"Poudat, Céline & Landragin, Frédéric (2017). Explorer un corpus textuel : Méthodes – pratiques - outils. Louvain-la-Neuve : De Boeck Supérieur." Note de lecture

Eva Schaeffer-Lacroix

Poudat, Céline & Landragin, Frédéric (2017). *Explorer un corpus textuel : Méthodes – pratiques - outils*. Louvain-la-Neuve : De Boeck Supérieur.
Note de lecture proposée par Eva Schaeffer-Lacroix (juillet 2017)[1].

Remarques générales

L'ouvrage *Explorer un corpus textuel : Méthodes – pratiques – outils* de Céline Poudat et de Frédéric Landragin a paru en mars 2017. Les auteurs visent principalement un public composé d'étudiants en sciences du langage, niveau master (p. 10), mais selon eux, ce livre peut également convenir à tout étudiant ou chercheur d'une autre discipline intéressé par l'analyse des données textuelles à l'aide des outils et méthodes de la linguistique de corpus. Enseignante-chercheuse intervenant au niveau master depuis plusieurs années, didacticienne spécialiste des corpus, je lis cet ouvrage non seulement pour compléter mes connaissances, mais aussi pour évaluer la possibilité de l'utiliser lors d'un module d'enseignement dédié à l'introduction à la linguistique de corpus.

Les auteurs ont pour objectif "d'expliquer de manière progressive et pédagogique ces méthodes de la linguistique outillée qui sont orientées vers l'exploration de corpus, et donc vers l'accès aux données, leur visualisation, le calcul de statistiques" (p. 10). Il ne s'agit pas d'une annonce vaine : le sommaire du livre est claire et logique : *Introduction – Exploration par l'annotation – Exploration de la structure d'un corpus – Exploration d'une hypothèse en corpus – Conclusion – Annexes* (dont *Outils d'exploration de corpus*). Le style est, lui aussi, limpide et abordable tout en restant scientifiquement valide. Certains éléments ou caractéristiques sont évoqués de façon brève parce que la connaissance de leur existence est considérée comme une condition pour la bonne compréhension de la suite ; rentrer dans les détails aurait probablement comporté le risque de noyer le lecteur novice en la matière dans une masse d'informations ou de l'ennuyer s'il s'y connaît déjà un tant soit peu en linguistique de corpus. Citons comme exemple le langage *SQL* (*Structured Query Language*) dont on se contente de mentionner l'existence et l'importance (p.23) ce qui est suffisant dans le cadre posé par les auteurs.

On peut également lire que le propos des auteurs "ne relève pas de la technique : il est méthodologique avant tout et vise à fournir des repères sur le fonctionnement des outils en général, et non sur l'utilisation de tel ou tel outil particulier" (p. 10). Il ne s'agit visiblement pas de présenter toutes les fonctionnalités d'un outil d'exploration de corpus en particulier. On peut tout de même constater qu'au fil du texte, des noms d'outils sont cités, certains principalement pour illustrer un point précis (*Glozz* ou *Annis*). D'autres comme *TXM* ou *IRaMuTeQ* apparaissent de façon récurrente ; de plus, en annexe, un paragraphe descriptif leur est dédié.

Au fil des chapitres

Introduction

Le chapitre nommé *Introduction* pose les jalons susceptibles de permettre aux lecteurs et lectrices novices en matière d'exploration de corpus de comprendre les notions et démarches fondamentales du domaine. Si certains paragraphes présentent des informations basiques (Qu'est-ce que c'est qu'un corpus ? Comment rechercher un mot ou une suite de mots dans un texte ?), d'autres sont plus spécifiques : ils concernent, entre autres, la façon dont on peut visualiser les données et dont on peut formuler des requêtes complexes.

[1] Une version réduite de cette note de lecture a été proposée pour publication dans la revue de linguistique et de didactique des langues *Lidil*.

Ce chapitre rappelle également les distinctions fréquemment faites en linguistique de corpus telles que celles présentées au tableau 1.

approche quantitative	approche qualitative
approche inductive	approche déductive
corpus-driven	*corpus-based*

Tableau 1 – Juxtaposition d'approches en linguistique de corpus.

Les auteurs ont fait le choix de consacrer le chapitre 3 à l'approche inductive (*corpus-driven*) et aux fonctionnalités qui la soutiennent, comme l'AFC (analyse factorielle des correspondances)[2]. Selon cette approche, on "laisse parler" les données pour découvrir des usages linguistiques. Le chapitre 4 présente l'approche déductive (*corpus-based*) : les chercheurs formulent une hypothèse de travail, par exemple, sur les différentes façons d'exprimer la négation en une langue donnée, et ils la testent en parcourant le corpus. L'usage du concordancier ou la recherche de cooccurrents[3] se prêtent bien à une approche déductive. Poudat et Landragin insistent sur le fait qu'une alternance entre l'usage de ces deux approches peut faire sens au sein d'une même recherche (p. 31).

