Jan GOES **Luis MENESES-LERIN**

Jean-Marc MANGIANTE **Françoise OLMO**

Carmen PINEIRA-TRESMONTANT

Jan GOES **Luis MENESES-LERIN**

Jean-Marc MANGIANTE **Françoise OLMO**

Carmen PINEIRA-TRESMONTANT

APPORTS ET LIMITES DES CORPUS NUMÉRIQUES EN ANALYSE DE DISCOURS ET DIDACTIQUE DES LANGUES DE SPÉCIALITÉ

Editura Universitaria
Craiova, 2019

Descrierea CIP a Bibliotecii Naţionale a României
Apports et limites des corpus numériques en analyse de discours et didactique des langues de spécialité / Jan Goes, Luis Meneses-Lerin, Jean-Marc Mangiante, ... - Craiova: Universitaria, 2019
Conţine bibliografie
ISBN 978-606-14-1550-2
I. Goes, Jan
II. Meneses-Lerin, Luis
III. Mangiante, Jean-Marc
81

Publié avec le concours de l′Equipe d′Accueil Grammatica, (EA 4521), Laboratoire de Recherche, Université d′Artois – 62000 Arras, France.

AVANT-PROPOS

Jan Goes
Jean-Marc Mangiante
Luis Meneses-Lerin
Françoise Olmo
Carmen Pineira-Tresmontant

Cet ouvrage franco-espagnol s'inscrit dans le programme de collaboration scientifique mis en place depuis 2012 par les laboratoires Grammatica et Textes et Cultures (axe CoTraLiS) de l'université d'Artois et le laboratoire GALE de l'université Polytechnique de Valence (Espagne) portant sur le croisement des notions d'analyse de discours spécialisés (professionnels, politiques et universitaires) et de didactique des langues sur objectif spécifique.

Les études menées par les différents partenaires sont échangées et valorisées au sein de colloques internationaux, franco-espagnols organisés tous les deux ans alternativement à Arras et à Valence (Espagne). Le premier colloque, organisé à Arras en septembre 2012, traitait de l'argumentation en langues de spécialité ; le deuxième, en septembre 2014 à Valence, s'interrogeait sur l'apport des outils numériques en analyse de discours et didactique du français.

Au croisement de ces deux thématiques complémentaires, l'apport des corpus numériques à l'analyse des discours et à la didactique des langues de spécialité constitue la thématique centrale de l'ouvrage.

L'analyse de discours politiques, professionnels et universitaires, exploités par les laboratoires partenaires, constitue une étape importante dans la conception de programmes de formation notamment en français spécialisé ou sur objectif spécifique. Les linguistes et didacticiens doivent décrire et analyser des phénomènes et mécanismes langagiers observables dans ces discours produits au sein de situations auxquelles ils n'ont pas toujours accès.

Le recours à des corpus numérisés constitue un moyen d'accéder à un matériel linguistique exploitable par le biais notamment d'outils numériques d'analyse de discours (*Tropes*, *Alceste*, *Trameur*, *Lexico3*, …) permettant de dégager des occurrences significatives, des caractéristiques discursives, des fréquences lexicales, des collocations et des figements particuliers…

Cet ouvrage se propose de réunir les articles des chercheurs travaillant sur différents corpus constitués ou en cours de constitution, afin d'en analyser les pratiques d'utilisation, les prolongements didactiques et les différents apports à l'analyse de discours.

Ces articles mettent aussi en évidence les limites des corpus numériques existants et s'interrogent sur leurs apports aux travaux de recherche en analyse de discours et didactique des langues, et sur leurs limites

En particulier la décontextualisation des corpus existants constitue leur lacune principale. Les discours collectés, traités, indexés, etc. dans les différents corpus sont dégagés du contexte énonciatif et communicatif, spécifique, dont ils sont issus, ce qui constitue un frein à leur analyse et à leur exploitation didactique. En effet, la connaissance des données contextuelles est nécessaire à la compréhension des phénomènes linguistiques à l'œuvre dans ces discours.

