Humanités numériques

9 | 2024
Varia

À la recherche des réseaux intertextuels : défis de la recherche littéraire à grande échelle

In Search of Intertextual Networks: the Challenges of Literary Research at Scale

Valentina Fedchenko, Dario Maria Nicolosi et Glenn Roe

Édition électronique
URL : https://journals.openedition.org/revuehn/3940
DOI : 10.4000/11wmw
ISSN : 2736-2337

Éditeur
Humanistica

Référence électronique
Valentina Fedchenko, Dario Maria Nicolosi et Glenn Roe, « À la recherche des réseaux intertextuels : défis de la recherche littéraire à grande échelle », *Humanités numériques* [En ligne], 9 | 2024, mis en ligne le 01 juin 2024, consulté le 24 janvier 2026. URL : http://journals.openedition.org/revuehn/3940 ; DOI : https://doi.org/10.4000/11wmw

humanités
numériques

9 | 2024
Varia

À la recherche des réseaux intertextuels : défis de la recherche littéraire à grande échelle

In Search of Intertextual Networks: the Challenges of Literary Research at Scale

Valentina Fedchenko, Dario Maria Nicolosi et Glenn Roe

Résumés

Cet article expose certains des défis qui ont émergé au cours des premières phases du projet *Modern*, programme de recherche financé par l'ERC (European Research Council, ou Conseil européen de la recherche) pour cinq ans, qui adopte une nouvelle approche partant des données (*data driven*) pour étudier l'histoire littéraire du siècle des Lumières. À partir d'un grand corpus de textes français du début de la période moderne, les auteurs détaillent les diverses étapes de la construction de réseaux intertextuels en se servant des résultats d'algorithmes de réutilisation de textes. De l'harmonisation du corpus et des métadonnées à l'entraînement d'un réseau neuronal pour filtrer les passages « bruités », cet article propose une chaîne de traitement pragmatique pour les projets similaires travaillant sur d'importantes collections de textes numérisés, tout en mettant en lumière les promesses ainsi que les périls de la recherche littéraire à grande échelle.

This article outlines some of the challenges that have arisen during the first phases of the ERC-funded *Modern* project, a five-year research programme that takes a new "data driven" approach to the literary history of the French Enlightenment. Drawing on a large curated corpus of French texts of the Early Modern period, the authors describe in detail the various steps for building intertextual networks using the output of text reuse algorithms. From corpus and metadata cleaning to training a neural network for filtering 'noisy' passages, this article provides a pragmatic

technical pipeline for similar projects working with massive collections of digitised text, highlighting both the promise and perils of conducting literary research at scale.

Entrées d'index

MOTS-CLÉS : littérature, intertextualité, analyse de réseaux, réutilisation textuelle, apprentissage automatique

KEYWORDS: literature, intertextuality, network analysis, text reuse, machine learning

Introduction

1 Le projet européen *Modern* (*Modelling Enlightenment : Reassembling Networks of Modernity Through Data-Driven Research* [1]) met en place un programme de recherche qui vise à identifier et à analyser les « traces intertextuelles » – emprunts, citations, mentions, références, allusions, etc. – du siècle des Lumières et les réseaux qu'elles cocréent à une échelle inédite. Jusqu'à présent, la plupart des études consacrées aux Lumières ont été circonscrites sur le plan de l'étendue ainsi que de l'envergure, se fondant soit sur une vaste quantité de données pour pallier un manque de profondeur interprétative, soit, à l'inverse, sur des études de cas singuliers servant à extrapoler des phénomènes d'une nature plus globale. *Modern* cherche à surmonter cette impasse méthodologique en élargissant la portée et l'échelle de l'archive littéraire numérique, d'une part, tout en incorporant, d'autre part, des méthodes de l'analyse de réseau et de la science des données pour interroger intelligemment cette nouvelle archive.

2 La construction du corpus *Modern*, afin d'y inclure des textes éphémères tels que des correspondances privées, des pamphlets, des journaux et d'autres œuvres non canoniques, nous offre l'occasion de retracer une palette bien plus étendue de pratiques intertextuelles que celles qui étaient traditionnellement envisageables (Kristeva 1969 ; Barthes 1984 ; Genette 1992). À titre d'exemple, les philosophes du XVIII^e^ siècle ont fait preuve d'une grande agilité dans leur appropriation et réappropriation d'auteurs, tant anciens que modernes, de manière largement implicite mais tacitement admise, en se basant sur les connaissances culturelles partagées du lecteur pour identifier ces emprunts (Edelstein, Morrissey et Roe 2013). La majeure partie de ces références demeure voilée aux yeux des lecteurs contemporains, pour qui ces informations ne font plus partie intégrante de leur culture générale.

