Notes de lecture

Rubrique préparée par Denis Maurel

Université François Rabelais Tours, LI (Laboratoire d'informatique)

Pierre M. NUGUES, An Introduction to Language Processing with Perl and Prolog, *Springer*, 2006, 513 pages, ISBN 3-540-25031-X.

Lu par **Guy LAPALME**

Département d'informatique et de recherche opérationnelle, Université de Montréal

Ce livre réussit le tour de force de présenter dans un cadre unifié les bases de la linguistique informatique du traitement symbolique et statistique de la langue. Cet ouvrage couvre un large éventail de techniques et montre que Prolog peut servir de formalisme unificateur exécutable pour toutes les étapes du traitement depuis la manipulation des caractères jusqu'à la sémantique. Même si l'ouvrage est en anglais, l'auteur n'y limite pas ses exemples linguistiques, il en donne aussi en français (surtout), en allemand ou même en italien; c'est donc un livre très européen. Ce livre plaira tant aux linguistes-informaticiens qu'aux informaticiens-linguistes, ceux qui aiment les langues, pas seulement les maths.

Il y a longtemps que n'était pas parue une « vraie » introduction au traitement automatique de la langue. Ce livre de plus de 500 pages couvre l'ensemble des problèmes liés au TAL : traitement de corpus, morphologie, analyse syntaxique et même sémantique, et analyse du discours bien que ces deux derniers points ne soient qu'esquissés. Je placerais cet ouvrage dans la foulée et au même niveau que Manning et Shutze (1999) ou Jurafsky et Martin (2000). Le sous-titre du livre annonce « un survol des théories, de l'implémentation et de l'application avec une attention spéciale à l'anglais, le français et l'allemand ». L'auteur, un chercheur français faisant maintenant carrière en Suède, a bien relevé le défi. Il est très rafraîchissant de voir enfin des exemples qui ne sont pas seulement *américains*.

Le livre comprend 15 chapitres organisés selon une progression *classique* présentée dans le premier chapitre qui motive brièvement le domaine et décrit quelques applications; il fournit de multiples références à d'autres livres en anglais, mais aussi en français ou en allemand. L'ouvrage est complété par un site Web donnant accès aux *transparents*, aux programmes ainsi qu'à plusieurs liens bien choisis. On y retrouve même un chapitre supplémentaire de 45 pages sur le traitement de la parole. On aurait apprécié y trouver une archive contenant les sources de tous les programmes Prolog et Perl en un seul fichier; mais il est possible de les récupérer un à un en cliquant sur les liens de chaque chapitre.

Le deuxième chapitre présente le traitement des corpus avec des expressions régulières et des automates. Prolog et Perl sont utilisés à la fois pour programmer un système de concordance ou pour calculer la distance d'édition entre deux chaînes. Ces programmes Prolog et Perl sont très bien écrits, comme tous les autres, et utilisent les particularités de chaque langage à leur avantage.

Les chapitres 3 et 4 sont consacrés à des aspects trop souvent occultés, mais combien « *suants* » du traitement de la langue : le codage et le balisage des corpus ainsi que la notion de mot. Étant donné l'omniprésence de XML, il aurait été intéressant d'y donner plus de détails, mais au moins ce sujet est abordé dans un cadre de traitement de la langue. On traite d'entropie et de perplexité non seulement mathématiquement, mais aussi en faisant passer l'intuition linguistique. Tout un chapitre pour compter les mots pourrait sembler excessif, mais l'auteur en profite pour introduire la notion de n-gram et de modèle de langage afin de développer une approche à l'extraction de collocations. Comme Perl est bien adapté à ce type de calculs statistiques sur les chaînes de caractères, on retrouve là les derniers programmes dans ce langage dans tout le livre. Par la suite, tout est expliqué à un bon niveau d'abstraction avec Prolog.

