TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Enrichir la notion de patron par la prise en compte de la structure textuelle - Application à la construction d'ontologie

Marion laignelet[1] Mouna Kamel[2] Nathalie Aussenac-Gilles[2]
(1) CLLE-ERSS, Université de Toulouse 2, 5 allée A. Machado, 31058 Toulouse Cedex 9
(2) IRIT, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse Cedex 9
marion.laignelet@univ-tlse2.fr, kamel@irit.fr, aussenac@irit.fr

Résumé. La projection de patrons lexico-syntaxiques sur corpus est une des manières privilégiées pour identifier des relations sémantiques précises entre éléments lexicaux. Dans cet article, nous proposons d'étendre la notion de patron en prenant en compte la sémantique que véhiculent les éléments de structure d'un document (définitions, titres, énumérations) dans l'identification de relations. Nous avons testé cette hypothèse dans le cadre de la construction d'ontologies à partir de textes fortement structurés du domaine de la cartographie.

Abstract. Matching lexico-syntactic patterns on text corpora is one of the favorite ways to identify precise semantic relations between lexical items. In this paper, we propose to rely on text structure to extend the notion of pattern and to take into account the semantics that the structure (definitions, titles, item lists) may bear when identifying semantic relations between concepts. We have checked this hypothesis by building an ontology via highly structured texts describing spatial, i.e. geographical information.

Mots-clés : Construction d'ontologie, patron lexico-syntaxique, structure textuelle.

Keywords: Ontology engineering, lexico-syntactic patterns, textual structure.

1 Introduction

La projection de patrons lexico-syntaxiques sur corpus, utilisée pour l'extraction d'informations, s'applique également sur des corpus spécialisés pour la construction d'ontologies de domaines : on s'attend à trouver des traces linguistiques de concepts et de relations sémantiques binaires entre ces concepts. Des travaux en linguistique ont mis en évidence le rôle de la structure dans l'interprétation d'un texte (Luc & Virbel, 2001; Pascual & Péry-Woodley, 1997; Rebeyrolles *et al.*, 2009). Nous supposons que des informations issues de la structure textuelle peuvent être intégrées aux patrons de recherche de relations sémantiques, en plus des éléments lexicaux et syntaxiques. Nous avons testé cette hypothèse dans le cadre de la construction d'ontologies à partir de textes, en nous focalisant sur des éléments textuels particulièrement favorables à la recherche de relations ontologiques : les titres, les zones définitoires et les énumérations. Ces éléments permettent non seulement d'extraire des traces linguistiques pour définir des concepts et des relations pertinents pour notre domaine d'application mais également de résoudre certaines situations elliptiques, fréquentes dans nos données.

Le domaine d'étude de nos travaux s'inscrit dans le traitement automatique de textes structurés dans lesquels l'organisation même des éléments textuels reflète celle des concepts du domaine. Dans ces types de textes, la structure peut être explicite comme c'est le cas avec les versions électroniques de thésaurus ou de dictionnaires (Hearst, 1992) ou inférable à partir de leur mise en forme. En ce qui nous concerne, c'est à partir de la mise en forme matérielle des documents que les différentes classes (ou concepts) du domaine sont nommées et mises en relation les unes avec les autres. On trouve de tels types de textes dans des domaines spécifiques, comme la botanique qui se prête naturellement à la représentation de taxinomies : dans les travaux de (Role & Rousse, 2006), la structure des titres reflète un découpage en genres et en espèces et permet d'initialiser une hiérarchie de classes de l'ontologie.

Les textes de notre corpus appartiennent au domaine géographique [1] : ils décrivent les objets susceptibles d'intégrer des bases de données géographiques et cartographiques. Dans un premier temps, nous nous intéressons à

1. Projet Géonto, http ://geonto.lri.fr/

deux bases, BDTopo et BD Carto. Les documents associés constituent ainsi une ressource privilégiée, particulièrement bien adaptée à l'approche que nous avons définie. Ces guides de saisie de bases de données de l'IGN [2], appelés documents de spécification, indiquent les types d'objets prédéfinis pour décrire les objets du monde réel à saisir dans la base. Cet ensemble constitue un corpus de 23 884 mots (17 069 pour BDCarto et 6 815 pour BDTopo). Initialement disponibles au format doc, ils utilisent un style et une mise en forme spécifique à chaque type d'information qui ont été manuellement traduits au format xml.

