Les domaines du droit se reflètent-ils dans le langage juridique ?

Leonhard Voltmer et Oliver Streiter

La recherche du droit passe par la langue.

Mots-clés: branches de droit, domaines de droit, langage spécialisé, langage technique, langage juridique, terminographie

Abstract :
This article discusses in a theoretical part the meaning of « fields of law » inside law itself and in legal language. In the second part the authors present computer-linguistic experiments in automatic field-of-law recognition of legal texts. It is argued that the results show that the conceptual « fields of law » shape the legal language in such a strong way, that the discourse of different legal fields can be recognized by a computer.

Introduction :
Il y a peu de dyslexiques parmi les juristes, mais beaucoup d'hommes de lettres. Le droit est véhiculé presque uniquement par la langue, et qui cherche le droit doit commencer par la langue. Le traitement automatique des langues peut-t-il être utile pour une telle recherche ? Vu le peu de résultats on dirait non. Il semblait que seulement l'homme a la capacité de comprendre, d'arriver au plein sens derrière les mots. Dans la construction d'ontologies on essaie de spécifier le sens de chaque mot et de chaque construction linguistique. Les auteurs proposent une méthode beaucoup plus simple et notre objectif sera plus modeste. Nous essayons de reconnaître le domaine d'un texte à partir d'une collection de textes classifiée par domaine.
L'article se compose d'une première partie théorique sur les domaines du droit (A) et d'une deuxième partie sur la recherche linguistique (B). Nous divisons la partie théorique dans un part a) pour la signification des notions utilisées pour aborder dans b) la portée de la question de recherche. La partie pratique (B) décrit d'abord l'expérience (a) et s'occupe ensuite de son interprétation (b).

A) Qu'est-ce qu'une branche de droit ?

a) Théorie: le droit et sa langue

Commençons par définir les notions « domaine de droit » et « branche de droit ». La domaine donne une idée de possession, puissance, autorité et souveraineté d'un territoire ou bien foncier. Un **domaine de droit** indique une autorité exclusive dans un monde bidimensionnel. Si c'est du droit public, ce n'est pas du droit civil.

La branche porte l'idée de généalogie, ramification et différenciation d'un seul être. Une **branche de droit** indique que le droit est un seul système cohérent, possédant diverses

spécialisations et sous-spécialisations. La prescription trentenaire est un principe général du droit et s'applique également au droit public et au droit civil.[1]

Il est également nécessaire de distinguer le discours juridique du discours linguistique.

Dans le droit, les domaines de droit sont définis de manière précise et exclusive l'un de l'autre. Leur structure cartésienne reflète la différenciation du droit même. Dans le Droit des anciens Romains, on distinguait les dispositions d'intérêt public et des règles d'ordre privé (*ius civile*) pour les citoyens romains. Aujourd'hui cette classification des domaines nous semble fondamentale,[2] mais en fait certains systèmes juridiques contemporains l'ignorent, ou mieux dit, suivent une autre différenciation. Dans les systèmes juridiques anglo-saxons, ce n'est pas l'opposition entre le Droit public et le Droit privé qui est retenue, mais celle entre le Droit des personnes et le Droit des choses.[3]

Dans la linguistique, le droit se présente comme langue de spécialité très bien développée. La langue de spécialité est l'ensemble des moyens linguistiques dans un domaine spécifique assurant la communication des personnes de ce domaine.[4] D'après cette définition, les moyens linguistiques de divers domaines spécifiques peuvent se recouper et en réalité ils se recoupent en grande partie. La difficulté de la recherche linguistique dans les langages techniques réside plutôt dans la tâche de trouver et de décrire les peu de critères (souvent limité au lexique) qui permettent la différenciation.

Dans le droit et dans le langage, il y a des champs de spécialité, mais l'accentuation est diverse. Le droit est comme une peinture : Il y a du rouge du bleu et du vert, et si on examine attentivement le rouge, on discerne rouge carmine, rouge cerise, rouge de corail et rouge cuivré. Le langage est comme une image de télévision de cette peinture : Elle est composé de beaucoup de points colorés, et pour représenter le bleu et le vert, on a besoin d'utiliser aussi une certaine quantité de ponts rouges. L'accentuation du droit est donc sur l'extension d'un champ, et celle du langage sur l'hérédité et la racine commune.

Si on veut, la structure du droit est exclusive comme celle des domaines, la structure du langage juridique est héréditaire comme celle des branches d'un arbre. Dans le droit, confondre les domaines revient à confondre les idées. Dans la langue, la branche de langue juridique ne peut vivre qu'en connexion avec le tronc ferme de la langue générale.

