UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus

Didier Bourigault

Equipe de Recherche en Syntaxe et Sémantique
CNRS – Université Toulouse le Mirail
Maison de la Recherche
5, allées Antonio Machado
31058 Toulouse Cedex 1
didier.bourigault@univ-tlse2.fr

Résumé – Abstract

Nous présentons un module mettant en oeuvre une méthode d'analyse distributionnelle dite "étendue". L'analyseur syntaxique de corpus SYNTEX effectue l'analyse en dépendance de chacune des phrases du corpus, puis construit un réseau de mots et syntagmes, dans lequel chaque syntagme est relié à sa tête et à ses expansions. A partir de ce réseau, le module d'analyse distributionnelle UPERY construit pour chaque terme du réseau l'ensemble de ses contextes syntaxiques. Les termes et les contextes syntaxiques peuvent être simples ou complexes. Le module rapproche ensuite les termes, ainsi que les contextes syntaxiques, sur la base de mesures de proximité distributionnelle. L'ensemble de ces résultats est utilisé comme aide à la construction d'ontologie à partir de corpus spécialisés.

We present a software that implements a method of "extended" distributional analysis. The corpus syntactic analyser SYNTEX yields a dependency syntactic analyse of each sentence of the corpus. It builds a network of words and phrases in which each phrase is connected to its head and its expansion. The distributional analysis module UPERY relies on this network to associate to each term in the network a set of syntactic contexts. Syntactic contexts as well as terms may be simple or complex. The UPERY module calculates distributional proximities between terms as well as between contexts. The results are used for the building of ontological resources from specialized corpora.

Keywords – Mots Clés

analyse syntaxique automatique, analyse distributionnelle, corpus, ontologie, terminologie.
parsing, distributional analysis, corpus, ontology, terminology

1 Analyse distributionnelle et construction d'ontologies

L'analyse distributionnelle « à la Harris » (Harris 1968) est une technique bien connue dans le milieu du Traitement Automatique des Langues. Dans la communauté française d'Ingénierie des Connaissances, cette technique est exploitée depuis une dizaine d'année pour des applications de construction de ressources terminologiques ou d'ontologies à partir de textes (Assadi, Bourigault 1995) (Habert, Nazarenko 1996) (Faure, Nédellec 1998). Les rapprochements de mots effectués sur la base de contextes syntaxiques partagés s'avèrent être des amorces le plus souvent très utiles pour l'analyste chargé de construire un modèle de connaissances à partir d'un corpus spécialisé.

Le travail présenté dans ce papier constitue la suite des études entamées avec H. Assadi sur l'utilisation en acquisition de connaissances à partir de textes de l'outil d'extraction de termes LEXTER (Bourigault 1994) et de procédures d'analyse distributionnelle exploitant les résultats de cet analyseur (Assadi, Bourigault 1998) (Bourigault, Assadi 2000). Ces études ont montré à la fois l'intérêt de l'analyse distributionnelle pour la construction de ressources terminologiques à partir de textes, et aussi la nécessité d'utiliser, en amont, des outils d'analyse syntaxique large, qui prennent en compte en particulier les relations de dépendance syntaxique autour des verbes. Nous présentons dans cet article, une méthode et un outil d'analyse distributionnelle étendue (section 2), qui s'appuie sur le réseau de dépendance syntaxique construit par l'analyseur syntaxique de corpus SYNTEX. Par rapport aux travaux classiques, la méthode que nous proposons étend les fonctionnalités habituelles en ce qu'elle prend en compte des unités complexes, à la fois du côté des termes classés que de celui des contextes syntaxiques classificateurs. Dans la section 3, nous précisons comment se situe notre approche par rapport à l'état de l'art.

2 Analyse distributionnelle étendue

2.1 Analyse syntaxique de corpus et construction d'un réseau de dépendance syntaxique

La méthode d'analyse distributionnelle que nous présentons dans cette section s'appuie sur les résultats fournis par l'analyse syntaxique d'un corpus, sous la forme de relations de dépendance entre mots au sein des phrases du corpus. Dans les expériences décrites ici, nous avons utilisé les résultats de l'analyseur syntaxique de corpus SYNTEX (Bourigault, Fabre, 1999). A partir des résultats de l'analyse syntaxique des phrases du corpus, un module d'extraction de syntagmes (ES) construit un réseau de mots et syntagmes, calculé à partir des relations de dépendance identifiées dans chacune des phrases. Nous décrivons dans cette section comment est construit ce réseau, qui fournira les données de base à l'analyse distributionnelle étendue.