Le chapitre 5 décrit (enfin !) le traitement de la morphologie et des parties du discours avec plusieurs exemples appropriés en anglais, français et en allemand. On ébauche même la structure d'un analyseur morphologique à base de transducteurs. Le chapitre 6 aborde l'étiquetage des parties du discours avec des règles et une implantation simple, mais très claire, de l'algorithme de l'étiqueteur de Brill que trop de gens utilisent les yeux fermés. L'approche plus statistique à l'étiquetage automatique, basée sur la notion de canal bruité, est présentée dans le chapitre 7. L'auteur réussit à présenter de façon assez intuitive, tout en ne sacrifiant pas la formalisation, les modèles de Markov cachés et l'algorithme de Viterbi. En consultant le chapitre supplémentaire sur le site Web, on peut même en trouver une implantation Prolog, mais dont le code source se semble pas accessible directement.

Le chapitre 8 traite de l'analyse syntaxique *classique* avec les *Definite Clause Grammar* de Prolog, tant pour la séparation d'une chaîne de caractères en mots que pour le traitement sémantique basé sur le lambda calcul. Le chapitre 9 montre ensuite comment gérer les analyses partielles qui sont trop souvent le lot du traitement avec de *vraies* phrases. Il est donc possible de ne pas se contenter d'un « *no* » comme réponse à une analyse Prolog d'une phrase. Les chapitres 10 et 11 montrent qu'il n'est nullement nécessaire de se contenter du mécanisme d'analyse descendante implicite en Prolog et qu'il est possible d'obtenir des analyseurs par unification des grammaires de dépendance et même des analyseurs ascendants et probabilistes.

Les quatre derniers chapitres sont consacrés à la sémantique et à la logique des prédicats, à la sémantique lexicale, à l'analyse du discours et au dialogue. Ils sont beaucoup plus succincts et ils se limitent à ne présenter que des généralités et à fournir quelques pointeurs. Ceci est d'autant plus décevant qu'il y avait eu jusqu'ici

un grand soin apporté à une présentation complète des sujets. Il faut avouer que traiter de tous ces sujets en 100 pages est tout un défi, surtout que ces domaines sont beaucoup plus *en développement* que les précédents.

Le livre se termine sans un chapitre de conclusion qui aurait permis de faire ressortir les grandes tendances du domaine, on y trouve cependant un appendice de 50 pages d'introduction à Prolog, mais rien sur Perl. Cet appendice n'étant de toute façon pas suffisant pour apprendre Prolog, il aurait été, me semble-t-il, préférable de le remplacer par le chapitre supplémentaire qu'on retrouve sur le site Web de l'auteur. Le livre aurait alors eu une meilleure unité et élargi sa couverture du domaine.

En résumé, j'ai beaucoup apprécié ce livre qui est très rafraîchissant et qui couvre une large gamme de sujets actuels dans le traitement de la langue naturelle. L'auteur paraît être un passionné des langues et il réussit à nous faire partager son enthousiasme. Pour compléter notre bonheur, il ne lui resterait qu'à le (faire) traduire en français !!!

Jean-Paul HATON, Christophe CERISARA, Dominique FOHR, Yves LAPRIE, Kamel SMAÏLI, Reconnaissance automatique de la parole. Du signal à son interprétation, *Dunod*, 2006, 392 pages, ISBN 2100058428.

Lu par **Chantal ENGUEHARD**

LINA - FLE CNRS 2729, université de Nantes

Cet ouvrage collectif passe en revue les sciences et techniques relatives à la reconnaissance automatique de la parole, il balaie des domaines très différents allant du traitement de signal à la représentation de grands vocabulaires. La lecture de l'ensemble de l'ouvrage est particulièrement enrichissante puisqu'elle permet à une personne intéressée par une facette de la reconnaissance automatique de la parole d'en découvrir les autres aspects qui concernent des domaines situés hors de sa propre compétence. L'importante bibliographie permet d'approfondir ces connaissances. L'ouvrage est constellé de multiples exemples (issus de la langue française), figures et illustrations qui éclairent le propos.

Les textes sont remarquablement clairs et complets. La mise en perspective de différents modèles les uns par rapport aux autres permet au lecteur d'en comprendre les spécificités et limitations.

