Vérification sémantique pour l'annotation d'entités nommées

Caroline Brun (1), Caroline Hagège (2)

(1) Xerox Research Centre Europe – 6, chemin de Maupertuis, 38240 Meylan France
Caroline.Brun@xrce.xerox.com
(2) Xerox Research Centre Europe – 6, chemin de Maupertuis, 38240 Meylan France
Caroline.Hagege@xrce.xerox.com

Résumé Dans cet article, nous proposons une méthode visant à corriger et à associer dynamiquement de nouveaux types sémantiques dans le cadre de systèmes de détection automatique d'entités nommées (EN). Après la détection des entités nommées et aussi de manière plus générale des noms propres dans les textes, une vérification de compatibilité de types sémantiques est effectuée non seulement pour confirmer ou corriger les résultats obtenus par le système de détection d'EN, mais aussi pour associer de nouveaux types non couverts par le système de détection d'EN. Cette vérification est effectuée en utilisant l'information syntaxique associée aux EN par un système d'analyse syntaxique robuste et en confrontant ces résultats avec la ressource sémantique WordNet. Les résultats du système de détection d'EN sont alors considérablement enrichis, ainsi que les étiquettes sémantiques associées aux EN, ce qui est particulièrement utile pour l'adaptation de systèmes de détection d'EN à de nouveaux domaines.

Abstract In this paper we propose a new method that enables to correct and to associate new semantic types in the context of Named Entity (NE) Recognition Systems. After named entities (and more generally proper nouns) have been detected in texts, a semantic compatibility checking is performed. This checking can not only confirm or correct previous results of the NER system but also associate new NE types that have not been previously foreseen. This checking is performed using information associated to the NE by a robust syntactic analyzer and confronting this information to WordNet. After this checking is performed, final results of the NER system are better and new NE semantic tags are created. This second point is particularly useful when adapting existing NER systems to new domains.

Mots-clés : Entités nommées, Analyse syntaxique robuste, Types sémantiques
Keywords: Named Entities, Robust Parsing, Semantic Types

1 Introduction

Dans cet article, nous proposons une méthode permettant d'enrichir les résultats obtenus par un système de détection automatique d'entités nommées en utilisant les relations syntaxiques qui leur sont associées par un analyseur syntaxique robuste, XIP (Xerox Incremental Parser), et en vérifiant les types sémantiques des arguments en relation syntaxique à l'aide de la ressource ontologique WordNet (Fellbaum, 1998). L'utilisation de relations syntaxiques pour raffiner la tâche de détection d'EN n'est pas nouvelle, voir par exemple (Ehrmann et Jacquet, 2006) ou (Brun et Hagège, 2004). De même, l'utilisation de WordNet dans le cadre de la détection d'EN est décrite dans plusieurs travaux, comme par exemple dans (Magnini et al., 2002). Les travaux de (Benetti et al., 2006) montrent aussi l'enrichissement d'un système de reconnaissance d'EN grâce à l'utilisation de Wikipedia. La nouveauté de notre méthode est de coupler information syntaxique et information sémantique sur les résultats de cette analyse syntaxique fine. Cela va nous permettre non seulement de valider ou d'invalider les résultats préalablement fournis par le système de détection d'EN, mais aussi de considérer de nouveaux types sémantiques via WordNet et de les associer à des noms propres non typés par le système initial, créant ainsi de nouvelles catégories d'entités nommées.

Après avoir présenté l'analyseur syntaxique robuste XIP, ainsi que le système de détection d'EN, basé également sur XIP, nous décrivons en détail cette méthode et son implantation. Puis, nous donnons les résultats de différentes expérimentations réalisées avec le prototype que nous avons développé. Enfin, nous tirons les conclusions relatives à cette méthode.

2 Détection d'entités nommées

2.1 Généralités

La détection d'entités nommées fait l'objet d'un intérêt certain pour le TALN (voir les conférences MUC http://www.itl.nist.gov/iaui/894.02/related_projects/muc/, les campagnes ACE http://www.nist.gov/speech/tests/ace/, ESTER, etc.), en particulier pour la tâche d'extraction d'information, mais aussi pour beaucoup d'autres applications spécifiques. De nombreux systèmes, symboliques ou statistiques, détectent et catégorisent les NE avec de bonnes performances (environ 90 de f-mesure). Cependant, certaines applications requièrent plus de précision et la méthode que nous proposons vise à améliorer a posteriori les résultats d'en système d'extraction d'EN.

