Comparaison diachronique de motifs récurrents dans deux encyclopédies

Alice Brenon, Denis Vigier, Ludovic Moncla, Frederique Laforest

Comparaison diachronique de motifs récurrents dans deux encyclopédies

Alice Brenon[1, 2], Denis Vigier[1], Ludovic Moncla[2] et Frédérique Laforest[2]
[1] ICAR, CNRS, UMR 5191, F-69342 [2] Univ Lyon, INSA Lyon, CNRS, UCBL, LIRIS, UMR 5205, F-69621
alice.brenon@insa-lyon.fr

Introduction

En choisissant le titre *Encyclopédie ou dictionnaire raisonné des sciences, des arts et des métiers* (*EDdA* dans ce qui suit), Diderot et d'Alembert inscrivent leur œuvre dans une double lignée. Ils reconnaissent d'abord l'héritage de la *Cyclopædia* de Chambers parue en 1728 à Londres et dont l'*EDdA* ne devait être initialement qu'une traduction (Kafker & Loveland, 2016, p. 111 et seq.). Ils reconnaissent ensuite celle des dictionnaires universels comme celui de Furetière qui inaugura cette lignée en 1690 ou comme le *Dictionnaire Universel François et Latin*, dit « *de Trévoux* » qui fut au long de ses éditions successives un grand rival de l'*EDdA* dont il critique le projet philosophique. On remarque toutefois qu'à la différence de la *Cyclopædia* qui arbore l'expression *Universal Dictionary* dans son sous-titre, les encyclopédistes français lui préfèrent l'adjectif *raisonné*, reflet de l'approche progressiste de ses auteurs, emblématique du siècle des Lumières.

Parmi les domaines que traite l'*EDdA*, la géographie figure en bonne position : elle est le domaine qui comporte le plus d'articles (environ 20% du total). Et pour cause, au XVIII^{ème} siècle, grand siècle d'exploration, la géographie est une science à la mode, encore largement rattachée aux mathématiques (ce qui est visible jusque dans le « Système figuré des connoissances humaines », placé en tête de l'ouvrage).

Plusieurs études dont Vigier et al. (2022) ont contribué à montrer qu'à cette époque les articles encyclopédiques traitant de géographie se différencient peu des autres productions du même domaine, qu'elles soient issues d'ouvrages généralistes comme les dictionnaires universels ou plus spécialisés comme par exemple *Le Grand Dictionnaire géographique* de Bruzen de La Martinière (1726-1739) ou le *Dictionnaire géographique-portatif* de Vosgien (1747).

Tout porte à croire en revanche que le profil générique et discursif des articles de géographie a changé sensiblement au cours du XIX^{ème} siècle, sous l'influence directe d'une évolution majeure des discours scientifiques de tous les domaines à cette époque. Il a en effet été montré (Jacquet-Pfau, 2022) que ce siècle — et surtout sa seconde moitié — a connu une « disciplinarisation » des savoirs du fait de l'importance grandissante des universités dans leur production. Ce fait peut s'observer par exemple dans la composition du groupe des rédacteurs de *La Grande Encyclopédie ou Inventaire raisonné des Sciences, des Lettres et des Arts* (ci-après *LGE*), une œuvre majeure de la fin du XIX^{ème} siècle (la parution débute en 1885 pour prendre fin en 1902) qui s'inscrit dans la tradition de l'*EDdA* (Jacquet-Pfau, 2015, p. 88 et seq.), ce qui est reflété jusque dans son titre complet.

Dans cette communication, qui s'inscrit dans le contexte du projet GEODE[1], nous voulons donc explorer l'hypothèse selon laquelle les articles encyclopédiques de géographie diffèrent

sensiblement à la fin du XIX^{ème} siècle des articles du même domaine publiés un siècle plus tôt dans l'*EDdA*.

Corpus d'étude

Pour tester cette hypothèse, nous nous proposons de comparer l'*EDdA* à cette seconde encyclopédie plus récente qu'est *LGE*. Nous utiliserons pour les articles de l'*EDdA* la version publiée par l'ARTFL[2] et diffusée sur son site web (Morrissey & Roe, 2022). Cette partie de notre corpus représente 17 tomes (les 17 volumes de texte de l'*EDdA*, auxquels s'ajoutent 11 volumes de planches que nous laissons de côté pour notre étude) et comprend au total 19,6 millions de mots dans 74 190 articles.