Le livre comprend 373 pages. Il est structuré en onze chapitres que complètent la bibliographie très étendue (environ 800 références), un index des mots-clés et un index des auteurs cités. Chaque chapitre débute par un résumé qui permet à l'utilisateur de pratiquer éventuellement une lecture rapide, ciblée sur ses centres d'intérêt.

Le premier chapitre constitue une introduction. Il situe le domaine de la reconnaissance de la parole. Un bref historique précède une présentation des applications et un panorama rapide des modèles mis en œuvre.

Le second chapitre est articulé par le cheminement de la parole : émission, réception puis perception. Il décrit les processus mis en jeu dans la production de la parole par l'appareil phonatoire ainsi que les formes d'ondes et spectrogrammes qui en résultent. Le système auditif est abordé *via* une description physiologique détaillée. La compréhension est conditionnée par plusieurs étapes, quelquefois en parallèle ou avec des mécanismes de rétroaction, qui font intervenir des traitements aux niveaux auditif, phonétique, phonologique, lexical, syntaxique et sémantique.

Le troisième chapitre est consacré à la première étape de tout système de reconnaissance de la parole, l'acquisition et la paramétrisation du signal. Il s'agit donc de traitement de signal. Les grands principes sont rappelés ainsi que les équations fondamentales comme l'analyse fréquentielle par la transformée de Fourier ou les analyses fondées sur un modèle de production, comme le modèle linéaire ou l'analyse cepstrale. Ces analyses peuvent être complétées par des techniques bien connues en analyse de données et quantification vectorielle. Enfin, il existe encore des approches plus récentes par paramètres fréquentiels filtrés ou encore analyse par ondelettes. Ces différents modèles d'analyse sont comparés, les mérites et difficultés de chacun sont clairement établis.

Le quatrième chapitre détaille l'étape fondamentale de transformation de l'onde sonore en une suite d'unités phonétiques ou lexicales. Il s'agit de mettre en œuvre les principes de la reconnaissance de formes qui permettent la classification. Les approches probabilistes comme les modèles de Markov, les réseaux bayesiens ainsi que d'autres modèles stochastiques alternatifs sont largement présentés avec leurs problèmes et limites. Ce chapitre passe également en revue les modèles neuromimétiques et les machines à vecteur support.

Le chapitre 5 traite des différents aménagements qu'il est possible d'apporter aux modèles existants afin d'améliorer leurs performances, comme le regroupement des sons en diphones ou triphones. Les approches théoriques sont confrontées à des problèmes réels comme la représentation d'un grand vocabulaire, la traduction automatique ou la nécessité de tests de confiance. Des pistes de recherche sont évoquées.

Le chapitre 6 est entièrement consacré aux modèles statistiques de la langue qui sont précisément exposés. Il est illustré de résultats issus d'un système de dictée vocale.

La compréhension de la parole est traitée au chapitre 7. Elle est assimilée à une tâche de résumé ou de réponse à des requêtes. Les approches linguistiques et statistiques sont rappelées ainsi que les méthodes d'évaluation et leurs difficultés.

Le chapitre 8 se penche sur la nécessaire robustesse de tout système de reconnaissance de la parole. Il s'agit de résister aux variations du signal causées par

le bruit ou les variations inter ou intra locuteurs en effectuant une première étape de débruitage et en adaptant les modèles acoustiques existants.

Le neuvième chapitre présente des mises en œuvre réelles de système de reconnaissance automatique de la parole. Il détaille les étapes, justifie les choix effectués et présente des exemples de résultats réels.

Le chapitre 10 traite d'une approche spécifique du problème de la variabilité qui consiste à modéliser le canal articulatoire de la parole afin de pouvoir prédire les déformations des sons en fonction des variations de ce canal articulatoire. Il souligne les importants problèmes scientifiques et techniques qui restent en suspens.

Le onzième chapitre passe en revue différentes applications pratiques : saisie de données, commandes, transcription, recherche d'informations. Des systèmes existants sont présentés.

