

# Comment valider automatiquement des relations syntaxiques induites

Nicolas Béchet\*, Mathieu Roche\*, Jacques Chauché\*

\*LIRMM - UMR 5506, CNRS - Univ. Montpellier 2 - 34392 Montpellier Cedex 5 - France  
{bechet,mroche,chauche}@lirmm.fr

**Résumé.** Nous présentons dans cet article des approches visant à valider des relations syntaxiques induites de type Verbe-Objet. Ainsi, nous proposons d'utiliser dans un premier temps une approche s'appuyant sur des vecteurs sémantiques déterminés à l'aide d'un thésaurus. La seconde approche emploie une validation Web. Nous effectuons des requêtes sur un moteur de recherche associées à des mesures statistiques afin de déterminer la pertinence d'une relation syntaxique. Nous proposons enfin de combiner ces deux méthodes. La qualité de nos approches de validation de relations syntaxiques a été évaluée en utilisant des courbes ROC.

## 1 Introduction et contexte

L'acquisition de connaissances sémantiques est une importante problématique en Traitement Automatique des Langues (TAL). Ces connaissances peuvent par exemple être utilisées pour extraire des informations dans les textes ou pour la classification de documents. Les connaissances sémantiques peuvent être obtenues par des informations syntaxiques (Fabre et Bourigault (2006)). Comme nous allons le montrer dans cet article, les connaissances sémantiques acquises via la syntaxe permettent de constituer des classes conceptuelles (regroupement de mots ou termes sous forme de concepts). Par exemple, les mots *hangar*, *maison* et *mas* sont regroupés dans un concept *bâtiment*. De plus, ces concepts peuvent être organisés sous forme hiérarchique formant ainsi une classification conceptuelle.

Deux types d'informations syntaxiques peuvent être utilisés pour construire les classes sémantiques : les relations "classiques" issues d'une analyse syntaxique (Lin (1998), Wermter et Hahn (2004)) et les relations dites "induites" à partir des textes. Cet article s'intéresse plus particulièrement à ces dernières. La définition d'une relation induite est présentée ci-dessous. La méthode d'ASIUM consiste à regrouper les objets des verbes déterminés comme proches par une mesure de qualité (Faure (2000)). D'autres approches utilisent également ce principe, comme le système UPERY (Bourigault (2002)) qui regroupe les termes par des mesures de proximité distributionnelle. Par exemple, dans la figure 1, si les verbes *consommer* et *manger* sont jugés proches, des objets pouvant être obtenus par le biais d'informations syntaxiques sont regroupés (dans notre cas, les objets *essence*, *légume*, *nourriture* et *fruit*). Cependant, en considérant ce groupe d'objets, nous pouvons intuitivement exclure le mot *essence*. Notons que les objets *essence*, *légume* et *nourriture* appartiennent à un même contexte en tant qu'objets