2.2 L'analyseur syntaxique robuste XIP

XIP (Ait-Mokhtar et al., 2002) est un analyseur dont l'objectif est d'extraire des dépendances syntaxiques de façon robuste. Cet analyseur accepte en entrée n'importe quel document ou partie de document au format texte ou XML et produit en sortie une représentation grammaticale du contenu de ce document.

Le formalisme proposé par XIP nous permet d'exprimer un large éventail de règles qui vont de la désambiguïsation catégorielle à la construction de dépendances, en passant par la constitution de syntagmes noyaux : XIP permet de relier par des relations des éléments linguistiques qui peuvent être des éléments lexicaux, mais aussi des éléments non lexicaux correspondant à des syntagmes noyaux.

Dans le cadre du travail présenté dans cet article, nous utilisons la version la plus complète de la grammaire de l'anglais développée au sein de XIP, que nous désignons par grammaire «de normalisation ». Cette grammaire est construite sur la base des résultats obtenus lors de l'analyse grammaticale générale de l'anglais.

Grammaire générale : La grammaire générale de l'anglais permet le « chunking » (analyse en syntagmes noyaux) et réalise une extraction des dépendances standards (Sujet, Objet, modifieurs, attributs etc.).

A titre d'exemple, voici une phrase analysée par cette grammaire :

McMurphy was successful in changing many of the rules that were imposed upon them by Nurse Ratched.

Analyse par XIP (grammaire générale):

```
MAIN(was)
NUCL_SUBJCOMPL(was,successful)
SUBJ_PRE(was,McMurphy)
MOD_POST_GERUND(was,changing)
OBJ_POST(changing,rules)
QUANTD(rules,many)
MOD_POST_SENTENCE_RELATIV(rules,im
posed)
NUCL_VLINK_PASSIVE(were,imposed)
SUBJ_PRE_RELATIV(were,that)
AGENT(imposed,Nurse Ratched)
MOD_POST(imposed,them)
MOD_POST(them,Nurse Ratched)
PERSON(Nurse Ratched)
PERSON(McMurphy)
```

Grammaire de normalisation (Hagège, Roux, 2003)**:** Une couche supplémentaire de développement a été rajoutée à cette grammaire de base, l'objectif applicatif visé étant l'extraction d'information. Ces développements permettent d'avoir, après l'analyse, une représentation commune pour des suites de signifiants qui ne sont pas identiques mais qui véhiculent une information similaire. A l'heure actuelle, ce travail de normalisation s'effectue selon trois axes :

- L'exploitation des relations syntaxiques mises en évidence lors de l'analyse générale : L'analyse par la grammaire générale est tout d'abord raffinée afin de considérer les sujets et objets de verbes non finis et les antécédents des relatives dans le calcul du sujet et de l'objet, de normaliser la forme passive en forme active et de typer certains compléments. Ensuite, certaines alternances verbales telles qu'elles sont définies dans (Levin, 93) sont exploitées.

- La hiérarchisation des propositions dans une phrase : la grammaire normalisée permet de reconnaître les degrés d'enchâssements des verbes par rapport au verbe principal.

- L'exploitation d'information de morphologie dérivationnelle : cette information permet d'exprimer des équivalences entre verbe-compléments et nom-arguments.

L'analyse de la phrase précédente avec cette version de la grammaire donne alors :

Analyse par XIP (grammaire normalisée):

```
ATTRIB(McMurphy,successful)
MAIN(was)
MOD_POST_GERUND(was,changing)
SUBJ-N_PRE(was,McMurphy)
OBJ-N(changing,rules)
EMBED_PROG(changing,was)
SUBJ-N(changing,McMurphy)
MOD_POST_SENTENCE_RELATIV(rules,im
posed)
SUBJ-N(imposed,Nurse Ratched)
OBJ-N(imposed,rules)
MOD_POST(imposed,them)
```

```
MOD_POST(them,Nurse Ratched)
PERSON(Nurse Ratched)
PERSON(McMurphy)
SUBJ-N(succeed,McMurphy)
```

Nous pouvons remarquer dans l'analyse effectuée par la grammaire de normalisation qu'une relation de type « sujet normalisé » (SUBJ-N) est établie entre le verbe « succeed » et « McMurphy » (à partir de la suite « McMurphy was successful »). L'identification de l'antécédent de la relative ainsi que la normalisation entre forme passive et forme active permettent d'extraire une relation « sujet normalisé » entre le verbe « impose » et « Nurse Ratched » et une relation « objet normalisé » entre ce même verbe et le nom « rule ». Enfin, une relation « sujet normalisé » est également extraite entre le verbe « change » et le nom « McMurphy ».