3 De façon plus marquée, l'évolution et le raffinement de nombreux outils d'analyse textuelle, conjugués à la croissance continue de la disponibilité de vastes corpus numériques en libre accès, suscitent l'émergence de nouveaux projets en humanités numériques comme le nôtre [2]. Face à une telle abondance de données exploitables, de nombreuses questions de recherche émergent presque naturellement : quels textes du XVIII^e^ siècle étaient les plus fréquemment « cités » ? Quels auteurs se révèlent les plus

« influents », c'est-à-dire des auteurs dont les mots résonnent et circulent le plus dans les textes de cette époque ? Devant un corpus si vaste, peut-on remettre en question le canon des personnalités et des œuvres que deux siècles d'histoire littéraire nous ont transmis ? Parallèlement, d'authentiques défis méthodologiques se posent : à travers quels paradigmes d'organisation de la connaissance peut-on appréhender la complexité d'un phénomène tel que l'intertextualité [3] ? Quelles implications se cachent derrière certains choix de formalisation et de représentation ?

4 Guidés par les mots-clés qui dirigent nos intérêts, à savoir la diffusion et la circulation des idées (exprimées ici sous la forme spécifique du réemploi textuel), nous nous sommes naturellement orientés vers l'application des méthodologies de modélisation et d'analyse de réseau (SNA, *social network analysis*) dans le domaine des études littéraires. Depuis quelques années, plusieurs groupes de recherche ont en effet commencé à exploiter les potentialités de l'application de cette discipline au domaine des lettres. Des études portant sur les réseaux de correspondances jusqu'à la représentation graphique des interactions entre personnages dans un roman ou une pièce de théâtre, l'idée que de nombreux phénomènes, qu'ils soient en marge de la littérature ou qu'ils s'inscrivent dans ses structures essentielles, peuvent être décrits et analysés grâce aux outils heuristiques offerts par la SNA s'est progressivement répandue [4].

5 Malgré le succès de ces approches et leur adoption croissante, leur caractère relativement récent fait que, jusqu'à présent, nous n'avons pas identifié de procédures standardisées ni de réflexion partagée sur les implications épistémologiques de telles modélisations. Le chercheur, captivé par la perspective de travailler avec des corpus d'une telle ampleur et rassuré par les avancées en puissance de calcul, se retrouve cependant face à une terre inexplorée. Chaque pas en avant, chaque nouvelle intuition risque de remettre en question tout ce qui a été accompli précédemment, parfois au prix d'une grande dépense en ressources et en temps. C'est précisément pour cette raison que, fort de notre expérience, nous avons choisi de rendre compte, au sein de cet article, des défis techniques et des implications théoriques qui se posent lorsque l'on décide de se lancer dans un projet de repérage à grande échelle et d'étude de réemplois littéraires au sein de vastes corpus textuels, parfois issus de différentes campagnes de numérisation et présentant une qualité très variable.

6 Étant donné que notre travail est encore en cours, nous ne prévoyons pas de fournir des réponses techniques ou théoriques définitives. Notre intention est plutôt de décrire les problèmes les plus courants, les solutions que nous avons adoptées et leurs répercussions sur un plan théorique. Par exemple, nous interrogeons la nature même du réemploi, la possibilité de le concevoir en tant que réseau, ainsi que les implications que cela pourrait avoir. Dans l'article qui suit, nous aborderons la préparation d'expériences préliminaires visant à évaluer le corpus et les outils de repérage. Nous discuterons de la nécessité d'effectuer un filtrage sémantique des résultats lorsque l'on se confronte à un grand nombre de cooccurrences, ainsi que des implications sous-jacentes à une modélisation en réseau des échanges intertextuels. Notre objectif est de contribuer aux débats interdisciplinaires au sein de la communauté universitaire concernant ce genre d'études en proposant des procédures facilement

reproductibles avec d'autres logiciels et sur d'autres corpus, tout en mettant en lumière les concepts que nous considérons actuellement comme les plus complexes à démystifier.

Construction d'un corpus d'échantillonnage

7 Le corpus de textes sur lequel repose notre projet global résulte de la fusion de plusieurs collections indépendantes, issues de campagnes de numérisation distinctes menées à des moments historiques différents [5]. De la forme et du contenu des métadonnées jusqu'aux formats des textes et à la structure des fichiers XML-TEI, chaque collection était unique et ne suivait aucun standard : avant même d'entamer toute recherche, il fallait créer un échantillon de corpus sur lequel tester nos hypothèses et mettre en place une chaîne de traitement robuste, à déployer ensuite, avec les précautions requises, sur la totalité du corpus.