De plus les outils numériques d'analyse des discours réunis présentent des limites : ils ne prennent en compte ni les phénomènes discursifs comme la polysémie, l'implicite, l'argumentation, etc. ni les références culturelles, et n'exploitent pas suffisamment l'arrière-plan discursif des discours (interdiscursivité, intertextualité…).

L'ouvrage collectif permet également d'envisager, à la lumière notamment des travaux de recherche actuellement entrepris en analyse des discours spécialisés et didactique du FOS, des propositions d'amélioration des corpus en termes de contextualisation, d'outils complémentaires comme des référentiels de compétences en langue appliqués à des domaines professionnels ou universitaires.

Ainsi il interroge les différents apports des corpus numériques de discours spécialisés et leurs limites, ainsi que les améliorations ou outils complémentaires à leur adjoindre. Il se déclinera en 3 chapitres correspondant à 3 axes d'études :

Chapitre 1 : Corpus numériques utilisés en analyse de discours : politiques, professionnels, médiatiques, universitaires :

- Analyse des pratiques et usages.
- Apports des outils numériques.
- L'intérêt des méthodologies linguistiques (théories) dans le domaine des « humanités numériques »

Chapitre 2 : Exploitation des corpus numériques en didactique des langues spécialisées ou sur objectif spécifique :

- Quels usages pour quels publics d'apprenants ?
- Les outils numériques d'exploitation des corpus (Moodle, Scenari…) et leurs apports didactiques.
- Limites des corpus en didactique, quelles pistes d'amélioration ?

Chapitre 3 : Analyse critique des corpus numériques : prolongement de la démarche de construction des corpus et de traitement numérique :

- Comment améliorer le traitement numérique des corpus ?
- La re-contextualisation des corpus : quelle démarche, quels moyens ?
- La formation des linguistes et des didacticiens à l'exploitation des corpus numériques.

CHAPITRE 1

Corpus numériques utilisés en analyse de discours : politiques, professionnels, médiatiques, universitaires

ANALYSE DES COOCCURRENCES VERBE ET ADVERBE DANS L'ÉCRIT SCIENTIFIQUE ET RÉFLEXIONS DIDACTIQUES AVEC UNE APPROCHE INDUCTIVE

Rui Yan
Université Grenoble Alpes, *Lidilem*
Thi Thu Hoai Tran
Université d'Artois, *Grammatica*

Résumé
Les recherches de la linguistique de corpus portant sur les combinaisons des mots dans l'écrit académique/scientifique montrent l'existence de patrons préfabriqués spécifiques au domaine. Ces éléments préfabriqués constituent une grande difficulté pour les apprenants d'une langue étrangère. L'objectif de cette étude est d'analyser les cooccurrences du type [V+ADV] dans le corpus *Scientext* en se basant sur le modèle *Corpus Pattern Analysis*. La description linguistique permettra de proposer des activités didactiques adéquates pour l'enseignement de ces patrons avec une approche inductive.
Mots-clés : Approche inductive, cooccurrence, linguistique de corpus, lexique scientifique transdisciplinaire (LST)

Abstract
Research in corpus linguistics has demonstrated the existence of prefabricated word patterns in academic/scientific writings. These prefabricated elements constitute a great difficulty for foreign language learners. The objective of this study based on the Corpus Pattern Analysis is to analyze [V+ADV] co-occurrences in *Scientext* corpus. After a linguistic description, we propose pedagogic activities which should be suitable for teaching these patterns with an inductive approach.
Keywords: inductive approach, co-occurrence, corpus linguistics, academic vocabulary

1. Introduction

Sur le champ des « littéracies avancées (universitaires) », de nombreuses études montrent que les enjeux de la maitrise de l'écrit, chez les étudiants natifs comme chez les non-natifs, sont notamment liés au discours scientifique, par exemple, développer une argumentation à partir de données, établir une structure discursive cohérente, construire une « posture réflexive » (Rinck, 2011) au sujet des

savoirs d'une discipline (Delcambre et Lahanier-Reuter, 2012 ; Cavalla, 2014). L'objectif de ces études consiste à caractériser les besoins des étudiants afin de développer des ressources didactiques pour y répondre.

