Retour d'expérience sur la détection automatique de métaphores dans des textes de Géographie

Max Beligné**, Aleksandra Campar*, Jean-Hugues Chauchat*, Mélanie Lefeuvre*,
Isabelle Lefort**, Sabine Loudcher*, Julien Velcin*

*Université de Lyon, Lyon 2, ERIC EA 3083
{sabine.loudcher, julien.velcin, jean-hugues.chauchat}@univ-lyon2.fr
**Université de Lyon, Lyon 2, EVS UMR 5600
{max.beligne, isabelle.lefort}@univ-lyon2.fr

Si les recherches sur les métaphores en Géographie ne sont pas nouvelles, il n'existe pas d'étude de grande ampleur sur cette question. Pourtant les recherches et les réflexions existantes menées de façon qualitative sur des petits corpus ont montré que les métaphores constituent une entrée particulièrement intéressante pour réfléchir sur la scientificité de la Géographie. Dans le contexte des Humanités Numériques, un projet de recherche regroupant des chercheurs en Géographie, Informatique et Linguistique a donc identifié l'intérêt d'étudier ce trope sur un large corpus et cherche par conséquent à détecter automatiquement des métaphores dans des textes de Géographie. Cette communication présente un premier retour d'expérience de l'application de la méthode de Heintz et al. (2013). Il s'agit d'une première étape de travail dont les résultats sont pour l'instant mitigés. Par conséquent, l'objectif est de présenter le processus de recherche en insistant sur les choix qui ont été faits et sur leurs conséquences permettant de mieux comprendre les résultats obtenus et d'envisager des améliorations à venir.

La métaphore est un trope qui peut être défini comme un système de correspondances partielles entre un domaine source et un domaine cible. Le choix de la méthode de détection automatique de métaphores résulte d'un état des lieux sur la question. Le travail de Roy et al. (2006) présentant des évolutions diachroniques de métaphores conceptuelles dans des corpus textuels a été identifié comme le plus proche des attentes des géographes. Pourtant, la volonté de travailler sur un large corpus avec des thématiques diverses conduit à vouloir automatiser le processus de recherche. Dans ce cadre, le travail de Heintz et al. (2013) utilisant un modèle de thématiques latentes LDA (*Allocation de Dirichlet Latente*) est choisi comme répondant le mieux aux objectifs attendus. Il permet de travailler sur un large spectre de domaines sources de métaphores avec comme seule contrainte d'établir pour chaque domaine source quelques mots représentatifs. Concernant le domaine cible des métaphores, il est décidé dans un premier temps de cibler un seul domaine, celui de la Géographie.

La méthode choisie s'appuie sur l'utilisation de la méthode LDA sur un large corpus, ici la moitié des articles de Wikipédia choisi de manière aléatoire, pour extraire 100 thématiques. Les domaines sources choisis par les spécialistes sont ensuite articulés à ces thématiques par l'intermédiaire de mots représentatifs. Dans chaque phrase, la présence de chaque domaine source et du domaine cible est calculée par l'intermédiaire des fréquences d'apparition des mots dans les thématiques. Si une phrase contient le domaine cible et un domaine source sous représenté dans l'article (car la méthode fait l'hypothèse qu'une forte représentation est souvent le