8 Parmi plusieurs critères de choix et de sélection, celui qui semble le plus efficace pour vérifier la méthode d'extraction de réemplois textuels est certainement la différenciation des exemplaires numérisés en fonction de leur qualité : face aux biais potentiels de l'OCR (reconnaissance optique de caractères, en anglais *optical character recognition*) lors du traitement de textes, nous avons préféré pour cet article nous limiter à un corpus restreint de textes au format XML-TEI, exempt d'erreurs de transcription automatique, car il était soumis à une supervision manuelle [6]. Ce corpus restreint représente environ 10 % des textes dont le projet dispose et il est composé de 3 385 documents pour un total de 87 721 080 mots.

9 Il convient de noter incidemment que, pour des raisons inhérentes aux domaines d'intérêt de notre discipline, les collections qui suivent cette pratique éditoriale (très coûteuse en temps et en ressources) conservent, dans la plupart des cas, des textes qui sont généralement reconnus comme faisant partie du canon littéraire, ce qui va à l'encontre de l'orientation de notre projet. Cependant, bien que ce choix se soit imposé pour des raisons pragmatiques et qu'il ne soit en fin de compte qu'un choix opérationnel, qui sera appliqué ultérieurement aux textes océrisés, il est toujours important de s'efforcer de garantir une certaine représentativité au sein du corpus d'échantillonnage.

10 Nous avons, par exemple, maintenu une certaine proportion entre les textes d'auteurs masculins et féminins. Parmi les 648 auteurs uniques qui composent notre corpus, 588 sont des hommes et 50 sont des femmes. Cette distribution, avec les 7,8 % de textes écrits par des autrices sur l'ensemble des données, correspond aux taux de répartition de la production littéraire entre les deux genres tels qu'indiqués dans le catalogue de la Bibliothèque nationale de France (BNF), lesquels n'excèdent pas 3,63 % pour la période 1685-1800 (figure 1), et que nous avons pris comme référence statistique [7].

Figure 1. Nombre d'auteurs masculins et féminins dans le catalogue général de la BNF entre 1685 et 1800

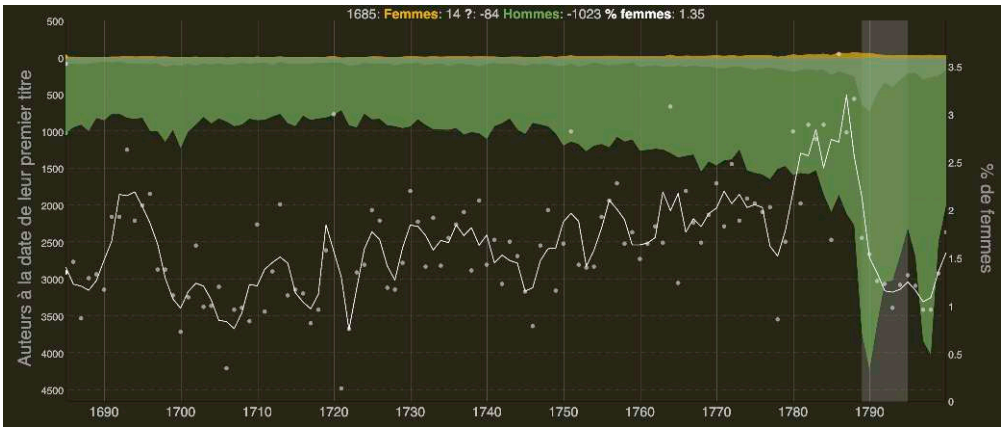

Image produite par les auteurs avec l'outil Cataviz (http://modern.huma-num.fr/cataviz/)

11 Par ailleurs, nous avons observé que l'une des principales explications du déséquilibre au sein de notre corpus, à savoir la prépondérance exceptionnelle des textes de Voltaire, en raison des choix éditoriaux de la Voltaire Foundation, source de nos textes (tableau 1), trouve également une correspondance dans les données statistiques du catalogue de la BNF.

