Néologie sémantique : modélisation, expérimentations automatiques multilingues dans le cadre de Néoveille

mots-clés : néologie sémantique, cycle de vie des lexies, profil combinatoire, profil distributionnel, traitement automatique des langues

Cette présentation proposera un modèle des changements lexicaux qui inclut la néologie comme l'un de ses aspects. Ce modèle s'appuie sur l'hypothèse distributionnelle qui énonce que, pour décrire les langues, il ne faut prendre appui que sur les matérialisations écrites ou orales des langues, et ne pas recourir à des hypothèses sur la pensée sous-jacente ou sur la relation de référence qui lie le langage au monde (Harris, 1954). On reformulera ainsi les hypothèses harrisiennes :

Hypothèse 1 : les unités linguistiques sont repérables par leur répétition en corpus.

La première hypothèse donne la clé d'identification des unités linguistiques. La reprise de ces hypothèses par les statistiques lexicales, les linguistiques de corpus, et les grammaires de construction, a permis de revisiter la notion d'unité lexicale, et d'étendre la notion à celle de construction et de forme-sens (Goldberg, 2013).

Hypothèse 2 : leur usage-sens est repérable au moyen de la répétition des contextes dans lesquels ils sont pris.

La seconde hypothèse fonde une étude du sens basée sur la combinatoire autour des lexies-cibles. (Ramish, 2015) montre que la combinaison entre les schémas syntaxiques de syntagmes nominaux les plus fréquents dans une langue donnée et un calcul de répétition de mots graphiques permet de récupérer un grand nombre de locutions. (Kilgarriff, 2004 ; Gries, 2010) déduisent, sur cette base, les structures argumentales les plus fréquentes des verbes, qu'ils appellent *Word Sketches ou Behavioral profile*. Nous détaillerons ces notions pour aboutir à la notion de *profil combinatoire*.

Hypothèse 3 : deux unités lexicales partageant un grand nombre de contextes sont dans une relation de similarité sémantique.

Cette hypothèse est celle qui fonde une étude sémantique sur des bases distributionnelles, et a donné lieu à de nombreux travaux en TAL (Baroni et Lenci, 2010 ; Mikolov et al., 2013), montrant qu'il est possible, en se basant sur une distribution similaire, d'identifier des lexies en relation de *similarité sémantique*. Nous appellerons cette approche *profil distributionnel*.

Ces trois éléments formels du sens doivent également être combinés avec d'autres paramètres, liés à la variation (Coseriu, 2006; Koch et Oesterreicher, 2011; Gadet, 2003), pour délimiter la portée d'une association forme-sens : notamment, l'inscription diastratique, diatopique et diaphasique des lexies, puisqu'une lexie peut appartenir à un groupe social et géographique défini, être propre à un individu, ou bien valoir pour l'ensemble de la communauté linguistique.

Nous montrerons, au travers d'une étude multilingue (français, portugais et italien) sur gros corpus dynamique contemporain, issu de la plateforme Néoveille, en nous concentrant sur une centaine de noms, que ces quatre aspects peuvent être opérationnalisés pour détecter automatiquement des changements de fréquence, de profil combinatoire, de profil distributionnel et de diastratie. Ces changements *formels* sont autant de signaux de changements sémantiques.

Références :

Baroni M., Lenci A. (2010) Distributional Memory : A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36-4 (2010), 50

Cartier, E. (2016) Neoveille, système de repérage et de suivi des néologismes en sept langues. *Neologica : revue internationale de la néologie* , (10).

Cartier, E. (2017) Neoveille, a Web Platform for Neologism Tracking. In *Proceedings of European*

Chapter of the Association for Computational Linguistics 2017, Valencia, 3-7 avril 2017 .

Coseriu E. (2006) *Lenguaje y discurso* (Eunsa, 2006).

Gadet F. (2003) *La variation sociale en français,* Paris, Ophrys, Coll. « L'essentiel ».

Goldberg A. (2013) Constructionist Approaches. *The Oxford Handbook of Construction Grammar*, Edited by Thomas Hoffmann and Graeme Trousdale, Oxford University Press.

Gries, S. T. (2010) Behavioral pro files: A fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon* , 5(3):323–346.

Harris Z. S. (1954). Distributional structure. *Word*, 10(23), 146–162. [trad. 1970 : La structure distributionnelle", *Langages*, No.20, 14-34. Paris]

Kilgarriff A. et al. (2004). The Sketch Engine. *Proceedings of Euralex*, p. 105–116, Lorient.

Koch P. et Oesterreicher W. (2011). *Gesprochene Sprache in der Romania. Französisch, Italienisch, Spanisch*. Berlin/New York: De Gruyter. (=Romanistische Arbeitshefte, 31 ; 1ère éd. 1990)

Ramisch C. (2015) Multiword Expressions Acquisition: A Generic and Open Framework. *Theory and Applications of Natural Language Processing* series XIV, Springer, ISBN 978-3-319-09206-5, 230 p., 2015.