Dans un premier temps, pour chaque phrase, le module ES procède à l'identification des constituants syntaxiques maximaux (verbaux, nominaux, adjectivaux) que détermine la structuration en relation de dépendance. Pour chaque mot recteur, il construit un syntagme maximal en parcourant toutes les relations de dépendance syntaxique dont ce mot est la cible jusqu'à aboutir à des mots qui soit ne sont pas recteurs, soit sont tête d'un syntagme maximal déjà construit. La caractérisation de la structure d'un syntagme est la suivante :

- une tête, qui est constituée du mot recteur avec sa catégorie ;
- une liste de couples (relation, expansion), chaque expansion étant (le lemme d') un mot régi ou (la forme normalisée d') un syntagme dont la tête est un mot régi, et la relation étant la relation de dépendance syntaxique ;
- une forme normalisée, constituée à partir du lemme de la tête et de la séquence des lemmes ou formes normalisée des expansions[1].

Dans un second temps, le module ES construit le réseau de dépendance en ajoutant pour chaque syntagme maximal différent rencontré : (1) un nœud dont le label est la forme normalisée du syntagme, (2) des liens vers les nœuds correspondant à ses expansions, étiquetés par le nom de la relation de dépendance.

Le réseau est ensuite enrichi suite à des opérations de *réduction* et de *simplification* sur les syntagmes maximaux (des exemples illustratifs seront donnés dans le tableau 1 de la section suivante).

- L'opération de *réduction* opère sur des syntagmes qui ont au moins deux expansions. Elle consiste à générer, à partir d'un syntagme donné, autant de syntagmes réduits qu'il y a d'expansion : chaque syntagme réduit est constitué de la tête du syntagme maximal, et d'une seule expansion, un couple (relation, expansion) extrait de la liste des expansions du syntagme maximal. La réduction est totale dans le sens où l'on extrait des syntagmes réduits à une seule expansion[2].
- L'opération de *simplification* opère sur des syntagmes, maximaux ou réduits, dont au moins une expansion est un syntagme. Elle consiste à générer, à partir d'un syntagme donné, des syntagmes dans lesquels les expansions syntagmes ont été remplacées par leur tête. La simplification est totale dans le sens où l'on réduit chaque expansion syntagme à sa tête[3].

Ces opérations de réduction et de simplification visent d'une part à établir des liens directs dans le réseau entre des syntagmes correspondant à des variations syntaxiques par expansion ou par insertion, et d'autre part à multiplier, de façon contrôlée, le nombre de contextes syntaxiques qui vont être exploités par l'analyse distributionnelle. L'opération de simplification s'apparente à celle effectuée par Habert et Fabre [1999] dans l'outil d'analyse distributionnelle ZELLIG pour obtenir des contextes élémentaires.

---

[1] Notons que c'est à ce niveau que nous avons choisi d'opérer la normalisation actif/passif.

[2] Nous travaillons à définir une opération de réduction plus complète, telle que, par exemple, pour un syntagme à trois expansions, on extraie des syntagmes partiellement réduits à deux expansions.

[3] Nous travaillons à définir une opération de simplification plus complète, de telle sorte que, par exemple, l'on remplace un syntagme expansion ayant lui-même deux expansions pas uniquement par sa tête, mais aussi par ses syntagmes réduits.

2.2 Les données de l'analyse distributionnelle : des contextes syntaxiques complexes et des termes complexes

Le module d'analyse distributionnelle UPERY exploite l'ensemble des données présentes dans le réseau pour effectuer un calcul des proximités distributionnelles entre les mots et syntagmes du réseau. Ce calcul s'effectue sur la base des contextes syntaxiques partagés. Il s'agit d'une mise en œuvre du principe de l'analyse distributionnelle "à la Harris". Les données de l'analyse sont constituées ainsi :

(1) pour chaque syntagme du réseau ayant *une seule expansion*, le module construit une information élémentaire pour le calcul distributionnel. Celle-ci se formalise sous la forme d'un couple (contexte, terme) :

- le *contexte* est le couple constitué de la tête et de la relation de dépendance ; Il s'agit d'un contexte *simple*.
- le *terme* est l'expansion.

