Studii de lingvistică 7, 2017, 7 - 12

Présentation

1. Corpus : applications et paradigmes de recherche

Ce volume de *Studii de lingvistică* comprend des contributions présentées lors de la Journée d'étude « Corpus et référentiels en ligne : un nouveau patrimoine langagier ? », ayant eu lieu le 9 octobre 2015 à l'Université d'Artois, Arras, auxquelles s'ajoutent des contributions retenues suite à un appel à contributions lancé par le comité de rédaction de la revue. Tous les articles réunis dans le dossier thématique de ce numéro sont spécifiquement consacrés aux aspects applicatifs des études sur corpus. Cette nouvelle collaboration de la revue *Studii de lingvistică* avec le laboratoire *Grammatica* (EA 4521) de l'Université d'Artois s'inscrit dans les axes de recherche du laboratoire, à savoir : l'axe linguistique, l'axe didactique et l'axe humanités numériques.

Actuellement, les Sciences Humaines et Sociales (SHS) sont confrontées à des documents numériques de façon permanente, au même titre que les autres disciplines et notamment les sciences dures. Cette réalité ouvre de nouvelles perspectives en SHS, s'appuyant sur des approches empiriques. C'est ainsi que l'exploitation des corpus numériques est devenue un « vivier » pour les travaux de recherche dont l'objet n'est pas seulement de proposer de nouvelles applications basées sur corpus mais aussi de remettre en question les présupposés théoriques et méthodologiques adoptés jusqu'à maintenant. Les travaux qui sont présentés dans ce numéro abordent les différentes problématiques liées aux corpus aussi bien en tant qu'« objets de recherche » qu'en tant qu'« objets exploités » pour des tâches précises.

Du point de vue de la recherche, plusieurs interrogations perdurent, qui concernent :

- la notion même de « corpus », qui soulève des débats;
- la démarche que les chercheurs doivent adopter pour appréhender le ou les corpus en tant qu'« objet(s) construit(s) » : démarche inductive, déductive, empirique, etc. ;
- l'impact en Sciences Humaines et Sociales des avancées théoriques, pratiques, descriptives, applicatives et expérimentales ;
- la constitution même du corpus : profilage, nettoyage, étiquetage, outillage, format(s), etc. ;
- la nature des données : (écrite, orale, iconique, visuelle, textuelle, thématique, unilingue, bilingue, multilingue, etc) et leur traitement.

Du point de vue de l'exploitation des corpus, on constate l'existence d'une multiplicité d'applications issues de la constitution

et de l'analyse de corpus, lesquelles exercent une influence directe en SHS, jusqu'au point de changer nos pratiques d'enseignement, d'apprentissage, d'assimilation ou de réflexion. Pour ne citer que quelques exemples d'applications en rapport avec les domaines abordés dans ce volume, nous pouvons mentionner l'exploitation de trois catégories de ressources :

- des ressources linguistiques présentant des unités lexicales très variées : collocations, unités phraséologiques, proverbes ; des étiquettes sémantiques, des dictionnaires électroniques, etc. ;
- des ressources didactiques permettant de créer une nouvelle génération d'outils pédagogiques : des exercices, des manuels, des dictionnaires, des dictionnaires de phraséologismes, des référentiels de compétences, etc. ;
- des ressources exploitables dans le domaine de la traduction et de la traductologie : des bases de données, des dictionnaires électroniques, des corpus parallèles, des corpus multilingues, des logiciels d'extraction basés sur corpus, des logiciels de traduction, etc.

En tirant profit de l'examen de ces trois catégories de données, les différents auteurs des contributions réunies dans ce volume ont pu faire émerger une problématique d'actualité, à savoir : l'articulation entre les aspects théoriques et pratiques de l'analyse de corpus, qui conduit à la mise en place de nouveaux paradigmes de recherche.

2. Présentation des contributions

Dans la contribution qui ouvre le volume, **Jean-Pierre Colson** présente les obstacles que la didactique de la phraséologie, et notamment l'apprentissage des langues, doit surmonter lors de l'exploitation de corpus. Parmi ces obstacles, trois sont mentionnés : premièrement, la faible fréquence des unités lexicales étudiées, deuxièmement, les contradictions issues des simples concordances dans l'analyse des unités phraséologiques, et, troisièmement, la difficulté à utiliser les outils existants pour manipuler de grandes « collections de textes » qui exigent souvent des connaissances en informatique et/ou en méthodes statistiques. En effet, les corpus web ne permettent pas aux apprenants de distinguer véritablement le sens littéral du sens idiomatique de certains blocs lexicaux et leur manipulation n'aide pas à l'identification de nombreux contextes où figurent les unités phraséologiques. À travers l'exemple de *tomber à l'eau*, l'auteur montre la difficulté que les étudiants natifs et non natifs peuvent éprouver face au sens littéral de cette expression : un natif dirait plutôt *tomber dans l'eau*, s'il s'agit du sens littéral, tandis qu'un non natif s'interrogerait sur la capacité de cette unité phraséologique à tolérer ou non une interprétation littérale. En tenant compte des problématiques liées

à la reconnaissance/apprentissage des unités phraséologiques, Jean-Pierre Colson propose un outil en ligne, de type expérimental (*IdiomSearch*), présentant la faculté d'être facilement manipulable et pouvant s'avérer être un excellent outil pour les linguistes, les didacticiens et les traducteurs, dans l'enseignement/ apprentissage de la phraséologie, quel que soit le domaine.

