Lexique(s) et corpus en perspective comparative. Présentation

Depuis plusieurs années, les recherches linguistiques fondées sur les données textuelles se multiplient et s'orientent vers de nouveaux horizons, en exploitant différents types de corpus : corpus monolingues, bilingues ou multilingues, corpus bruts ou annotés, corpus officiels ou « maison », corpus parallèles ou comparables, corpus écrits ou parlés entre autres (*cf.* Mellet, 2002 ; Williams, 2005 ; Condamines, 2005 ; Ballard & Pineira-Tresmontant, 2007 ; Azzopardi, 2010 ; Kraif, 2014 ; Tutin & Grossmann, 2014). Or, s'il est évident que ces différents types de données textuelles facilitent l'analyse et la description des langues et que, donc, on ne saurait surestimer leur intérêt en sciences du langage, les manières dont les corpus peuvent être utilisés ne font pas l'unanimité. Si certains exploitent les corpus afin d'illustrer leurs affirmations théoriques (approche *corpus-based*), d'autres les utilisent afin d'en tirer des conclusions, appuyées souvent sur des méthodes statistiques, et valider les résultats (approche *corpus-driven*) (Tognini-Bonelli, 2001). Quant au corpus utilisé pour l'étude du lexique, les questions que se sont posées Cappeau et Gadet (2007) tiennent toujours d'actualité : « doit-il [c'.-à-d. le corpus] être illustratif, fournissant des exemples qui prendront place dans les rubriques prévues par l'analyse, ou moteur de description, servant à construire le cadre d'analyse : en un mot, heuristique ? ». La question est d'autant plus complexe qu'il s'agit de voir comment les corpus peuvent enrichir les analyses comparatives (*cf.* Granger, Lerot, & Petch-Tyson, 2003).

Le numéro 49/4 de la revue *Studia Romanica Posnaniensia* est consacré à l'étude du lexique français en perspective comparative avec les lexiques d'autres langues, notamment non romanes (allemand, anglais, grec, polonais, suédois et tchèque), et du rôle que jouent dans ces études différents types de corpus : à commencer par des listes préétablies (tel un dictionnaire), pour passer aux textes collectés à la main (de langue générale ou de spécialité, littéraires, administratifs ou de presse) jusqu'aux (grands et petits) corpus électroniques annotés, souvent personnalisés, et le Web.

Les approches adoptées sont diverses et leurs visées ne vont pas toutes dans le même sens. Ainsi le lexique est-il étudié en différentes configurations : en rapport avec la syntaxe et/ou la sémantique, en association avec la grammaire ou encore le

discours, configurations que complète la problématique de la formation du lexique et de la créativité lexicale (dérivation, composition et phraséologie). Les études présentées s'inscrivent dans différents courants de recherche (FLE, lexicographie, terminologie comparée, analyse du discours et traduction) et font preuve, ainsi, de la complexité de la problématique posée dans le titre du volume.

Stéphane Carsenty (France) présente le rôle que peut jouer un corpus spécialisé dans la modélisation du système des concepts d'un domaine. En s'appuyant sur un corpus de textes en trois langues (allemand, anglais et français) élaboré pour le domaine de la balance des paiements, il s'intéresse notamment aux caractéristiques du corpus de textes collectés. Son analyse débouche sur une typologie des textes regroupés dans le corpus qui sera utilisé pour le travail terminologique et l'élaboration et l'enrichissement de l'ontoterminologie du domaine.

Małgorzata Izert et Ewa Pilecka (Pologne) étudient les phraséosomatismes à valeur intensive en français avec *pied(s)* / *jambe(s)* et leurs correspondants en polonais avec *noga(i)*. Les auteures démontrent que le recours aux grands corpus monolingues, parallèles et comparables, permet de soigneusement décrire les propriétés syntaxiques et sémantiques des phrasèmes étudiés en ayant pour objectif leur inclusion dans un futur dictionnaire bilingue électronique. Ce dernier pourra servir, dans l'avenir, aussi bien aux apprenants qu'aux traducteurs.

Agnieszka Kaliska et Kaja Gostkowska (Pologne) utilisent les outils d'analyse de corpus pour effectuer une analyse lexicométrique du parler écologique. En se basant sur un corpus bilingue, constitué à partir des discours initiés par les plus importantes organisations de protection de l'environnement en France et en Pologne, les auteures élaborent des listes de fréquence des mots de l'écologie et les interprètent sous différents angles. Leur étude est donc également qualitative : elles nous font découvrir les différences au niveau du lexique écologique employé dans les deux langues, et par la suite, quels sont les sujets d'actualité dans les deux pays.

Christina Lindqvist et Mårten Ramnäs (Suède) discutent de l'élaboration et l'utilisation des listes de vocabulaire dans l'enseignement des langues en se focalisant notamment sur l'enseignement du français langue étrangère (FLE) dans le contexte universitaire suédois pour présenter le travail qui a été fait afin de créer une liste de vocabulaire (*Riksprovsordlistan*, env. 4000 mots), utilisée dans toutes les universités suédoises. Leur analyse se concentre sur les défis méthodologiques tels que le choix de l'unité de comptage (lemme vs. famille de mots), le rôle de la fréquence, le vocabulaire thématique, ainsi que les caractéristiques des corpus écrits par rapport aux corpus parlés.