(2) pour chaque syntagme du réseau ayant *plus d'une expansion* (N expansions, N supérieur ou égal à 2), le module construit N(N-1) informations élémentaires pour le calcul distributionnel. Pour chaque expansion E, il construit N-1 couples (contexte, terme), un pour chacune des autres expansions E' :

- le *contexte* est le couple constitué du syntagme réduit construit avec la tête et l'expansion E, et de la relation de dépendance R' correspondant à l'expansion E'; il s'agit d'un contexte *complexe*.
- le *terme* est l'expansion E'.

Nous sommes en mesure maintenant de dérouler un exemple complet, pour illustrer en quoi on peut parler d'analyse distributionnelle « étendue ». Les données extraites pour l'analyse distributionnelle à partir de l'analyse syntaxique de la phrase « Les roches cristallines résistent à l'érosion », sont présentées dans le tableau 1. Le syntagme *roche cristalline* apparaît dans le contexte simple « sujet de *résister* » ainsi que dans le contexte complexe « sujet de *résister à érosion*». De même, le mot *érosion* apparaît dans le contexte syntaxique simple « complément de *résister à* » et dans les contextes complexes « complément de *roche résister à* » et « complément de *roche cristalline résister à* ».

Il est donc possible d'avoir des unités complexes à la fois du côté des contextes syntaxiques classificateurs et de celui des termes classés. A titre d'illustration, nous donnons quelques exemples et des résultats numériques obtenus sur les 4 corpus suivants : le code civil français (CCIV, 145 000 mots), un recueil d'article scientifiques dans le domaine de l'ingénierie des connaissances (IC, 200 000 mots), un ouvrage de géomorphologie (GEOM, 210 000 mots), un corpus de compte-rendus d'hospitalisation dans le domaine de la réanimation chirurgicale (REA 178 000 mots). Les résultats sont obtenus avec des valeurs de seuils de proximité donnés dans la section suivante. On constate sur le tableau 2 en particulier que sur les diffférents corpus entre 60 et 100 syntagmes nominaux ont été rapprochés d'autres noms ou syntagmes nominaux par l'analyse distributionnelle. Le tableau 3 illustre le fait que les types de couples de catégories rapprochés par l'analyse distributionnelle se répartissent de façon sensiblement différente d'un corpus à l'autre.

Terme	Contexte
cristalline	roche_ADJ
roche	résister_SUJ
roche	résister_à_érosion_SUJ
roche cristalline	résister_SUJ
roche cristalline	résister_à_érosion_SUJ
érosion	résister_à
érosion	résister_SUJ_roche_à
érosion	résister_SUJ_roche cristalline_à

Tableau 1 : Données extraites pour l'analyse distributionnelle étendue à partir de l'analyse syntaxique de la phrase « Les roches cristallines résistent à l'érosion ».

	CCIV			IC			GEOM			REA		
Adj	82	955	9 %	176	1482	12 %	271	2140	13 %	264	2004	13%
Adv	16	470	3 %	29	622	5 %	31	701	4 %	10	181	6 %
Nom	267	2220	12 %	356	2923	12 %	287	4264	7 %	301	5737	5 %
SN	60	10343	0,5 %	97	22764	0,5 %	44	24198	0,2 %	100	21236	0,5 %

Tableau 2 : Nombre de mots ou syntagmes rapprochés par l'analyse distributionnelle, par catégorie. Pour chaque corpus, dans la première colonne figure le nombre de mots de la catégorie qui ont au moins un voisin, dans la deuxième le nombre total de mots de la catégorie, et dans la troisième le pourcentage correspondant

		CCIV		IC		GEOM		REA	
Nom	Nom	690	38,94%	1346	56,39%	524	36,16%	690	38,94%
Nom	SNom	125	7,05%	222	9,30%	46	3,17%	125	7,05%
SNom	SNom	72	4,06%	16	0;67	4	0,28%	72	4,06%

Tableau 3 : Types de couples de catégories (nominales) rapprochés par l'analyse distributionnelle

