Mots. Les langages du politique

116 | 2018

Dire ou ne pas dire la « race » en France aujourd'hui

Céline Poudat et Frédéric Landragin, *Explorer un corpus textuel. Méthodes – pratiques – outils*

Louvain-la-Neuve et Paris, De Boeck Supérieur, 2017

Michèle Monte

Édition électronique

URL : <https://journals.openedition.org/mots/23101>

DOI : 10.4000/mots.23101

ISSN : 1960-6001

Éditeur

ENS Éditions

Édition imprimée

Date de publication : 20 mars 2018

Pagination : 145-148

ISSN : 0243-6450

Référence électronique

Michèle Monte, « Céline Poudat et Frédéric Landragin, *Explorer un corpus textuel. Méthodes – pratiques – outils* », *Mots. Les langages du politique* [En ligne], 116 | 2018, mis en ligne le 23 février 2018, consulté le 24 avril 2022. URL : <http://journals.openedition.org/mots/23101> ; DOI : <https://doi.org/10.4000/mots.23101>

Ce document a été généré automatiquement le 24 avril 2022.

© ENS Éditions

Céline Poudat et Frédéric Landragin, *Explorer un corpus textuel. Méthodes – pratiques – outils*

Louvain-la-Neuve et Paris, De Boeck Supérieur, 2017

Michèle Monte

RÉFÉRENCE

Céline Poudat et Frédéric Landragin, *Explorer un corpus textuel. Méthodes – pratiques – outils*, Louvain-la-Neuve et Paris, De Boeck Supérieur, 2017, 240 p.

- 1 Ce livre est destiné par ses auteurs aux linguistes qui souhaitent s'initier à l'exploration outillée de corpus numériques, afin de décrire des usages de la langue en discours en s'appuyant sur de nombreuses occurrences contextualisées. Il intéresse également les linguistes qui développent le traitement automatique des langues et les analystes de discours qui travaillent sur des corpus numériques. Il s'appuie sur le groupe de travail « Exploration de corpus » du consortium « Corpus écrits », coordonné par Céline Poudat. Son objectif est de présenter les méthodes de façon à permettre à l'utilisateur de ne plus choisir une fonctionnalité dans un logiciel au hasard, mais en connaissant les prérequis nécessaires à son exploitation pertinente, les méthodes statistiques qui lui sont sous-jacentes et les interprétations qu'il pourra tirer des résultats obtenus (p. 10).
- 2 L'introduction distingue trois types d'exploration d'un corpus muni de ses métadonnées et enrichi d'annotations : la visualisation du corpus annoté, la requête ciblée et les calculs statistiques. Elle évoque ensuite les requêtes basées sur les expressions régulières, et celles qui portent à la fois sur le corpus et les annotations. Puis les branches disciplinaires qui travaillent sur corpus (lexicométrie, textométrie, linguistique fondée sur un corpus et guidée par un corpus) sont présentées rapidement ainsi que les deux approches déductive et inductive et les fonctionnalités qui leur sont liées. Le livre se divise ensuite en trois grands chapitres, de 50 à 60 pages chacun,

consacrés successivement à l'exploration d'un corpus par l'annotation, par l'étude de sa structure et par la vérification/infirmation d'une hypothèse préalable sur le fonctionnement en contexte de ses unités linguistiques. À chaque étape, les auteurs mentionnent les logiciels libres permettant de réaliser telle ou telle exploration, leurs atouts et leurs limites. Une annexe les rassemble qui référence AntConc, Dtm-Vic, Hyperbase, Iramuteq, Le Trameur, Lexico, TXM et Unitex.

- 3 Le premier chapitre aborde les annotations – automatisées ou manuelles – qui explicitent la structure d'un corpus (fin de chapitre, tours de parole, fin de phrase, vers ou strophes), et celles qui enrichissent le corpus avec des interprétations : catégories morphosyntaxiques, fonctions syntaxiques, actes de langage, rôles ou traits sémantiques (agent, patient, animé, abstrait, etc.), repérage des connecteurs, des cadratifs, des relations sémantiques entre énoncés, etc. Le chapitre développe un peu plus le marquage des expressions référentielles (sujet sur lequel Frédéric Landragin a beaucoup travaillé). La question de la partageabilité des annotations au sein de la communauté des chercheurs est évoquée. Sont ensuite abordées la structure des annotations en différentes couches, la visualisation des relations entre unités, la construction de schémas regroupant des annotations de types variés. Le chapitre étudie ensuite les limites des modèles et trace la marche à suivre quand on veut annoter un corpus (il est en particulier recommandé de travailler préalablement sur un bref corpus test). Il se clôt sur la recherche de régularités dans un corpus par le repérage de suites de mots consécutifs (« N-gramme »), permettant par exemple d'observer la récurrence de schémas phrastiques, de types de chaînes référentielles, et de mettre au jour des corrélations entre traits.
- 4 Le deuxième chapitre est consacré à la façon dont l'analyse outillée met au jour la structure d'un corpus en rapprochant ou en contrastant des parties de corpus. Les auteurs mettent en garde sur le risque de circularité entre hypothèses et résultats si les hypothèses ne sont pas explicitées au départ et invitent à travailler *a minima* sur trois niveaux d'observation : mots-formes, lemmes, parties du discours. L'observation préalable du vocabulaire du corpus permet éventuellement d'écarter des textes trop singuliers, présentant des hapax qui, s'ils sont très fréquents dans ces textes, peuvent fausser les calculs. Les auteurs présentent ensuite l'analyse factorielle des correspondances (AFC), l'analyse en composantes principales, les classifications hiérarchiques ascendante et descendante (CAH et CDH), les analyses arborées. Les principes de chacune de ces méthodes sont expliqués dans un souci de démystifier les résultats et les auteurs mettent en garde sur la confusion fréquente entre corrélation et causalité. Des encadrés proposent l'analyse de quelques exemples concrets : AFC de récits de témoins de la Shoah faisant apparaître un groupement chronologique des textes similaires, CHD des professions de foi des femmes députées de la V^e République, analyse arborée de récits de voyage.
- 5 Le troisième chapitre développe la façon dont on peut, de façon déductive, chercher la vérification d'une hypothèse en corpus. Il pointe les deux principaux écueils qui guettent le chercheur : être débordé par la richesse des résultats et donner un sens à tout rapprochement ou opposition. Il soulève la question de la définition du texte, qui peut inclure ou non le périphrase, et l'importance des annotations pour pouvoir créer un sous-corpus, tel que les introductions d'articles, le dernier paragraphe d'un chapitre, etc. Il souligne l'importance des métadonnées relatives aux dates, aux auteurs, aux genres et types de discours, lesquelles vont conditionner les partitions. Le chapitre