Tableau 1. Les auteurs les plus représentés dans le corpus échantillonné

Auteur	# textes
Voltaire (1694-1778)	1 054
Carmontelle (1717-1806)	73
Dancourt (1661-1725)	51
Denis Diderot (1713-1784)	47
Pierre de Marivaux (1688-1763)	41
Jean-Jacques Rousseau (1712-1778)	40
Pierre Corneille (1606-1684)	39
Claude-Louis-Michel de Sacy (1746-1794)	36
Molière (1622-1673)	33
Paul Henri Dietrich baron d'Holbach (1723-1789)	28
Jean-François Regnard (1655-1709)	26
Philippe Quinault (1635-1688)	26
Marie-Catherine Le Jumel de Barneville baronne d'Aulnoy (1650-1705)	25
Stéphanie-Félicité Du Crest comtesse de Genlis (1746-1830)	23
Bernard de Fontenelle (1657-1757)	23
Jacques Bénigne Bossuet (1627-1704)	20

12 Durant les phases analytiques d'un projet, il est essentiel de prendre en considération les inégalités résultant des principes ayant guidé la constitution des corpus utilisés. Il convient de veiller à ce que ces inégalités ne soient pas simplement perçues comme un biais de construction, mais plutôt comme une disparité justifiée par la réalité éditoriale de l'époque, comme c'est le cas dans notre situation, en raison de l'immense succès qu'a connu le philosophe dès le XVIII^{e} siècle (figure 2).

Figure 2. Nombre d'éditions par auteur dans le catalogue général de la BNF entre 1685 et 1800

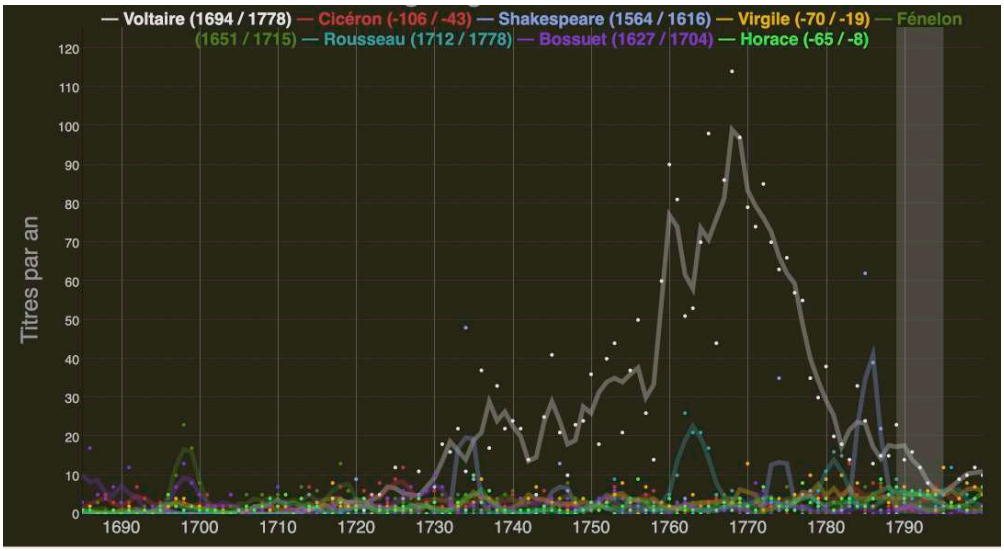

Image produite par les auteurs avec l'outil Cataviz (http://modern.huma-num.fr/cataviz/)

13 Enfin, le lecteur pourra être surpris de la présence de plusieurs écrivains du XVII^e^ français dans la liste des auteurs présents dans notre sous-corpus : comme tout projet de ce type, la définition des limites chronologiques à envisager est l'un des problèmes les plus épineux. En réalité, il s'agit d'une question qui touche l'ensemble des projets s'intéressant à la détection de réemplois textuels : bien que, parfois, la source d'une phrase puisse être facilement retracée, nous verrons plus en détail que très souvent, nous sommes confrontés à des cas douteux ou insaisissables, notamment lorsque, en restreignant trop les limites du corpus, nous excluons les œuvres des époques antérieures. Évidemment, l'exhaustivité est impossible et le choix d'exclure chronologiquement les textes en dehors de nos intérêts de recherche reste toujours justifiable. Cependant, si ces textes sont disponibles, leur exclusion demeure déconseillée, car souvent, en raison de leur importance reconnue ou d'un processus de canonisation déjà en cours à l'époque, ils contribuent de manière significative à la circulation et à la diffusion des idées. À titre d'exemple, bien que notre domaine de recherche principal reste le XVIII^e^ siècle et ses réseaux de cooccurrences, sans *Le Misanthrope* de Molière, deux textes qui s'y réfèrent apparaîtraient dans notre modèle comme s'entrecitant, alors qu'en réalité, ils font référence à une œuvre précédente.