L'ouvrage souligne que des progrès importants ont été effectués, mais que les approches mises en œuvre ont maintenant atteint leurs limites. Une amélioration des performances ne peut être envisagée qu'avec la définition de nouveaux modèles.

Muriel BARBAZAN, Le temps verbal. Dimensions linguistiques et psycholinguistiques, *Presses Universitaires du Mirail,* 2006, 470 pages, ISBN 2-85816-849-0.

Lu par **Armelle JACQUET**

LISAA – EA 4120, Université Paris Est – Marne-La-Vallée

Cet ouvrage de 470 pages, comportant une bibliographie nourrie, est d'emblée une synthèse intéressante sur une question fondamentale du langage et de la cognition, au sens kantien du terme : « comment l'on acquiert des connaissances » (Dic. Littré, t. 2, p. 440). Il est axé sur la compréhension du discours, à travers la dimension du « temps verbal ».

Le chapitre 1 fixe des « repères impératifs cognitifs » : abstraction, représentation et conceptualisation, nécessaires à tout apprentissage. Ensuite, les catégories grammaticales et catégories lexicales sont envisagées, en relation avec la notion de mémorisation et disponibilité. Ce premier chapitre se conclut sur un récapitulatif des concepts de base nécessaires à la compréhension de l'ouvrage.

Le chapitre 2, consacré à « la description des formes verbales », avec ses postulats et démarches spécifiquement linguistiques, reprend diverses théories dont celles de Reichenbach (1947), avec ses points forts, ses paradoxes et sa part de subjectivité. Les définitions et limites descriptives d'un cadre onomasiologique temporel sont largement développées et débouchent sur une « perspective dégrammaticalisante » et son gain théorique et didactique (p. 94 et sq.).

Les alternatives au temps et à l'aspect sont revues, « la valeur temporelle passant du statut de signifié fondamental au statut de désigné contextuel » (p. 119). Une réflexion sur les « emplois récalcitrants » est menée, avec parfois des prises de position fermes et argumentées. Sur un plan strictement sémantique, la structure du signifié, dans ses rapports au désigné, fait référence à Le Ny (1979), avec sa définition du concept et du mode d'activation des sèmes en contexte (pp. 133-139).

Le chapitre 3 traite de la question du passé simple (PS) et du passé composé (PC), dans le cadre de l'interaction sémantique des dimensions énonciative et référentielle. La problématique est d'abord définie, avant de proposer des solutions, à propos des dimensions *temps* et *aspect*, en référence au trait [+/–accompli] de Reichenbach, puis à l'ouverture apportée par les registres histoire/discours de Benveniste (1966) et disciples.

Compte tenu d'une argumentation sur l'impasse explicative de la référence aux types de texte, l'auteur prend « un engagement franc dans le sens énonciatif » (pp. 159-164) qui débouche sur « les espaces sémantiques des dimensions énonciatives et référentielles » et leur complémentarité structurelle et matérielle (pp. 169-177). Dans une perspective strictement énonciative, le PS et le PC se distinguent alors dans les oppositions PS [–allocutif], [autonome] et le PC [allocutif], [accompli] (pp. 164 et sq.).

Le chapitre 4 évoque la problématique de l'imparfait, par rapport au PC et PS, en français, étudiés au chapitre précédent, en adoptant le point de vue de l'apprenant. La « mise en relief » (Weinrich, 1964) renvoie au paradoxe d'une validité cognitive partielle seulement si l'on se réfère à la compréhension, d'où la proposition d'une vision psycholinguistique qui permet de lever en partie, les ambiguïtés de la traduction de textes, sur ce point, sans rejeter complètement le principe de « mise en relief » lui-même (pp. 200-212). Il en ressort que le récepteur est « réellement acteur de la construction des représentations mentales », autrement dit, de sa compréhension/interprétation (p. 203).