Ceci dit, on ne peut pas compter uniquement sur la structure du langage pour rendre la structure du droit. Exclu de la voie analogue, le langage doit utiliser une autre technique pour rendre les idées d'un système normatif et idéal.

---

[1] Conseil d'État statuant au contentieux N° 247976, séance du 1er juillet 2005, lecture du 8 juillet 2005, juge que l'article 2262 du code civil français exprime un principe général du droit. Un autre principe général de droit empêche l'abus de droit (l'exercice d'un droit de manière injuste).

[2] La classification du droit belge: http://www.thesaurusuniversel.be .

[3] « Les branches du droit » Faculté de Droit de l'Université Lyon 3, http://www.facdedroit-lyon.com/modules/ivd/04_branches.php: 11/08/2005.

[4] Lothar Hoffmann „Vom Fachwort zur Fachtextsorte", en: „Vom Fachwort zum Fachtext", (Narr Tübingen 1988), 131-144, 133.

Les juristes disputent souvent, mais ils vont d'accords que la solution de droit se trouve dans des textes. Le droit est rendu par le langage juridique, quoiqu'il soit leur liaison. Les notions du droit doivent se traduire en langue pour faire effet, c'est-à-dire pour juger la qualité normative des événements et pour former la réalité.

En renversant l'argument, le droit se trouve en recherchant la langue. De tout temps les agents compétents pour prononcer ce qui est droit commençaient à le chercher dans des textes écrits. Aujourd'hui des machines peuvent rechercher dans des textes, et elles sont capables de traiter une énorme quantité de texte en peu de temps. Leur fouille se borne pourtant au médium qui porte le message, à la langue juridique, tout en ignorant les notions juridiques. La solution serait proche si on pouvait établir le lien manquant entre la parole et la notion.
Est-ce qu'il est possible de révéler une telle relation ? Est-ce qu'on retrouve la structure cartésienne du droit dans le langage juridique ?

b) pratique : le classement formel de textes
Une répartition évidente de textes est la répartition en grandes codifications. C'est la solution du Centre de Recherche en Informatique de l'École des Mines de Paris.[5]
Malgré l'attractivité de cette approche, soit de classer les documents selon leur codification, la réalité n'est point si simple. A titre d'exemple voyons le premier article du Code Civil Français, qui selon la codification devrait faire partie du Droit Civil : « Les lois et, lorsqu'ils sont publiés au Journal officiel de la République française, les actes administratifs entrent en vigueur à la date qu'ils fixent ou, à défaut, le lendemain de leur publication «. Il s'agit d'une règle concernant des actes publics, d'une part, la loi et d'autre part, l'acte administratif publié dans le JO. Sans doute cette règle se réfère à des objets de Droit Public, et aussi la sanction prescrite est de Droit Public. En fait, tout le Livre Premier du Code Civil ne traite pas du Droit Privé ou du Droit Civil, mais du Droit Public.
Tout généralement les règles d'un Code traitent de divers domaines de droit, par exemple: une activité privée (droit civil) connaît des limites d'ordre public (droit public) et ces limites sont défendues par des peines (droit pénal). La répartition des textes en codes ne tient conte ni du contenu des règles, ni de la langue juridique utilisée, mais de leur qualification sommaire et formelle. Cette approche à la classification ignore le lien entre notion et langue juridique qui est par contre l'objet de notre recherche.

B) Branches : idée ou réalité ? Une recherche linguistique

a) Description de la recherche : identification du domaine juridique d'un texte dont le domaine est encore inconnu

Le lien entre le droit et la langue juridique s'enregistre dans des dictionnaires terminologiques. Ces dictionnaires décrivent une notion avec sa définition et ses termes. On remarque que ce n'est point le terme qui est défini, mais la notion. Normalement la définition se réfère implicitement ou explicitement à un domaine de droit. En outre beaucoup de dictionnaires donnent des exemples de contextes.

[5] „Une ontologie du droit français à fins documentaires", Centre de Recherche en Informatique, École des Mines de Paris, http://ontologie.w3sites.net/index.html .

L'idée principale est que tous les exemples de contextes et définitions qui se réfèrent à des notions du même domaine de droit se ressemblent. Ces textes sont orientés vers les notions d'un champ sémantique commun.[6] L'hypothèse est que cette « direction commune » se reflète suffisamment dans le langage pour permettre de classifier des textes dont le domaine est encore inconnu.[7] Par conséquent nous allons falsifier l'hypothèse contraire : »La similitude du langage juridique n'est en aucune relation avec la classification en domaines de droit ».

