# EXIT : EXtraction Itérative de la Terminologie

Mathieu Roche, Thomas Heitz, Oriane Matte-Tailliez, Yves Kodratoff

LRI, Université Paris-Sud, F91405 Orsay Cedex
roche@lri.fr, thomashe@firtech.lri.fr, oriane@lri.fr, yk@lri.fr

Les travaux que nous présentons se rapportent à l'extraction de la terminologie à partir de textes spécialisés. Nous travaillons à partir de quatre corpus de tailles, de langues et de spécialités différentes. La première étape de notre travail a consisté à normaliser le corpus en utilisant des règles de nettoyage puis à apposer des étiquettes grammaticales à chacun des mots en utilisant l'étiqueteur de Brill [Brill, 1994]. Nous pouvons alors extraire les collocations *nom-nom*, *adjectif-nom*, *nom-adjectif* et *nom-préposition-nom*. L'étape suivante consiste à sélectionner les collocations les plus pertinentes selon une mesure statistique [Roche *et al.*, 2003]. Pour comparer les différentes mesures, nous nous appuyons sur la précision évaluée par un expert du domaine et la courbe d'élévation qui donne la précision en fonction de la proportion de collocations.

Une des contributions essentielles de notre travail tient dans le caractère itératif de la méthode. En effet, les termes extraits à chaque itération sont réintroduits dans le corpus avec un trait d'union afin qu'ils soient reconnus comme des mots à part entière. Nous pouvons ainsi effectuer une nouvelle recherche terminologique à partir du corpus avec prise en compte de la terminologie du domaine acquise aux itérations précédentes. Notre méthode permet alors de détecter des termes très spécifiques composés de plusieurs mots. Ceci est essentiel, par exemple dans le domaine de la biologie moléculaire, où les termes les plus pertinents sont les termes composés de nombreux mots. Après l'acquisition des termes nominaux, nous avons extrait, de la même manière, les termes verbaux et, grâce au logiciel FASTR [Jacquemin, 1996], les termes variants.

Nous avons, de plus, ajouté des paramètres pour extraire des termes plus pertinents. Un des paramètres consiste à privilégier les termes présents dans de nombreux textes différents du corpus qui sont considérés comme davantage représentatifs du domaine.

Afin de faciliter le travail de l'expert en prenant en compte l'ensemble des caractéristiques de notre processus, nous avons développé une interface graphique qui permet de modifier certains paramètres (types de collocations à extraire, mesures à utiliser, élagage, etc.) et de visualiser les termes dans leur contexte (phrases).

## Références

[Brill, 1994] E. Brill. Some Advances in Transformation-Based Part of Speech Tagging. In *AAAI, Vol. 1*, pages 722–727, 1994.

[Jacquemin, 1996] C. Jacquemin. A symbolic and surgical acquisition of terms through variation. In *Lecture Notes in Computer Science*, pages 425–438, 1996.

[Roche *et al.*, 2003] M. Roche, O. Matte-Tailliez, J. Azé, et Y. Kodratoff. Extraction de la Terminologie du Domaine : Étude de Mesures sur un Corpus Spécialisé Issu du Web. In *Actes des Journées Francophones de la Toile 2003*, pages 279–288, 2003.