Élaboration semi-automatique d'une ressource de patrons verbaux

Sylvain Hatier[1] Rui Yan[1]
(1) LIDILEM – EA 609 – Université Grenoble Alpes
sylvain.hatier@univ-grenoble-alpes.fr, rui.yan@univ-grenoble-alpes.fr

RÉSUMÉ

Nous présentons dans cet article l'élaboration semi-automatique d'une ressource de patrons verbaux. Nous nous intéressons aux verbes du Lexique Scientifique Transdisciplinaire, lexique essentiel dans l'argumentation et l'organisation du discours dans les écrits scientifiques. Ce travail se base sur un corpus arboré d'articles scientifiques à partir duquel sont extraits les cadres de sous-catégorisation. Ces cadres sont alors manuellement regroupés en patrons, sur le modèle CPA, puis associés à une acception définie pour chaque verbe. La ressource résultante, élaborée dans un but d'aide à la rédaction et à la compréhension scientifique, répertorie les constructions verbales les plus fréquentes dans ce genre. L'utilisateur a ainsi accès, pour chaque acception verbale, aux cooccurrents et constructions syntaxiques préférentiels.

ABSTRACT

Semi-automatic elaboration of a verbal patterns dictionary

This study focuses on the semi-automatic elaboration of verbal patterns resource. We explore more specifically Cross-Disciplinary Lexicon verbs, a lexicon which plays an essential role in reasoning and discourse organizing in academic writing. This work is based upon a parsed corpora of French scientific articles, from which we automatically extracted subcategorization frames. These frames are then manually gathered, according to CPA model, and combined with a verbal meaning. The verbal patterns resource, designed for the didactic purpose of academic writing and reading, lists the most frequent verbal constructions in this genre. User thus has access, for each verbal meaning, to preferential syntactical collocates and constructions.

MOTS-CLES : écrit scientifique, sous-catégorisation verbale, linguistique de corpus, patron lexico-syntaxique.

KEYWORDS : academic writing, verbal subcategorization, corpus linguistics, lexico-syntactic pattern.

1 Cadres et objectifs

Ce travail d'élaboration d'une ressource lexicale dédiée aux constructions verbales dans l'écrit scientifique répond à plusieurs observations. Nous avons constaté dans de précédents travaux (Hatier, Yan, 2015) la difficulté des étudiants à maîtriser les patrons verbaux dans l'écrit scientifique, à la suite des travaux de Nesselhauf (2005), Hyland (2008) ou Granger, Paquot (2009). Ces difficultés sont d'ailleurs partagées par les apprenants et les étudiants francophones natifs (Hatier, Yan, à paraître). Notre objectif, d'ordre didactique, est ainsi de proposer une ressource des

patrons verbaux permettant de remédier à ces lacunes, afin d'améliorer la rédaction et la compréhension de textes scientifiques. La ressource est destinée aux enseignants, afin de le permettre la mise en place de séquences didactiques adaptées au genre de l'écrit scientifique. Nous adoptons pour cela une approche de linguistique de corpus, en faisant émerger du corpus d'analyse, représentatif du genre étudié, les constructions correspondant à l'usage.

Nous nous intéressons plus spécifiquement aux verbes du lexique scientifique transdisciplinaire (LST). Ce lexique méta-discursif et méta-scientifique est associé au genre de l'écrit scientifique. Nous le définissons, à la suite de Tutin (2007), comme un lexique renvoyant au discours sur les objets et les procédures scientifiques. Le LST (*hypothèse, analyser, qualifier,* etc.) est un lexique essentiel à maîtriser dans l'argumentation, l'organisation textuelle et l'expression de l'opinion.

1.1 Lexique Scientifique Transdisciplinaire

Dans une étude précédente (Hatier et al., 2014), nous avons procédé à l'extraction d'une liste de 1312 mots du LST : 274 adjectifs, 202 adverbes, 493 noms et 342 verbes. L'ensemble a ensuite été organisé en une classification sémantique en classes et sous-classes transcatégorielles, en se basant sur les propriétés sémantiques et lexico-syntaxiques des unités lexicales. Ces classes sont des ensembles de co-hyponymes, mots du LST, confirmant une définition et un test lexico-syntaxique définitoire de la classe. Ainsi, la sous-classe {document} de la classe {communication}, ayant notamment pour membre *article, ouvrage, texte*, a pour test d'appartenance : *Ce N présente*. En disposant d'une telle classification, nous pouvons intégrer dans l'analyse des constructions verbales le niveau sémantique, au niveau du verbe analysé et de ses arguments. De plus, cette typologie nous permet d'étudier les patrons verbaux au sein des classes sémantiques homogènes et de proposer une entrée onomasiologique dans notre ressource adaptée pour l'aide à la rédaction scientifique. L'identification des acceptions verbales du LST a été effectuée en prenant comme référence la ressource *Les Verbes Français*[1] (LVF, cf Dubois, Dubois-Charlier, 1997).