Chapitre 2 – *Exploration par l'annotation*

Le chapitre 2 présente les facettes multiples de l'annotation, qui est présentée comme l'une des méthodes possibles d'exploration d'un corpus. Une distinction éclairante est faite entre l'étiquetage automatique, se servant d'une liste finie d'étiquettes (cf. ce que permet de faire l'outil *Treetagger* pour créer des annotations morphosyntaxiques) et l'étiquetage fait par des personnes. Ce dernier peut être motivé par les besoins spécifiques d'un projet de recherche.

La promesse initiale d'un discours qui ne relèverait pas de la technique n'est pas parfaitement tenue dans ce chapitre, mais cela est probablement lié à la nature de l'objet et à sa complexité. Je constate que l'effort qu'implique la mise en place d'étiquettes est parfois titanesque, surtout s'il s'agit d'étiquettes à plusieurs traits tenant compte de la complexité, voire de l'ambiguïté d'une forme ou d'un groupe de formes. Lors de telles entreprises d'étiquetage, il semble nécessaire d'évaluer le gain potentiel pour la communauté scientifique au-delà des résultats obtenus par les personnes qui négocient et choisissent les annotations dont ils ont besoin dans le cadre d'un projet particulier.

Chapitre 3 – *Exploration d'une hypothèse en corpus*

Le chapitre 3 s'intéresse au travail partant de la structure du corpus (approche *corpus-driven*). Après la description des différentes manières de structurer un corpus, le fonctionnement et les apports de l'analyse factorielle sont présentés. Là aussi, un discours non technique est recherché, mais les lecteurs peu formés en mathématique et statistique peuvent avoir du mal à suivre les explications toutefois concises et bien présentées. Il faut peut-être tout simplement admettre que l'on ne peut se passer de connaissances de base dans ces domaines et qu'il est nécessaire de s'y initier si on veut utiliser les outils d'exploration à bon escient. Ceci dit, même si la lecture de certains des passages de ce chapitre peut être éprouvante pour les novices en

[2] Quelques éléments définitoires : "(...) l'ACF (...) est une méthode qui s'applique à des (...) **tableaux de contingence**. Ces tableaux ventilent une population selon deux variables catégorielles (deux catégorisations) et permettent de visualiser les liens, les *correspondances* entre les catégories" (Poudat & Landragin 2017 : 115).

[3] Définition de "cooccurrence" sur le portail lexical *CNRTL* (2012) : "Co-occurrence, cooccurrence, subst. fém.,ling. Apparition simultanée de deux ou plusieurs éléments ou classes d'éléments dans le même discours ; p. méton., ces mêmes éléments. Telle pourrait être notre nouvelle définition de la co-occurrence: une suite d'enchevêtrements à probabilité infime (P. Lafon, M. Tournier, Une Nouvelle approche lexicométrique des cooccurrences dans un texte. ds Travaux de lexicométrie et de lexicol. pol., nov. 1978, p.137)". http://www.cnrtl.fr/definition/cooccurrent

statistique textuelle, il est précieux de savoir que l'on pourra trouver dans ces pages des informations utiles le jour où l'on en aura besoin, par exemple quand on aura l'occasion de participer à un projet nécessitant des connaissances spécifiques dans ce domaine.

J'apprécie particulièrement les remarques suivantes, présentées à la fin du chapitre 3, qui placent le travail de structuration de corpus dans un contexte méthodologique plus vaste :

> *Ce qui importe ici de notre point de vue, c'est la pertinence de la démarche adoptée, qui doit être explicitée de bout en bout, et le fait que chaque choix doit être associé à une réflexion accompagnée d'hypothèses (p. 140).*
>
> *(...) en matière d'exploration de corpus, la méthode a pour vocation première d'objectiver et non pas d'illustrer l'analyse (p. 141).*
>
> *(...) toute visualisation produite doit être considérée avec précaution (...). On voit ici les limites de l'intuition, et la nécessité de bien s'approprier les principes de chaque méthode avant de se lancer dans une démarche exploratoire (p. 141).*

Dans les dernières lignes du chapitre (p. 142), les auteurs dressent le tableau des évolutions à venir par rapport aux démarches de structuration et d'exploration des données.