Inscrite dans la lignée de ces travaux, notre étude porte sur la pratique de l'écrit scientifique/académique, et ceci à travers un lexique transdisciplinaire « partagé par la communauté scientifique mis en œuvre dans la description et la présentation de l'activité scientifique » (Tutin, 2007 : 6). Ce « lexique transdisciplinaire des écrits scientifiques » (LST) (Tutin, 2007 ; Drouin, 2007), largement étudié en anglais dans le cadre de *l'English for academic purposes* (Howarth, 1996 ; Biber, 2006 ; Gledhill, 2000), se caractérise par des éléments linguistiques préfabriqués (ex. : *c'est-à-dire, point de vue, contredire une théorie*).

Dans une étude précédente (Tran et Yan, 2016), nous avons constaté que certains verbes scientifiques transdisciplinaires entrent en cooccurrence avec des adverbes dans l'écrit scientifique, par exemple, *ces travaux s'appuient principalement sur*, *les résultats présentés précédemment*, *ce type se distingue nettement du précédent*. Ce phénomène de cooccurrence nous intéresse plus particulièrement pour deux raisons principales. D'abord, il n'a pas – à notre connaissance – fait l'objet de travaux dans le domaine de l'analyse du discours scientifique. Ensuite, comme le montrent les exemples ci-dessus, les cooccurrences entre les verbes et les adverbes remplissent souvent des fonctions rhétoriques importantes : structurer les procédures d'analyse (*examiner d'abord*), indiquer un fait évident (*montrer clairement*), atténuer une concession (*sembler cependant*), etc. Sur le plan didactique, nous suivons plusieurs chercheurs (Larivière, 1998 ; Cavalla, 2015) en considérant que l'utilisation du verbe approprié dans l'écrit scientifique, accompagnant un cooccurrent (de nom ou d'adverbe) témoigne d'une connaissance d'un niveau de langue élevé ainsi que d'une bonne maîtrise des formules discursives. Dans le présent article, nous nous focalisons sur l'étude de deux verbes (*montrer* et *noter*) et leurs cooccurrents d'adverbes. Le choix de ces deux verbes s'explique par le fait qu'ils se caractérisent non seulement par leur fréquence élevée dans l'écrit scientifique mais aussi par leur rôle important lié à l'argumentation et à l'exposition des faits scientifiques. Par ailleurs, les premières observations sur des cooccurrences *verbes et adverbes* (ci-après [V+ADV]) dans le corpus montrent que ces verbes sont particulièrement intéressants en ce qu'ils apparaissent fréquemment avec des adverbes.

Ce travail de recherche se situe à l'intersection de plusieurs domaines : la didactique du Français sur Objectif Universitaire (Mangiante et Parpette, 2010) (désormais FOU), la linguistique de corpus et l'analyse du discours. L'objectif principal de cette étude est de développer les compétences rédactionnelles chez les étudiants non-natifs du français et de les sensibiliser au genre du discours scientifique à travers l'étude des cooccurrences du type [V+ADV]. Il est à noter que notre public ciblé se trouve en général au niveau B1-B2 du CECRL. Nous nous intéressons à l'enseignement/apprentissage des structures et routines langagières, des éléments qui posent souvent des difficultés aux étudiants allophones (Gonzalez-Rey, 2007). Nous établirons en premier lieu un bref bilan des études qui portent sur

l'enseignement/apprentissage de ces éléments en milieu universitaire. En second lieu, nous décrirons notre méthodologie de travail, ainsi que les corpus de travail. Nous présenterons ensuite les résultats de l'analyse linguistique des cooccurrences *verbes et adverbes*. Les recommandations didactiques dans la dernière partie peuvent être intégrées dans une formation d'appropriation des normes universitaires destinée aux étudiants allophones.