1.2 Patrons verbaux dans l'écrit scientifique

Nous nous intéressons à l'étude des patrons lexico-syntaxiques des verbes du LST, selon une approche contextualiste (Sinclair, 1991 ; Hunston, Francis, 2000 ; Hanks, 2008). Un patron consiste en « une structure syntaxique intégrant des collocations privilégiées » (Hanks, 2013 : 92, traduit par l'auteur). Il est défini dans le cadre du modèle Corpus Pattern Analysis (CPA) de Hanks (*ibid.*) et se caractérise par 1) l'association entre le sens et l'usage réel du mot 2) l'étiquetage sémantique au niveau des arguments. Un des patrons du verbe *to execute* est ainsi représenté comme suit :

> Pattern[2]: [[Human | Institution]] execute [[Plan | Command | Activity]]
> Implicature : [[Human | Institution]] does work in order to put [[Plan | Command | Activity]] into effect.
> Ex : The **student** has the opportunity to formulate and **execute a search**. (L'étudiant a la possibilité de concevoir et d'exécuter une recherche)

Dans ce patron, le verbe *to execute* sélectionne de préférence des sujets noms renvoyant à un humain / une institution et des objets noms renvoyant à un plan / un ordre / une activité. Le sens

1 Consultable en ligne : http://rali.iro.umontreal.ca/rali/? q=fr/node/1237/ [consulté le 25/04/2016]
2 http://www.pdev.org.uk/#browse ? q=; f=C [consulté le 25/04/2016]

associé à ce patron est indiqué dans l'*implicature*, au travers d'une paraphrase. Le patron permet ainsi de faire le lien entre sens mobilisé et constructions verbales. Les verbes étant polysémiques, y compris dans le contexte de l'écrit scientifique, l'ambiguïté peut être ainsi levée en associant les valeurs sémantiques des arguments aux structures syntaxiques. Le modèle CPA accorde un rôle primordial aux contextes et aux usages en corpus, ce qui nous paraît adapté à notre perspective didactique.

2 Méthodologie

L'élaboration de la ressource de patrons verbaux s'effectue en trois phases principales. Dans un premier temps, nous constituons et annotons le corpus d'analyse. De ce corpus sont ensuite extraits automatiquement les cadres de sous-catégorisation des verbes du LST. Ces cadres sont alors regroupés sous forme de patrons lexico-syntaxiques modélisés manuellement.

2.1 Corpus d'analyse

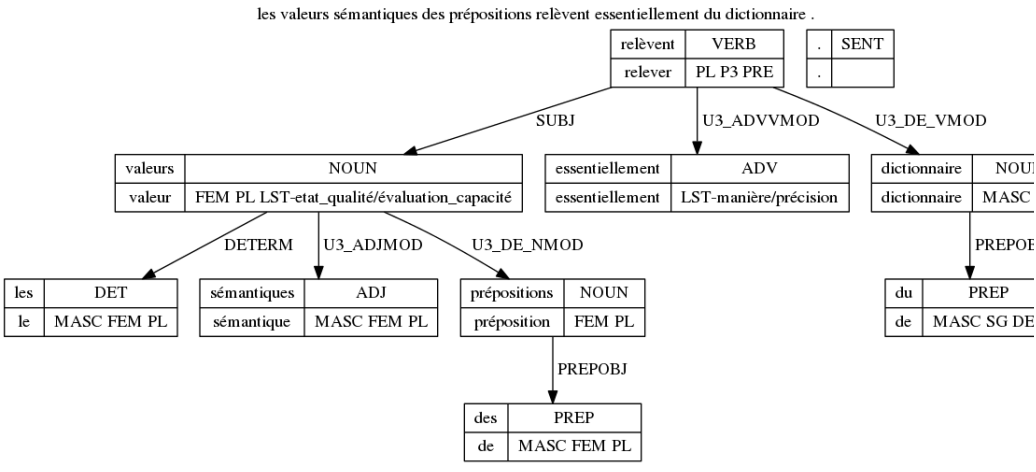
Le corpus d'analyse, de 5 millions de mots, issu du projet scientext, est composé de 500 articles de recherche en français (Tran, 2014), de 10 disciplines des sciences humaines et sociales. À l'aide de l'analyseur en dépendances XIP[3] (Aït-mokhtar et al., 2002), il a ensuite été annoté en lemmes, traits morpho-syntaxiques et relations syntaxiques. Les traits de classes et sous-classes sémantiques ont ensuite été projetés dans le corpus afin d'intégrer ces informations sémantiques dans les patrons verbaux extraits. Nous avons également procédé à un post-traitement du corpus en définissant des grammaires locales en vue de l'amélioration de l'annotation en dépendances. Nous souhaitons ainsi améliorer l'analyse à travers notamment la propagation du sujet pour les structures avec verbes de contrôle, ou lorsque le sujet est en situation de coréférence[4] ou de coordination.