La version de *LGE* que nous verserons à notre corpus est issue des résultats du projet DISCO-

LGE[23] (Vigier & Brenon, 2021). La segmentation de ses 31 tomes est encore imparfaite mais elle comprend dans sa version actuelle 134 820 articles pour 56,6 millions de mots. Elle ne comporte pas de métadonnées sinon le titre des entrées, le tome dont elles proviennent et leur position dans la séquence des articles. Les logiciels développés pour réaliser ce découpage sont disponibles publiquement[24].

Afin de ne mesurer que les différences dans la manière d'écrire et pas dans les sujets traités, relevant davantage de la variation du périmètre même du domaine géographique que de changements stylistiques, nous nous concentrerons sur des articles décrivant les mêmes objets dans l'*EDdA* et dans *LGE*. Dans ce but, nous avons constitué un sous-corpus de notre corpus d'étude initial, que nous nommerons *Corpus Parallèle*, qui comprend les articles communs à l'*EDdA* et à *LGE*, c'est à dire les articles ayant la même vedette, unique à la fois dans l'*EDdA* et dans *LGE*, afin de limiter les ambiguïtés et donc les erreurs d'appariement. Ce corpus parallèle contient 7 215 paires d'articles de chacune des deux encyclopédies (soit 14 430 articles au total) pour un total de 13,2 millions de mots environ, avec un déséquilibre en faveur de *LGE* légèrement moins important que pour le corpus complet (8,7 millions contre 4,5 millions).

Transversalement, pour pouvoir également tenir compte de la variation diachronique des frontières entre domaines dans les deux encyclopédies, un deuxième sous-corpus a été défini en intégrant tous les articles portant sur une même thématique mais des objets potentiellement différents. Vigier et al. (2020, p. 7) ont repéré des motifs syntaxiques apparaissant préférentiellement en début d'articles et permettant d'identifier des « mots classifieurs ». À leur suite, nous avons donc constitué un corpus des hydronymes en choisissant les articles pour lesquels ce mot appartient au lexique des cours et plans d'eau (rivière, fleuve, lac, etc.). Ce deuxième sous-corpus contient 1354 articles de l'*EDdA* et 1917 de *LGE*, représentant 99k mots pour l'*EDdA* et 261k pour *LGE*.

Méthodologie

Afin de pouvoir contraster les articles de géographie avec ceux des autres domaines, il est nécessaire d'avoir un étiquetage des domaines de connaissances des articles. Un modèle de classification entraîné de manière supervisée sur l'*EDdA* pour un ensemble de domaines

[23] . https://www.collexpersee.eu/projet/disco-lge/

[24] . https://gitlab.huma-num.fr/disco-lge

communs aux deux encyclopédies et s'appuyant sur un modèle de langue pré-entraîné (Brenon et al., 2022, p. 6) pour classer les articles a été appliqué sur *LGE*. Le croisement de cette information avec les deux sous-corpus permettra à la fois d'interroger la pertinence de leur définition et les évolutions propres qu'a pu subir chaque domaine : si tous les hydronymes devraient relever de la géographie, le corpus parallèle contient probablement en revanche un certain nombre d'homonymes dont l'appariemment accidentel était indésirable mais sans doute aussi des examples d'objets dont une discipline s'est emparée au détriment d'une autre entre les deux époques.

La possibilité de partitionner le corpus d'étude complet par œuvre et par domaine ouvre la perspective de calculs de spécificités à l'aide d'outils comme TXM (Heiden, 2010). L'étude des coocurrences présentes dans les différentes parties du corpus révélera si les motifs présents dans l'*EDdA* le sont encore dans *LGE*.

Mais, faisant l'hypothèse que la « disciplinarisation » dont il est question plus haut s'accompagne d'une normalisation de la production écrite de chaque science passant possiblement par des routines discursives spécifiques, nous utiliserons également le Lexicoscope (Kraif, 2016) pour détecter des motifs syntaxiques plus profonds et peu visibles en restant au niveau de la surface des phrases.

Pour travailler ainsi au niveau de la syntaxe aussi bien que des parties de discours, nous avons annoté le corpus avec la librairie Python *Stanza* (Qi et al., 2020) pour le modèle French-GSD[25] utilisant le jeu d'étiquettes des *Universal Dependencies* (Marneffe et al., 2021).

Premiers résultats

Une étude préliminaire du corpus des hydronymes a permis de mettre en évidence dans *LGE* des tournures de phrases déjà apparue l'*EDdA*, comme le motif suivant utilisé fréquemment dans les articles : « ... prend sa source [...] et se jette [...] ».