C'est cette version de la grammaire que nous avons utilisée dans pour construire le prototype de validation et découverte d'EN.

2.3 XIP et la détection d'EN

Un système de détection des EN a été développé au sein de l'analyseur XIP. Il permet de détecter les types « standards » d'entités nommées à savoir : dates, pourcentages, monnaies, lieux, personnes, organisations. Il s'agit d'un système à base de règles, consistant en un ensemble de règles locales qui utilisent de l'information lexicale combinée à de l'information contextuelle sur les catégories syntaxiques. Ces règles locales sont très similaires à des règles de « chunking » (identification des syntagmes noyaux), sauf qu'elles opèrent au niveau du nom.

Voici un exemple d'analyse[1]:

Margaret Sinclair Trudeau, born September 10, 1948 in Vancouver, British Columbia, Canada, was the wife of the late Canadian Prime Minister Pierre Trudeau. The daughter of James Sinclair, a former Liberal member of the Parliament of Canada and fisheries minister, she attended Simon Fraser University where she obtained a degree in English literature.

```
PERSON(Margaret Sinclair Trudeau)
DATE(September 10 , 1948)
LOC_CITY(Vancouver)
LOC_REGION(British Columbia)
LOC_COUNTRY(Canada)
PERSON(Canadian Prime Minister  Pierre Trudeau)
PERSON(James Sinclair)
ORGANISATION(Parliament of Canada)
ORGANISATION(Simon Fraser University)
```

Le système a été évalué en interne, sur un corpus de dépêches d'environ 87000 mots, et montre une f-mesure de 90 tous types d'entités confondus (Erhmann 2004). Ce système de détection des entités nommées est intégré aux différentes grammaires présentées dans le paragraphe précédent. La validation et découverte d'EN se fait sur la base des résultats obtenus par ce système initial.

[1] Les entités extraites sont présentées sous forme de « dépendances unaires ».

3 Validation et découverte d'EN

3.1 Détection des relations attributives

La première étape réalisée par notre prototype est de détecter les entités nommées ainsi que les entités potentielles non étiquetées sémantiquement (noms propres) à l'aide du système présenté au paragraphe 2.3. Notre système va considérer comme nom propre, toute suite de mots (éventuellement non reconnus par l'analyseur lexical) présentant des particularités typographiques comme une majuscule initiale et ne rentrant dans aucune des catégories d'entités nommées reconnues par le système.

Nous appliquons ensuite sur le même texte la grammaire normalisée. Cette grammaire normalisée a pour particularité d'extraire une relation que nous désignons par relation attributive. Une relation attributive relie des suites de caractères lorsque des indicateurs textuels et des constructions syntaxiques permettent d'affirmer qu'ils entretiennent une relation de type IS-A

Les exemples suivants illustrent la notion de relation attributive telle que nous l'entendons.

(1) J*ohn Smith was an inventor.*

(2) *John Smith, the inventor, made a presentation.*

(3) *John Smith is expected to be an inventor.*

(4) *They consider John Smith an inventor.*

(5) *The inventor John Smith was awarded.*

(6) A*n inventor called John Smith was awarded.*

(7) *John Smith, who is a great inventor, was awarded.*

(8) *John Smith, as the inventor of the process, was awarded.*

Dans tous ces exemples, "John Smith" est en relation attributive avec "inventor".

A titre d'exemple, l'analyse donnée par XIP pour la phrase (5), en utilisant la grammaire de normalisation, est la suivante :

```
SUBJ-N_PRE(consider,They)          OBJ-N(consider,Smith)
PREPD(inventor,as)                 ATTRIB(John Smith,inventor)
PERSON(John Smith)
```

Nous nous intéresserons uniquement aux relations attributives mettant en jeu les entités nommées et nom propres extraites à l'aide de XIP : elles correspondent à une relation de type IS-A du point de vue sémantique, et vont ainsi nous permettre de typer sémantiquement les EN.

