

---

# Collocations scientifiques transdisciplinaires : de l'extraction à la modélisation

Agnès Tutin\*<sup>1</sup>, Sylvain Hatier\*<sup>1</sup>, and Thi Thu Hoai Tran\*<sup>2</sup>

<sup>1</sup>Université Grenoble Alpes – Laboratoire LIDILEM EA 609 – France

<sup>2</sup>Université d'Artois – laboratoire grammatica – France

## Résumé

Cette proposition de communication s'inscrit dans le domaine de la combinatoire lexicale et du Français pour Objectifs Universitaires (FOU).

Les discours scientifiques regorgent d'expressions préfabriquées, de façons de dire spécifiques qui permettent de considérer la communauté des chercheurs comme une communauté de discours, au sens de Swales (1990). Parmi ces expressions, au-delà de la terminologie propre aux disciplines, on observe des associations compositionnelles, généralement binaires, qui sont transversales aux domaines et renvoient à la démarche scientifique, au raisonnement, à la structuration textuelle du discours : *débat scientifique*, *contraster fortement*, *enjeu principal*, *opérer une distinction*. Ces collocations du "lexique scientifique transdisciplinaire" (Drouin 2007 ; Tutin 2007 ; Hatier 2016a ; Jacques & Tutin 2018), typiques de ce genre, apparaissent centrales. Dans plusieurs perspectives applicatives (enseignement du FOU et de l'écrit scientifique, étude de l'épistémologie des disciplines, applications du traitement automatique des langues), il apparaît pertinent de recenser et de modéliser ces collocations.

Nous présenterons dans cette communication une ressource de ces collocations transdisciplinaires des écrits scientifiques de sciences humaines, élaborée dans le cadre du projet ANR TermITH[1]. Cette base lexicale contient à peu près 1500 collocations transdisciplinaires, constituée à partir de la base des mots simples du Lexique Scientifique Transdisciplinaire de Hatier et collègues (Hatier 2016b). Le repérage des expressions a été effectué semi-automatiquement à partir d'un corpus d'articles de sciences humaines analysé automatiquement au plan syntaxique. L'extraction des collocations a été réalisée à partir des bases de la collocation, l'élément stable de ces expressions (Hausmann & Blumenthal 2006 ; Tutin 2013), en exploitant un outil développé par O. Kraif, le Lexicoscope (Kraif 2016), recourant à un ensemble de relations syntaxiques et de mesures statistiques classiques pour l'extraction des collocations. Ces expressions ont ensuite été filtrées et désambiguïsées, puis accompagnées d'exemples dans les écrits scientifiques et pour certaines d'entre elles, d'un codage de Fonctions Lexicales (Mel'čuk 2013)[2]. Dans cette communication, nous exposerons quelques-unes des difficultés rencontrées pour effectuer ces traitements et présenterons quelques applications didactiques envisageables pour l'enseignement de la rédaction scientifique.

Drouin, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, 12(2), 45-64.

Hatier, S. (2016a). *Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche outillée sur un corpus d'articles de recherche en SHS* (Thèse de doctorat,

---

\*Intervenant

Université Grenoble Alpes).

Hatier, S., Augustyn, M., Yan, R., Tran, T. T. H., Tutin, A., & Jacques, M. - P. (2016b). French cross-disciplinary scientific lexicon: extraction and linguistic analysis. *Proceedings of the XVII EURALEX International congress Lexicography & Linguistic diversity* (p. 355–365). Présenté à XVII EURALEX International congress, Tbilisi: Ivane Javakhishvili Tbilisi State University.

Hausmann Franz Josef & Blumenthal Peter (2006). Presentation : collocations, corpus, dictionnaires, *Langue française*

2006/2 (no 150), 3-13.

Jacques, M. P., & Tutin, A.(eds) (2018). *Lexique transversal et formules discursives des sciences humaines*. Londres, ISTE Editions.

Kraif, O. (2016). Le lexicoscope: un outil d'extraction des séquences phraséologiques basé sur des corpus arborés. *Cahiers de lexicologie*, (108), 91-106.

Mel'čuk, I. (2013). Tout ce que nous voulions savoir sur les phrasèmes, mais.... *Cahiers de lexicologie*, 102, 129-150.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Tutin, A. (2007). Autour du lexique et de la phraséologie des écrits scientifiques. *Revue française de linguistique appliquée*, 12(2), 5-14.

<http://www.atilf.fr/ressources/termith/>

Une première version de la base, en cours de réalisation, est consultable en ligne sur : [http://sylvain.hatier.me/unites\\_lst/collocations.html](http://sylvain.hatier.me/unites_lst/collocations.html)