière...). Ce corpus de référence s'exprime au travers du même langage de domaine (le langage juridique normatif) que les documents du J.O. et constitue une ressource terminologique importante, les concepts du domaine y étant fortemment présents. Ce corpus des codes a également la particularité de diviser la discipline en sous-domaines (droit civil pour le Code civil, droit pénal pour le Code pénal...). Nous pouvons dès lors envisager les différentes parties de notre ontologie : concepts du droit civil, concepts du droit pénal... Les termes du domaine juridique étant les termes exprimant des concepts du droit ("contrat de travail" ou "salarié" par exemple), nous ne pouvions occulter le fait que cette discipline évolue, que de nouveaux concepts sont créés par le législateur. Notre corpus de référence se doit donc d'être *évolutif*, il doit pouvoir intégrer les nouveaux concepts du droit. Nous proposons donc d'inclure méthodiquement les documents du J.O. dans notre corpus de référence. Chaque code, et plus précisemment chaque sous-division de code, est considérée comme une sous-catégorie du domaine juridique et les documents du J.O. y sont automatiquement classés. Le processus de classement des documents du J.O. dans des catégories définies sur la base du corpus Code nous permet de proposer un accès thématique aux documents du J.O. et d'envisager nos documents par sous-domaines, ce qui permet d'entrevoir les différentes parties qui constitueront notre ontologie et faciliter la nécessaire mise à jour de cette ontologie. Nous travaillons donc dans un premier temps (Section 2) à l'élaboration d'un langage d'indexation avec lequel nous menons la description terminologique des catégories de classement. Le processus de classement des documents du J.O. dans ces catégories est décrit dans la Section 3. La Section 4 étant dédiée à l'analyse du corpus *catégorisé* dès lors constitué, vers l'élaboration d'une ontologie.

2. Elaboration d'un langage d'indexation

Le choix a donc été fait de considérer les codes comme des catégories dans lesquelles les documents du J.O. vont être classés (2.1). L'objectif ici est d'élaborer un langage d'indexation de façon à obtenir des représentations des catégories. Une étude terminologique du domaine est menée (2.2) qui combine une analyse des connaissances terminologiques (2.2.1) et une analyse statistiques (2.2.2) pour détecter des termes du domaine à partir du corpus Code.

2.1. Les codes comme catégories

Nous choisissons donc de classer nos documents du corpus J.O. dans des catégories définies sur la base du corpus Codes. Le droit français est en effet divisé thématiquement en domaines et sous-domaines. Cette catégorisation est prescrite par le législateur lui-même au travers de ces recueils thématiques que constituent les Codes. Nous avons donc décidé d'utiliser ces Codes, une soixantaine sont disponibles actuellement en ligne, pour définir les catégories dans lesquelles nos documents du J.O. vont se classer.

Chaque Code est divisé en Livres, Titres, Chapitres et Sections. Considérer chaque Code comme une catégorie finale du droit reviendrait à distinguer nos documents selon qu'ils traitent de droit civil ou de droit pénal par exemple. L'exercice n'est pas inutile mais la

catégorisation serait somme toute assez grossière. Les Livres ne réalisant pas de divisions sémantiquement pertinentes, et les Chapitres constituant des séparations parfois trop fines, nous envisageons les Titres des Codes comme catégories. Les Titres des Codes sont dès lors chacuns considérés comme un même document sur lequel est menée l'étude terminologique nous permettant de constituer des représentations des catégories. Ces représentations présentent comme attributs les descripteurs sélectionnés ci-dessous.