L'EURAC[8] détient un dictionnaire juridique de langue italienne.[9] Environ 15.000 notions sont décrites avec leurs terme(s), domaine de droit,[10] définition et contexte (voir image 1 première colonne).

A fin de nos expériences, nous intégrons les contextes et définitions d'un même domaine dans un corpus de textes de domaine (voir image 1 deuxième colonne). Ensuite nous coupons chaqu'un des 24 corpus ainsi obtenus en dix tranches de à peut près la même taille (voir image 1 troisième colonne). En prenant une tranche de chaque corpus, nous pouvons former un 25ème corpus de textes mixtes. La tâche est de reconnaître le domaine des textes du corpus 25 sur la base des 24 corpus d'apprentissage.
Le système classifie des textes qui sont déjà classifiés, et nous pouvons donc facilement vérifier sans influences subjectives ou possibilité d'interprétation la réussite du système.
Pour obtenir des résultats valides nous répétons cette expérience dix fois avec des données différentes, c'est-à-dire avec une autre des dix tranches à former le 25ème corpus de textes mixtes.[11]

---

[6] Le champ sémantique est l'„ensemble d'unités-mots appartenant à un même champ de signifiés et partageant un certain nombre de leurs constituants sémantiques (sèmes). " http://www.med.univ-rennes1.fr/sisrai/dico/R262.html : 20/09/2005.

[7] Autrement dit, nous cherchons les différences entre des champs sémantiques en utilisant des sèmes, quoique nous ignorions l'identité des sèmes. Utilisant des n-grams, nous n'utilisons pas les sèmes mêmes, mais leur fragments, parfois trop petits ou trop grands. Nous avons établit lors d'une pré-expérience que la longueur plus adaptée pour ces fragments est de 8 signes (voir plus bas).

[8] Académie Européenne de Bolzano/Bozen, www.eurac.edu .

[9] Le dictionnaire se trouve ici: www.eurac.edu/bistro et contient aussi allemand et deux variantes de ladin, (ladino en italien, ladin en ladin), une langue rhéto-romane.

[10] Il y en a 24, le même nombre comme le premier niveau de l'hiérarchie dans: « Archivio DoGi - Dottrina Giuridica - Abstract di articoli pubblicati in riviste italiane », http://nir.ittig.cnr.it/dogiswish/consistenze/class2000.htm : 20/09/2005.

[11] Cette technique s'appelle validation croisée.

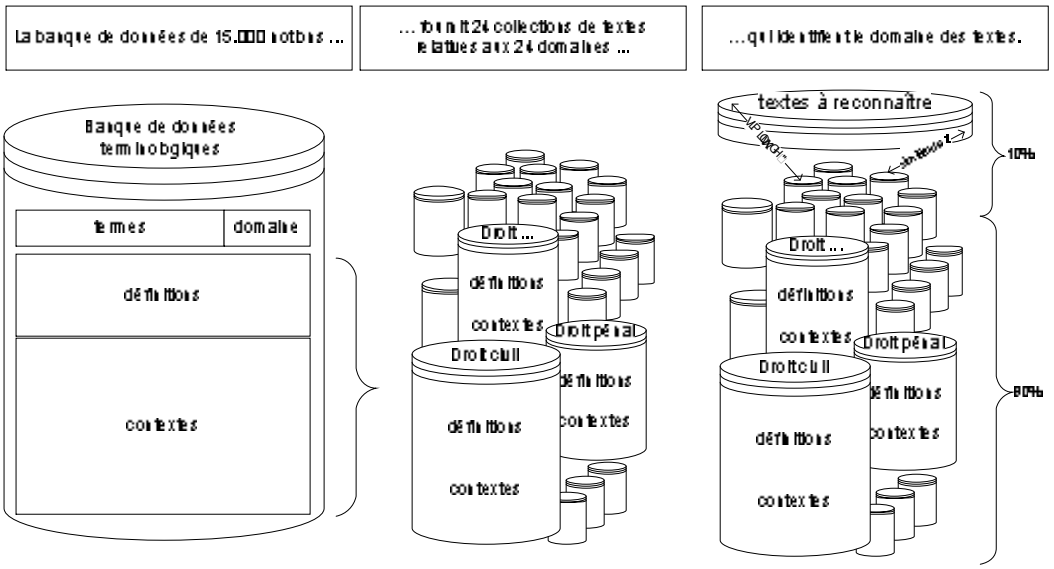

Image 1 : Schéma de l'identification du domaine d'un texte utilisant les textes descriptifs d'un dictionnaire de droit