Après avoir présenté les approches par patrons pour la construction d'ontologie et montré l'intérêt de la prise en compte de la structure des documents, nous présentons le principe de fonctionnement de nos patrons ainsi qu'une évaluation qualitative des traitements automatiques mis en oeuvre.

2 Approches par patrons : prendre en compte les éléments structurants

Différentes approches, statistiques ou linguistiques, ont été proposées pour extraire des relations à partir de textes dans la perspective de construire une ressource terminologique ou ontologique (Maedche, 2002). Ces méthodes et les logiciels associés assistent plus ou moins l'identification et la représentation des relations. Elles peuvent soit localiser des contextes en corpus (Knowledge Rich Context au sens de (Meyer, 2001), soit aider à identifier les termes en relation et la nature des relations, comme le font Prométhée (Morin, 1999) ou Caméléon (Aussenac-Gilles & Jacques, 2008), soit encore, définir des concepts et des relations sémantiques formalisées dans des ontologies existantes (Scarlett, (Sabou *et al.*, 2008)).

Nous nous focalisons sur la question de l'outillage du repérage en corpus d'indices linguistiques pouvant donner lieu à la création de concepts et de leurs relations dans une ontologie en cours de construction. Notre proposition est de s'intéresser, en plus des aspects lexicaux et syntaxiques, aux éléments structurants d'un texte traduits en xml à partir de la mise en forme des documents. Cela permet la prise en compte de situations textuelles diversifiées ce qui étend la portée des patrons. Nous proposons un travail sur les titres, les définitions et les énumérations car ces objets bénéficient de régularités syntaxiques et de mise en forme facilitant leur traitement automatique.

Bien que les relations trouvées par les patrons soient des relations lexicales, nous les représentons au niveau conceptuel afin de réaliser une évaluation sur les résultats produits. Nous sommes conscients de gommer ainsi une phase délicate de l'interprétation, qui consiste à décider si un terme doit être associé à un concept existant ou donner lieu à la naissance d'un concept ou d'une instance de concept. *A priori*, nous générons un nouveau concept systématiquement.

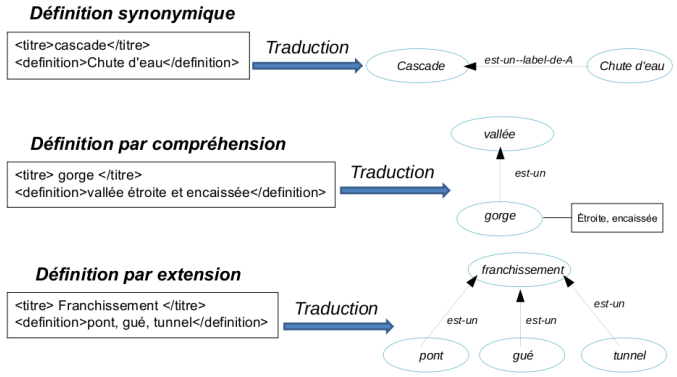

FIGURE 1 – Types de définitions propices à la création de fragments d'ontologie

Une définition pose *« une équivalence de sens entre un terme et une périphrase structurée selon la logique des classes »* (EscoubasBenveniste, 2010). Beaucoup étudiées en linguistique, les définitions sont des objets textuels au sein desquels des relations sémantiques entre les mots sont relativement explicites (Cartier, 1998; Condamines & Rebeyrolles, 2000; Pascual & Péry-Woodley, 1997) et stables : les principales sont la synonymie, l'hypéronymie et la méronymie ; les relations spécifiques aux domaines d'application sont la causalité ou le but (Montiel-Ponsada *et al.*, 2009; Malaisé *et al.*, 2005). De plus, l'ordre canonique des termes dans les définitions explicitent

2. Institut Géographique National

des relations précises (inclusion d'un terme, terme superordonné ou attribut(s) précisant un terme) et facilitent leur traitement automatique. En effet, les énoncés définitoires présentent des propriétés textuelles suffisamment stables pour qu'un repérage automatique puisse être envisagé dans différents discours (Hearst, 1992; Rebeyrolles & Tanguy, 2000). La figure 1 exemplifie les fragments d'ontologie susceptibles d'être extraits à partir de types de définition simples. Contrairement aux travaux de (Malaisé *et al.*, 2005), nous ne travaillons pas sur des définitions intégrées au flux textuel mais sur des champs xml prétaggés, ce qui facilite naturellement leur traitement.