Dans les processus de traitement des textes, une confrontation des théories aspectuelle ou logico-temporelle est proposée, faisant référence à Pollak (1976), Kamp (1981), Kamp et Rohrer (1983), associée à l'application de Confais à la didactique du français langue étrangère (FLE) ; elle apporte un complément descriptif, avec l'interaction sémantique entre aspectualité lexicale et grammaticale. Les exemples portent sur le PS [successif], l'imparfait (IMP) [coréférentiel] qui pose le problème de la surgénéralisation du fonctionnement des contextes narratifs.

Des cas de PS [– successif] et d'IMP [– coréférentiel] sont exposés alors, suscitant des propositions d'analyse et de gestion plus cognitive de la part du sujet comprenant. Les unités propositionnelles linguistiques ne correspondent pas forcément aux « micro-patrons » et « micro-unités cognitives » perçus, sur le plan de la cognition, « ce qui conduit à une distribution des tiroirs verbaux différente » (p. 249) de celle des contextes narratifs déclarés types. Cependant, d'Aristote à Adam ou Ricoeur, le principe de causalité reste présent et dépend essentiellement du

processus narratif. Kintsch et Van Djik (1978), Denhière et Baudet (1992) développent des positions assez convergentes sur ce point. Diverses contraintes macrostructurales et microstructurales sont évoquées ensuite, avec des degrés d'acceptabilité, des contraintes syntaxiques et conceptuelles, etc. Les exemples distinguent des PS [– successif], des IMP narratifs, 'de clôture' et en cascade (pp. 294-304). La reprise des analyses linguistiques débouche sur l'hypothèse d'une différence de gestion cognitive entre « séquence narrative » et « séquence descriptive », évoquant une complémentarité, déjà soulignée à d'autres niveaux. Dans la macrostructure, la notion de séquence ou période débouche sur celle d'épisode. Ces charnières textuelles permettent de mettre en lumière la distribution des « tiroirs » verbaux, dans une optique à la fois linguistique et cognitive, du point de vue du sujet (pp. 317-356).

Le chapitre 5 traite de l'imparfait (IMP) et du présent (PRES), plus spécifiquement des emplois énonciatifs et modaux de l'imparfait. Le discours est traité comme « polyphonique », dans son décodage, ce qui rappelle la théorie psychanalytique de Lacan (1966), même si cette référence est absente de l'ouvrage. Le mécanisme qui, de « la véric onditionnalité des contenus discursifs » conduit « à la validation signée de l'énonciateur », est posé à travers l'étude parallèle des énoncés subjectifs et objectifs.

Dans le récit fictionnel, d'abord défini, un statut particulier est attribué au narrateur dont la voix est décodée, avec divers degrés de matérialisation : ironie, marqueurs de subjectivité, etc. L'émergence du discours indirect libre (DIL) est évoquée comme non porteuse de marqueurs spécifiques, bien qu'aisément décelée dans la grande majorité des cas (Vuillaume, 2000). La référence à *Madame Bovary*, de Flaubert (pp. 376-388) est présentée comme un cas de « narration extradiégétique et omnisciente » (p. 381) dont les signaux précurseurs de la catégorisation des éléments du discours sont clairement distincts de ceux du narrateur objectif. Des nombreuses séquences de *Madame Bovary*, analysées ici, trois à quatre ancrages essentiels saillent : a) ouverture par un lexème marquant un changement de voix, b) éventuelle séquence de récit de « vie psychique », c) énoncés au DIL, d) retour à l'énonciation du narrateur (p. 388). L'auteur s'intéresse ensuite au trait [inactuel] de l'IMP, dans le DIL, puis à l'alternative possible avec le PRES et une corrélation avec le trait [actuel].

Le chapitre 6 est une synthèse de l'ouvrage qui débouche sur les définitions émergentes de l'IMP, du PC et du PS, avec l'adoption, pour l'IMP, des traits [inactuel] et [anaphorique] et l'articulation du PS et du PC à l'IMP.

