# Acquisition de relations sémantiques pour un Web Biomédical

**Laurent Alamarguy[1] et Rose Dieng-Kuntz[1]**

[1] *Projet Acacia, INRIA Sophia Antipolis, France*
*{Laurent.Alamarguy, Rose.Dieng}@sophia.inria.fr*

**Résumé**
Nous décrivons une méthodologie d'acquisition de relations sémantiques à partir de corpus, basée sur une analyse des fonctions syntaxiques. Cette méthodologie tend à construire et à enrichir des ontologies et des annotations sémantiques de manière coopérative, afin de pouvoir optimiser la recherche d'informations sur le web au travers d'un moteur d'inférences.
***Mots-clés*** : Ontologies ; relations sémantiques ; sémantique lexicale ; acquisition de connaissances sur corpus ; TALN ; RDF ; génétique.

## 1. Introduction

La capitalisation des connaissances dans le domaine biomédical devient un enjeu important ; devant la masse considérable de données sur le domaine, le besoin de structurer et d'ordonner ces connaissances nécessite une automatisation au moins partielle de la tâche. L'objectif est d'aider à l'élaboration d'une mémoire de communauté, en recensant des corrélations entre gènes et maladies du système nerveux central. Pour ce faire, il est nécessaire que la recherche d'information s'appuie sur un web construit sémantiquement, reposant sur des ontologies du domaine et des documents annotés sémantiquement. Pour cela le moteur d'inférences CORESE [1] traite des métadonnées en RDF(S) qu'il traduit en Graphes Conceptuels [2] améliorant ainsi la recherche d'informations.

## 2. Contexte

Afin de transformer des données hétérogènes pour les intégrer dans un web sémantique, nous proposons une approche d'analyse linguistique de corpus visant à construire semi-automatiquement une ontologie ou des annotations sémantiques. Cela se traduit par une extraction de connaissances faisant émerger des hiérarchies consensuelles de types conceptuels et relationnels qui doivent expliciter les idées implicites des documents du web. Comme il est souligné dans [3], il est important de s'attacher à faire émerger les relations sémantiques dans le domaine de la génomique puisque la fonction des gènes est l'information essentielle à découvrir. Nous nous focalisons donc plus particulièrement sur l'extraction de relations sémantiques reposant sur une analyse des fonctions syntaxiques. Nous partons de l'hypothèse que plusieurs schémas syntaxiques puissent référer à une même structure conceptuelle. Ainsi si l'on prend exemple sur les objectifs des projets MetaMap et SemRep [4] nous devrions pouvoir acquérir des connaissances du style : hypoxaemia TRIGGER cardiovascular events IN dialysis patients. En revanche pour parvenir à ce degré de granularité des connaissances, nous devons pouvoir distinguer la fonction de chaque argument d'une fonction génomique. Cette granularité est également nécessaire pour désambiguiser les expressions telles que *inhibitor protein* et *enzyme inhibitor* puisqu'ici l'enzyme et la protéine ne jouent pas le même rôle dans la fonction d'inhibition. Par exemple nous devrions pouvoir reconnaître pour une fonction d'inhibition quel est l'argument *inhibant*, l'argument *inhibé*, la localisation où a lieu cette inhibition, *etc*.

## 3. Méthode

Dans un premier temps nous nous appuyons sur des ontologies et des outils de structuration des connaissances déjà existants : la GeneOntology, et l'UMLS par l'intermédiaire de