14 Sans grande surprise, le choix d'un corpus d'échantillonnage doit être guidé par des impératifs de représentativité et de cohérence par rapport à l'ensemble des textes disponibles. Dans le cas de projets de cette nature, les conditions idéales impliquent la disponibilité d'un ensemble de textes de qualité fiable, qui permettent de vérifier les paramètres des outils de fouille, ainsi que la possibilité d'intégrer les résultats avec des œuvres d'époques antérieures. Cela permet immédiatement d'évaluer leur impact sur l'ensemble des résultats et d'en mesurer la spécificité.

Repérage des réemplois

15 Après avoir constitué notre jeu de données, il fallait donc choisir et tester des outils capables d'identifier les passages similaires au sein des textes de notre corpus. Aujourd'hui, plusieurs logiciels sont disponibles à cet effet, exploitant des techniques d'analyse de séquences développées dans des domaines très divers et ayant des applications allant du séquençage du génome à la détection du plagiat. Parmi les outils d'extraction de réemplois textuels disponibles en libre accès, nous avons considéré ceux qui utilisent des langages de programmation tels que R (R textreuse package [8]), Java (TRACER [9]), PHP/Perl (Tesserae [10]) et Python (Passim [11], BLAST [Basic Local Alignment Search Tool], un outil conçu pour l'étude des séquences d'ADN [12], et Text-PAIR [Pairwise Alignment for Intertextual Relations [13]]). Bien que ces outils offrent des fonctionnalités similaires, nous avons opté pour Text-PAIR, car il a été spécialement conçu pour répondre aux besoins de la recherche en études littéraires et s'appuie sur l'environnement PhiloLogic [14]. Celui-ci indexe et harmonise les métadonnées des fichiers XML-TEI, les organisant dans une base de données facilement interrogeable à l'aide d'un moteur de recherche. En sus, Text-PAIR est relativement facile à paramétrer et bien adapté à l'extraction des réemplois textuels partiels, ce qui est important pour le traitement des données océrisées.

16 Finalement, les défis liés à l'installation d'un environnement de travail auxiliaire nous ont semblé acceptables, car ils sont compensés par la capacité à lire et à extraire des informations, même lorsque la structure des fichiers est hétérogène ou présente des anomalies. Malheureusement, comme nous le verrons plus loin, c'est souvent le cas lorsque l'on traite un corpus issu de la fusion de différentes sources. Bien entendu, nos réflexions peuvent être appliquées à d'autres outils de détection de réemplois, tous fondés sur cette même logique.

17 Le principe de fonctionnement de Text-PAIR est le suivant : il divise chaque chaîne de mots qui compose un texte en plusieurs séquences de n-grammes [15], qui peuvent se chevaucher entre elles. Ensuite, il les stocke et les indexe en vue d'une comparaison systématique avec celles qui sont présentes dans les autres textes. Prenons par exemple l'incipit du *Contrat social* de Jean-Jacques Rousseau : « L'homme est né libre, et partout il est dans les fers. Tel se croit le maître des autres, qui ne laisse pas d'être plus esclave qu'eux. »

18 Supposons que nous souhaitons le diviser en séquences de trois mots (3-grammes) ; après la lemmatisation, la suppression des accents et l'élimination des mots-outils, nous obtiendrons plusieurs triplets de mots prêts pour la comparaison :

- homme_libre_partout
- libre_partout_fer
- partout_fer_croire
- fer_croire_maitre
- croire_maitre_laisser
- maitre_laisser_esclave

19 Text-PAIR identifie alors la coprésence, dans deux textes, d'au moins trois n-grammes similaires, sans prendre en compte de critères sémantiques ou contextuels. Cependant, bien que son fonctionnement soit intuitif, ses paramètres doivent tous être définis : la taille des n-grammes, leur flexibilité (par exemple, est-ce qu'une cooccurrence est établie même lorsqu'un ou plusieurs mots de deux chaînes ne coïncident pas ?), la taille du contexte relevé, etc. Tous ces éléments doivent être précisés pour les adapter à la typologie de documents et aux objectifs de recherche. Plus spécifiquement, le choix d'une taille de n-gramme appropriée semble déterminant, car il influence significativement la capacité à récupérer les réemplois significatifs : lorsque l'on utilise des 4-grammes ou des 5-grammes, leur dimension affecte la détection des occurrences les plus brèves. Avec des 2-grammes ou des 3-grammes, on repère des cas moins probants, des formules langagières qui, par conséquent, apparaissent très fréquemment. Au sein du corpus d'échantillonnage, nous avons sélectionné des textes que, grâce à notre connaissance de la réalité littéraire du XVIII^e^ siècle, nous savions mieux adaptés à une analyse ponctuelle des paramètres à choisir.