2. Littéracies universitaires et enseignement des structures langagières

Dans le domaine de l'*English for Academic purposes*, des travaux ont été menés sur les phénomènes phraséologiques dans les écrits d'apprenants qui montrent que les apprenants d'une langue étrangère peinent à manier ces éléments d'une manière fluide et experte (Granger et Paquot, 2009). Ces lacunes langagières se révèlent d'une part par des erreurs liées à une mauvaise utilisation d'un élément (ex. : *come to a conclusion*, *as a consequence*) (Nesselhauf, 2005 ; Narita & Sugiura, 2006), et d'autre part par des cas de sous-emploi (ex. : *it's possible that*) (Hyland, 2008) ou de suremploi (ex. : *as far as I am concerned*) (Granger, 2008). Néanmoins, faute de réels corpus d'apprenants de FLE en France, les études sur l'utilisation des phénomènes phraséologiques chez les apprenants de FLE sont beaucoup moins nombreuses, hormis quelques études récentes effectuées dans le cadre du projet *Scientext* : erreurs liées à l'utilisation des connecteurs argumentatifs (ex : *cependant, ainsi*) (Le, 2013) et des collocations verbales (ex : *mettre comme hypothèse*) (Cavalla, 2015), sous-utilisation des marqueurs de reformulation (ex : *en d'autres termes*) (Tran, 2014) et des constructions verbales (ex : *cela s'explique par*) (Hatier et Yan, 2017). À l'instar de ces travaux, nous souhaitons relever en premier lieu les difficultés des étudiants allophones qui se situent notamment au niveau des cooccurrences *verbes et adverbes* afin de mieux cerner leurs besoins langagiers et de proposer par la suite une approche pédagogique appropriée.

Dans cet article, nous insistons sur l'importance de l'utilisation du corpus en classe de langue comme un outil permettant à l'apprenant de découvrir les phénomènes linguistiques, ceci à travers l'approche inductive pour l'enseignement/apprentissage des cooccurrences [V+ADV]. Alors que l'approche déductive va des règles aux exemples, l'approche inductive suit une démarche inversée en mettant en avant la découverte. L'approche inductive est fréquemment utilisée pour enseigner les phénomènes grammaticaux (Chartrand, 1996). Depuis ces dernières années, elle suscite l'intérêt des chercheurs en linguistique de corpus (Chambers, 2010 ; Boulton et Tyne, 2014), notamment pour l'enseignement des langues. En effet, cette approche correspond aux principes du *Data-Driven Learning* (désormais DDL) (Johns, 1991). Les apprenants se trouvent au cœur de l'apprentissage et peuvent endosser un nouveau rôle, celui de « détectives » comme l'a décrit Johns. Dans le cadre de cet article, nous insistons tout particulièrement sur le lien entre l'approche inductive et l'introduction du corpus en classe de FLE ; nous nous appuyons donc sur un apprentissage fondé sur l'observation des phénomènes

langagiers pour développer le métalangage des apprenants. Nous souhaitons sensibiliser les apprenants au sens véhiculé par les verbes et à la relation sémantique qui lie les verbes et les adverbes. En outre, nous souhaitons également attirer l'attention des enseignants de français sur l'apport du corpus pour l'enseignement d'un élément grammatical qui est spécifique dans les écrits universitaires car celui-ci permet de refléter le point de vue de l'auteur.

3. Méthodologie

Notre étude est fondée sur deux corpus, à savoir un corpus d'apprenants et un corpus d'experts. Dans les parties suivantes, nous allons décrire en détail chaque corpus de travail.

3.1. Corpus d'apprenant : identification et interprétation des difficultés

Notre corpus d'apprenants se compose de mémoires de recherche d'étudiants sinophones et vietnamophones (environ 85 000 mots) relevant des disciplines en sciences humaines et sociales (désormais SHS) (économie, géographie, littérature, traduction, linguistique et psychologie). Il s'agit de mémoires de quatrième année de licence (correspondant à la Licence 3 en France), rédigés en langue française et corrigés partiellement par des enseignants locaux par manque de temps. Notre intérêt, qui se porte sur les disciplines en SHS, s'explique par la forte présence du métalangage dans ces écrits (Grossmann et Tutin, 2010). À l'aide de ce corpus, nous envisageons d'analyser les difficultés ainsi que les pratiques d'écriture des étudiants allophones qui rédigent un écrit universitaire.