Pour donner un exemple, la banque de données terminologique contient une entrée pour le terme « atto amministrativo », classifié manuellement comme appartenant au droit administratif. La même entrée contient la définition : « Qualsiasi manifestazione di volontà posta in essere da una autorità amministrativa nell'esercizio di una funzione amministrativa per un caso concreto ». Nous intégrons ce texte dans le corpus spécifique nommé « droit administratif ». Ce corpus est partitionné en dix tranches. N'importe quelle tranche contient notre texte, il est utilisé en neuf répétitions dans le corpus d'apprentissage « droit administratif » et en une répétition dans le corpus de teste. Si le système assigne automatiquement à notre texte la classe « droit administratif », nous contons une réussite, sinon un échec.

La similitude entre un texte à classifier et les 24 corpus d'apprentissage se calcule techniquement comme suit : Nous coupons tous les textes en chaînes de 8 caractères (ici « qualsias », « ualsiasi », « alsiasi », « lsiasi m », etc.).[12] Nous obtenons ainsi des vecteurs de chaîne de caractères pour le texte à classifier et pour les 24 corpus d'apprentissage. Nous calculons ensuite les proximités du vecteur du texte à classifier avec chaqu'un des vecteurs des 24 corpus. Le vecteur plus proche gagne et le texte est classifié comme l'indique ce corpus. Pour retourner à notre exemple : Si le vecteur de droit administratif est le plus similaire, le texte est classifié comme tel. Nous pouvons contrôler ce résultat en révélant d'où le texte à reconnaître a été extrait.

Le texte ici reporté est composé de seulement 150 caractères. Dans l'expérience les textes à reconnaître avaient en moyenne 388 caractères. Pour comprendre l'influence de la lon-

[12] Une description plus détaillée des pré-expériences pour trouver la longueur idéale des n-grams se trouve en Leonhard Voltmer, « Werkzeuge für Rechtsdatenbanken », dissertation LMU München 2005, p. 52-58, http://edoc.ub.uni-muenchen.de/archive/00003716/01/Leonhard_Voltmer.pdf : 20/09/2005.

gueur du texte à reconnaître, nous avons répété l'expérience avec 2, 4, 8 et 16 textes collés ensemble. Ainsi nous avons atteint des textes de 776, 1552, 3104 et 6208 caractères. Le texte de cet article conte jusqu'ici un peu plus de 6000 caractères. Avec cette augmentation graduelle et proportionnelle, nous avons pu examiner la performance en dépendance de la quantité de texte, à partir de quelques lignes jusqu'à plusieurs pages de texte. Nous devons cependant concéder que ce rassemblement de texte n'est pas un texte dans le sens classique, parce que ce sont plusieurs textes autonomes (définitions et contextes) réunis, collés l'un après l'autre. Les répétitions sont probablement plus fréquentes dans des textes autonomes collés que dans un texte complet.

Les résultats :
Notre procédé a reconnu correctement le domaine de droit de 60 % des textes. En augmentant le nombre de caractères, le taux de réussite augmente. Avec 776 caractères, nous atteindrons déjà 90 %, avec 1552 environ 95 % et avec plusieurs pages de texte nous pouvons presque toujours reconnaître correctement le domaine de droit.
Ce bon résultat[13] était surprenant même pour nous et requiert une interprétation.

b) Evaluation et conclusions: les idées forment la réalité

Notre hypothèse était : »La similitude du langage juridique n'est en aucune relation avec la classification en domaines de droit ». Nos résultats ne sont pas aléatoires, par conséquent l'hypothèse est falsifiée. En rebours notre hypothèse positive est nourrie : nous continuons à supposer que la ressemblance au niveau du langage juridique correspond à une familiarité sémantique. La réponse à la question du titre est alors : oui, les branches de droit se reflètent dans le langage juridique.

La seule caractéristique qui distingue les textes est leur domaine de droit. Par conséquent cette expérience démontre que les domaines de droit se reflètent dans le langage juridique. Les définitions et contextes d'un même domaine de droit possèdent des caractéristiques lexiques communes qui les distingue de tous les autres domaines de droit. Il semble peu, mais cette hypothèse est la base pour toute recherche de caractéristiques dans les domaines de droit.