2.2 Extraction des cadres de sous-catégorisation

À l'instar de Messiant et al. (2010), nous procédons alors à l'extraction des cadres de sous-catégorisation « sans a priori, pour faire émerger du corpus [ceux] correspondant à l'usage ». Nous commençons par extraire autant de cadres que d'occurrences pour chaque verbe. Les cadres sont ensuite regroupés, lors d'une phase de factorisation, à la manière de Kupsc (2007). Certaines relations de dépendance sont rassemblées sous une même étiquette (telles les relations avec un semi-modal : *devoir, sembler, aller*). Chaque occurrence d'un verbe du LST est ainsi représentée en termes d'ensemble de relations (impliquant le verbe en tant que gouverneur ou dépendant) pour aboutir aux cadres de sous-catégorisation, tel que l'illustre la figure ci-après :

3 Nous remercions Claude Roux pour ces précieux conseils dans le paramétrage et l'utilisation de XIP

4 Considérons la phrase « *Nous analysons la préposition qui introduit le nom* ». Si l'analyseur syntaxique définit une relation de coréférence entre *qui* et *préposition* et une relation sujet entre *introduit* et *qui*, alors nous propageons la relation sujet entre *introduit* et *préposition*.



Étant donné que nous ne retenons pas les relations avec les adverbes, le verbe *relever* est, dans cet exemple, impliqué dans deux relations :

- gouverneur dans la relation “sujet” avec pour dépendant *valeurs* (nom du LST, dont la classe et la sous-classe sont renseignées dans les traits) ;
- gouverneur dans la relation 'complément prépositionnel' avec pour dépendant *dictionnaire.*

La phrase illustrée ci-dessus a alors pour cadre de sous-catégorisation :

(SUBJ {document}) relever_VERB (COMPLEMENT PREP.).

Le verbe *relever*, dans cette construction transitive indirecte, prend pour sujet un nom du LST de la classe sémantique {document] et pour complément un syntagme nominal introduit par une préposition.

Pour chaque cadre, sont indiqués les lemmes et les classes sémantiques les plus fréquents en tant qu’argument. Le tableau suivant présente deux des cadres de sous-catégorisation les plus fréquents dans notre corpus.

Cadre – *Exemple*	Fréquence	Sujet	Objet
(Subj – hum) montrer (Complétive QUE) *Les résultats montrent que les effets sont significatifs*	762	résultat, il, analyse, étude, travai	-
(Subj – hum) constituer (Obj – hum) *Ils constituaient un groupe solidaire et homogène*	616	il, pratique, type, celui-ci, espace	élément, enjeu, point, forme

TABLE 1: Exemples de cadres de sous-catégorisation

Ces cadres résultent d'un premier regroupement, opéré automatiquement, d'occurrences d'un même verbe et permettent le passage à la phase de modélisation manuelle des patrons, étape essentielle pour le traitement de la polysémie.

2.3 Modélisation des patrons verbaux

Afin de modéliser les patrons verbaux, nous nous basons dans un premier temps sur les cadres de sous-catégorisations extraits pour repérer les acceptions verbales spécifiques à l'écrit scientifique, utilisées dans l'argumentation et la présentation de l'activité scientifique. Nous mettons ainsi en correspondance un cadre avec un sens défini, présent dans notre classification du LST. Dans un second temps, l'observation du corpus et l'analyse des cadres de sous-catégorisations nous permettent d'aboutir aux patrons verbaux du LST.

