Étude sur corpus et annotation du lexique scientifique transdisciplinaire

Sylvain Hatier[1], *Evelyne Jacquey*[2], *Agnès Tutin*[1], *Laurence Kister*[2], *Mario Marcon*[2]
1 – LIDILEM – Linguistique et Didactique des Langues Etrangères et Maternelles
2 – ATILF – Analyse et Traitement Informatique de la Langue Française

Mots-clés : linguistique de corpus – écrit scientifique – lexique scientifique transdisciplinaire – annotation – sémantique

Nous présentons dans cette étude un travail de linguistique de corpus centré sur un lexique associé au genre de l'écrit scientifique, le lexique scientifique transdisciplinaire (désormais LST). Le LST intègre les mots qui sont partagés par les disciplines scientifiques et qui sont propres au genre du discours scientifique (Drouin 2007; Tutin 2007; Hatier 2016). Il s'agit de mots simples comme *hypothèse, analyser, épistémologique* ou d'expressions comme *dans un premier temps, faire référence, point de vue*.

L'annotation consiste à isoler les mots de ce lexique dans les textes en les différenciant des termes des domaines de spécialité. Dans cette expérimentation, nous avons opéré sur un ensemble de cinq textes du genre académique une annotation des occurrences des unités lexicales du LST en contexte, en procédant simultanément à la désambiguïsation sémantique, i.e en associant à chaque occurrence une acception précise de la ressource du LST élaborée par Hatier & al (2016). Cette ressource, organisée en classes et sous-classes sémantiques, a plus de 1850 entrées et recouvre les quatre principales catégories : verbe, nom, adverbe et adjectif. Ce sont sur ces catégories que nous effectuons ce travail d'annotation. Par exemple, le mot *conclusion* a deux acceptions : conclusion_1 renvoie à la sous-classe des sections textuelles, dans la classe 'communication_support' alors que conclusion_2 est une 'implication' dans la classe générique des 'relations'.

Les candidats-LST sont projetés automatiquement et par la suite validés ou invalidés par quatre annotateurs dont l'accord est ensuite calculé. La projection du LST est effectuée à l'aide du logiciel Nooj (Silbertzein, 2015) puis une vérification est effectuée à l'aide d'un éditeur XML pour désambiguïser les différentes annotations. Les annotateurs effectuant cette tâche sont familiarisés avec le concept de l'écrit scientifique et plus précisément du LST[1]. L'annotation humaine permet de valider des mots potentiellement membres du LST mais sujets à plusieurs types d'ambiguïtés, difficiles à traiter automatiquement et pour lesquelles une évaluation humaine experte est nécessaire. Ces ambiguïtés sont de plusieurs ordres :

1 Ils ont ainsi tous participé au projet TermITH, ANR-12-CORD-0029, dont le sujet était l'indexation en termes de textes en sciences humaines et sociales.

- au niveau des catégories syntaxiques : au niveau de la forme, *analyse* peut être nom ou verbe, *alternative* nom ou adjectif;
- au niveau de la segmentation, surtout pour les expressions polylexicales. Ainsi, *par exemple* est une entrée adverbiale polylexicale. Le nom *exemple,* autre entrée de la ressource ne doit alors pas être annoté seul;
- au niveau sémantique : certaines formes potentiellement LST recouvrent des acceptions non transdisciplinaires (pouvant appartenir à la terminologie, à la langue générale). Par exemple, *sujet,* entrée LST au sens de 'thème' peut se réaliser comme terme linguistique. De plus, certaines unités lexicales ont plusieurs acceptions transdisciplinaires dans la ressource du LST,. Ainsi, le verbe *présenter* a notamment une première acception 'faire part de/exposer' réalisée dans l'exemple 1 mais recouvre également une autre acception 'avoir quelque chose/comporter' réalisée dans l'exemple 2.

1. *Dans le tableau suivant, nous présentons les résultats de…*

2. *Les opinions des répondants présentent deux caractéristiques majeures…*

La tâche d'annotation du LST permet alors d'étudier en contexte les relations entre ce lexique et les autres présents dans le genre académique, en particulier la terminologie (Jacquey *et al.*, à paraître). En effet, LST et terminologie partagent de multiples contextes, allant de la simple cooccurrence à l'intégration d'un mot du LST comme tête de terme complexe (*analyse syntaxique*). L'intrication LST-terminologie apparaît également dans les cas où un élément du LST se spécialise dans une discipline pour y prendre une valeur terminologique (tel *corpus* en linguistique). La présence de traits sémantiques dans la ressource nous permet alors de déterminer quelles classes sont les plus à mêmes de se « terminologiser », de devenir têtes de termes complexes, etc.

Le calcul de l'accord inter-annotateurs nous permet ensuite d'évaluer quelles dimensions de l'annotation posent le plus de problèmes et à l'inverse quels choix sont les plus partagés. Plus précisément, nous verrons que si la question de l'appartenance d'un mot au LST engendre de manière générale peu de divergence chez les annotateurs, le choix de l'acception en contexte se révèle plus problématique.

Ce travail de désambiguïsation lexicale, portant sur un lexique associé à un genre particulier, l'écrit scientifique, nous permet d'envisager diverses applications :

- au niveau description linguistique du LST, et également au niveau didactique de l'écrit scientifique, l'étude du LST en contexte, ainsi que la création d'un corpus de textes annotés permettent de disposer d'exemples attestés pour un grand nombre d'acceptions de la ressource du LST;
- au niveau de l'étude des interactions entre terminologie et LST, et notamment de l'utilisation de certains éléments, voire certaines classes du LST dans la construction de termes complexes. De telles observations peuvent améliorer l'identification automatique des termes.

Bibliographie

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

Hatier, S., Augustyn, M., Yan, R., Tran, T. T. H., Tutin, A., & Jacques, M.-P. (2016). *French cross-disciplinary scientific lexicon: extraction and linguistic analysis*. In G. Meladze (Éd.), Proceedings of the XVII EURALEX International congress (p. 355-365). Tbilisi

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

Jacquey, E., Kister, L. Marcon, M., Barreaux, S. (à paraître), Lexique scientifique transdisciplinaire, terminologies et langues de spécialité en SHS, in Jacques, M.P. & Tutin, A. (eds)., *Lexique transversal et formules discursives des sciences humaines*. ISTE, Londres.

Silberztein, M. (2015). *La formalisation des langues: l'approche NooJ*. ISTE, Londres.

Tutin, A. (2007c). *Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques*. In Actes de TALN 2007. Toulouse.