3.2 Confrontation avec WordNet

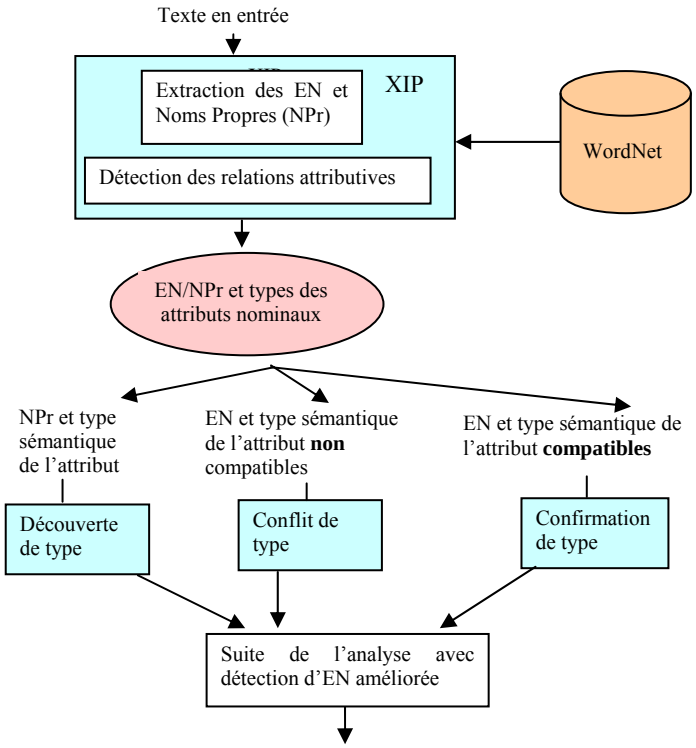
Nous utilisons l'information sémantique fournie par la base de données lexicale WordNet[2] (Fellbaum 1998), en particulier les classes sémantiques de plus haut niveau associées aux synsets.

Une fois les relations attributives extraites entre entités et attribut nominal, le système utilise WordNet pour associer un type sémantique à l'attribut nominal. Pour ce faire, nous avons extrait de WordNet tous les noms dont le « super type » (person, artifact, substance, etc.) n'est pas ambigu, par exemple :

Girl [5 sens] → noun.person Ship [1 sens] → noun.artifact
Drug [1 sens] → noun.artifact Liquidation [3 sens] → noun.act

En résulte un vocabulaire de 44406 noms accompagnés de leur « super type » sémantique qui est intégré à des lexiques au sein de XIP. Un appariement entre « super type » sémantique de WordNet et type d'entités nommées reconnues par le système initial est également effectué (par exemple le type WordNet « person » correspond au type « PERSON » de notre système de reconnaissance de EN.

Le schéma suivant décrit l'architecture du prototype que nous avons développé :



Le résultat de la confrontation entre entité nommée ou nom propre extraits par le système initial et type sémantique de l'attribut nominal selon WordNet peut produire les cas de figures suivants :

[2] http://wordnet.princeton.edu/

1) Conflit de type

Dans ce cas, le système initial a associé à une EN un certain type sémantique qui n'est pas compatible avec le type que WordNet assigne à l'attribut de cette entité, comme par exemple dans la phrase (où l'EN est indiquée en gras) :

*The warship is called the **Armando Diaz***

Le type de l'EN extrait par le système initial est « PERSON » alors que le type de « warship » selon WordNet est « artifact ». Ces deux types sont contradictoires.

2) Confirmation de type

C'est le cas pour lequel le type de l'entité nommée est compatible avec le type WordNet de son attribut. Dans ce cas, il s'agit d'une information supplémentaire qui vient confirmer le choix du système initial, comme par exemple dans :

*If one man has done more than any other to keep Old Labour behind Tony Blair it is surely his deputy, **John Prescott.***

Une relation attributive est détectée entre « deputy » et « John Prescott », préalablement identifié comme nom de personne par le système initial. Selon WordNet, « deputy », n'appartient qu'au « super type » « person ». Les deux types sont parfaitement compatibles, donc le système confirme le type de l'entité.

3) Découverte d'un nouveau type

Ce cas s'applique uniquement aux noms propres non typés extraits par le système initial. Grâce à l'association d'un attribut à ce nom propre, et au typage par WordNet de cet attribut, nous pouvons proposer d'attribuer ce type sémantique WordNet au nom propre. Par exemple dans la phrase suivante :

***Activia** is yogurt, but not just any yogurt.*

"Activia" n'est pas détecté comme une EN par le système initial de XIP. Il est cependant en relation attributive avec le nom « yogurt », dont le super type dans WordNet est « food ». Le système associe donc la nouvelle étiquette sémantique « food » à « Activia ».