L'énoncé définitoire est souvent elliptique et anaphorique (EscoubasBenveniste, 2010) : le recours à son contexte discursif et à son environnement textuel permet de l'interpréter correctement. Pour gérer les situations elliptiques, nous exploitons les informations présentes dans les titres. Les situations anaphoriques sont plus délicates à traiter et leur résolution, même partielle, n'est pas encore implémentée.

Les titres explicitent la structure logique du document. Pour (Ho-Dac *et al.*, 2004), l'emboîtement ou le parallélisme de titres de sous-sections d'une section donnée reflète des relations de subordination ou de juxtaposition existant entre ces sections : lorsque les sous-titres présentent la même structure discursive ou syntaxique, il est possible d'instaurer une relation sémantique entre le titre et chacun des sous-titres, ou entre les sous-titres (Kamel & Aussenac-Gilles, 2009). La figure 2 illustre le type de fragment d'ontologie pouvant être généré à partir d'un ensemble de titres. Les travaux de (Role & Rousse, 2006) vont également dans ce sens.

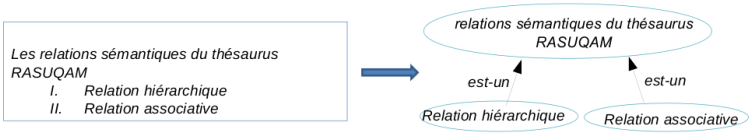

FIGURE 2 – Exemple de titres propices à la création de fragments d'ontologie

Une énumération consiste à énoncer un à un les éléments d'un même champ conceptuel, ces éléments entretenant un lien hiérarchique direct ou indirect avec un élément classifieur (Luc & Virbel, 2001). Sur le plan textuel, cet acte se traduit par une structure hiérarchique dite structure énumérative, composée d'une amorce, d'une liste d'items (introduite par l'amorce) et éventuellement d'une conclusion. (Luc & Virbel, 2001) définit comme paradigmatique une structure énumérative » au sein de laquelle les items n'ont aucun lien de dépendance syntaxique et sont considérés comme fonctionnellement équivalents. La figure 3 illustre une énumération et la structure hiérarchique des concepts ontologiques correspondants.

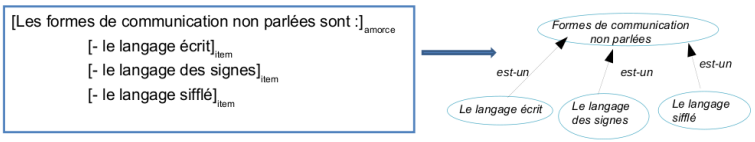

FIGURE 3 – Interprétation d'une structure énumérative

3 Proposition de patrons de textes

Nous appellons "patron de texte" un patron lexico-syntaxique qui intègre à la fois des éléments lexico-syntaxiques et des éléments de structure. Les patrons que nous présentons ont pour but de créer des fragments d'ontologies à même de venir compléter et enrichir une ontologie existante. Nous avons développé trois modules différents : le premier exploite uniquement des informations de type lexical et syntaxique, le second combine structure, lexique et syntaxe, enfin le troisième exploite uniquement la structure et se rapproche en cela des wrappers (Kushmerick *et al.*, 1997).

Approche par patron lexico-syntaxique « classique »

Un premier module exploite les zones en langage naturel présentes dans les définitions. Des patrons lexico-syntaxiques « classiques » permettent de rechercher des traces linguistiques de relations sémantiques et de concepts potentiels. Ici, la relation identifiée dépend du marqueur et du contexte d'interprétation. Nous traitons les relations

d'hypéronymie, de méronymie, d'artefact et de fonction. Ce module a été développé à l'aide de la plateforme LinguaStream, sous forme de grammaires locales en Prolog. Les patrons ont été enrichis sur la base des travaux de (Aussenac-Gilles & Jacques, 2008) et des problématiques du projet Géonto (par les relations de fonction (*sert à*) et d'artefact (*est représenté par*)). Au total, ce module repère 18 situations de méronymie, 48 relations d'hyperonymie, 14 d'artefact et 82 cas de fonction.