Cette étude du temps verbal révèle les dimensions textuelles et énonciatives du discours de façon novatrice et efficace, dans l'analyse des textes proposés. La corrélation entre les plans strictement linguistique, psycholinguistique et didactique est d'autant plus intéressante que l'intégration de données de la psychologie cognitive en enrichit l'argumentation linguistique. Sur ce plan, l'ouvrage ouvre de nouvelles pistes d'analyse des textes et le choix judicieux de centrer la réflexion sur

la dynamique cognitive de la fonction prédicative, au sens logique du terme, est un apport notoire, car il permet de vivifier la dimension du sujet, « acteur » de ses propos, mais aussi de sa compréhension du langage.

Fidelia IBEKWE-SANJUAN, Anne CONDAMINES, M. Teresa CABRE CASTELLVÍ, éditeurs, Application-Driven Terminology Engineering, *John Benjamins*, 2007, 203 pages, ISBN 978 90 272 2232 9.

Lu par **Lina F. SOUALMIA**

Université Paris XIII, laboratoire LIM&Bio

Le but de cet ouvrage est d'analyser l'influence d'une application sur le choix de la définition d'un terme et de son traitement. Les applications peuvent être de deux types. Les applications intermédiaires sont relatives à la construction de ressources terminologiques et elles constituent des données pour des applications terminales impliquant l'utilisateur, comme la recherche d'information ou la construction d'index. Les différentes contributions de l'ouvrage sont indépendantes les unes des autres et concernent principalement des applications intermédiaires comme la modélisation des connaissances d'un domaine. Elles ont toutes comme point de départ un corpus de textes du domaine pour l'application visée que ce soit la construction de dictionnaire, d'ontologie ou la structuration de terminologie. Différentes méthodes et expérimentations sont décrites pour le français, avec le domaine de la médecine ou de l'ingénierie des connaissances, pour l'anglais avec des textes de nanotechnologie ou de génétique et pour l'espagnol avec le domaine du droit. Les techniques employées pour la construction de ces ressources sont relatives au TAL mais également à sa combinaison avec des approches d'apprentissage statistique.

L'ouvrage, en langue anglaise, est la version livre du numéro spécial de la revue internationale Terminology 11 : 1 parue en 2005 qui fait lui-même suite au workshop international « Terminologies, Ontologies et Représentation des Connaissances » organisé conjointement par l'ATALA et le groupe TIA à Lyon en 2004. Il est composé d'une introduction et de sept chapitres de volumes plus ou moins équilibrés. Quatre chapitres sont relatifs à la modélisation des connaissances du domaine comme la construction d'ontologie et la structuration de terminologie. Ils évoquent également les relations entre terminologie et ontologie. Un chapitre traite de la construction de dictionnaires de collocations entre termes et un autre est consacré à un état de l'art sur les variations terminologiques. Enfin, on y trouvera la description d'une application terminale aboutie permettant la construction d'index de fin de livre.

L'*introduction*, rédigée par les éditeurs, rappelle la notion d'application et d'unité terminologique qu'est le terme. Une application est définie comme étant l'utilisation finale pour laquelle les termes traités sont destinés. Un terme désigne la plus petite unité lexicale compréhensible pour un domaine et une application et

possédant une structure morphologique ou syntaxique. Les auteurs dressent un état de l'art très complet sur les approches d'acquisition de relations sémantiques à partir de corpus qui sont de deux types : descendante, impliquant l'existence d'une ressource externe, ou montante, utilisant soit des outils TAL, soit des outils statistiques. On y trouvera des définitions sur les ontologies et un résumé des différentes contributions de l'ouvrage. Les auteurs concluent sur une perspective de recherche intéressante de l'ingénierie terminologique : la prise en compte de la typologie des utilisateurs en complément de la prise en compte de l'application.

Le chapitre 1 étudie la modélisation de définitions de termes dans le but de construire des ontologies différentielles à partir de corpus. Leur méthode est fondée sur des patrons lexico-syntaxiques et concerne principalement des relations d'hyperonymie. Les expérimentations sont réalisées sur des textes en français relatifs au domaine de la nutrition et de l'enfance. Les différentes évaluations donnent de bons résultats que l'on peut justifier par la taille et l'homogénéité des corpus. Ils montrent enfin comment des co-hyponymes peuvent être déterminés sur la base des mots en commun dans leurs définitions.