20 Nous avons opté pour une sélection du corpus tragique de Voltaire, comprenant quinze tragédies de *Mérope* (1743) à *Irène* (1778) et *Agathocle* (1778), au format XML-TEI corrigé [16]. Pour faciliter l'analyse du paramétrage de Text-PAIR, nous avons inclus le tome X du *Lycée, ou Cours de littérature ancienne et moderne* (1799) de Jean-François de La Harpe. Ce texte, fortement intertextuel, consiste en une critique étendue de la production littéraire de son époque, agrémentée d'une abondance de citations de longueurs variables. Le tome X est spécifiquement dédié au théâtre de Voltaire [17]. Ces tragédies étant en vers, la dimension des cooccurrences détectées entre les deux ensembles peut varier d'un simple hémistiche à plusieurs lignes, ce qui nous permet de mesurer aisément l'impact de la variation de la taille des n-grammes sur l'ensemble des résultats.

21 Lors de nos tests avec les tailles de séquence les plus courantes (2-, 3-, 4-, 5-grammes), nous avons obtenu les résultats suivants. Les 5-grammes ont été immédiatement exclus, car la plupart des citations d'un vers n'étaient pas détectées. De même, les 2-grammes ont été écartés, car ils introduisent trop de résultats peu utiles : malgré une augmentation exponentielle des cooccurrences retrouvées (+ 30 % par rapport au 3-grammes), les citations voltairiennes repérées de cette manière et non retracées par d'autres tailles de n-grammes se limitent à deux seuls hémistiches. La différence entre les 3-grammes et les 4-grammes apparaît beaucoup plus faible. Les 3-grammes ont détecté 13 % de cooccurrences en plus, dont toutefois seulement 33 % étaient de véritables citations, tandis que les autres s'avéraient des cas peu probants, des tournures langagières similaires. Face à ce choix, qui est en soi arbitraire, entre une plus grande exhaustivité et la génération de nombreux « faux positifs », nous avons décidé de réintroduire une variable significative : la présence d'un texte en OCR non corrigé, à savoir le tome X du *Lycée*, téléchargeable depuis Gallica [18]. En effet, disposant d'un corpus composé de textes numérisés et corrigés, mais également en version OCR non corrigée, il a fallu évaluer l'impact de la taille des n-grammes sur les résultats finaux. Or, indépendamment des n-grammes choisis, lorsque nous avons affaire à un texte uniquement en OCR, nous constatons une diminution du taux de détection global de 20 %. En sus, plus de la moitié des cooccurrences repérées

exclusivement par les 3-grammes étaient composées de citations de Voltaire, que nous aurions sinon perdues si nous avions utilisé les 4-grammes [19]. En fin de compte, malgré le grand nombre de résultats peu pertinents nécessitant un filtrage, les 3-grammes se sont avérés le compromis le plus efficace pour la détection des cooccurrences textuelles, compte tenu de la qualité inégale de notre corpus.

22 Une fois ce paramètre essentiel de Text-PAIR défini, la taille du contexte sélectionnée, etc., nous avons procédé à l'alignement et à la comparaison des textes du corpus d'échantillonnage entre eux. Dans 3 385 documents, ce processus a identifié 76 761 cooccurrences, extraites au format JSON et accompagnées de leurs métadonnées. Évidemment, la logique des réemplois suggère qu'un texte soit considéré comme la « source » et l'autre comme la « cible » d'un échange intertextuel, orienté intuitivement selon un critère chronologique, de l'œuvre la plus ancienne à la plus récente. Ainsi, 1 187 documents, soit 35 %, ont été identifiés comme sources d'au moins un réemploi, tandis que 1 024 ont été répertoriés comme cibles. Si le nombre relativement restreint de textes impliqués dans un échange intertextuel peut déjà être en soi une source d'intérêt (seulement un tiers des textes dans un corpus de bonne qualité, à l'encontre des théories d'intertextualité les plus extrêmes), la quantité de réemplois détectés les rend très difficilement analysables dans leur intégralité et la fiabilité de nos résultats demeure encore à tester. Pour y parvenir, la transformation des résultats sous forme de réseau, objectif ultime de cette typologie de recherche, peut également s'avérer un puissant instrument d'analyse préliminaire.