Combiner structure, lexique et syntaxe : analyse des titres et des définitions

Ces traitements exploitent à la fois les titres, leur forme syntaxique et celle des définitions sous la dépendance du titre. Nous nous basons sur l'idée qu'il y a une relation sémantique différente selon la structure syntaxique de la définition. Nous considérons deux situations :

– un terme-concept est défini par un syntagme nominal dont la structure syntaxique est simple, *i.e.* qui ne comprend qu'un seul nom. Dans ce cas, c'est une relation sémantique de type synonymie qui s'instancie entre les deux termes.

```
(patron syn) <synonymie> {type:titreDebut} {type:snSimple} {type:titreDebut}
        {type:definitionDebut} {type:phraseDebut} {type:snSimple} {type:phraseFin} {type:definitionFin} </synonymie>
```

  Par exemple dans une situation où le titre est <*titre*>*Cascade* </*titre*> et la définition associée est <*definition*>*Chute d'eau.* </*definition*>, une relation de synonymie relie "cascade" et "chute d'eau".

– un terme-concept est défini par un nom/terme auquel sont adjoints des adjectifs/propriétés. Dans ce second cas, une relation hyperonymique s'établit entre les deux termes.

```
(patron hyper) <hyperonymie> {type:titreDebut} {type:snSimple} {type:titreDebut}
        {type:definitionDebut} {type:phraseDebut} {type:snSimple} ( ({type:sAdj}|{type:sAdv})+ | {type:pronRel} ) </hyperonymie>
```

  Par exemple, si le titre est <*titre*> *piste d'aérodrome*</*titre*> et la définition <*definition*>*aire située sur un aérodrome* </*definition*>, le fragment d'ontologie proposé sera : #piste_d'aérodrome *est-un* #aire.

Le titre sert également à gérer les situations elliptiques dans lesquelles un des éléments de la relation (souvent le sujet) est absent ou incomplet dans la zone de définition. Par exemple, si la définition est <*definition*>*Permet de distinguer les cours d'eau naturels des cours d'eau artificiels* </*definition*> alors un syntagme capable de jouer le rôle de sujet pour le verbe en question est recherché dans le titre. En l'occurrence <*titre*> *artificialisé*</*titre*> est le sujet de *permet de*. Au total, ce module repère et analyse 359 situations où les titres et les définitions requièrent un traitement spécifique.

Traiter la structure : analyse des énumérations

Bien que possédant des propriétés lexico-syntaxiques, typographiques et dispositionnelles différentes des structures énumératives classiques, les énumérations paradigmatiques de notre corpus en conservent les caractéristiques sémantiques. L'amorce est marquée par certaines des balises de titre (<*className*>, <*attributeName*> ou <*valueName*>), la liste d'items par les balises <*description type="extensionalDefinition"*> ou <*enumeratedValues*>. Le lien hiérarchique entre l'amorce et les items est traduit par les niveaux d'imbrication des balises. Par exemple, le patron ci-dessous repère les objets balisés par <*className*> comme amorces de l'énumération, les objets balisés par <*description type="extensionalDefinition"*> comme les items de la liste étant séparés par des caractères délimiteurs. Au total, 228 énumérations sont exploitées.

```
(P) {type:class}
        {type:className} <Amorce> {type:sentence} </Amorce> {type:className}
        {type:geometryType}   {type:sentence}{type:geometryType}
        {type:description type="definition"}{type:sentence}{type:description}
        {type:description type:"extensionalDefinition"} (<Item> {type:string} </Item> ({" | "} | {" / "}) )* <Item> {type:string} </Item> {type:decription}
        ({type:description type:"selectionPrinciple"} {type:sentence}{type:description})+
        ({type:description type:"geometryDescription"} {type:sentence}{type:description})+
        {type:attributes} {type:paragraph}{type:attributes}
    {type:class}
```

4 Résultats et évaluation qualitative

L'évaluation des ontologies en termes de performance est un problème reconnu et délicat (Schutz & Buitelaar, 2005). Dans l'idéal, disposer d'une ontologie faisant référence permettrait de fournir une évaluation quantitative intéressante. Mais construire ce type d'ontologie nécessite un investissement humain coûteux que nous n'avons pas pu mettre en place.