4 Expérimentations sur corpus

4.1 Corpus général

Le corpus général, de 5500 mots, est constitué d'un ensemble de dépêches de provenance diverse (Herald Tribune, The Guardian, The Observer, The New York Times) ainsi que de quelques courtes biographies. Le système initial de reconnaissance d'EN détecte 7441 entités nommées.

L'application postérieure de notre méthode nous permet de considérer 115 entités, parmi lesquelles 78 sont des compatibilités entre types, 19 sont des découvertes de nouveau type, 18 sont des conflits ;

Compatibilité :

Seuls des entités de type PERSON et ORGANISATION sont concernées. Les 78 cas de compatibilités extraites sont justes, comme par exemple pour la phrase suivante :

The Democratic challenger, John Kerry, has called on the White House to turn words into action.

John Kerry a été détecté comme EN de type personne par le système général ("John" étant une suite codée comme prénom dans le lexique). Cela est confirmé dans un deuxième temps grâce à l'identification du lien attributif entre "John Kerry" et "challenger" qui lui-même est considéré comme de type personne par WordNet.

Découverte :

Les types découverts et proposés par le biais de WordNet dans ce texte sont les types COMMUNICATION et ARTIFACT, comme par exemple dans :

For Tom Hanks, it's his maiden voyage to Cannes where he will be publicising the Coen brothers ' remake of the classic British comedy The Ladykillers.

```
ENTITE(Cannes)
ENTITE(Ladykillers)
PERSON(Tom Hanks)
PERSON(Coen)
LOCORG_CITY(Cannes)
ATTRIB(Ladykillers,comedy)
DISCOVERY_WN_COMMUNICATION(Ladykillers)
```

"Ladykillers" qui avait simplement été repérée comme nom propre par le système général, se voit attribuer le type "COMMUNICATION" grâce à la relation attributive qu'il entretient avec le nom "comedy".

Sur les 19 entités découvertes, 5 sont erronées et 14 sont correctes. Les erreurs sont dues a des erreurs de détection de la relation attributive provenant en général d'une erreur dans la détection de la tête d'un groupe nominal complexe. Cette erreur de détection de la tête est à son tour souvent liée à des erreurs de désambiguisation de la partie du discours.

On peut remarquer que le faible nombre d'entités découvertes s'explique par le fait que le corpus traité est en harmonie avec le type d'entités que le système général prévoit (langue générale, presse).

Conflit :

Nous obtenons pour ce corpus 17 cas de conflit. Il est intéressant de noter que parmi ces conflits, certains relèvent de cas d'emploi métonymique des entités nommées, comme par exemple dans la phrase suivante :

While Schröder was saying that D-Day signaled the starting point for today's new Europe and maintained that the EU was the best guarantor of peace in Europe,...

Une relation attributive entre l'entité nommée « EU » et le nom « guarantor » est extraite. Le système de base de détection des EN indique que « EU » est de type ORG. Par le biais de notre méthode, dans la mesure où le nom « guarantor » est de type PERSON, nous obtenons un conflit de type.

Or, dans le cas présent, c'est bien un usage métonymique de l'entité « EU » dont il s'agit dans cette phrase, même si par nature, « EU » correspond à un lieu ou à une organisation.

Parmi les 18 cas de conflit, nous obtenons 6 cas de conflits liés à des usages métonymiques des entités nommées détectées. Nous pouvons donc considérer que ces 6 cas sont pertinents. Nous obtenons également 3 cas pertinents de conflits pour lesquels le système initial de détection des EN a fait des erreurs que nous pouvons corriger grâce à notre méthode.

Enfin, les 8 cas restants sont des détections de conflit erronées, ces erreurs étant essentiellement liées, comme pour les erreurs de découverte à des erreurs d'analyse syntaxique et de désambiguïsation.

4.2 Corpus spécialisé

Nous avons également testé notre méthode sur un corpus[3] spécialisé de biologie, d'environ 92000 mots (il s'agit de 422 articles extraits de pubmed). Nous nous sommes intéressées aux noms de gènes, annotés dans ce corpus, et qui correspondent aux classes sémantiques « BODY » et « SUBSTANCE» de Wordnet. Sur ce corpus, le système « découvre » 505 noms de gènes, et détecte 8 cas de conflits avec le système initial. Voici une illustration de ces résultats :

(1) *Structural basis of multidrug recognition by BmrR, a transcription activator of a multidrug transporter.*

ATTRIB(activator,BmrR)
DISCOVERY_WN_SUBSTANCE(BmrR)
0>TOP{NP{AP{Structural} basis} PP{of NP{multidrug recognition}} PP{by NP{BmrR}} , NP{a transcription activator} PP{of NP{a AP{multidrug} transporter}} .}

Le nom propre « Bmr» est découvert comme une entité de type « Substance » (gène) par notre système car l'attribut « activator» appartient à la classe noun.substance dans WordNet.

