Lectures
Les comptes rendus

Émeline Comby, Yannick Mosset, Stéphanie de Carrara (dir.), *Corpus de textes : composer, mesurer, interpréter*

Adrien Mathy

revues.org

Édition électronique
URL : http://lectures.revues.org/22553
ISSN : 2116-5289

Éditeur
Centre Max Weber

Référence électronique
Adrien Mathy, « Émeline Comby, Yannick Mosset, Stéphanie de Carrara (dir.), *Corpus de textes : composer, mesurer, interpréter* », *Lectures* [En ligne], Les comptes rendus, 2017, mis en ligne le 22 mars 2017, consulté le 22 mars 2017. URL : http://lectures.revues.org/22553

Émeline Comby, Yannick Mosset, Stéphanie de Carrara (dir.), *Corpus de textes : composer, mesurer, interpréter*

Adrien Mathy

1 À l'heure des humanités numériques[1], la problématisation de l'objet et de la méthodologie du corpus de textes est plus que jamais d'actualité : défini par les auteurs comme étant un « ensemble a priori homogène de documents (textuels, mais aussi iconographiques, sonores, etc.) réunis dans un but scientifique » (p. 178), le corpus est un objet primordial de la linguistique et des sciences sociales, qui y trouvent un terrain d'exploitation empirique. Aussi, l'ouvrage coordonné par Émeline Comby, Yannick Mosset et Stéphanie de Carrara répond-il aux attentes du champ de la recherche en sciences humaines et sociales. Il est néanmoins laborieux de traiter d'une problématique aussi large avec exhaustivité : non seulement les sciences humaines constituent un ensemble épars de théories et de méthodes, mais de surcroit la pratique du corpus de textes connaît de nombreuses approches divergentes.

2 C'est pourquoi les auteurs prennent le parti de recueillir des contributions variées dont la ligne directrice est l'émergence d'une métaréflexion sur le corpus, afin de mettre à disposition du lecteur une synthèse représentative des diverses branches des sciences humaines, de différents types de corpus et surtout de ses approches variées. Sont ainsi représentés la linguistique et l'analyse du discours, les sciences politiques et médiatiques, l'architecture, la littérature ou encore la cartographie linguistique. Selon le type d'étude, la nature du corpus diffère quant au genre (discours médiatique, littéraire, ordinaire), au code (discours oral ou écrit), à la circonscription spatio-temporelle, ou encore à la méthode de constitution et d'exploitation, qui varie en fonction de l'outil utilisé (analyse de contenu ou analyse de textes) et de la conceptualisation du corpus (corpus-*based* ou corpus-*driven*).

3 L'approche textuelle, qui repose sur une analyse statistique, notamment à l'aide de logiciels (présentés dans les contributions), est l'approche dominante dans ce recueil. *A contrario*, l'analyse de contenu ne s'intéresse pas tant aux occurrences statistiques de lexèmes (ou de locutions) qu'aux réseaux de sens qui se tissent dans et par le discours. Elle permet donc regrouper des entités descriptives, presque sémiotiques, comme le montre avec succès l'article de Nadia Zidelmal et Azeddine Belakehal. Quant à la conceptualisation du corpus, elle se particularise selon l'usage qui en est fait : le corpus est utilisé tantôt afin de répondre à des hypothèses avancées (corpus-*based*), tantôt comme un terrain d'exploration à part entière (corpus-*driven*).

4 Compte tenu de cette diversité heuristique, les contributeurs essaient de présenter un regard critique sur leurs usages de corpus. Les contributions sont cadrées et organisées selon trois thèmes. La première partie, « corpus linguistique », s'intéresse spécifiquement à la composition d'un corpus ; la seconde, « le temps du corpus », considère l'étape de l'analyse et de la mesure statistique du corpus, en envisageant plus spécifiquement la dimension temporelle de celui-ci ; enfin, la partie « corpus, productions de discours et représentations sociales » se concentre sur l'interprétation des données, par le truchement des représentations sociales véhiculées dans et par le discours. Chacune de ces parties est ouverte par un bref exorde.

5 Le premier thème de l'ouvrage est introduit par un article de Catherine Pinon. Elle s'intéresse aux valeurs du verbe *kâna* en arabe contemporain et cherche à les mettre en relation avec deux variables : le critère géographique et le critère générique, c'est-à-dire relatif au genre du texte. Éminemment linguistique, cette problématique nécessite de développer un corpus à même de dépasser plusieurs écueils fréquents propres à l'étude de la variation linguistique. L'intérêt de l'article se situe notamment dans la phase de conclusion critique, dans laquelle l'auteure formule un paradoxe qui exemplifie parfaitement l'importance des choix de constitution pour l'étape d'analyse et d'interprétation. Elle explique ainsi que « plus l'on cherche à être représentatif, plus on est contraint à faire des choix qui s'avèrent être totalement subjectifs » (p. 37).