Nous avons procédé à une évaluation manuelle qualitative des concepts et des relations sémantiques extraits à l'issue de la deuxième étape. Pour chaque fragment d'ontologie extrait [3], la pertinence de la relation et des concepts X et Y renvoyés par nos modules a été évaluée. Cinq valeurs sont proposées :
– la valeur est **valide** si la relation est jugée valide ainsi que les deux concepts X et Y reliés par cette relation,
– la valeur est **inverse** si la relation est valide mais est inversée (relations anti-symétriques comme l'hyperonymie et la méronymie),
– la valeur est **approximative** lorsque la relation est valide mais les concepts sont approximatifs (concepts mal extrait, présence d'un pronom ou d'une situation de co-référence non resolue),
– la valeur est **incertaine** lorsqu'une expertise particulière est requise (nécessité de recourir à des experts en cartographie),
– la valeur est **fausse** quand la relation est erronée (dans ce cas, les concepts sont également souvent non valides).
Les résultats de cette évaluation [4] sont indiqués dans le tableau suivant. Ils mettent en avant de grandes disparités qualitatives selon les types de relations sémantiques.

Relation	Artefact	Fonction	Hyperonymie	Synonymie	Méronymie	Holonymie	Total
Valide	16,7 %	14,3 %	38,8 %	25 %	19 %	58,3 %	33,1 %
Inverse	0 %	0 %	1 %	0 %	43 %	0 %	2,7 %
Approximative	16,7 %	23,8 %	12 %	25 %	19 %	8,3 %	15,4 %
Incertaine	66,6 %	55,5 %	19,6 %	0 %	0 %	16,7 %	22,2 %
Fausse	0 %	6,4 %	28,5 %	50 %	19 %	16,7 %	26,5 %

FIGURE 4 – Evaluation qualitative des fragments d'ontologie issus du traitement des titres et des définition

Si d'un point de vue linguistique, elles sont tout à fait acceptables, les relations d'artefact et de fonction présentent le plus fort pourcentage d'incertitude sur de la connaissance extraite, pour la tâche visée. Cette incertitude est liée à leur spécificité pour le domaine et au recours nécessaire à des experts géographes pour leur validation effective. La relation d'hyperonymie présente des résultats acceptables où presque 70 % des fragments extraits sont corrects ou approximatifs. (Maynard *et al.*, 2009) constatent également, sur leurs données, une situation de surgénération des patrons d'hyperonymie. Le traitement actuel de la synonymie n'est pas convainquant : exploiter la structure et la syntaxe n'est pas suffisant pour distinguer l'hyperonymie de la synonymie. Sur ce point, (Malaisé *et al.*, 2005) constatent le même phénomène qui semble donc aller au delà du genre : *"The pattern "N(N) introduced "hypernymic relation", as well as "synonymic" or "meronymic" ones [...]"*. En revanche, les résultats concernant la méronymie et l'holonymie sont encourageants (moins de 20 % d'erreur).

5 Bilan et Perspectives

Partir des textes pour construire une ontologie n'est pas une tâche triviale : la non linéarité des segments textuels complexifie l'extraction de connaissances et leur mise en relation. Nous avons présenté dans cet article une approche prenant en compte la combinaison d'informations lexicales, syntaxiques et structurelles dans des patrons. La prise en compte de ces éléments discursifs s'avère productive et utile pour la construction et l'enrichissement d'ontologie. Notre démarche se veut généralisable à d'autres textes à condition que leur mise en forme soit explicite et structurée. Le domaine importe peu, c'est l'objectif taxonomique des informations dans les documents qui doit être privilégié.

Cependant, un certain nombre de problèmes ne sont pas résolus, notamment à cause de la complexité des liens de sens entre les différents objets textuels. Par exemple, le fonctionnement par patron impose un ordre d'application des règles qui rend délicat le traitement de certains phénomènes comme les anaphores et les pronominalisations, ou encore la négation. Sur ce dernier point, se pose la question de la gestion des informations négatives dans une ontologie. Par exemple, dans la définition de l'objet *réservoir* (« tous les réservoirs de plus de 10m de haut sont inclus sauf : les réservoirs souterrains et les citernes qui sont exclus »), comment les objets *réservoirs souterrains* et *citerne* doivent-ils être organisés ?

3. Au total, 441 fragments ont été évalués : 6 pour la relation d'artefact, 63 pour la relation de fonction, 291 pour la relation d'hyperonymie, 48 pour la synonymie et enfin 33 pour la méronymie/holonymie.
4. Effectuée par Marion Laignelet, l'une des auteures de ce papier.