Le chapitre 2 traite également d'hyperonymie avec une méthode linguistique et statistique. La méthode statistique permet de construire une hiérarchie de concepts qui est complétée grâce à une analyse linguistique. Les quatre étapes de leur algorithme y sont détaillées et appliquées sur un grand corpus de nanotechnologie. On pourra noter le grand nombre de définitions sur les ontologies et les terminologies. L'état de l'art est assez complet avec un bon positionnement par rapport à d'autres travaux existants.

Le chapitre 3 présente un travail de terminographie intitulé *Termontography* qui fait partie du projet européen FFPOIROT traitant de la prévention de la fraude sur la TVA. Les auteurs montrent comment les utilisateurs et les applications sont déterminants pour l'inclusion de l'information textuelle dans la base terminologique multilingue et pour sa structuration. L'approche proposée comprend une phase de spécification des connaissances qui assiste le processus de sélection du corpus et des critères spécifiques pour la sélection des termes.

Le chapitre 4 traite de l'acquisition et de la structuration d'ontologies. Les auteurs y rappellent les mesures standard pour la recherche d'information. Ils proposent d'utiliser les techniques de similarité distributionnelle pour organiser une terminologie médicale en anglais et s'intéressent au problème que pose l'évolution des connaissances, à savoir comment de nouvelles classes peuvent-elles être ajoutées à une ontologie de type terminologique ? L'approche décrite consiste à calculer quatre mesures de similarité distributionnelle entre termes du corpus et de faire l'hypothèse que les termes reliés par similarité appartiennent à la même classe sémantique. L'efficacité des mesures dans la prédiction d'un type sémantique d'une classe de termes est évaluée par rapport à une ontologie existante considérée comme « gold standard ». Les expérimentations donnent de très bons résultats en terme de précision. Les auteurs proposent en perspectives de tester d'autres mesures et

d'utiliser les classes sémantiques construites dans l'expansion des termes des requêtes pour augmenter le rappel en recherche d'information.

Le chapitre 5 aborde les techniques d'apprentissage pour l'extraction automatique de collocations à partir de textes. L'objectif de ces travaux est la construction d'un dictionnaire de collocations de termes d'un domaine de spécialité pour permettre la traduction automatique pour des non-professionnels. Les collocations sont des liens lexico-sémantiques entre termes et lexèmes. Les tests sur le même corpus de lois espagnoles de la méthode des *k* plus proches voisins donne de meilleurs résultats que celle du réseau bayesien.

Le chapitre 6 dresse une synthèse des différents types de variations de termes présents fréquemment dans les corpus de textes spécialisés. La prise en compte du type de variation va dépendre du domaine, du corpus et de l'application. Deux types de variations sont caractérisées : celles définies pour la construction de ressources terminologiques ou pour des applications liées à la langue, et celles destinées à des applications comme la recherche d'information, l'indexation ou la veille technologique. En conclusion, l'auteur discute des avantages et inconvénients de les prendre en compte dans plusieurs applications.

Enfin, le chapitre 7 considère l'indexation comme une application terminale de l'ingénierie terminologique à partir de textes, même si l'objectif n'est pas d'établir le vocabulaire d'un domaine mais de naviguer dans des documents de type livre. La méthode utilisée est basée sur des outils TAL. Les trois modules du système opérationnel IndDoc sont détaillés et différentes expériences sur des corpus sont évaluées avec de bons résultats. L'application est aboutie. Comme elle est de type terminale, contrairement à toutes celles décrites dans les autres contributions qui sont de type intermédiaire, la validation de l'index par un expert demeure nécessaire. Le chapitre est bien équilibré et structuré mais on déplorera de nombreuses erreurs dans la numérotation des paragraphes tout au long de sa lecture.