2.2. Etude terminologique du domaine

Le processus d'indexation consiste [Moens, 2000] à extraire des documents des mots (chaînes de caractères à l'exclusion des espaces ou de signes de ponctuation) ou groupes de mots (ensembles de mots identifiés par une analyse soit statistique soit syntaxique) qui dès lors constituent les éléments du langage d'indexation, les descripteurs. Nous avons décidé de considérer à la fois les mots simples et les groupes de mots dans notre processus d'indexation. Il s'avère en effet que de nombreux concepts juridiques s'expriment au travers de groupes de mots et plus précisément de syntagmes nominaux [Catala, 1994]. Pour extraire les mots et groupes de mots de nos documents, nous utilisons un outil de traitement automatique de la langue, Lexter [Bourigault, 1994]. Cet outil mène des analyses syntaxiques sur les documents qui lui sont soumis et en extrait les syntagmes. Il liste également les noms, verbes et adjectifs apparaissant dans les documents. Nous considérons les mots et syntagmes extraits de nos documents par Lexter comme les termes candidats de notre langage d'indexation. Lexter élimine lui-même les prépositions et articles. Une sélection doit être menée sur les candidats termes qui permette de ne considérer que les plus intéressant, ceux qui caractérisent sémantiquement les documents, ceux qui relèvent de la terminologie du domaine[1]. Une première façon d'identifier les termes les plus pertinents d'un document consiste à considérer que plus un terme apparaît dans un document, plus il est significatif du contenu du document. Cette approche a été décrite comme *pauvre* d'un point de vue des connaissances [Grefenstette, 1994] dans le sens où elle repose sur une simple analyse statistique des fréquences d'apparition des termes dans un document. Cette approche comporte par ailleurs un défaut majeur qui est celui de ne pas considérer ces termes qui, en n'apparaissant que relativement peu dans un document, peuvent néanmoins être fortement chargés sémantiquement et constituer ainsi de bons indicateurs de contenu. Une deuxième approche d'identification des termes saillants d'un document est celle qui repose sur l'utilisation de ressources terminologiques extérieures aux documents traités. Des dictionnaires de domaines (listes de termes du domaine chacun suivi d'une définition) sont alors utilisés comme dans [Agostini, 1999] et [Saadoun, 1997]. L'utilisation de ces dictionnaires revient à mener une analyse des connaissances de domaines sur les documents. La méthode a l'avantage de détecter dans un document les termes significatifs d'un domaine. Cependant, le principal inconvénient de cette méthode est de reposer sur des dictionnaires de domaines qui sont le plus souvent élaborés à la main et qui souffrent par là même de défauts de réactualisation et de complétude.

Nous décidons donc de combiner ces deux approches, souvent envisagées séparément, pour constituer à partir de nos documents un langage d'indexation constitué des descripteurs juridiques et contextuels. La figure ci-dessous (Figure 1) illustre le processus qui du corpus

---

[1] Pour une analyse similaire mais par d'autres moyens des concepts exprimés dans des documents juridiques, voir [Poblet, 2000] ; pour une étude consistant à extraire les concepts juridiques de documents, voir [Osborn, 1999], [Schweighofer, 1995].

nous amène, grâce à différentes analyses et différentes sources extérieures, aux descripteurs constitutant le langage d'indexation.