Il est difficile d'estimer si ce procédé se prête à une **commercialisation**. Pouvoir reconnaître le domaine de droit d'un texte automatiquement serait très utile pour la classification des textes juridiques. On classifie dans la terminologie,[14] la lexicologie, la linguistique de corpus et en général dans la gestion d'informations juridiques des systèmes d'information. Le champ d'application est vraiment vaste.

---

[13] En comparaison avec d'autres recherches comme Guiraude Lame, « A Categorization Method for French Legal Documents », http://citeseer.ist.psu.edu/640479 : 21/09/2005 et *id*., « Knowledge acquisition from texts towards an ontology of French law », http://citeseer.ist.psu.edu/lame00knowledge.html : 21/09/2005.

[14] En terminologie juridique, il paraît particulièrement intéressant de pouvoir générer automatiquement des contextes toujours actualisés dans le domaine correct. Il suffirait de définir le contexte idéal comme „une phrase de 150 à 300 signes qui contient le terme et qui appartient au domaine de droit du terme" et le programme cherche toujours dans les textes actuels. En choisissant le contexte manuellement, les contextes vieillissent et risquent de citer des exemples caducs.

Pour **reproduire** des résultats de cette qualité, il faudra une base très solide de textes de même répertoire que le texte à classifier. Il est possible que le texte d'un manuel de droit administratif ressemble plus au texte d'un manuel de droit civil qu'au texte produit par un tribunal administratif.
Deuxièmement, les résultats de très bonne qualité (90% de réussite et plus) se réfèrent à des textes virtuels, qui n'existent pas sous la forme utilisée. Dans un texte de plusieurs pages, le sujet change et normalement les répétitions sont évitées. Dans l'addition de textes courts et précis, les répétitions sont systématiques et leur provenance est donc plus facile à reconnaître.

En conclusion nous voulons présenter une **expérience de suite**, ou peut-être plutôt un jeu d'idées. Vu que nous avons 24 ensembles de textes et une méthode pour comparer la similitude de textes, nous sommes dans la position de calculer la proximité des (textes de) domaines de droit. La matrice de 24x24 est de difficile interprétation, et ainsi nous avons transformé les chiffres en un cluster.

Nous prenons le domaine qui est en moyenne le plus proche à tous les autres domaines comme racine. Ce domaine s'appelle « juris » dans l'image 2 et contient des textes pour termes comme « validité », « abus de droit » et « assistance judiciaire ». Ensuite nous attachons les domaines qui sont plus proches à « juris » qu'ils ne le sont à un autre domaine. Nous continuons de la même manière et arrivons après la quatrième niveau d'hiérarchie à l'arbre généalogique de l'image 2.