Nous avons utilisé le logiciel *Anatext*[1], développé par Olivier Kraif du *Lidilem*, afin de chercher les verbes du LST fréquents et d'observer les adverbes cooccurrents employés par les apprenants. Les maladresses que nous avons relevées se trouvent essentiellement sur le plan sémantique comme le montrent les exemples ci-dessous. Les disciplines des étudiants sont indiquées entre crochets et les corrections sont proposées entre parenthèses.

(1) *Dans la première partie de cette œuvre, l'auteur décrit beaucoup le cauchemar de la veille au soir pour découvrir la source. [littérature] (Suggestion de correction[2] : *décrit à plusieurs reprises*).

(2) *Alors pour les produits on doit noter clairement que ce sont les produits touristiques fabriqués par les fins artisans de Cơ Tu. [géographie] (Suggestion de correction : *noter au passage*).

(3) *Le point commun le plus net qu'on peut observer immédiatement grâce à ces deux tableaux. [linguistique] (Suggestion de correction : *observer dans ces deux tableaux*)

[1] Accessible en ligne : http://phraseotext.univ-grenoble-alpes.fr/anaText/

[2] Les suggestions de correction s'appuient sur les résultats des requêtes sur le corpus d'experts (cf. 3.2)

Comme nous pouvons le constater, deux cas de figure se distinguent : soit les étudiants ont du mal à choisir l'adverbe cooccurrent approprié pour exprimer une idée (1) (2), soit l'emploi de l'adverbe nous semble inutile (3). À notre requête dans le corpus d'experts (cf. 3.2), *noter* est souvent utilisé avec les adverbes qui ajoutent une information complémentaire (*par ailleurs*, *au passage*, *également*) ou ceux qui marquent une énumération (*tout d'abord*, *enfin*). Le verbe *décrire* est fréquemment utilisé avec les adverbes qui renvoient à une partie textuelle (*précédemment*, *ci-dessous*, *ci-dessus*) ou les adverbes modaux (*brièvement*, *précisément*). Même si dans ces exemples l'emploi inapproprié des adverbes n'empêche pas la compréhension de la phrase, l'utilisation des routines langagières, en l'occurrence les cooccurrences [V+ADV], peut être considérée comme un des critères qui différencie les niveaux des étudiants.

Au vu de ces maladresses relevées, il nous semble important d'étudier les cooccurrences [V+ADV] dans les écrits des experts pour ensuite réfléchir à la manière d'aborder ces éléments phraséologiques en classe de FLE. Nous considérons ces écrits comme un modèle vers lequel les étudiants vont s'orienter pour construire leur écrit.

3.2. Corpus Scientext : repérage des cooccurrents

Notre corpus d'experts, d'une taille d'environ 5 000 000 de mots, est composé de 500 articles en SHS (linguistique, sciences de l'éducation, économie, psychologie, histoire, géographie, sociologie, anthropologie, sciences politiques, sciences de l'information et de la communication). Ce corpus a été constitué dans le cadre du projet TermITH[3]. Le corpus a été étiqueté morphosyntaxiquement et annoté semi-automatiquement, ce qui permet des extractions d'exemples authentiques. Il est intégré dans la plateforme Lexicoscope[4], un outil en ligne développé par Olivier Kraif du LIDILEM (Kraif et Diwersy, 2012). Lexicoscope permet d'extraire à la fois des concordances et des lexicogrammes, c'est-à-dire des tables de cooccurrences, ce qui est pratique pour étudier le profil combinatoire des unités phraséologiques (Blumenthal, 2005). Dans notre recherche, nous examinons les cooccurrences des [V+ADV] pour analyser leur comportement syntaxique et sémantique (Figure 1).