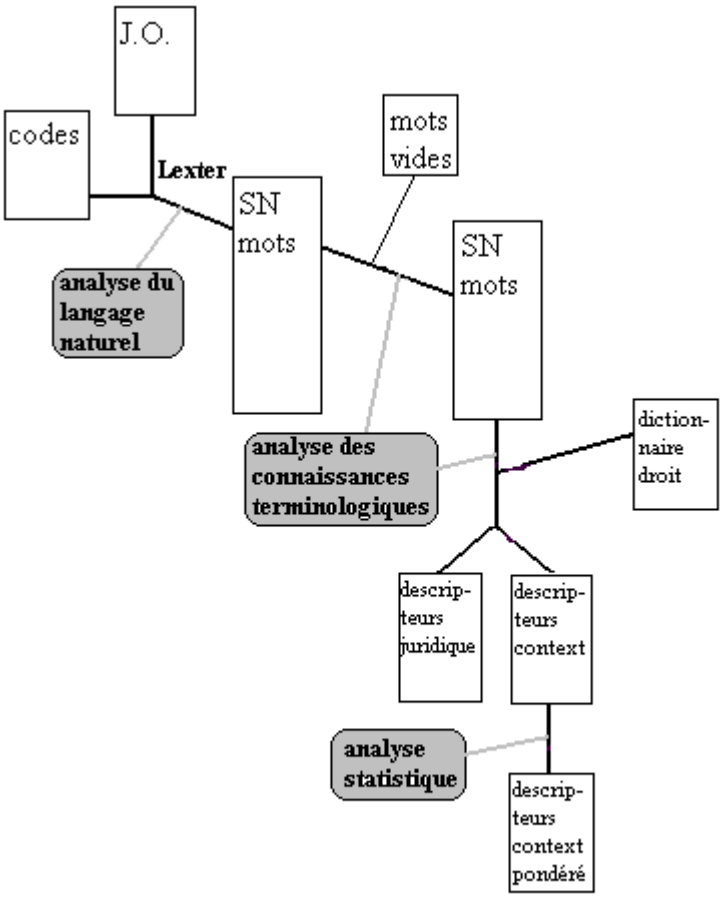

Figure 1 : le langage d'indexation

2.2.1 Analyse des connaissances terminologique

Dans l'objectif qui est le nôtre de sélectionner parmi les sorties Lexter les termes qui conviennent le mieux en tant que révélateurs du contenu des documents, une étude de la terminologie du domaine du droit est menée de deux façons. La première consiste à élaborer une liste des mots vides relativement à nos documents et la seconde consiste à utiliser des ressources terminologiques extérieures à nos documents. Sachant que la plupart des concepts juridiques s'expriment au travers de noms ou de groupes nominaux, nous décidons d'éliminer de la liste des candidats termes tous les mots identifiés par Lexter lui-même comme étant des adjectifs ou des verbes. Nous considérons par ailleurs que les nombres n'ont pas d'intérêt dans notre processus de sélection. Nous éludons donc de la liste des candidats termes tous les termes contenant des caractères numériques. Certains termes, même spécifiques du domaine juridique, ne sont que de peu d'utilité dans un processus d'indexation. Nous avons donc élaboré une liste de 81 termes vides du domaine. Cette liste constitue une analyse des connaissances du domaine. Des termes tels que "*ministre*", "*décret*" ... y sont rassemblés. Il s'avère en effet que la plupart de nos documents du J.O. font mention de leurs signataires respectifs qui sont des ministres ou des secrétaires d'état. Les termes correspondants ne permettent dès lors pas de distinguer ces documents. De la même façon, nos documents font référence à leur date de publication dans laquelle le nom du mois est mentionné en toutes lettres, les mois sont donc inclus dans notre liste de mots vides. Les termes extraits par Lexter appartenant à cette liste ou contenant des mots de cette liste de mots vides sont donc éliminés,

et ce en tenant compte des formes plurielles et des mots composés. Une première sélection est donc réalisée parmi la liste des candidats termes extraits par Lexter, par élimination des termes identifiés comme vides ou supposés tels, les adjectifs ou verbes. Un deuxième processus de sélection des termes spécifiques du domaine est réalisé par l'utilisation de ressources terminologiques extérieures au corpus. Nous utilisons les termes définis dans les dictionnaires de droit français disponibles sur l'Internet. Une liste de termes du domaine juridique est ainsi élaborée semi-automatiquement. Elle rassemble à ce jour 1471 termes (mots et groupes de mots). Cette liste permet d'effectuer une analyse des connaissances du domaine sur nos documents. Elle nous permet de détecter dans la liste des candidats termes extraits par Lexter les termes qui sont ceux du droit. De la même façon que pour les mots vides, le processus d'identification prends en considération les formes plurielles des mots et les mots composés. Cette étude terminologique menée sur nos documents nous permet d'identifier les termes juridiques parmi ceux-ci, de constituer ce que nous appelons les *descripteurs juridiques* de nos documents. Le processus est un simple processus de reconnaissance morphologique. Dès lors, les ambiguités telles que la polysémie ne sont pas résolues. Le terme « avocat » est ainsi détecté, que son sens soit celui du magistrat ou du fruit. Les *descripteurs juridiques* sont alors ceux qui portent éventuellement une sémantique juridique. Cependant, le contexte de leur extraction, le corpus J.O. ou le corpus Code, nous permet d'avancer que ces termes ont de grandes chances d'être entendus dans leur sens juridique. L'ambiguité fondamentale du langage ne peut toutefois être ignorée [Frath, 2000], nous y consacrons la Section 3 de cet article.