6 Le second article, écrit par Soufiane Lanseur, s'intéresse à la formation morphologique et sémantique des néologismes dans le discours médiatique algérien relatif à l'économie, néologismes qui représentent un état de langue peu voire non attesté dans la littérature lexicographique. À l'instar de Pinon, l'auteur circonscrit son corpus, composé d'articles de presse, de chroniques et d'entretiens, selon divers critères : codique, géographique et générique, et il prend en compte quelques variables qui affinent la typologie (domaine de profession des locuteurs, etc.). Au terme de son article, l'auteur est capable de fournir des données relatives aux occurrences des néologismes, qu'il organise en plusieurs catégories selon leur formation : siglaison, emprunt lexical, xénisme, dérivation, composition[2], etc. L'auteur n'apporte, toutefois, aucune interprétation, puisqu'il s'agissait uniquement d'interroger la constitution du corpus et la classification des données.

7 Le troisième article, coécrit par Natacha Souillard, Pierre Ratinaud et Pascal Marchand, introduit la seconde partie de l'ouvrage. Les auteurs analysent un corpus de presse composé de dépêches AFP traitant des révolutions arabes, dans une approche lexico-chronologique, à l'aide du logiciel IRaMuTeQ. Cette méthode consiste à corréler, dans un corpus de discours, la présence de lexèmes avec la dimension temporelle de l'événement auquel se rapportent ces discours. Autrement dit, il s'agit d'étudier l'évolution du choix des mots utilisés dans les dépêches AFP pour traiter les révolutions arabes. L'article montre la pertinence d'une telle approche pour décrire la narration journalistique et

mettre en évidence les grilles conceptuelles sur lesquelles elle repose. Grâce à cette analyse, les auteurs révèlent que ces grilles conceptuelles sont inopérantes pour rendre compte de la chronologie des événements.

8 Émeline Comby s'intéresse, quant à elle, à l'analyse d'un corpus d'archives textuelles et des bases de données numériques. Elle étudie la couverture médiatique de la pollution du Rhône par les polychlorobiphényles (PCB), un agent très polluant. L'étude se fonde sur l'analyse de trois quotidiens régionaux entre 2005 et 2010. L'intérêt de cette contribution est double. D'une part, l'auteure interroge le fonctionnement deux logiciels différents, IRaMuTeQ et TXM, et la logique qui leur est sous-jacente, afin de mettre au jour leurs divergences techniques, méthodologiques et, dans une certaine mesure, leurs divergences de résultats[3]. D'autre part, elle interroge finement et comparativement les enjeux relatifs à l'usage des bases de données numériques et aux archives traditionnelles, en attirant notamment l'attention sur les limites propres aux bases de données : le tri reste particulièrement chronophage, les doublons ne sont pas détectés et le chercheur n'a pas toujours accès à la donnée initiale.

9 Cette seconde partie se ferme sur l'article coécrit par Nadia Zidelmal et Azeddine Belakehal, dont l'originalité est de mêler analyse d'un corpus littéraire et histoire de l'architecture. Les auteurs constatent que les opérations de préservation du patrimoine ignorent souvent la vie originale des édifices, en l'occurrence celle des maisons kabyles : leur identité culturelle, leurs dimensions sensorielles, etc. Ainsi, pour y remédier, ils se donnent comme objectif de reconstituer l'ambiance de ces maisons traditionnelles, à partir d'un corpus de quatre romans. Les auteurs exemplifient leur méthode sur l'un des quatre et regroupent plusieurs ensembles sémiotiques, tels que la description sensorielle (odeur, couleur, son) ou l'usage de lieux communs relatifs aux traditions kabyles (métier à tisser, étable, etc.). Ils concluent que le procédé consistant à mettre en relation un vécu réel avec une description statique est positif, en ce qu'il permet de reconstituer, peu ou prou, l'ambiance des maisons kabyles, en considérant la dynamique de la vie qui s'y déroulait afin d'éviter une muséification du patrimoine.

10 L'article coécrit par Yves-François Le Lay, Serge Heiden, Luc Merchez et Bénédicte Pincemin ouvre la dernière partie de l'ouvrage. Les auteurs ont recueilli l'avis de tous les pêcheurs en activité au Lac Léman (du côté français) concernant leur profession, et ont ainsi constitué un corpus dit clos, c'est-à-dire comprenant la totalité des discours produits par des locuteurs donnés sur un sujet donné. D'un point de vue méthodologique, la contribution est remarquable par sa combinaison de l'analyse de corpus et de la cartographie linguistique. L'étude met en lumière la représentation que les locuteurs ont de leur pratique professionnelle et la manière dont ils se situent par rapport aux mutations contemporaines qui affectent leur profession : changements techniques et modernisation (filet en nylon, géolocalisation, etc.), changement de la situation économique et politique du pêcheur (augmentation du chiffre d'affaire, opposition entre petits artisans et « barons » (p. 130), nouvelles lois et normes, etc.), transmission du savoir-faire (importance de l'héritage familial) et patrimonialisation, etc.