Nous détaillons dans la section suivante ces étapes d'analyse et, en partant d'exemples issus de notre corpus, illustrons la modélisation des patrons.

2.3.1 Repérage des acceptions

Le repérage des acceptions constitue l'étape de base de notre analyse étant donné la polysémie des verbes. Prenons par exemple le verbe *postuler,* dont un des cadres extraits est :
(SUBJ +hum) postuler_VERB(COMPLETIVE_QUE)

Ce cadre s'actualise dans l'exemple ci-dessous :

Ex : *Dans cet article,* ***nous postulons que*** *l'observation des phénomènes non-verbaux telle que nous la concevons est susceptible d'enrichir une approche didactique.*

Le sens mobilisé correspond à l'entrée 2 du verbe *postuler* dans le LVF, dont la définition est « supposer » (ex. : *On postule la bonne foi des joueurs/que l'accord est possible*.). Cette acception apparaît dans une construction avec un sujet humain et soit un objet chose soit une complétive.

2.3.2 Exemple de patrons lexico-syntaxiques

Après avoir relié cadres de sous-catégorisation et acception, nous procédons à la modélisation des patrons verbaux. Nous détaillons dans cette partie des exemples de patrons pour des verbes renvoyant à une ou plusieurs acceptions dans le corpus d'analyse.

Lorsqu'un verbe est monosémique dans le corpus, ce sens peut correspondre à plusieurs patrons. Par exemple, le verbe *constater*, appartenant à la sous-classe {analyse_info/constat}, prend pour sujet un nom humain et comme objet soit un complément nominal soit une complétive conjonctive. Ici, il prend le sens « remarquer » (entrée *constater 01* dans le LVF). Cette acception est réalisée à travers les deux patrons suivants :

1. [[humain=auteur, chercheur]] constate [complétive *que*]
   Ex : ***On constate que*** *le taux d'actualisation optimal est en général au-dessus* […].

2. [[humain=auteur, chercheur]] constate [[phénomène|relation]]
   Lexset [[phénomène|relation]] : *différence, effet, écart, inégalité*
   Ex : *Deuxièmement,* ***nous constatons trois problèmes*** *qui concernent uniquement* […].

Pour chaque cadre sont renseignées les informations sur les cooccurrences statistiquement significatives au niveau des actants. Ceci nous permet d'indiquer le(s) type(s) sémantique(s) approprié(s). Notons qu'il ne s'agit pas de représenter l'ensemble des cooccurrents possibles, mais de présenter un usage prototypique permettant de clarifier l'acception.

Dans les cas où un verbe renvoie à plusieurs acceptions, nous vérifions si le type de construction syntaxique permet de désambiguïser. Ainsi, le verbe *considérer* possède trois sens correspondant à trois constructions différentes :

1. [[humain=auteur, chercheur]] (peut) considérer [[entité abstraite|éventualité]] comme
   SENS : Quelqu'un donne son jugement par rapport à quelque chose
   EX : *On peut dire la même chose des niveaux d'action, c'est-à-dire des niveaux que* ***l'on considère comme*** *pertinents* […]

2. [[humain=auteur, chercheur]] (peut) considérer [complétive que]
   SENS : Quelqu'un se forme une opinion autour d'un fait scientifique
   EX : *Sur le plan symbolique,* ***on pourrait considérer que*** *la figure de professeur renvoie à une forme hiérarchique plus explicite* […]

3. Si/Lorsque [[humain=auteur, chercheur]] considère [[concept|éventualité]]
   SENS : Si/Lorsque quelqu'un donne son attention à quelque chose pour un examen attentif ou critique, introduisant ainsi un nouveau thème
   EX : ***Si l'on considère*** *par exemple la notion de passé,* […]

La polysémie peut se situer au niveau des arguments, l'attribution des types sémantiques joue ainsi un rôle important pour distinguer les différents sens. Par exemple, le verbe *caractériser* apparaît dans la construction transitive directe. Au niveau du sujet, nous pouvons distinguer deux types sémantiques distincts : [[humain]] (*nous, on,* etc.) et [[entité abstraite]] (*propriété, élément,* etc.). Le premier type est associé au sens « indiquer le trait de », tandis que le second active le sens « constituer le trait de », comme illustré dans les deux patrons suivants :

1. [[humain]] caractérise [[évènement|entité abstraite]]
   Ex : *Dans la proposition suivante,* ***nous caractérisons la probabilité*** *de contrôle optimale.*

2. [[entité abstraite 1=trait]] caractérise [[évènement|entité abstraite 2]]
   Ex : ***L'indécidabilité*** *qui* ***caractérise ces faits*** *finit alors par se conjuguer […].*

L'analyse des patrons permet ainsi de représenter les propriétés syntaxiques et sémantiques du verbe, dans un usage spécifique à l'écrit scientifique, assurant l'adaptation de la ressource de patrons à l'aide à la rédaction scientifique.