2.3 Trois mesures de proximité

L'analyse distributionnelle rapproche d'abord deux à deux des termes qui partagent les mêmes contextes. L'analyse distributionnelle est symétrique, en ce sens qu'elle peut rapprocher aussi les contextes, en fonction des termes qu'ils partagent. Nous travaillons actuellement sur trois mesures qui permettent d'appréhender la proximité entre deux unités (termes ou contextes). Ces mesures ont l'avantage d'être simples à appréhender par l'utilisateur final, et de recouvrir des aspects différents et complémentaires des conditions dans lesquelles deux unités peuvent être jugées plus ou moins proches. Notre application cible est l'aide à la construction de ressources terminologiques ou ontologiques à partir de textes. Notre objectif est de fournir à l'utilisateur, avec ces quelques mesures de proximité, différents outils pour l'aider à trouver au plus vite les relations qu'il jugera les plus intéressantes.

On dispose pour un contexte donné de l'ensemble des termes (mots ou syntagmes) qui apparaissent dans ce contexte, et pour un terme donné l'ensemble des contextes (simples ou complexes) dans lesquels il apparaît. On définit la productivité d'un contexte et la productivité d'un terme ainsi :

- la *productivité d'un contexte* est égale au nombre de termes qui apparaissent dans ce contexte ;
- la *productivité d'un terme* est égale au nombre de contextes dans lesquels ce terme apparaît.

Les trois mesures de la proximité sont les suivantes :

Le coefficient *a*. Soient deux termes t_1 et t_2. Le coefficient *a* est égal au nombre de contextes syntaxiques partagés par les deux termes. Cette mesure donne une première indication de la proximité entre deux termes, facile à interpréter. Mais l'expérience montre que cette mesure reflète de façon insatisfaisante la proximité : il faut tenir compte, d'un côté, de la productivité des contextes partagés (coefficient *prox*), d'un autre côté, du nombre de contextes que chaque terme a en propre (coefficients j_1 et j_2).

Le coefficient *prox*. Avec ce coefficient, nous visons à formaliser le fait si un contexte partagé par deux termes est très productif, sa contribution au rapprochement des deux termes est a priori plus faible que celle d'un contexte peu productif. Le coefficient prox est calculé ainsi :

$$prox = \sum_{c \in C} 1/ prod(c)^{1/2}$$

où C est l'ensemble des contextes partagés par t_1 et t_2, et prod(c) la productivité du contexte c

Les coefficients j_1 et j_2. Pour évaluer la proximité entre deux unités, il est important de tenir compte non seulement de ce qu'elles partagent, mais aussi de ce qu'elles ont en propre. Un certain nombre de mesures statistiques implémentent cette idée, sous des formes diverses (e.g. information mutuelle, Jaccard, Anderberg). Ces mesures présentent presque toujours la particularité de "symétriser" la relation de proximité. Cette propriété, qui dans beaucoup de contextes d'application, constitue un avantage, voire une nécessité, nous est apparue finalement comme masquant un phénomène marquant à l'œuvre dans les corpus : la dissymétrie de la relation de proximité. Quand deux termes partagent un certain nombre de contextes en commun, il arrive le plus souvent que l'un des deux termes possède un nombre

élevé de contextes, tandis que l'autre en possède beaucoup moins et en partage l'essentiel avec le premier. C'est pourquoi, nous caractérisons la proximité entre deux termes à l'aide de deux indices, simples et eux aussi faciles à interpréter : rapport entre le nombre de contextes partagés et le nombre total de contextes

$j_1 = a / prod(t_1)$

$j_2 = a / prod(t_2)$

Le module d'analyse distributionnelle UPERY calcule pour chaque couple de termes l'ensemble de ces coefficients. Dans l'interface, on ne présente à l'utilisateur que les couples dont les coefficients dépassent certains seuils. Ces seuils sont définis de façon empirique et varient en fonction d'une part de l'homogénéité et de la redondance du corpus et d'autre part du contexte dans lequel doivent être exploités les résultats de l'analyse distributionnelle. Pour les exemples présentés dans ce papier, les différents corpus ont été traités avec les seuils suivants :

- le nombre de contextes partagés *a* doit être supérieur ou égal à 3
- le coefficient *prox* doit être supérieur à 0.75
- l'un des deux coefficients j_1 ou j_2 doit être supérieur à 0.25 (l'un des deux termes doit partager au moins le quart de ses contextes avec l'autre)