présente ensuite les principales possibilités d'exploration : calcul des spécificités, concordance, segments répétés, cooccurrences et poly-cooccurrences. Les différentes méthodes statistiques de calcul des spécificités (écart réduit et F^2 fondés sur la distribution normale, distribution hypergéométrique, rapport de vraisemblance ou *log likelihood*) sont présentées, suivies d'une brève interprétation des spécificités des différentes tranches chronologiques du corpus Shoah. Un tableau résume de façon très claire les différents types de concordance et ce qu'on peut en attendre, et l'étude des poly-cooccurrences est illustrée par l'étude d'un réseau reliant un connecteur concessif, le pronom personnel *nous* et un prédicat de sentiment dans des rapports de travailleurs sociaux. Le chapitre débouche donc sur la question des motifs et des patrons qui permettent d'étudier à nouveaux frais la phraséologie et les routines discursives.

- 6 La conclusion revient sur l'ancienne réticence de la lexicométrie à la lemmatisation et à l'annotation, dont on doit trouver des échos dans *Mots*, puisque Michel Tournier défendait une approche rigoureusement objective et le refus de catégories imposées de l'extérieur au corpus. Ce rappel historique permet aux auteurs de conclure sur la vigilance à observer dans toute procédure d'annotation tout en plaidant pour un travail conjoint des acteurs de l'annotation (plutôt présents en TAL), de la linguistique de corpus et de l'analyse textométrique.
- 7 Ce livre rendra de grands services aux personnes, nombreuses, qui se sont lancées dans l'analyse outillée des données textuelles en découvrant par elles-mêmes les logiciels et en tâtonnant empiriquement dans leurs fonctionnalités. Il leur permettra de rationaliser leurs choix et de comprendre un peu mieux les calculs statistiques sous-jacents à certains résultats. Très rigoureux dans sa démarche et ses formulations, le livre est un peu décevant sur les analyses, souvent assez sommaires. Plutôt que de multiplier les exemples empruntés à des recherches différentes, il aurait sans doute été préférable de se borner à deux ou trois corpus et de montrer au fil des pages, en développant la partie dévolue aux analyses, ce qu'une exploration systématique fait surgir au niveau interprétatif. Les mises en garde auraient pu être illustrées de quelques exemples d'erreurs naïves. Les copies d'écran ne sont pas toujours très lisibles et il aurait peut-être mieux valu stocker les visualisations sur un site universitaire où on aurait pu recourir à la couleur et agrandir les tableaux. Ce sont là des reproches mineurs. Ce qui est plus regrettable, c'est que l'ouvrage nous semble – mais nous espérons nous tromper – peu accessible, en raison de son degré élevé d'abstraction, à des chercheurs qui n'auraient pas encore mis les mains dans le cambouis. De ce point de vue, un cours en ligne tel que celui proposé par Véronique Magri sur le site de l'université ouverte des humanités (UOH) pour la prise en main d'Hyperbase est sans doute plus approprié qu'un ouvrage et aurait pu être indiqué en introduction¹. Nous avons beaucoup apprécié en revanche les recommandations salutaires alertant sur les risques de surinterprétation : le livre constitue à cet égard une importante contribution à l'analyse réflexive des pratiques en matière d'exploration outillée dont nous avons tous et toutes besoin. Il s'adresse donc plutôt, à notre avis, à des personnes déjà un peu familiarisées avec l'analyse outillée. Pour les lecteurs souhaitant s'y initier, une lecture en parallèle de cet ouvrage et de celui de Jean-Marc Leblanc recensé dans ce même numéro serait une bonne solution pour à la fois acquérir une vue d'ensemble des analyses outillées et en voir les applications à un corpus spécifique.

NOTES

1. Textométrie pour le texte littéraire : le logiciel Hyperbase. En ligne : [<http://www.uoh.fr/front/notice?id=a17612b3-8ef3-43c5-b5ad-d01dd52e97dc>] (consulté le 03/10/2017)

AUTEURS

MICHÈLE MONTE

Université de Toulon