Une différence mineure a cependant été constatée dans les réalisations de ce motif entre les deux œuvres : le sujet en est plus fréquemment le pronom relatif «qui» dans l'*EDdA*, ce qui permet de l'introduire dans la phrase nominale, souvent unique, qui définit le cours d'eau, alors qu'il fait davantage l'objet dans *LGE* d'une phrase séparée où il commence par le pronom personnel « il » ou «elle». De plus, certains articles dans lesquels le motif précédent n'a pas été trouvé possèdent toutefois une structure voisine où la description du cours d'eau a bien lieu de sa source à son embouchure mais au travers d'une chaîne de coréférence plus complexe pouvant occuper tout un paragraphe avec une plus grande variabilité lexicale au niveau des verbes employés.

Ces premiers résultats suggèrent qu'une évolution a eu lieu et amènent à s'interroger sur l'existence possible de tournures de phrases propres à l'une ou l'autre des deux œuvres du corpus, ou sur ce qui pourrait distinguer leur emploi dans celles qui leur sont communes. Sur le plan quantitatif il serait également intéressant de comprendre comment sont redistribués les volumes de mots entre les domaines de connaissance. Le phénomène se limite-t-il aux hydronymes ou concerne-t-il toute la géographie?

[25] . https://universaldependencies.org/treebanks/fr_gsd/

Remerciements

Les auteurs remercient le LABEX ASLAN (ANR-10-LABX-0081) de l'Université de Lyon pour son soutien financier dans le cadre du programme français "Investissements d'Avenir" géré par l'Agence Nationale de la Recherche (ANR).

Références bibliographiques

Brenon, A., Moncla, L., & Mcdonough, K. (2022). Classifying encyclopedia articles : Comparing machine and deep learning methods and exploring their predictions. Data and Knowledge Engineering, 102098. https://doi.org/10.1016/j.datak.2022.102098

Heiden, S. (2010). The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In R. Otoguro, K. Yoshimoto, K. Ishikawa, H. Umemoto, & Y. Harada (Eds.), 24th Pacific Asia Conference on Language, Information and Computation (Vol. 2, p. 389-398). Institute for Digital Enhancement of Cognitive Development, Waseda University. https://halshs.archives-ouvertes.fr/halshs-00549764

Jacquet-Pfau, C. (2022). Actualiser l'Encyclopédie de Diderot et d'Alembert au XIXe siècle : Le Grand Dictionnaire universel (1866-1890) et La Grande Encyclopédie (1885-1902) : Langue Française, N° 214(2), 95–109. https://doi.org/10.3917/lf.214.0095

Jacquet-Pfau, C. (2015). Élaboration et destinée d'une encyclopédie la fin du XIXe siècle : Les trente- et-un volumes de la grande encyclopédie, inventaire raisonné des sciences, des lettres et des arts par une société de savants et de gens de lettres. Éla. Études de Linguistique Appliquée, 177, 85–100. https://doi.org/10.3917/ela.177.0085

Kafker, F. A., & Loveland, J. (2016). André-François Le Breton, initiateur et libraire en chef de l'Encyclopédie. Recherches Sur Diderot Et Sur l'Encyclopédie, 51, 107–125. https://doi.org/10.4000/rde.5390

Kraif, O. (2016). Le lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus arborés. Cahiers de Lexicologie, 1(108), 91–106. https://doi.org/10.15122/isbn.978-2-406-06281-3.p.0091

Marneffe, M.-C. de, Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. Computatio- nal Linguistics, 47(2), 255–308. https ://doi.org/10.1162/coli$_{a0}$0402 Morrissey, R., & Roe, G. (2022). Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers, etc. University of Chicago : ARTFL Encyclopédie Project. https://encyclopedie.uchicago.edu/

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza : A Python natural language pro- cessing toolkit for many human languages. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations. https://nlp.stanford.edu/pubs/qi2020stanza.pdf Vigier, D., & Brenon, A. (2021). La Grande Encyclopédie ou Inventaire raisonné des Sciences, des Lettres et des Arts. 264458212, 185163346, 398759061, 1371168210. https://doi.org/10.34847/NKL.74EB1XFD

Vigier, D., Moncla, L., Brenon, A., Mcdonough, K., & Joliveau, T. (2020). Classification des entités nommées dans l'Encyclopédie ou dictionnaire raisonné des sciences des arts et des métiers par une société de gens de lettres (1751-1772). In Actes du 7ème Congrès Mondial de

Linguistique Française, Jul 2020, Montpellier, France. https://doi.org/10.1051/shsconf/20207811008

Vigier, D., Moncla, L., Lefort, I., Joliveau, T., & McDonough, K. (2022). Les articles de géographie dans le Dictionnaire Universel de Trévoux et l'Encyclopédie de Diderot et d'Alembert : Langue Française, N° 214(2), 59–80. https://doi.org/10.3917/lf.214.0059