2.2.2 Analyse statistique

L'analyse terminologique ci-dessus décrite nous permet d'identifier les termes sémantiquement chargés de notre liste de candidats termes. Cependant, la méthode reposant que des listes construites plus ou moins à la main, il est possible que des termes importants exprimés dans les documents lui échappe. Et ce pour des raisons d'incomplétude ou de non réactualisation des listes. Nous remédions à cet inconvénient en doublant l'analyse des connaissances terminologiques d'une analyse statistique. Cette analyse statistique est menée sur ce qui reste de la liste des candidats termes dès lors que l'*index juridique* rassemblant les termes du droit en est séparé. Ces termes restant sont pondérés en fonction de leur fréquence d'apparition conjointe à un *descripteur juridique* dans un même contexte. Un contexte étant constitué par un article de code. Ainsi pour chaque *descripteur juridique* de chaque titre de code (de chaque catégorie donc), les articles contenant ce descripteur sont isolés, et dans ces articles sont identifiés les termes non éléments de l'*index juridique* qui apparaissent le plus. Ces termes sont pondérés selon leur fréquence d'apparition dans chacun de ces articles, fréquence relativisée par le nombre d'articles dans lesquels se trouve le descripteur juridique concerné et ne nombre d'articles totaux du titer en question. Ces termes sont dès lors les *descripteurs contextuels*, dans le sens où ils apparaissant en nombre autours d'un *descripteur juridique*, qu'ils en constituent une contextualisation en quelque sorte. Nous envisageons ainsi les contextes et les termes contextuels d'une façon différente que [Nazarenko, 1997]. Ainsi, dans la catégorie correspondant au titre du Code civil *de la filiation adoptive*, le descripteur juridique « *abandon* » et contextualisé par les descripteurs contextuels de la façon suivante : abandon {enfant, 0.36 ; parent, 0.14 ; aide, 0.11 ; aide social, 0.11 ; délai, 0.11 ; demande en déclaration de abandon, 0.11 ; enfance, 0.11 ; particulier, 0.11}

Aucun des termes de l'*index contextuel* ne sont éliminés car nous considérons que chacun, relativement à son poids peut être utile à définir chaque catégorie, à les différentier des autres. Finalement, chaque catégorie K est représentée par deux listes (voir tableau ci-dessous) : la

liste des *descripteurs juridique* L_i tels qu'ils sont identifiés à partir de ressources terminologiques extérieures aux documents, et la liste des termes pondérés de l'*index contextuel* C_i spécifiant le contexte d'apparition des termes de L_i. Cette méthode combinant analyse des connaissances terminologiques et analyse statistique permet nous semble-t-il de détecter l'ensemble des termes pertinents en vue de la sélection de termes d'indexation de documents.

index juridique L_i (descripteurs juridiques)	index contextuel C_i (descripteurs contextuels)
l : un *descripteur juridique* p : nombre de descripteurs juridiques dans le document considéré $L_i = [(l_{i1}),(l_{i2}),...,(l_p)]$	c : *descripteur contextuel* considéré w_c : poids attribué à c n : nombre d'articles dans lesquels l apparaît A_{il} : article dans lequel l apparaît $f(c, A_{il})$: fréquence d'apparition de c dans A_{il} q : nombre totaux d'articles dans le titre considéré r : nombre de descripteurs contextuels dans le document donné $w_c = \left(\sum_{i=1}^{n}(f(c, A_{il}))\right) \times n \,/\, q$ $C_i = [(c_{i1}, w_{i1}),(c_{i2}, w_{i2}),...,(c_{ir}, w_{ir})]$

A la suite de cette analyse terminologique des titres des Codes, nous obtenons des représentations des catégories dans lesquelles les documents du J.O. vont être classés.