Au niveau de la ressource des patrons, une entrée correspond à une acception verbale qui peut être mobilisée par un ou plusieurs patrons, comme l'illustre l'entrée d'une acception du verbe *montrer* ci-dessous :

[[objet scientifique]] (sembler) montre [complétive que ou comment]
SENS : Quelque chose révèle ou atteste qu'un fait scientifique est juste et le met en évidence, ceci constitue un argument important de l'auteur.
LEXSET : [[objet scientifique]] : résultat, exemple, tableau, enquête, donnée, entretien
COMMENTAIRE : le gérondif est aussi fréquent

ROUTINES : *ce résultat tend à montrer que, cela montre que*
EX : *Le tableau V montre par ailleurs que l'apprentissage du solfège profite aux enfants issus des classes sociales favorisées quelle qu'en soit la durée alors que seul un apprentissage durable (d'au moins deux ans) a un effet sur les résultats des élèves issus des classes sociales non favorisées*. (scienceseducation.xml-s1945)

3 Conclusion

La combinaison de traitement automatique sur les cadres de sous-catégorisation et de modélisation manuelle de patrons lexico-syntaxiques nous a permis d'élaborer une ressource lexicale des patrons verbaux dans l'écrit scientifique. L'intégration de traits sémantiques au niveau des arguments permet de faciliter l'identification des sens mobilisés et des emplois adaptés. Nous avons pour perspective l'amélioration de la phase de factorisation afin d'alléger la tâche d'observation des cadres qui précède la modélisation des patrons. Ainsi, bien que ces regroupements réduisent de moitié les cadres à analyser (le verbe *considérer,* 2559 occurrences, passe de 1119 cadres à 569, le verbe *montrer,* 3114 occurrences, passe de 1137 cadres à 522), le nombre de cadre correspondant à une unique occurrence reste trop élevé (569 pour *considérer* contre 864 avant factorisation, 355 pour *montrer* contre 857 avant factorisation). Une autre amélioration se situe au niveau de l'identification des erreurs d'analyse syntaxique qui donnent lieu à des cadres dont un ou plusieurs arguments sont manquants. Ainsi, dans la phrase suivante, le verbe *correspondre* n'entre dans aucune relation selon l'annotation syntaxique : *à chaque niveau de segmentation institutionnalisé correspond un doyen*. Il résulte de ce genre d'erreur du bruit dans la liste de cadre (en intégrant un cadre tronqué) et du silence (en ne faisant pas correspondre ces occurrences avec le bon cadre).

En termes d'utilisation de la ressource, nous en distinguons deux principales. La première, d'ordre didactique, se situe dans le cadre de l'aide à la rédaction. La ressource peut faciliter alors l'encodage et/ou le décodage de constructions verbales typiques de l'écrit scientifique. Ainsi, au niveau du décodage, l'enseignant peut proposer à l'apprenant de repérer les différentes acceptions du verbe *considérer* pour un ensemble d'exemples sélectionnés (voir le tableau 2). L'enseignant choisit des exemples correspondant aux patrons sur lesquels il veut faire travailler les apprenants, pour que ceux-ci identifient les différentes acceptions mobilisées selon la construction employée.

Cela est vrai, également, si	l'on considère	l'immense importance qu'ont prise les ONG dans la vie du monde…	sociologie.xml-s10035
Patrick Charaudeau propose de	considérer	tout fait humoristique comme un acte de langage.	scinfo.xml-s10057
En effet, leurs pratiques d'information se sont révélées très disparates, certains de ces jeunes	considérant	le Net bien plus comme un moyen de communication et de jeu...	scinfo.xml-s10171
En ce qui concerne la dénomination des types de mobilité, nous	considérons	qu'il existe trois critères de distinction : la dimension spatiale (mobilité locale / régionale / nationale / internationale)… [4].	geo.xml-s11184
Premièrement, lorsqu'on	considère	le type d'interventions mises en place après le diagnostic, on constate qu'il s'agit principalement d'interventions centrées sur les conditions de travail.	psycho.xml-s592

TABLE 2: Concordancier du verbe *considérer*

À l'issue de ce type de séquence didactique, l'apprenant peut être amené à réinvestir ces connaissances en employant à son tour les patrons dans des exercices de rédaction scientifique.