Les premiers graphes : l'incohérence constructive

23 Pour rendre nos résultats accessibles et analysables au sein d'un réseau, nous avons fait usage de la bibliothèque Python NetworkX [20], reconnue pour sa simplicité en matière de manipulation de données, de création et de visualisation de graphes. Une fois le graphe élaboré, nous avons calculé et extrait les mesures les plus couramment utilisées dans l'analyse de réseaux, évaluant ainsi leur cohérence. Bien que ces mesures ne soient pas l'objet principal de cet article, leur maîtrise dans ce type de projet s'avère cruciale en tant qu'outil analytique, ainsi qu'en tant que moyen de détection et de correction des éventuels problèmes susceptibles d'affecter le corpus et, par conséquent, les résultats obtenus.

24 En un mot, la modélisation des échanges intertextuels sous forme de graphe implique que les œuvres ou les auteurs soient représentés en tant que nœuds (dans le cas des auteurs, le nœud les symbolise en tant qu'ensemble de leurs textes) ; la cooccurrence textuelle est représentée par un lien qui relie ces deux nœuds ; ce lien aura une direction, car il indique une source et une cible ; de plus, il est pondéré, sa valeur variant en fonction du nombre de réemplois entre les deux nœuds concernés. Notre modèle se présente ainsi comme un graphe orienté et pondéré, dans lequel la direction et la valeur des liens ont un impact sur les mesures obtenues.

25 Lors de notre première analyse du graphe, il est apparu que la mesure de PageRank ne correspondait en aucun cas à nos attentes. Pour résumer, le PageRank d'un nœud est déterminé en tenant compte de « l'importance » des nœuds qui lui sont reliés, plus précisément dans un graphe

orienté, des nœuds qui pointent vers lui : un nœud aura un PageRank élevé s'il est la cible de nombreux liens et si les nœuds à partir desquels ces liens partent sont, à leur tour, ciblés par d'autres nœuds dotés d'un PageRank significatif [21].

26 Cependant, étant donné que le PageRank est l'une des mesures les plus couramment utilisées pour évaluer la valeur d'un nœud au sein d'un réseau, il est nécessaire de se questionner sur son application dans ce modèle spécifique. En effet, nous avons intuitivement défini un réemploi comme un contenu textuel (représenté par un lien orienté) qui « se propage » d'une source chronologiquement antérieure vers une cible qui le réutilise. Cela signifie que les nœuds représentant des textes et des auteurs avec un PageRank élevé sont ceux qui, étant la cible de nombreux liens, démontrent une tendance marquée à réutiliser des contenus textuels. Par exemple, si nous avions construit un graphe des réemplois entre Voltaire et La Harpe, le PageRank de ce dernier aurait été considérablement plus élevé que celui de Voltaire.

27 Néanmoins, cela contredit les objectifs de ce genre de recherche : au sein d'un réseau de circulation intertextuelle, une valeur accrue devrait être attribuée aux textes qui servent de point de référence pour les autres, ceux à partir desquels les idées, concepts et phrases se propagent. Il est impératif d'envisager une inversion graphique, bien que celle-ci puisse paraître insolite. Dans la représentation graphique d'un échange intertextuel, la source textuelle doit être perçue comme la cible du lien dans le modèle, à concevoir comme le nœud vers lequel tous les regards convergent, d'où son importance. À l'inverse, l'auteur réemployant deviendra le point d'origine de la connexion graphique : c'est de là que naît le processus de valorisation de la source réemployée. Cette inversion conceptuelle au sein du modèle affecte manifestement les autres mesures du graphe, car elles ressentent la direction des liens et s'ajustent en conséquence. Mais, une fois leur nouvelle signification définie, cela permet de rétablir le rôle central de PageRank dans la détermination des nœuds les plus influents.

28 Néanmoins, même avec cette inversion de la direction des liens, les résultats obtenus continuaient à poser des problèmes (tableau 2).

Tableau 2. Classement des auteurs par PageRank, résultats non filtrés

Auteur	PageRank
CORNEILLE, Pierre (1606-1684)	0,053321962
Jacques GRÉVIN (1538-1570)	0,043764032
DU RYER, Pierre	0,029404436
ROTROU, Jean (1609-1650)	0,02436457
Ronsard, Pierre de, 1524-1585	0,02290324
SCUDERY, Georges de (1601-1667)	0,022283164
Voltaire	0,021039766
LA CALPRENEDE, Gautier Costes de	0,018106744
Montesquieu, Charles de Secondat, baron de, 1689-1755.	0,015043999
Balthasar BARO (1600?-1650)	0,015032163
Balzac, Jean-Louis Guez, seigneur de, 1597-1654.	0,013312553
MARESCHAL, André	0,012388469
Latin Vulgate Bible	0,011967069