3. Le processus de classement

La méthode de classement automatique de documents que nous développons dans ce travail est supervisée dans le sens où nous définissons les catégories dans lesquelles les documents sont classés indépendamment des documents eux-mêmes. Pour une approche comparable consistant à classer les documents sur la base d'une taxonomie préexistante voir [Aggarwal, 1999]. Il s'avère en effet que des techniques non supervisées de classification automatique de documents, telles que la clusterisation [Slonim, 2000], ne sont pas efficaces [Agostini, 1999], [Skarmeta, 2000] dans des cas comme le nôtre où les documents à classer appartiennent au même domaine et s'expriment dès lors au travers du même langage de domaine. L'objectif ici est de déterminer les catégories qui conviennent le plus à un document donné. Les attributs (descripteurs juridiques et contextuels) du document considéré et des catégories sont comparés au cours d'une analyse de similarité. Les catégories comportant les représentations les plus proches d'un document lui sont donc attribuée. Notre approche se déroule en deux temps : le premier consiste à comparer les *descripteurs juridiques* respectifs et à alors identifier un sous-ensemble de catégories correspondant éventuellement au document (3.1) ; le deuxième consiste à ajuster la pré-décision de l'étape précédente sur la base des poids des *descripteurs contextuels* respectifs (3.2).

3.1 Une pré-décision sur la base des descripteurs juridiques

Les distributions des termes des *index juridiques* du document considéré et des catégories

sont comparées de façon à identifier les k-catégories les plus proches du document. Une mesure de similarité est alors calculée entre le document d et chaque catégorie K. Les techniques de mesures de similarités sont reconnues performantes dans le domaine du classement automatique de document [Yang, 1999]. Pour un exemple de classement automatique de document utilisant une méthode fondée sur des mesures de similarités, voir [Masand, 1992]. Pour ce qui est des descripteurs juridiques, la mesure de similarité entre une catégorie et un document donné correspond simplement au nombres de descripteurs juridiques communs à la catégorie et au document. Dans le cas où tous les termes de l' *index juridiques* L_i d'un document sont présents dans les *index juridiques* de une ou plusieurs catégories $L_i \in \{L_1, L_2, ..., L_K\}$, cet ensemble de catégories est envisagé pour l'étape suivante. Dans les autres cas, nous décidons - pour l'instant - de considérer les 40% des catégories qui contiennent le plus de descripteurs juridiques en commun avec le document donné. Ce seuil a été déterminé de façon temporaire et pour des raisons pratiques. Il est tout à fait réajustable mais il s'avère pour l'instant que garder les 40% de la totalité des catégories qui ont le plus de descripteurs juridiques en commun avec un document semble un bon choix.

3.2 Un ajustement sur la base des descripteurs contextuels

L'ensemble des catégories identifiées dans l'étape précédente est reconsidéré sur la base des descripteurs contextuels. L'objectif est de sélectionner parmi ce premier ensemble de catégories, celles qui correspondent le plus au document, celles qui en font une description sémantique. Nous considérons dès lors dans cette étape les descripteurs contextuels et leurs poids. Pour ce faire, nous attribuons des scores de similarité aux tuples catégories K-document d. Ce score est établit comme suit : pour chaque descripteur contextuel c commun au document d et à la catégorie K, le nombre d'occurrence de c dans d est multiplié par le poids attribué à c dans K (voir Section 2.2.2). Ces produits sont ensuite additionnés pour obtenir le score de similarité attribué au tuple K – d. Un deuxième processus de sélection de catégories est donc mené parmis les catégories présélectionnées sur la base des descripteurs juridiques. Ce processus en deux temps nous permet d'identifier les catégories qui correspondent le mieux au document donné. Nous pouvons dès lors envisager un accès thématique par catégorie à nos documents du J.O.