Références

AUSSENAC-GILLES N. & JACQUES M.-P. (2008). Designing end evaluating patterns for relation acquisition from texts with caméléon. *Terminology*, **14**(1), 45–73.

CARTIER E. (1998). Analyse automatique des textes : l'exemple des informations définitoires. In *Actes Rencontre Internationale sur l'Extraction, le Filtrage et le Résumé Automatiques*, p. 6–18, Sfax, Tunisie.

CONDAMINES A. & REBEYROLLES J. (2000). *Recent Advances in Computational Terminology*, chapter Searching for and identifying conceptual relationships via a Corpus-based approach to a Terminological Knowledge Base (CTKB) : Method and Results, p. 127–148. M.-C. L'Homme and C. Jacquemin and D. Bourigault.

ESCOUBASBENVENISTE M.-P. (2010). La définition dans le texte économique écrit de vulgarisation savante. In PUBLIFARUM, Ed., *Autour de la définition*, volume 11.

HEARST M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proccedings of the 14th International Conference on Computational Linguistics*, p. 539–545.

HO-DAC L.-M., JACQUES M.-P. & REBEYROLLE J. (2004). *Sur la fonction discursive des titres*, In S. PORHIEL & K. D., Eds., *L'unité texte*, p. 125–152. Pleyben.

KAMEL M. & AUSSENAC-GILLES N. (2009). Utiliser la structure du document dans le processus de construction d'ontologies. In *8th Intern. Conference on Terminology and Artificial Inelligence TIA'09*.

KUSHMERICK N., WELD D. & DOORENBOS B. (1997). Wrapper induction for information extraction. In *Proc. Int. Joint Conf. Artificial Intelligence*.

LUC C. & VIRBEL J. (2001). Le modèle d'architecture textuelle : fondements et expérimentation. *Verbum*, **XXIII**(1), 104–123.

MAEDCHE A. (2002). *Ontology learning for the Semantic Web*, volume 665. Kluwer Academic Publishing.

MALAISÉ V., ZWEIGENBAUM P. & BACHIMONT B. (2005). Detecting semantic relations between terms in definitions. *Terminolgy*, **11**(1), pp. 21–53.

MAYNARD D., FUNK A. & PETERS W. (2009). Using lexico-syntactic ontology design patterns for ontology creation and population. In S. BLOMQVIST, Ed., *Proc. of the WS on Ontology Patterns (WOP 2009)*, volume 516, Washington DC, USA.

MEYER I. (2001). *Recent Advances in Computational Terminology*, chapter Extracting Knowledge-rich Contexts for Terminography : A Conceptual and methodological Framework, p. 279–302. M.-C. L'Homme and C. Jacquemin and D. Bourigault.

MONTIELPONSADA E., AGUADODECEA G. & ALAREZDEMON I. (2009). From linguistic pattern to ontology structures. In *Inter. Conference on Terminology and Artificial Intelligence TIA'09*.

MORIN E. (1999). Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. *Traitement Automatique des Langues*, **40**(1), 143–166.

PASCUAL E. & PÉRY-WOODLEY M.-P. (1997). Modélisation des définitions dans les textes à consignes. In J. VIRBEL, J.-M. CELLIER & J.-L. NESPOULOUS, Eds., *Cognition, discours procédural, action.* Atelier "Texte et Communication" PRESCOT.

REBEYROLLES J., JACQUES M.-P. & PÉRY-WOODLEY M.-P. (2009). Titres et intertitres dans l'organisation du discours. *Journal of French Language Studies*, **19**(2), 269–290.

REBEYROLLES J. & TANGUY L. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire*, **25**, 153–174.

ROLE F. & ROUSSE G. (2006). Construction incrémentale d'une ountologie par analyse du texte et de la structure du document. *Document numérique*, **9**(1), 77–91.

SABOU M., D'AQUIN M. & MOTTA E. (2008). Scarlet : Semantic relation discovery by harvesting online ontologies. In *European Semantic Web Symposium - Conference - ESWS*, p. 854–858.

SCHUTZ A. & BUITELAAR P. (2005). Relext : A tool for relation extraction from text in ontology extension. In SPRINGER, Ed., *Proceedings of the 4th International Semantic Web Conference (ISWC)*, volume 3729, p. 593–606.