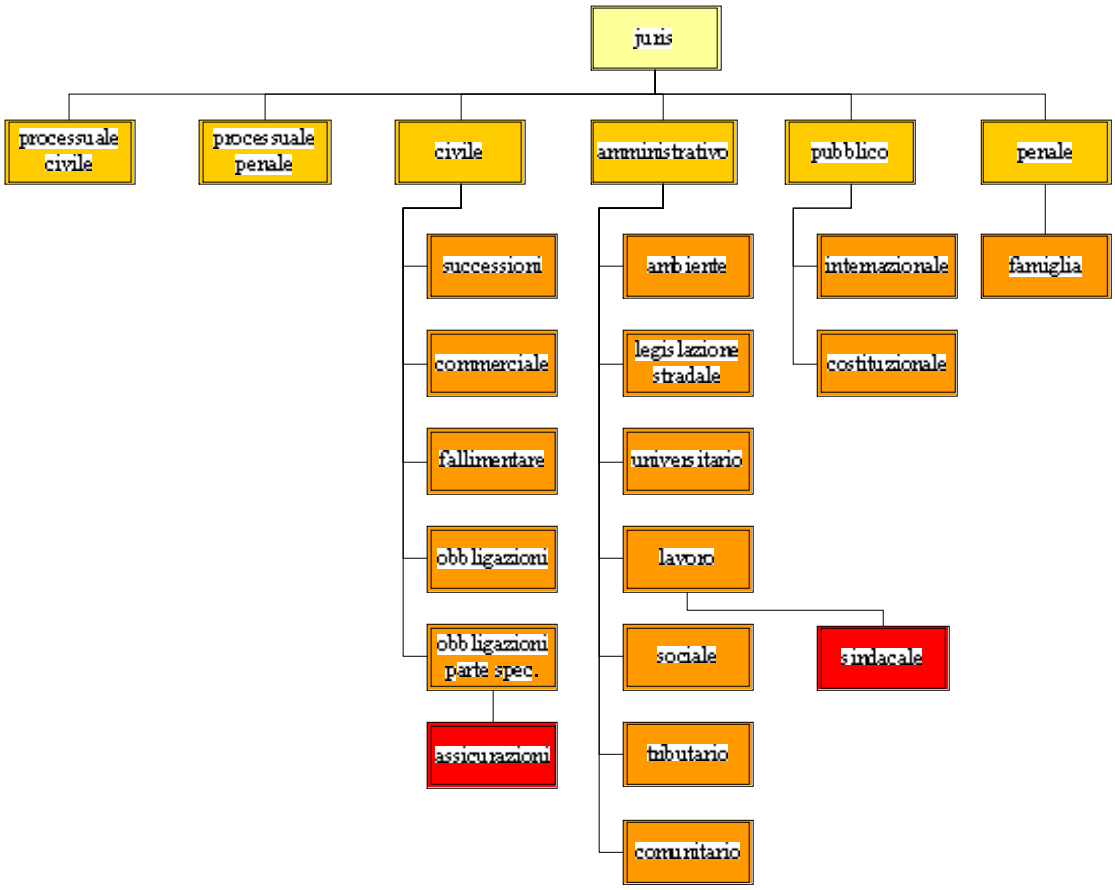

Image 2 : La hiérarchie implicite des domaines de droit reconstruit par l'ordinateur

Le résultat est de nouveau assez surprenant : toute la hiérarchie des domaines du droit est reproduit par l'ordinateur.

On remarque que les terminologues qui ont réalisé le dictionnaire de termes, définitions et contextes selon les domaines de droit ont fait ceci sans spécifier une hiérarchie quelconque des domaines. Aux fins du dictionnaire il suffit de distinguer entre les domaines. La connexion des thèmes au delà d'un domaine de droit n'est même pas implicitement contemplée. Cette expérience illustre que des liens assez notables existent entre le langage juridique et le sens juridique.

Conclusion :

La comparaison de droit a toujours été limitée à la méthode qualitative. Les limitations d'une analyse trop souvent subjective, basée sur seulement quelques notions de droit délibérément choisies, sont évidentes. Les résultats ressemblent aux hypothèses de travail et ne couvrent que des parties très spéciales des systèmes de droit comparés.
Nous avons obtenu des résultats qui encouragent l'application de la **méthode quantitative**. Nous pouvons comparer divers systèmes de droit au niveau de leurs conceptions de base. Notre procédé permet de visualiser les fondements de systèmes d'un système de

droit, et de les comparer. Une question de recherche particulièrement intéressante serait d'examiner le droit anglo-saxon (*common law*) et de comparer sa structure avec celle d'autres systèmes. Peut-être la division en Droit des Personnes et Droit des Choses au sommet de l'hiérarchie n'est qu'un autre couvercle pour des structures profondes très similaires au droit continental ?

La recherche quantitative sera d'intérêt pour la **traduction juridique**. Le traducteur peut déjà profiter d'instruments de la linguistique de corpus en recherchant la fréquence d'un certain terme. Pour le droit, cette forme de recherche avait toujours l'hypothèque que la fréquence absolue dévoyait, parce qu'il y a trop d'homographes dans le langage général et dans d'autres domaines de droit.
Prenons encore un exemple. Un traducteur juridique doit traduire « aide » en Allemand. Il trouve deux termes synonymes, « Zuschuss » et « Beihilfe ». Une recherche de fréquences montre que le terme « Beihilfe » est plus fréquent. Le traducteur se méfie et vérifie la fréquence des termes dans des textes juridiques, mais il reçoit le même résultat.[15] Avec la possibilité de chercher les cooccurrences dans les divers domaines de droit, il peut découvrir que le terme plus fréquent a deux significations, une en droit pénal (aide à un délit) et une autre en droit administratif (aide économique).[16] Le terme « Beihilfe » dans le sens de « aide économique » est moins fréquent que le terme « Zuschuss ». Le traducteur choisira, pour éviter une mécompréhension, le terme sans possibilité d'équivoque.

La classification automatique de texte dans domaines de droit est possible et promet un grand futur pour maintes applications.

[15] Quand on n'a pas de corpus équilibré, une recherche de cooccurrences donne déjà une très bonne idée : On dans tous les textes de l'Internet qui contiennent contemporanément le mot „Recht" (=droit).
[16] Cette information se trouve, aussi pour des textes français, dans le site http://wortschatz.uni-leipzig.de/ : 20/09/2005, sous « Sachgebiet » et « Synonyme ».