3.3 Exemple – évaluation

Pour illustrer le processus tel que décrit ci-dessus, nous envisageons les catégories correspondant aux titres du Code civil et la loi n° 99-944 du 15 novembre 1999 relative au pacte civil de solidarité[2].

descripteurs juridiques		**descripteurs contextuels**	
intitulé des catégories	**mesure de similarité**	**intitulé des catégories**	**mesure de similarité**
des privilèges et hypothèques	34	du divorce	3366
des successions	33	du mariage	3240
des donations entre vifs et des testaments	33	des conventions relatives à l'exercice des droits indivis	2935
du mariage	29	de la société	2681
du contrat de mariage et des régimes	26	des successions	2659

[2] la loi est disponible à l'URL suivante : http://admi.net/jo/19991116/JUSX9803236L.html

matrimoniaux			
des contrats ou des obligations conventionnelles en général	23	de la vente	2556
de la minorité, tutelle et émancipation	22	des contrats ou des obligations conventionnelles en général	2539
des actes de l'état civil	20	des donations entre vifs et des testaments	2452
du divorce	20	de la prescription et de la possession	2116
de la vente	20	de la majorité et des mineurs protégés par la loi	2066
du contrat de louage	20	de la minorité, tutelle et émancipation	1931
de la société	19	des actes de l'état civil	1277
de la majorité et des mineurs protégés par la loi	16	des privilèges et hypothèques	483
de la prescription et de la possession	16	du contrat de louage	303
des conventions relatives à l'exercice des droits indivis	15	du contrat de mariage et des régimes matrimoniaux	253

Le tableau ci-dessus montre dans sa première colonne les catégories sélectionnées et classées en fonction de la mesure de similarité des descripteurs juridiques. Il y en a quinze en application du seuil de 40%. La deuxième colonne procède à un reclassement de ces quinze catégories sur la base de la mesure de similarité des descripteurs contextuels. S'il est très difficile pour un expert juridique quel qu'il soit de quantitier les thèmes traités par un texte, on ne peut établir avec certitude que la loi relative au PACS évoque effectivement plus les *successions* que les *donations entre vifs et testaments*, notre méthode en deux temps permet tout de même d'avancer que la loi relative au PACS évoque plus les *conventions relatives à l'exercice des droits indivis* (similarité contextuelle 2935) que les *contrats de louage* (similarité contextuelle 303) alors même que les similarités juridiques de ces deux catégories sont proches (20 et 15).

La méthode présentée a été appliquée à l'ensemble des lois publiées en 2000 et les premières de 2001 (87 documents au total) pour les 38 catégories correspondant aux 38 titres du Code civil. Les résultats obtenus ont été examinés par un expert qui estime que pour 72 % de ces documents, les catégories identifiées conviennent. Cependant, cette expérience est partielle dans le sens où, ne considérant que le Code civil, seul l'aspect droit civil de nos documents est envisagé. Or toutes les lois ne traitent pas de cet aspect du droit. Ce point permet d'expliquer une partie des mauvais résultats. Il s'avère par ailleur que ne considérant que le Code civil, la distinction entre les catégories n'en est que plus difficile. Le lexique à l'intérieur même du Code civil étant en effet relativement homogène alors qu'il l'est moins entre le Code de la voirie routière et le Code de déontologie des sages-femmes par exemple. Dès lors, une expérimentation à grande échelle, comprenant tous les Codes, s'impose, elle est à ce jour en cours.

4. Analyse terminologique du corpus *catégorisé*

Notre méthode consiste donc à faire correspondre un ensemble de documents du J.O. à chaque catégorie définie sur la base des Codes. Il en résulte un corpus *catégorisé* qui constitue dès lors un champs d'investigation pour des analyses terminologiques de domaines telles que celles présentées par [Aussenac, 2000].