Chapitre 4 – *Exploration d'une hypothèse en corpus*

Le chapitre 4 donne des pistes pour ce que les auteurs appellent "l'exploration d'une hypothèse en corpus" ; il s'agit de se servir du corpus pour mettre à l'épreuve les idées que l'on se fait sur l'une ou l'autre caractéristique d'une langue.

La première partie est dédiée aux hypothèses de catégorisation des données. On y trouve des informations sur des méthodes permettant d'appréhender les caractéristiques générales d'un corpus. Il est possible de visualiser les données de diverses façons : on peut y appliquer une analyse factorielle des correspondances (AFC) ou faire une classification. Les deux fonctionnalités permettent d'avoir une idée des caractéristiques qui rapprochent ou séparent les différentes parties d'un corpus. La figure 1 fournit un exemple de classification appliquée au corpus *Erziehungsroman* [roman d'éducation] (Schaeffer-Lacroix, 2017). On peut y observer – sans surprise – un regroupement lexical des deux volumes de Johanna Spyri relatant les aventures de Heidi.

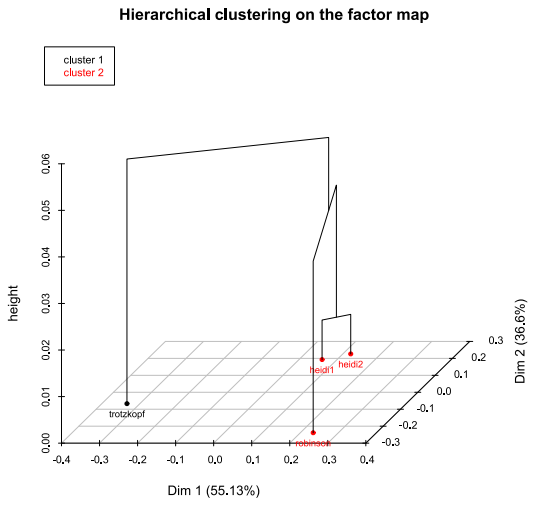

Figure 1 – Classification des mots à fréquence ≥ 10 dans le corpus Erziehungsroman (Schaeffer-Lacroix, 2017).

Dans le chapitre 4, l'arrière-plan statistique des différentes méthodes de catégorisation est expliqué en détail. Les auteurs abordent les questions suivantes : quels types de calcul pour quels types de recherches ? Comment interpréter les résultats ? Une distinction est faite entre l'approche des calculs en milieu francophone et en milieu anglo-saxon.

La deuxième partie du chapitre, plus abordable pour les non-statisticiens, présente les méthodes d'exploration d'une unité linguistique en corpus. Elle concerne la façon dont on peut faire des requêtes portant sur des formes uniques ou des groupes de formes. Des termes centraux sont expliqués de façon limpide, mis à part la distinction entre "cooccurrence", "collocation" et "colligation", que je ne trouve pas suffisamment claire dans ce chapitre (p. 205). Les auteurs mettent en avant le rôle primordial des segments répétés : ceux-ci seraient "les seules séquences adjacentes véritablement opératoires dans un parcours textométrique" (p. 197). J'ai découvert des informations intéressantes au sujet de ce que l'on appelle "motif". Le motif est défini comme un ensemble composé de plusieurs formes qui ne se suivent pas forcément et qui peuvent varier dans une certaine mesure (p. 200). Il peut en particulier intéresser les chercheurs travaillant sur des thématiques, par exemple en sociolinguistique.

La distinction entre recherches portant davantage sur la morphosyntaxe et recherches dans le domaine de la sémantique, mise en rapport avec l'observation de fenêtres de concordance plus ou moins grandes, est éclairante (p. 201).

Dans ce chapitre, j'apprécie, tout comme dans le précédent, la mention des évolutions futures, en l'occurrence dans le domaine de la représentation visuelle de corpus autres que simplement textuels et de la prise en compte plus universelle d'annotations autres que morphosyntaxiques.

Conclusion

En conclusion, les auteurs rappellent les principes qui sous-tendent la rédaction de l'ouvrage *Explorer un corpus textuel* et qui caractérisent leur positionnement général : des méthodes d'exploration de corpus répandues et ayant fait leurs preuves dans la communauté francophone et au-delà, le choix de logiciels libres, l'insertion dans des communautés de chercheurs comme le consortium *CORLI*, et une clarté dans leur propos pour donner aux lecteurs novices l'occasion de progresser dans un domaine en pleine expansion. J'apprécie le rappel de certains des débats encore actuels, en particulier celui autour du sens de l'annotation : "l'idée de décompter ce que l'on a annoté et déjà interprété, et de s'éloigner de la matérialité textuelle peut encore poser problème à certains" (p. 215 ; cf. *supra*). Les auteurs plaident pour une meilleure coopération entre communautés travaillant sur l'annotation et communautés spécialistes en exploration quantitative des données (p. 215). Le développement de formats standard, par exemple pour les annotations de type XML-TEI, est également considéré comme une urgence.