Mª Isabel González-Rey présente la méthode *PHRASÉOTEXT – Le Français idiomatique*, qui comprend un corpus d'expressions, toutes extraites d'un recueil d'œuvres littéraires, ayant permis la conception d'un dictionnaire phraséodidactique pour l'apprentissage et la didactisation des unités phraséologiques. À travers sa contribution, nous constatons que les outils du type dictionnaires, sites web, etc., ont conservé, et ce depuis les années 2000, une place importante dans le domaine de la lexicologie, qu'il s'agisse des théories de référence ou des approches didactiques. L'auteure souligne également que la « phraséodidactique » ou « didactique phraséologique » est une discipline en pleine expansion dans le domaine de la didactique des langues vivantes et qu'il ne s'agit pas d'une « jeune » discipline. Pour appuyer cette observation, Mª Isabel González-Rey renvoie au premier matériel phraséodidactique complet (manuel et dictionnaire), conçu en 1909 par celui que l'on peut considérer comme le précurseur de la phraséologie : Charles Bally. Ce premier matériel a été une source d'inspiration pour la conception du dictionnaire phraséodidactique proposé par González-Rey, qui en détaille tout au long de son article les objectifs et les spécificités. L'apport de ce nouveau dictionnaire phraséodidactique par rapport aux premiers (Bally, Galisson, Bárdosi) consiste dans le fait que, étant construit à partir d'un corpus de textes littéraires en français, il est placé au centre de la méthode, afin d'en extraire les unités phraséologiques autour desquelles s'articuleront ensuite les « séquences didactiques ». Un tel dictionnaire phraséodidactique n'aurait jamais vu le jour sans l'aide d'un corpus : les corpus permettent donc aujourd'hui de proposer de nouveaux types d'activités dans le domaine de l'enseignement des langues.

L'article de **Jan Goes** prend comme point de départ les théories conventionnelles sur la place de l'adjectif en français contemporain pour ensuite les confronter aux données relevées dans les corpus en ligne. Les théories en question ont été conçues à partir d'un petit nombre d'exemples non informatisés que l'on peut considérer comme formant un « corpus fermé ». Les règles générales concernant la place de l'adjectif avaient été établies sans avoir recours aux moyens techniques nécessaires à leur validation. L'auteur présente une synthèse pertinente des théories et des généralisations sur la place de l'adjectif et indique que le recours aux corpus permet d'aborder « une analyse modulaire et hiérarchisée du *déplacement* (vers l'antéposition ou la postposition) de l'adjectif, hors de sa place prépondérante ». Cet

article montre donc que les corpus s'avèrent être un excellent outil de recherche qui permet de remettre en question certains présupposés théoriques et d'avoir une meilleure compréhension du « phénomène adjectival ».

Dans sa contribution « Le corpus multilingue InterCorp : nouveaux paradigmes de recherche en linguistique contrastive et en traductologie », **Olga Nádvorníková** souligne l'importance de la linguistique de corpus, qui ouvre de nouveaux paradigmes d'investigation dans plusieurs disciplines, en particulier en linguistique contrastive et en traduction. Pour l'auteure, l'apport de la linguistique de corpus aux études effectuées dans ces deux domaines est double : d'une part, il est possible d'exploiter des « occurrences réelles » d'unités lexicales ou d'unités de traduction, et, d'autre part, de les analyser dans leurs « contextes authentiques ». À partir des résultats fournis par l'utilisation du corpus multilingue InterCorp, elle montre que les corpus parallèles ouvrent de nouvelles perspectives de recherche dans les domaines susmentionnés, en particulier l'analyse bi- et multidirectionnelle et l'étude des « universaux » de la traduction.

Christine da Silva Genest et **Caroline Masson** montrent que la valorisation et la diffusion de corpus oraux sont aussi importantes que leur constitution et leur exploitation. Leur travail s'inscrit dans le champ des études sur l'acquisition et les pathologies du langage et présente des résultats obtenus à partir de l'analyse d'un corpus réunissant des enregistrements audio et/ou vidéo. Les auteures exposent les principes et la méthodologie de la constitution d'un corpus, à savoir les conditions de recueil des données, le choix du support (audio et/ou vidéo), ainsi que les enjeux liés à l'utilisation des termes « données réelles » et « données écologiques ». L'analyse d'un corpus d'enregistrements vidéo permet de proposer non seulement un traitement plus riche des données, mais aussi une prise en compte de la dimension multimodale du langage, au travers de gestes, regards, contextes, etc. L'utilisation d'analyseurs automatiques (CLAN), ainsi que l'interprétation des résultats obtenus ouvrent des perspectives intéressantes dans l'étude des situations cliniques et de la prise en charge orthophonique et pathologique et montrent d'une façon plus générale l'importance de la linguistique de corpus dans ce domaine.