MetaMap [4] qui permet de trouver les concepts du domaine biomédical, des divers domaines de spécialité, ainsi que les concepts plus génériques correspondants.
Notre approche d'acquisition de relations sémantiques s'appuie essentiellement sur les schémas de sémantique lexicale basés sur les dépendances syntaxiques [5, 6]. Il y a plusieurs avantages à s'appuyer sur ce type de modèles : tout d'abord la structure sémantique est suffisamment granulaire, de même que l'interface syntaxe-sémantique est au cœur de la problématique de ce type de modèle, avec une recherche de correspondance entre les relations grammaticales (*e.g.*, Sujet, Objet, *etc*.) et les rôles sémantiques des arguments (*e.g.*, inhibant, inhibé, *etc*.). Enfin, plusieurs schémas syntaxiques peuvent référer à une même structure sémantique, en n'étant pas contraint par les catégories lexicales (*e.g.*, *inhibition, inhibit, inhibitor* vont tous trois référer à une seule structure sémantique).
Notre méthode est coopérative, *i.e.*, elle se compose de différentes étapes où éventuellement lors de chacune d'elle un expert du domaine peut intervenir pour valider les connaissances extraites. La méthode débute par la phase de constitution du corpus, où nous avons sélectionné environ 5000 résumés de la base Medline, traitant du type de corrélations recherchées. Ensuite, nous effectuons sur le corpus une extraction terminologique (nous utilisons en l'occurrence les outils Nomino [7], Fastr [8]) qui va nous servir à faire émerger les termes du domaine récurrents. Ainsi, *dialysis patient* sera considéré comme un terme à part entière. Pour chaque terme on cherche s'il est ou non déjà recensé dans une des ontologies. Parallèlement, on effectue sur le corpus une phase d'extraction des relations grammaticales en utilisant le shallow parser RASP [9]. Les dépendances syntaxiques qui en résultent seront exploitées pour l'élaboration de schémas lexico-syntaxiques. Ces schémas fonctionnent sur le principe que plusieurs structures syntaxiques renvoient à une même relation sémantique. Ainsi les deux structures syntaxiques *hypoxaemia triggers cardiovascular events in dialysis patients* et *hypoxaemia as a trigger of cardiovascular events in dialysis patients* sont conceptuellement identiques sur le fait qu'une action de déclenchement possède un rôle sémantique de *déclenchant*, de *déclenché* et de *domaine d'action*. Chaque document est annoté par un schéma représentant une relation sémantique prenant pour arguments les concepts des ontologies d'après les termes extraits. Le document de notre exemple aurait donc pour *déclenchant* le concept HYPOXAEMIA trouvé dans l'UMLS ; en revanche le *domaine d'action* resterait le terme extrait *dialysis patient* puisque recensé dans aucune des ontologies.
Ainsi nous voulons souligner dans ce papier que l'analyse linguistique reposant sur les dépendances syntaxiques est particulièrement bien adaptée pour l'émergence de fonctions géniques pour une annotation sémantique cohérente des données biomédicales du web.

**Références**

[1] Corby, O., Dieng, R., & Hébert, C. (2000). A Conceptual Graph Model for W3C Resource Description Framework. In *ICCS 2000*. Springer Verlag (LNAI 1867).
[2] Sowa, J.F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley.
[3] Burgun, A., Bodenreider, O., Le Duff, F., Mounssouni, F., & Loréal, O. (2002). Representation of roles in biomedical ontologies: a case study in functional genomics. In *AMIA 2002*.
[4] Aronson, A.R., & Rindflesch, T.C. (1998). *Semantic knowledge representation project*. Report to the Board of Scientific Counselors. Lister Hill National Center for Biomedical Communications.
[5] Fillmore, C.J., & Atkins, B.T.S. (1998). FrameNet and lexicographic relevance. In *LREC'98*.
[6] Hudson, R. (2003). Word Grammar. In H. Cuyckens & D. Geeraerts (Eds.), *Handbook of Cognitive Linguistics*. Oxford University Press.
[7] Dumas, L., Plante, A., & Plante, P. (1997). *ALN: Analyseur Linguistique de ALN*, vers.1.0. ATO, UQAM.
[8] Jacquemin, C., Klavans, J.L., & Tzoukermann, E. (1997). Expansion of Multi- Word Terms for Indexing and Retrieval Using Morphology and Syntax. In *ACL-EACL'97*.
[9] Briscoe, T., & Carroll, J. (2002). Robust accurate statistical annotation of general text. In *LREC 2002*.