Dans cette publication, les auteurs privilégient la présentation de méthodes par rapport à la présentation d'outils. Le lecteur désireux d'approfondir ses connaissances sur l'un ou l'autre outil est invité à lire l'un des ouvrages et articles cités dans Poudat et Landragin (2017), par exemple, Widlöcher et Mathet (2009) concernant la plateforme *Glozz*. Pour le compte du groupe de travail *Exploration de corpus* du consortium *CORLI* (Corpus, Langues et Interactions), Céline Poudat est responsable de publication de synthèses descriptives attestant des usages d'outils d'exploitation de corpus par la communauté francophone[4]. Afin de compléter les suggestions des auteurs, je proposerais la lecture de James (2015) qui peut convenir aux personnes désirant connaître en profondeur le système de gestion de corpus *Sketch Engine* afin de l'utiliser avec des étudiants. Il s'agit, certes, d'un logiciel payant ;

[4] http://explorationdecorpus.corpusecrits.huma-num.fr/

toutefois, il a des vertus pédagogiques avérées (Schaeffer-Lacroix 2014 et 2015). Les chercheurs germanistes intéressés par la linguistique de corpus et la didactique peuvent consulter avec profit Káňa (2014). N'oublions pas de mentionner les autres moyens de s'informer sur l'un ou l'autre outil de façon exhaustive et parfois agréablement conviviale : je pense à des manuels en ligne (*CQPweb*), des listes de diffusion (*Corpus*, utilisateurs de *TXM*), à des chaînes *Youtube* (Corpus Workbench) et à des réseaux comme le wiki et le groupe *Facebook* de *TXM*.

Pour finir, je souhaite remercier Céline Poudat et Frédéric Landragin pour la publication de cet ouvrage qui soutient parfaitement une évolution que je considère comme nécessaire, à savoir la banalisation des méthodes et outils de corpus dans le monde de la recherche.

Références

James, Thomas (2017, 2ème édition). Discovering English with Sketch Engine. Brno : Versatile.

Káňa, Tomáš (2014). Sprachkorpora in Unterricht und Forschung DaF/DaZ. Brno : Masarykova univerzita. https://munispace.muni.cz/index.php/munispace/catalog/book/178 ; doi :10.5817/CZ.MUNI.M210-7748-201

Schaeffer-Lacroix, Eva (2014). "Utiliser des corpus numériques avec un public Lansad". Séminaire NumLangues. Université Montpellier, 12-13 décembre 2013. *Alsic*, vol. 17 | 2014. http://alsic.revues.org/2720 ; DOI : 10.4000/alsic.2720 13 pages.

Schaeffer-Lacroix, Eva (2015). "Analyse de trois systèmes de gestion de corpus pour l'enseignement-apprentissage des langues étrangères". Alsic, vol. 18, n° 1 | 2015. 24 pages. https://alsic.revues.org/2852

Widlöcher, Antoine & Mathet, Yann (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. Actes de la 16ème Conférence Traitement Automatique des Langues Naturelles (TALN'09), session posters, Juin 2009, Senlis. 10 pages. Disponible en ligne : https://hal.archives-ouvertes.fr/hal-01011969/document

Sites, outils et corpus

CNRTL (Centre National de Ressources Textuelles et Lexicales) (2012). Portail lexical. http://www.cnrtl.fr

Hardie, A. (2012). CQPweb - combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics, 17(3), 380-409.

CORLI (*CORpus, Langues, Interactions*). Consortium. https://corli.huma-num.fr/

Corpus Workbench (2012-2017). Chaîne *Youtube*. Tutoriels et aides à la prise en main des logiciels *CQP*, *CQPweb* et *BNCweb*. https://www.youtube.com/user/corpusworkbench

Kilgarriff, Adam, Rychly, Pavel & Pomikalek, Jan (nd). *Sketch Engine*. http://www.sketchengine.co.uk/

Schaeffer-Lacroix, Eva (2017). *Erziehungsroman*. Corpus en langue allemande contenant quatre romans d'éducation publiés sur *Gutenberg Project* :

- Spyri, Johanna (1880). *Heidis Lehr- und Wanderjahre* ;
- Spyri, Johanna (1881). *Heidi kann brauchen, was es gelernt hat* ;
- von Rhoden, Emmy (1885). *Trotzkopf* ;
- Schoppe, Amalia (1843). *Robinson in Australien*.