Ana Zisman aborde la problématique de la constitution des corpus d'un point de vue diachronique, dans le cadre d'une étude du comportement grammatical et pragmatique de quelques verbes parenthétiques en roumain. Cette étude s'inscrit dans le cadre d'un projet de thèse en cours, dont les objectifs ont imposé la constitution d'un « corpus numérisé » à partir d'un ensemble de documents au format papier classés par genres. L'auteure décrit la méthodologie adoptée, qui comprend deux volets : d'une part, la constitution du corpus DIY (*do it yourself*) et, d'autre part, l'exploitation des données

et en particulier l'étude des verbes parenthétiques. Puis, elle énumère certaines difficultés rencontrées au niveau méthodologique (critères de sélection, analyse des données, type de recherches) et technique (conversion de documents). Ana Zisman présente, enfin, les différents outils lui permettant de faire des recherches dans les documents numérisés.

David Brett exploite une collection d'articles de la rubrique « UK Crime » du journal britannique *The Guardian*, nommée « Guardian Crime Corpus » (GCC). L'auteur présente d'abord la méthodologie employée pour l'annotation du corpus (*Tree Tagger*). Ensuite, il présente la méthode d'extraction des données, qui est basée sur des scripts codés par lui-même, en perl, afin d'en retirer des « concgrams » (listes de collocations) qui sont intégrées ensuite dans un « réseau de collocations » moyennant le logiciel Gephi. Suite à quoi, il applique une stratégie alternative, qui gagne du terrain ces dernières années dans l'extraction de collocations, à savoir les procédures basées sur la *Social Network Analysis*. La contribution de David Brett nous fournit une excellente synthèse de la complexité du phénomène collocationnel et nous propose une nouvelle approche pour l'extraction et l'analyse des collocations, étant ainsi une illustration parfaite des nouveaux paradigmes de recherche, qui permettent l'implémentation d'une nouvelle méthode afin d'obtenir de « nouveaux observables ».

Quatre articles sont intégrés dans la section *Varia* de ce septième volume de *Studii de lingvistică*. **Samuel Bidaud** aborde différents faits de langue tchèque dans le cadre d'une linguistique dite « opérative ». Parmi les faits de langue analysés dans cette perspective « opérative », nous retrouvons : le système du genre, la catégorie du nombre, les déterminants et les pronoms possessifs, ainsi que le système du verbe.

Alicja Hajok, **Teresa Muryn**, **Małgorzata Niziołek** et **Katarzyna Gabrysiak** proposent une analyse des structures lexicales et syntaxiques du roman policier. Les auteures présentent le corpus utilisé, la méthodologie adoptée, ainsi que des analyses aux niveaux sémantique et lexico-syntaxique, qui les amènent à plaider en faveur d'une matrice lexico-syntaxique de type linéaire pour le genre textuel abordé.

L'article d'**Henry Tyne** s'inspire du célèbre roman *Le Parfum* de Patrick Süskind pour comparer l'utilisation, en langue maternelle (L1) et en langue étrangère (L2), des mots appartenant au vocabulaire de l'*odorat* en français. L'étude de leurs collocations relève des convergences mais aussi des divergences entre natifs et non-natifs.

Dans la contribution qui clôt le volume, **Daniela Francesca Virdis** examine les pratiques linguistiques de l'écrivain W. Wordsworth dans la "Preface" à la deuxième édition de *Lyrical Ballads* (1800). Son objectif est de montrer qu'il fait partie des écrivains que l'on peut considérer comme « novateurs », à travers une analyse qui relève de la pragmatique historique.

3. Conclusion et perspectives

Les diverses contributions réunies dans ce numéro thématique témoignent de l'importance des travaux exploitant des corpus pour la constitution de ressources numériques et leurs applications en linguistique, en didactique et en traduction. L'articulation des aspects théoriques et pratiques en linguistique de corpus est essentielle et ouvre de nouveaux paradigmes de recherche dans des domaines comme la phraséologie, la didactique, la constitution d'une nouvelle génération de dictionnaires, le traitement des collocations à l'aide de réseaux, les analyses linguistiques dans des domaines spécialisés, etc. Tous ces nouveaux paradigmes changent non seulement nos pratiques de recherche et d'observation mais aussi nos manières d'acquérir de nouvelles connaissances et, en particulier, grâce aux corpus et aux ressources numériques.

Luis Meneses-Lerín
Université d'Artois, *Grammatica* (EA 4521)