L'analyse présentée par Fabrice Marsac (France) et Witold Ucherek (Pologne) rend compte d'une partie de recherches menées dans le cadre du projet franco-polonais « On the translation of French perception structures into Polish » (n°PPN/BIL/2018/1/00181), et notamment du contenu et de l'avancement de la partie *éti-*

quetage multidimensionnel. Les auteurs décrivent le codage formel de structures, catégories et fonctions des items du grand corpus multi-étiquettes (français-polonais) qu'ils collectionnent depuis 2019. Ils ont choisi comme cadre théorique le protocole traductologique Ucherek 1982, établi, à l'origine, pour la description contrastive des prépositions, et repris, avec succès, pour l'analyse d'autres constructions.

Radka Mudrochová et Dagmar Koláříková (Tchéquie) étudient les néologismes formés en français et en tchèque à partir du fractolexème anglais *-gate*. Elles puisent leurs données dans les corpus Web offerts par SketchEngine (Czech Web 2017 et French Web 2017), ce qui leur permet d'analyser de manière contrastive la circulation des lexies formées avec *-gate* dans les deux langues, tout en abordant des questions théoriques pour cadrer le concept de fractocomposition dans les études linguistiques actuelles.

Mavina Pantazara, Eleni Tziafa et Angeliki Christopoulou (Grèce) proposent une étude terminologique comparée concernant le domaine du « travail ». À partir d'un corpus comparable grec-français constitué de textes administratifs de la période 2010–2018, elles étudient le discours autour du travail (lois, actions, réformes, mesures), mené dans les deux pays affectés par la crise économique. Leur analyse contrastive, à la fois quantitative (calcul des fréquences) et qualitative (interprétation des usages des termes), vise à mettre en évidence les aspects morphosyntaxiques, sémantiques et pragmatiques qui caractérisent les termes étudiés dans leur contexte lexical habituel.

Dans une perspective traductologique sont menées les recherches d'Alice Ray (France) qui s'intéresse, en revanche, à la traduction des innovations lexicales de la science-fiction. Ces dernières, appelées par l'auteure *termes-fictions*, possèdent des caractéristiques mixtes : d'un côté, elles font partie d'une terminologie spécialisée, science-fictionnelle, anglaise à l'origine, de l'autre ce sont des créations littéraires. Il est, en effet, intéressant d'observer les stratégies mises en œuvre pour les traduire en français.

L'étude présentée par Regina Solová (Pologne) s'inscrit dans le courant de la linguistique culturelle. L'auteure propose d'étudier les profils du concept de *Pologne* à partir d'un corpus de textes de propagande extérieure publiés en 1968 dans trois mensuels : *La Pologne*, *Polsko* et *La Revue Polonaise*, adressés à trois groupes de lecteurs correspondant au *premier*, au *second* et au *tiers monde*, afin de vérifier l'hypothèse selon laquelle les profils du concept analysé y sont différents et que les différences sont motivées par des facteurs extralinguistiques, à savoir la politique étrangère du pays.

Grażyna Vetulani (Pologne), dans son article, rend compte des problèmes de nature méthodologique rencontrés lors de la préparation d'un dictionnaire polonais des prédicats nominaux. Le projet décrit dans l'article date de plusieurs années, ce qui permet d'observer le rôle croissant de grands corpus informatisés dans les recherches en linguistique aussi bien pour le français que pour le polonais. L'attention du lecteur

est attirée également sur le fait que les usages de la langue, tels qu'ils se reflètent dans les corpus, ne sont pas toujours conformes aux modèles syntaxico-sémantiques des prédicats nominaux qu'on peut construire à partir des dictionnaires de langue, les deux (usages et modèles) étant justifiés pour autant.

Les études réunies dans ce volume démontrent la complexité de la problématique abordée sans pour autant épuiser le sujet des relations complexes entre le choix du corpus, son maniement et le regard porté vers le lexique, surtout dans une perspective comparative.

Nous remercions tous les auteurs de leurs précieuses contributions et nous espérons que la lecture des articles réunis dans ce numéro inspirera de nouvelles recherches sur le lexique de différentes langues et les corpus. Nous remercions également tous les évaluateurs d'avoir généreusement partagé leurs savoir et savoir-faire.

Agnieszka Kaliska
Kaja Gostkowska
Mavina Pantazara

BIBLIOGRAPHIE

Azzopardi, S. (ed.) (2010). *Cahiers de praxématique. Corpus, données, modèles*, vol. 54-55. DOI : 10.4000/praxematique.1102.

Ballard, M. & Pineira-Tresmontant, C. (eds.). (2007). *Les corpus en linguistique et en traductologie*. Arras : Artois Presses Université.

Cappeau, P. & Gadet, F. (2007). L'exploitation sociolinguistique des grands corpus : Maître-mot et pierre philosophale. *Revue française de linguistique appliquée*, XII/1, 99-110.

Condamines, A. (2005). Linguistique de corpus et terminologie. *Langages*, 157. DOI : 10.3917/lang.157.0036.

Granger, S., Lerot, J. & Petch-Tyson, S. (2003). *Corpus-based Approaches to Contrastive Linguistics and Translation*. Amsterdam : John Benjamins.

Kraif, O. (2014). *Corpus parallèles, corpus comparables : quels contrastes ?. Informatique et langage*. Poitiers : Université de Poitiers.

Mellet, S. (ed.) (2002). Corpus et recherches linguistiques. *Corpus*, 1. DOI : 10.4000/corpus.7.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work.* Amsterdam/Philadelphia : John Benjamins.

Tutin, A. & Grossmann, F. (eds.) (2014). *L'écrit scientifique : du lexique au discours. Autour du Scientext*. Rennes : Presses universitaires de Rennes.

Williams, G. (ed.) (2005). *La linguistique de corpus.* Rennes : Presses universitaires de Rennes.