Thésaurus pour la Sécurité Radiologique à partir d'un corpus de textes et d'outils linguistiques en ligne

Olena Goncharova-Orobinska[*,**], Jean-Hugues Chauchat[**], Natalya Sharonova[*]

[*] Kharkiv Polytechnical Institute KhPI, Ukraine
[**] Université de Lyon, Lyon 2, ERIC EA3083, France

Nous présentons un ensemble de méthodes pour créer un thésaurus d'un domaine spécialisé (la Sécurité Radiologique) à partir d'un corpus original de textes et d'outils linguistiques en ligne ; les résultats sont appliqués au français et au russe, ce qui permet de comparer les résultats sur des langues de structures grammaticales différentes et de mettre en évidence l'importance de la qualité des outils linguistiques disponibles pour chaque langue.

Deux corpus ont été créés à partir de documents de l'Agence Internationale pour l'Énergie Nucléaire : les Normes de Sûreté Radiologique et les rapports publiés par l'IAEA et les Commissions nationales de protection radiologique ; le corpus français contient environ 1 500 000 mots dans 63 documents ; le corpus russe contient 600 000 mots dans 48 documents. Ces deux corpus ont été étiquetés avec les balises morpho-syntaxiques avec les versions TreeTagger de chaque langue.

Nous proposons trois méthodes pour extraire des termes, installer un noyau d'ontologie, puis l'enrichir par la labélisation des concepts et par les relations qui les lient.

La première étape est la constitution de la liste initiale des concepts du domaine ; elle se base sur l'utilisation des patrons morpho-syntaxiques (nous les appelons "patrons terminologiques"). Une première analyse statistique des corpus a extrait les noms caractéristiques du domaine ; confrontée aux résultats du projet RISQUE, cette liste a permis d'élaborer un modèle conceptuel de la sécurité radiologique à l'aide d'un expert du domaine.

La deuxième méthode semi-automatique permet de compléter une ontologie de domaine. Cette méthode hybride combine différentes techniques (statistiques, linguistiques, lexicales). Elle permet de détecter les variations lexicales (les différentes entrées lexicales) des concepts initiaux du noyau de l'ontologie. Nous avons utilisé des dictionnaires de synonymes en lignes (CRISCO pour le français et DCS pour le russe), puis des classes de noms associés, dans les corpus, à des classes de verbes.

La troisième méthode permet d'établir les relations associatives entre les concepts de l'ontologie sous-jacente (noyau d'ontologie) au moyen des classes sémantiques des verbes.

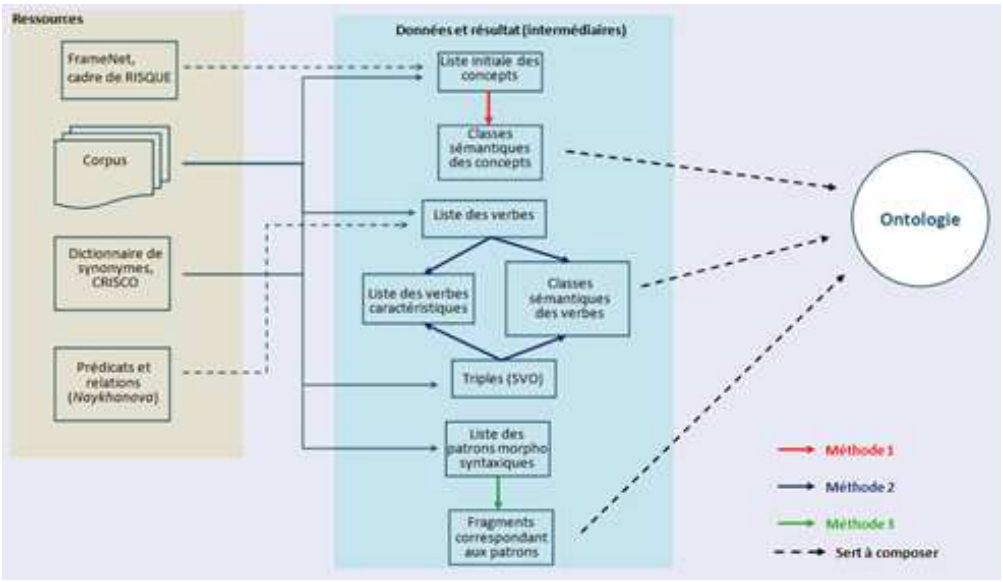

FIG. 1 – *Organisation de la chaîne de traitements*