4.1 Acquisition de connaissances terminologiques

Dès lors qu'un ensemble de documents est identifié comme appartenant à une catégorie, une évolution de la définition de cette catégorie est envisageable. Ce sont les nouveaux documents qui nous permettent d'acquérir de nouvelles connaissances terminologiques. Un nouveau document d étant identifié par notre système comme appartenant à la classe K, la définition de cette classe par la liste des *descripteurs juridiques* L_{iK} et la liste des *descripteurs contextuel* C_{iK} peut dès lors être complétée sur la base de ce document lui-même défini par deux listes L_{jd}, C_{jd}. On peut envigager qu'un terme $l_j \in L_{jd}$ ou $c_j \in C_{jd}$ soit intégré aux listes L_{iK} ou C_{iK} dès lors que son nombre d'occurrence dans le document est relativement élevé. Pour reprendre l'exemple de la loi relative au PACS, on peut envisager d'ajouter le terme "pacte civil de solidarité" (39 occurrences dans la loi pour une moyenne d'apparition des descripteurs juridiques de 4) à la définition des catégories identifiées par notre méthode. Le classement des documents du J.O. nous permet dès lors d'envisager la réactualisation de notre ontologie et plus spécifiquement d'indentifier les parties de l'ontologie à modifier au vu des nouveaux documents.

4.2 Détection de relations d'association entre termes

Notre corpus *catégorisé* constitue un champs d'investigation propre à la détection de relations d'associations entre termes de l'ontologie. La première méthode a consisté en une analyse morphologique [Zweigenbaum, 2000] menée sur les descripteurs juridiques. Chaque mot seul est ainsi relié aux syntagmes nominaux le comprenant. On obtient des résultats du type :

donateur {ayants cause de le donateur ; bien de le donateur ; bien présent de le donateur ; bien à le donateur ; charge de le donateur existant ; charge de le successeur de le donateur ; dernier enfant de le donateur …}.

Une deuxième méthode a consisté à lier les termes présents dans les intitulés des catégories (*contrat* par exemple présent dans *du contrat de mariage et des régimes matrimoniaux*, dans *des contrats ou des obligations conventionnelles en général* et dans *du contrat de louage*) aux descripteurs juridiques définissant cette catégorie. On obtient des résultats du type :

contrat {bien ; bail ; preneur ; mariage ; conjoint ; bailleur ; dette ; chose ; locataire ; louage …}

Notre système de classement automatique de documents est donc envisagé comme une première étape vers l'élaboration d'un corpus structuré de façon partiellement hiérarchique nous permettant d'en extraire une terminologie du domaine la plus fine possible. La première étape vers la conception d'un thesaurus est ainsi accomplie qui nous permettra d'élaborer un système à base de connaissances juridiques [Visser, 1998].

5. Conclusion

Une méthode supervisée de classement automatique de documents juridiques a donc été présentée. Elle repose sur la définition terminologique de catégories à partir d'un corpus constituant une taxonomie prédéfinie du domaine. Un langage d'indexation est élaboré qui permet de définir de façon homogène donc comparable les attributs de chaque catégorie et document. La construction de ce langage combine une analyse des connaissances terminologiques du domaine avec une analyse statistique, ce qui conduit à la représentation des documents et catégories par deux types de termes, les *descripteurs juridiques* et les *descripteurs contextuels*. Des mesures de similarités sont ensuite calculées entre les représentations des catégories et du document donné en considérant dans un premier temps les *descripteurs juridiques* et dans un deuxième temps, pour l'ajustement de la décision, les

descripteurs contextuels. Ce système de classement automatique de document fondé sur la terminologie est la première étape vers une étude linguistique plus approfondie. Les expériences menées jusqu'à présent n'ont été que partielles dans le sens où seules les catégories définies sur la base du Code civil étaient envisagées. La prise en compte de tous les Codes est la prochaine étape de nos travaux vers une ontologie du droit.

