Journée d'étude: Traitements Automatisés des Discours politiques.

Objets nouveaux, nouvelles méthodes.

CARTOGRAPHIE DE CORRESPONDANCES DANS LES CORPUS MULTILINGUES

Serge Fleury, Maria Zimina

SYLED-CLA2T, Université de la Sorbonne nouvelle – Paris 3

Les comparaisons isolées des unités textuelles recensées dans les différents volets de corpus multilingues sont souvent insuffisantes pour explorer les correspondances intertextuelles.

A côté des systèmes d'extraction automatique de lexèmes appariés présentés sous forme de listes (Fung, 2000 ; Déjean *et al*., 2002), il nous paraît indispensable de fournir à l'utilisateur des outils de navigation en corpus, modulables en fonction de besoins particuliers. Dans cette perspective, l'utilisation de la « cartographie de correspondances », fondée sur la représentation topographique du bi-texte, nous semble particulièrement adaptée à l'analyse de corpus multilingues (Zimina, 2004).

La topographie textuelle a pour objectif une localisation graphique des phénomènes mis en évidence par l'étude statistique (Lamalle et Salem, 2002). Dans le contexte de cette approche, l'expertise humaine de textes est appuyée par de nombreux outils de lecture et de visualisation qui offrent de nouveaux moyens d'investigation de l'espace textuelle (Lamalle *et al*., 2003).

Dans le contexte multilingue, la description cartographique du bi-texte peut être réalisée à l'aide d'outil dont l'interface graphique permet de construire et de visualiser l'alignement d'un couple de textes en modifiant au besoin la correspondance entre leurs segments respectifs (Fleury, 2005).

Ces nouvelles fonctionnalités de navigation textométrique à l'aide de la carte bi-textuelle seront illustrées au cours de notre communication par des exemples concrets d'explorations de corpus en français, anglais et russe.

Références :

Déjean (H.), Gaussier (É.), Sadat (F.) 2002 : « An approach based on multilingual thesauri and model combination for bilingual lexicon extraction », *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, 24 août-1 septembre 2002, Association for Computational Linguistics, Morristown, NJ, p. 1-7.

Fleury (S.) 2005 : « mkAlign », *documentation*. Paris : Université de la Sorbonne nouvelle – Paris 3 (Travaux du SYLED-CLA^{2}T) : http://tal.univ-paris3.fr/mkAlign/mkAlignDOC.htm

Fung (P.) 2000 : « A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. » dans Véronis (J.), éd., *Parallel Text Processing: Alignment and use of translation corpora*. Dordrecht: Kluwer Academic Publishers, p. 219-236.

Lamalle (C.), Martinez (W.), Fleury (S.), Salem (A.), Kuncova (A.), Maisondieu (A.) 2003 : « Lexico3 – Outils de statistique textuelle », *manuel d'utilisation*. Paris : Université de la

Journée d'étude: Traitements Automatisés des Discours politiques.

Objets nouveaux, nouvelles méthodes.

Sorbonne nouvelle – Paris 3 (Travaux du SYLED-CLA^{2}T) : http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW

Lamalle (C.), Salem (A.) 2002 : « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels », dans Morin (A.) et Sébillot (P.), éd., actes des 6es Journées internationales d'Analyse statistique des Données Textuelles, Saint-Malo, 13-14 mars 2002, INRIA, p. 403-412.

Zimina (M.), 2004 : « Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles », dans Purnelle (G.), Fairon (C.) et Dister (A.), éd., *Le poids des mots*, actes des 7es Journées internationales d'Analyse statistique des Données Textuelles, Louvain-la-Neuve, 10-12 mars 2004, Presses Universitaires de Louvain, p. 1195-1202.

PLAN :

1. Approche textométrique de corpus multilingues : (MZ)

- **a) choix d'unités**
- **b) segmentation parallèle**
- **c) variations contextuelles des traductions**

2. Exploration textométrique des traductions à l'aide de la topographie textuelle (MZ)

3. Cartographie de correspondances : *Lexico3* & *mkAlign* (SF)