(2) *ADP is an inhibitor of the phosphorylation by ATP.*

ORGANISATION(ADP)
ATTRIB(ADP,inhibitor)
CONFLICT_WN_SUBSTANCE(ADP)
0>TOP{SC{NP{ADP} FV{is}} NP{an inhibitor} PP{of NP{the phosphorylation}} PP{by NP{ATP}}}

Ici, un conflit est détecté car le système initial de reconnaissance considère ADP comme une organisation (ADP= Aéroport de Paris), alors que notre système lui associe le type SUBSTANCE.

Les 8 cas de conflits détectés sont tous corrects, c'est-à-dire que le système initial donne une annotation erronée pour un nom de gène. Concernant les noms de gènes découverts, l'évaluation sur ce corpus donne 85% de précision et 6,6 % de rappel.

Nous obtenons donc une très bonne précision pour un rappel faible, ce dernier point étant assez prévisible car le système n'utilise que des relations attributives pour détecter les EN. A titre d'information, 1900 relations attributives sont détectées sur l'ensemble du corpus.

---

[3] Disponible ici : http://mig.jouy.inra.fr/recherches/bibliome/linguistic-and-semantic-resources/corpora/manual_genes_annotation.tar.bz2/view

5 Conclusion

La méthode que nous présentons permet d'améliorer les résultats d'un système de reconnaissance d'entités nommées en validant ou invalidant les résultats produits par ce système et permet également de proposer de nouveaux types d'entités nommées pour des noms propres détectés mais non typés par le système. La méthode utilise les résultats d'une analyse syntaxique fine permettant d'extraire des relations de type IS-A entre noms propres et entités reconnues par le système initial et d'autres noms communs catégorisés sémantiquement par la ressource WordNet.

Les tests effectués sur les corpus montrent que l'apport de notre méthode est très variable selon le type de corpus sur lequel on travaille. Dans le cas de textes appartenant à un domaine particulier, c'est l'aspect découverte de nouvelles entités dont les types n'ont pas été préalablement prédéfinis qui semble le plus intéressant. En effet, même si la couverture reste basse elle peut constituer une amorce fiable pour des systèmes d'apprentissage. Dans le cas de textes généraux pour lesquels les types d'entités nommées prédéfinis sont couvrants, c'est l'aspect confirmation qui semble le plus prometteur.

Références

AÏT-MOKTHAR S., CHANOD J.P., ROUX C. (2002). Robustness beyond Shallowness : Incremental Dependency Parsing. *Special issue of NLE Journal*.

BENETI A., HAMMOUMI W., HIELSCHER E., MULLER M., PERSONS D. (2006). Automatic Generation of Fine-Grained Named Entity Classification, http://www.ifarm.nl/erikt/ltp2006/ltp2006.pdf

BRUN C. HAGÈGE H. (2004). Intertwining deep syntactic processing and named entity detection. *Actes de ESTAL 2004*, Alicante, Spain, October 20-22, 2004.

EHRMANN M. (2004). Evaluation d'un Système d'extraction d'Entités Nommées. *Rapport de stage DESS Texte*, Nancy, 2004.

EHRMANN M., JACQUET G. (2006). Vers une double annotation des entités nommées. *Revue TAL*, numéro 46, volume 3.

FELLBAUM C. (1998). *WordNet: An Electronical Lexical Database*. MIT Press, Cambridge, USA.

HAGÈGE C, ROUX C. (2003). Entre syntaxe et sémantique : normalisation de l'analyse syntaxique en vie de l'amelioration de l'extraction d'information. *Actes de TALN 2003*, Batz-sur-Mer, France.

LEVIN, B. (1993). English Verb Classes and Alternations – A preliminary Investigation. *The University of Chicago Press*.

MAGNINI B, NEGRI M, PREVETE R., TAEV H. (2002). A WordNEt-Based Approach to Named Entities Recognition. COLING-02 on SEMANET, Vol.11, pp. 1-7