11 L'article de Lise Jacquez et Audrey Arnoult met en contraste le traitement journalistique de deux thématiques *a priori* sans corrélation : l'expulsion des *sans-papiers* et les troubles de l'adolescence (anorexie, boulimie, automutilation, etc.), mais néanmoins porteuses d'enjeux semblables. Les auteurs cherchent à étudier la médiatisation de situations de souffrance psycho-sociale qui « constitu[e] aussi des transgressions par rapport à certaines normes de la société » (p. 135). Pour ce faire, ils ont développé une démarche

d'analyse de contenu, en utilisant notamment le logiciel Modalisa, afin de mettre au jour la similarité dans l'approche journalistique de ces deux problèmes. Au terme de leur analyse, ils font émerger la structure narrative propre à la problématisation de la souffrance sociale. Les récits journalistiques sont structurés en trois phases : description de la situation dénoncée comme problématique, mise en accusation de responsables, processus de réclamation et demande de solution. Cette narration met en scène plusieurs « actants incarnés par des acteurs précis ou des institutions (des victimes, des porteurs du problème ou dénonciateurs), [... et] un destinataire de la dénonciation – en général le public » (p. 137).

12 Enfin, la dernière contribution de l'ouvrage, de Konstantinos Delimitsos, s'inscrit dans une perspective sociologique et décortique les discours des nouveaux experts en sécurité urbaine. L'auteur formule une hypothèse selon laquelle ces discours – qu'il critique vertement – sont émaillés de références lexicales bellicistes. Pour le démontrer, Delimitsos dresse premièrement un inventaire des locutions bellicistes afin d'appliquer ensuite une analyse lexicométrique sur le corpus préalablement numérisé. Les occurrences sont finalement rapportées à une unité de référence *ad hoc* (la « page-type », qui compte en moyenne *x* occurrences). En conclusion, l'auteur porte un regard critique sur son étude et constate les limites de l'analyse quantitative. Il remarque, par exemple, que le lexème *guerre* apparaît dans des formulations non bellicistes. Cet article permet d'une part de mesurer l'intérêt croissant des outils informatiques et, d'autre part, de relativiser leur performance en considérant qu'en définitive, le chercheur doit toujours être présent afin d'interpréter les résultats.

13 Ce dernier article rend compte de la richesse du recueil et des nombreuses problématiques qu'il embrasse. En somme, parce que chaque corpus implique des méthodes et des utilisations différentes, il révèle la relation éminemment complexe entre l'objet étudié et le chercheur. Nous pouvons ainsi paraphraser Charaudeau[4] et considérer que le corpus de textes, plus qu'une méthode ou un outil, est un véritable miroir des réflexions et des hypothèses du chercheur : en effet, de sa composition à son interprétation et sa finalité, le corpus représente le cheminement théorique et critique du chercheur.

NOTES

1. Sur l'importance et la vitalité du concept, voir Jean-Paul Fourmentraux (dir.), *L'Ère post-média. Humanités digitales et cultures numériques*, Éditions Hermann, coll. « Cultures numériques », 2012 ; compte rendu de Christophe Premat pour *Lectures* : https://lectures.revues.org/9619.

2. La siglaison consiste à créer un sigle, la composition à assembler plusieurs lexèmes (*micro-entreprise*), la dérivation à créer un néologisme à partir d'un lexème existant (créer un nom à partir d'un verbe, etc.) Quant à l'emprunt et le xénisme, les typologies lexicologiques divergent, et la nuance est parfois subtile. L'auteur différencie ici les lexèmes morphologiquement intégrés à la langue (ce qu'il nomme *emprunt*, par exemple *timing* ou *business*) de ceux empruntés, mais pas encore inscrits dans le paradigme de la langue cible (ce qu'il nomme *xénisme*, par exemple *working* ou *banking*), et donc toujours perçus comme étrangers par le locuteur.

3. L'analyse extrêmement technique, difficile d'accès aux néophytes et à la vulgarisation, de Comby permet de comparer les deux logiciels : IRaMuTeQ propose une logique de lemmatisation, de segmentation et d'analyse des formes pleines (les unités sémantiques non grammaticalisées), et prend en compte certaines métadonnées. Il repose sur la méthode d'Alceste conçu par Reinert et se base sur le logiciel statistique R ; tandis que TXM, qui est *open source*, permet de subdiviser le corpus, de produire une annotation morphosyntaxique et facultativement de lemmatiser. Il permet d'ailleurs de produire des statistiques univariées, bivariées et multivariées, selon qu'elles n'utilisent qu'un seul, deux ou plusieurs paramètres. Enfin, TXM est lui aussi basé sur le logiciel R. La divergence mineure de résultat provient par conséquent non pas du logiciel de calcul, mais de la divergence méthodologique.

4. « Dis-moi quel est ton corpus, je te dirai quelle est ta problématique », *Corpus,* n° 8, 2009, p. 37-66 ; disponible en ligne : https://corpus.revues.org/1674.

AUTEUR

ADRIEN MATHY

Assistant au service de Linguistique française de l'Université de Liège.