2.4 Le concept de double clique

Le module d'analyse distributionnelle UPERY calcule des proximités entre couples de termes (et de contextes). Nous n'avons pas encore implémenté de calcul automatique de regroupement de termes, sous forme d'arbre de classification hiérarchique ascendante [Assadi 1998] ou de cliques et composantes connexes [Nazarenko & al. 2000]. A l'instar de [Faure 2000], nous laissons à l'utilisateur le soin de repérer et de valider des regroupements qu'il juge pertinents, grâce à une interface spécialement conçue pour cette tâche. Celle-ci guide l'utilisateur vers les regroupements a priori pertinents en lui permettant d'accéder rapidement à des structures que nous nommons *doubles cliques* : un double clique est constituée d'un ensemble de termes et d'un ensemble de contextes, tels que chacun des termes apparaît dans chacun des contextes. Il s'agit d'une clique de termes, car chaque terme est relié à chacun des autres termes par une relation de proximité établie sur le même ensemble de contextes partagés. Il s'agit aussi d'une clique de contextes, car chaque contexte est relié à chacun des autres contextes par une relation de proximité établie sur le même ensemble de termes partagés. Ces doubles cliques constituent des rapprochements directement interprétables, qui sont d'une grande utilité à l'utilisateur pour la constitution de classes sémantiques ou conceptuelles. Des exemples de doubles cliques sont donnés dans le tableau 4. La notion de double clique est à rapprocher de celles de *classe d'opérateurs* et de *classe d'arguments* de Z. Harris.

Termes	Contextes
Code civil	
époux, donateur, débiteur, créancier	profit_de, faveur_de, héritier_de, s'obliger_SUJ
tribunal, juridiction, cour, conseil de prud'homme	disposition_à, fonctionnement_de, compétence_de, procédure_devant
immeuble, bien, récolte, chose	partie_de, fruit_de, prix_de, quotité_de, portion_de
Réanimation chirurgicale	
réanimation chirurgicale, neurochirurgie, chirurgie cardiaque	transferer_OBJ_patient_en, service_de, transferer_en
détresse respiratoire, syndrome, insuffisance	présenter_OBJ, présenter_SUJ_patient_OBJ, apparition_de, tableau_de, développer_SUJ_patient_OBJ, développer_OBJ

Tableau 4 : Quelques exemples de doubles cliques construites "à la main" à partir des résultats de l'analyse distributionnelle fournis UPERY sur différents corpus. Chacun des termes (colonne de gauche) de la clique de termes apparaît dans chacun des contextes (colonne de droite) de la clique des contextes.

3 Travaux liés

Nous parlerons ici essentiellement des travaux de Gregory Greffenstette (Greffenstette 1994). Là où G. Greffenstette se contente volontairement d'une analyse syntaxique relativement rudimentaire, réalisée par l'analyseur SEXTANT, nous avons fait le choix d'une analyse, certes encore partielle, mais plus large et plus précise, réalisée par SYNTEX. De ce fait, les procédures statistiques d'analyse distributionnelle de Greffenstette ne concernent que des mots simples, alors que nous pouvons prendre en compte des entités complexes (contextes ou termes). Les mesures statistiques utilisées par Greffenstette sont beaucoup sophistiquées que les nôtres. Outre leur degré de complexité, et le fait que nous tenons à une mesure de proximité dissymétrique, la différence tient aussi à ce que nous avons fait le choix de ne pas tenir compte du tout des fréquences. Greffenstette introduit la fréquence des mots dans la mesure de pondération, alors que les mesures de proximité sur lesquelles nous travaillons négligent la fréquence (au profit de la productivité). Ce choix s'appuie sur le constat maintes fois confirmé que, dans le contexte de la construction de ressources terminologiques ou ontologiques à partir de corpus, l'utilisateur peut juger pertinents des phénomènes rares dans le corpus.

Par ailleurs, contrairement à Greffenstette, nous maintenons une distinction entre recteur et régi dans l'analyse distributionnelle. Par exemple, dans notre approche, les noms peuvent être rapprochés d'un côté par les contextes syntaxiques dans lesquels ils apparaissent (en tant que régis), et d'un autre côté, de façon indépendante, par les modifieurs qu'ils régissent (en tant que recteurs). Par exemple, dans le corpus sur la réanimation chirurgicale, *échographie* et *scanner cérébral* sont rapprochés à la fois en tant que régis car apparaissant tous les deux dans les contextes *résultat_de*, *réalisation_de*, *noter_SUJ*, *réaliser_OBJ*, *montrer_SUJ*, et en tant que recteur modifiés par les participe passés *réalisé*, *effectué*, *pratiqué*.