La seconde utilisation que nous envisageons permettrait d'interroger un corpus analysé en dépendances afin d'identifier automatiquement la réalisation d'un patron (et de son acception correspondante) en fonction des éléments le constituant (un lemme verbal, un ensemble de relations syntaxiques, des éventuels traits sémantiques pour les arguments).

Remerciements

Nous souhaitons remercier la région Rhône-Alpes pour le financement de nos travaux de recherche ainsi que les partenaires du projet ANR-Contint Termith [5] pour leur collaboration dans nos expérimentations sur le LST.

Références

AÏT-MOKHTAR S., CHANOD J.-P., ROUX C. (2002). Robustness beyond shallowness : incremental deep parsing. *Natural Language Engineering*, *8* (2-3), 121–144.

DUBOIS, J., DUBOIS-CHARLIER, F. (1997) : *Les verbes français*. Larousse.

GRANGER S., PAQUOT M. (2009). Lexical Verbs in Academic Discourse : A Corpus-driven Study of Learner Use. In : Charles, Maggie/Hunston, Susan/Pecorari, Diane (éds.) : *Academic Writing : At the Interface of Corpus and Discourse*. London : Continuum International Publishing Group :193-214.

HANKS P., PUSTEJOVSKY J. (2005). A Pattern Dictionary for Natrual Language Processing. *Revue française de Langue Appliquée*10 (2), 63-82.

HANKS P. (2008). Lexical Patterns : from Hornby to Hunston and beyond. In E. Bernal et J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress.* Barcelona : IULA : 89-129.

HANKS P. (2013). *Lexical Analysis : Norms and Exploitations.* MIT Press.

HATIER S., TUTIN A., JACQUES M.-P., JACQUEY E., KISTER L. (2014). Catégorisation sémantique des noms simples du lexique scientifique transdisciplinaire. Présentation à *ACFAS, colloque 330 : Étude de lexiques à vocation particulière : approches théoriques, méthodologiques, pédagogiques et multidisciplinaires*, Montréal.

HATIER S., YAN R. (2015). Comparaison de constructions verbales entre un corpus d'apprenants et un corpus d'articles de recherche. Présentation à *8es Journées Internationales de Linguistique de Corpus (JLC2015)*, Orléans.

HATIER S., YAN R. (à paraître). Analyse contrastive des patrons verbaux dans l'écrit scientifique entre scripteurs étudiants et experts.

HUNSTON, S., FRANCIS, G. (2000). *Pattern Grammar : a corpus-driven approach to the lexcial grammar of English*. Amsterdam/Philaelphia : John Benjamins.

5 TermITH (Terminologie et Indexation de Textes en sciences Humaines) : ANR-12-CORD-0029 CONTINT. ATILF, INIST, LIDILEM, LINA, INRIA NGE et Saclay. http://www.atilf.fr/ressources/termith/ (visité le 25 avril 2016)

HYLAND K. (2008). Academic clusters : text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, 18, (1) : 41-62.

JACQUES M.-P. (2011). Nous appelons X cet Y : X est-il un terme émergent ? In K. Kageura & P. Zweigenbaum (Éd.), *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence* (p. 31–37). Paris, France : INALCO.

KUPSC A. (2007). Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré. Présenté à *TALN* 2007.

MESSIANT C., GABOR K., POIBEAU T. (2010). Acquisition de connaissances lexicales à partir de corpus : la sous-catégorisation verbale en français. *Traitement automatique des langues*, *51*(1), 65–96.

NESSELHAUF N. (éd.) (2005) : *Collocation in a Learner Corpus*. Amsterdam / Philadelphia : John Benjamins Publishing Company.

PAQUOT M. (2010) : *Academic vocabulary in learner writing : From extraction to analysis*. Bloomsbury Publishing.

SINCLAIR J. (1991). *LexicalCorpus, concordance, collocation (Vol.1). : Norms and Exploitations*. Oxford University Press Oxford.

TRAN T. T. H. (2014). *Description de la phraséologie transdisciplinaire scientifique et réflexions didactiques pour l'enseignement à des étudiants non-natifs. Application aux marqueurs discursifs* (Thèse de doctorat). Université de Grenoble.

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