[3] TermITH Terminologie et Indexation de Textes en sciences humaines, le site du projet est consultable à l'adresse suivante : http://www.atilf.fr/ressources/termith/.

[4] Cet outil permet d'avoir, pour un mot donné, l'ensemble de ses co-occurrents les plus significatifs, il est accessible en ligne : http://phraseotext.univ-grenoble-alpes.fr/lexicoscope/index.php?

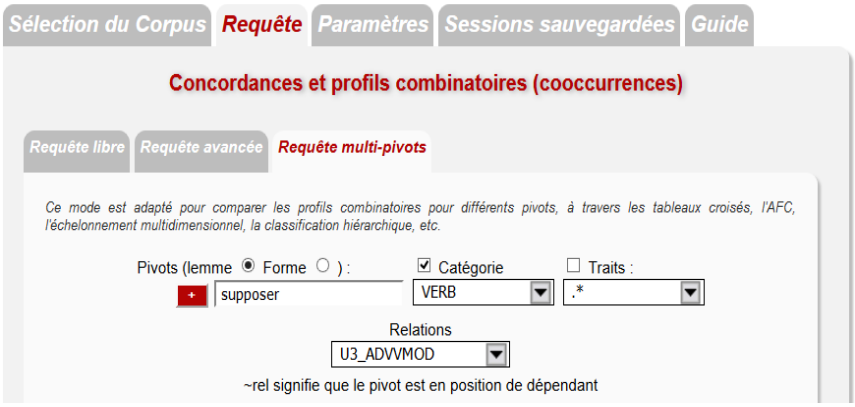

Figure 1 : Requête des cooccurrents des verbes+adverbes

À l'instar de Coxhead (2001) et Pecman (2004), nous nous appuyons sur des critères statistiques pour déterminer les éléments linguistiques à retenir. Ainsi, ceux-ci doivent apparaître 7 fois dans 3 disciplines au moins. Nous présenterons dans la partie suivante notre analyse linguistique qui se réalise en deux étapes : le classement sémantique et la modélisation avec le modèle *Corpus Pattern Analysis*.

4. Typologie des verbes et des adverbes

Afin d'analyser le phénomène de cooccurrence entre les verbes et les adverbes du LST dans l'écrit scientifique, il nous paraît important de constituer dans un premier temps la liste des verbes et des adverbes. Cette liste a été établie dans le cadre du projet TermITH (Hatier *et coll.*, 2016). Les unités lexicales (noms, verbes, adjectifs et adverbes) appartenant au LST ont été extraites sur des critères lexicométriques à l'instar de Coxhead (2000), Drouin (2007), Paquot (2010). À la suite de l'extraction des mots, nous avons obtenu 698 verbes et 757 adverbes (monolexicaux et polylexicaux).

Dans un deuxième temps, nous avons procédé au traitement sémantique des verbes et des adverbes. Concernant les verbes, nous nous sommes fondées sur la ressource de Dubois et Dubois-Charlier *les Verbes Français* (LVF[5]) (Dubois et Dubois-Charlier, 1997) pour repérer les acceptions verbales dans le corpus. Ce travail a abouti à un classement des verbes du LST basé sur leurs propriétés sémantiques et syntaxiques, par exemple, les verbes de constat (*constater*, *observer*, *noter*), les verbes d'hypothèse (*postuler*, *supposer*), les verbes de démonstration (*montrer*, *démontrer*), les verbes de relation (*amener*, *entrainer*, *lier*), etc. Il faut noter que les verbes sont polysémiques. Un verbe peut donc appartenir à plusieurs

[5] Le LVF couvre « 25 610 entrées verbales simples représentant 12 310 verbes différents dont 4 188 à plusieurs entrées ». Ce modèle propose une classification des verbes selon leurs propriétés sémantiques et syntaxiques et montre l'adéquation entre les constructions syntaxiques et l'interprétation sémantique par les classes sémantico-syntaxiques.