4 Conclusion

Les pistes de recherche sont nombreuses. Outre l'amélioration de l'analyse syntaxique, nous travaillons sur l'extension des opérations de réduction et de simplification, et sur l'affinement des mesures de proximité, avec le souci de vérifier la pertinence de pondérer ces mesures en fonction de la nature simple ou complexe des contextes et des termes. En ce qui concerne l'évaluation et la validation, nous privilégions un mode de validation par l'usage. Toute ressource terminologique ou ontologique est construite pour un usage spécifié, et c'est donc au sein de cet usage qu'elle peut être évaluée. Plutôt que de comparer les rapprochements effectués par le module d'analyse distributionnelle UPERY à des ressources déjà constituées, nous nous efforçons de multiplier les expériences dans lesquelles les résultats du logiciel UPERY sont exploités dans des tâches de construction de ressources terminologiques ou ontologiques. C'est sur la base des retours d'expérience que nous évaluons comment améliorer les différents modules de la chaîne de traitement. L'une des dernières expériences en date concerne la construction d'une ontologie dans le domaine de la réanimation chirurgicale. A partir des résultats d' UPERY sur un corpus de 600 comptes-rendus d'hospitalisation (corpus REA), un médecin a construit une ontologie d'environ 2000 concepts et 200 relations en un peu moins de 80 heures (Le Moigno & al 2002). Dans autre expérience (Bourigault, Lame 2002), une ontologie documentaire du Droit français codifié, construite à partir, entre autre, des résultats fournis par le module UPERY, est évaluée au sein de son environnement d'usage, à savoir en tant qu'outil d'aide à la reformulation et à l'expansion de requête sur le site d'accès à une base documentaire de textes juridiques (www.droit.org).

Références

Assadi H., Bourigault D. (1995). Classification d'adjectifs extraits d'un corpus pour l'aide à la modélisation des connaissances. *Actes des 3èmes Journées internationales d'Analyse statistique de Données Textuelles* (JADT 95), Rome

Assadi H. (1998) *Construction d'ontologies à partir de textes techniques. Application aux systèmes documentaires*. Thèse Université Paris VI, Paris, 1998

Bourigault D., Assadi H. (2000). Analyse syntaxique et analyse statistique pour la construction d'ontologie à partir de textes, in Charlet J, Zacklad M., Kassel G., Bourigault D. éds. *Ingénierie des connaissances. Tendances actuelles et nouveaux défis*. Editions Eyrolles/France Telecom, Paris

Bourigault D, Fabre C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, 25, 131-151, Université Toulouse le Mirail

Bourigault D., Lame G. (2002). Analyse distributionnelle et structuration de terminologie. Application à la construction d'une ontologie documentaire du Droit. *Revue Traitement automatique des langues*, n° 47:1, Hermès, Paris

Faure D. (2000) *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantique à partir de textes : le système ASIUM*, thèse de l'Université Paris XI Orsay

Faure D., Nédellec C. (1998). Apprentissage de cadres de sous-catégorisation et de restrictions de sélection à partir de textes. *Actes de la 5ème conférence annuelle sur le Traitement Automatique des Langues Naturelles* (TALN 98), 233-235

Grefenstette G. (1994). *Exploration in Automatic Thesaurus Discovery*, Londres, Kluwer Academic Publishers.

Habert B., Fabre C. (1999). Elementary dependancy trees for identifying corpus-specific semantic classes. *Computer and the Humanities* 33 :3, 207-219

Habert B., Nazarenko A. (1996). La syntaxe comme marche-pied cd l'acquisition des connaissances : bilan critique d'une expérience. *Actes des Journées d'acquisition des connaissances* (JAC 96), 137-149

Harris Z. (1968) *Mathematical Structures of Language*, New-York, John Wiley & Sons.

Le Moigno S., Charlet J., Bourigault D., Jaulent M.-C. (2002). Construction d'une ontologie à partir de corpus : expérimentation et validation dans le domaine de la réanimation chirugicale. *Actes des 13èmes journées francophones d'ingénierie des connaissances* (IC 2002), Rouen