Références

AGGARWAL C., Gates S., Yu P. (1999). On the merits of building categorization systems by supervised clustering. In *Int. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, San Diego, Ca, USA, Proceedings, p. 352

AGOSTINI BUENO T., GRESSE VON WANGENHEIM C., DA SILVA MATTOS E., HOESCHL H., BARCIA (1999). Jurisconsulto: Retrieval in Jurisprudencial Text Bases using Juridical Terminology. In *Proceedings of the 1999 Int. Conference on Artificial Intelligence and Law*, Oslo, p.147

AUSSENAC-GILLES, SZULMAN S. (2000). Modélisation du domaine par une méthode fondée sur l'analyse de corpus. In *Actes de la conférence IC 2000*, Toulouse, p. 93

BOURIGAULT D. (1994). Lexter, un Logiciel d'Extraction de TERminologie. Application à l'extraction des connaissances à partir de textes. Thèse, Mathématiques, Informatique appliquée aux sciences de l'homme. EHESS, Paris.

CATALA P. (1994*). Le droit à l'épreuve du numérique. Jus ex Machina*. Paris, Eds. Presses Universitaires de France.

FRATH P., OUESLATI R., ROUSSELOT F. (2000). *Identification de relations sémantiques par repérage et analyse de cooccurrences de signes linguistiques*. In CHARLET J., ZACKLAD M, KASSEL G., BOURIGAULT D., Paris, eds. *Ingéniérie des connaissances, évolutions récentes et nouveaux défis*. Eds Eyrolles.

GREFENSTETTE G. (1994). *Explorations in Automatic Thesaurus Discovery*. (Kluwer Academic Publishers)

LAME G. (2000). Acquisition de connaissances à partir de textes, vers l'élaboration d'une ontologie du droit. In *Actes de la conférence RJCIA 2000*, Lyon p. 211

MASAND B., LINOFF G., WALTZ D. (1992). Classifying news stories using Memory Based Reasoning. In *Proceedings of the 1992 Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 59

MOENS M.-F. (2000). *Automatic Indexing and Abstracting of Document Texts*. (Kluwer Academic Publishers), p. 50

NAZARENKO A., ZWEIGENBAUM P., BOUAUD J., HABERT B. (1997). Corpus-Based Identification and Refinement of Semantic Classes. In Daniel R. Marys Ed., In *Proceedings of the 1997 American Medical Informatics Association (AMIA) Annual Fall Symposium*, Nashville, Tenessee, octobre 1997

OSBORN J., STERLING L. (1999). JUSTICE: A Judicial Search Tool Using Intelligent Concept Extraction. In *Proceedings of the 1999 Int. Conference on Artificial Intelligence and Law*, Oslo, p.173

POBLET M., PASCUAL E., ROIG A., COMIN J., CASANOVAS P. (2000). La représentation de la pluridimentionalité pragmatique des discours juridiques : quelques problèmes. In BOURCIER D., HASSET P., ROQUILLY C. *Droit et Intelligence artificielle. Une révolution de la Connaissance Juridique*. Paris, Eds. Romillat, p. 178

SAADOUN A., ERMINE J.-L., BELAIR C., POUYOT J.-M. (1997). A Knowledge Engineering Framework for Intelligent Retrieval of Legal Case Studies. *Artificial Intelligence and Law*, vol. 5(3), pp. 179-205

SKARMETA A., BENSAID A., TAZI N. (2000). Data Mining for Text Categorization with Semi-Supervised Agglomerative Hierarchical Clustering. *Int. Journal of Intelligent Systems*, vol. 15, pp. 633-646

SLONIM N., TISHBY N. (2000). Document Clustering using Word Clusters via the Information Bottleneck Method. In *Proceedings of the 2000 Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 208

SCHWEIGHOFER E., WINIWARTER W. & MERKL D. (1995). Information filtering: The computation of similarities in large corpora of legal texts. In *Proceedings of the 1995 Int. Conference on Artificial Intelligence and Law*.

VISSER P.R.S, BENCH-CAPON T.J.M. (1998). A comparison of four ontologies for the design of legal knowledge systems. *Artificial Intelligence and Law*, vol. 6, pp. 27-57

YANG Y., LIU X. (1999). A re-examination of text categorization methods. In *Proceedings of the 1999 Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 42

ZWEIGENBAUM, P., GRABAR, N. (2000). Relations morphologiques et structuration de terminologie. In *Actes de la conférence IC 2000*, Toulouse, p. 325