29 Plusieurs anomalies ont immédiatement attiré notre attention. Tout d'abord, une évidence saute aux yeux : les métadonnées des auteurs présentent toutes des normes différentes (nom, prénom ; prénom, nom ; avec ou sans date de naissance et de décès). Pour ne prendre qu'un exemple, cela signifie que les résultats liés aux cinq entrées de notre corpus consacrées au dramaturge Carmontelle n'étaient pas regroupés sous une seule et même étiquette cohérente : « CARMONTELLE, Louis Carrogis, dit Louis de Carmontelle (1717-1806) » ; « CARMONTELLE, Louis Carrogis de (1717-1806) » ; « Louis Carrogis de CARMONTELLE (1717-1806) » ; « CARMONTELLE? Louis Carrogis de (1717-1806) » ; « CARMONTELLE, Louis Carrogis, dit Louis de Carmontelle ». Deuxièmement, bien que l'on puisse discuter de l'influence de Corneille sur l'esthétique dramaturgique des Lumières, son classement nous semble surprenant, tout comme celui des auteurs qui le suivent et dont l'impact sur le XVIII^e^ siècle est généralement considéré comme faible (par exemple, Rotrou ou Ronsard) ; et si l'objectif de ce type d'analyse est précisément de remettre en question les connaissances partagées, une divergence aussi significative par rapport aux idées reçues nécessite des vérifications supplémentaires.

L'harmonisation des métadonnées et les doublons

30 Notre corpus résulte de la fusion de textes provenant de plusieurs collections distinctes et il souffre d'un problème très répandu dans notre domaine : étant conçus à des moments différents et en réponse à des objectifs de recherche spécifiques, chaque projet de numérisation adopte sa propre logique de classification, ce qui introduit une variabilité significative dans les métadonnées. Parfois, la source de cette variabilité n'est pas due au non-respect des normes communes, mais plutôt à la complexité du monde éditorial et littéraire. Évidemment, dans le cas de « l'affaire Carmontelle », il aurait pu être facilement résolu en suivant les normes standards comme Dublin Core, par exemple [22]. Cependant, par exemple, face à une œuvre imprimée en plusieurs volumes, se pose la question de savoir s'il convient de les regrouper en un seul document ou de les maintenir séparés. De même, la question se pose de savoir si les paratextes doivent être inclus ou exclus d'une édition numérique.

31 Face à la diversité des options, à la fois arbitraires et justifiables, il devient nécessaire de rechercher des méthodes automatisées ou semi-automatisées pour harmoniser les métadonnées. L'approche que nous avons adoptée combine trois méthodes numériques qui évaluent la similitude entre des chaînes de caractères. Les méthodes classiques couramment utilisées en traitement automatique des langues (TAL), telles que la distance de Levenshtein et la similarité cosinus [23], ont produit de bons résultats. Cependant, plusieurs cas incertains ont nécessité le recours à des techniques plus simples, comme la comparaison des deux mots les plus longs ou la suppression systématique de toute date. Si au moins deux de ces méthodes indiquaient que deux étiquettes étaient similaires, les notices correspondantes étaient fusionnées. Grâce à cette approche combinée, les métadonnées « Joseph Marie Loaisel de Tréogate » et « Loaisel de Tréogate, Joseph Marie, 1752-1812 » ont, par exemple, été identifiées comme faisant référence au même auteur et uniformisées.

32 Toutefois, il demeure essentiel de maintenir une surveillance rigoureuse. Pour rester dans le domaine de l'onomastique, si l'assimilation de « François de Salignac de La Mothe-Fénelon » avec « François Louis de Salignac de La Mothe-Fénelon » peut aisément être détectée et corrigée lors des analyses successives, étant donné l'importance considérable du premier dans l'histoire littéraire française par rapport à son neveu, la fusion des étiquettes des deux frères, « Jean Castilhon » et « Jean-Louis Castilhon », est plus difficile à détecter, avec des conséquences sur nos résultats. Finalement, et bien que le prétraitement automatique des métadonnées puisse grandement faciliter ce genre de tâches préliminaires, la validation par un expert disciplinaire se confirme comme un élément incontournable de toute recherche numérique.

33 Le même processus d'uniformisation a été aussi employé pour les titres des textes, regroupant sous une même étiquette les volumes d'une même œuvre, souvent distingués à travers la mention du numéro du tome. Pourtant, plusieurs titres coïncidaient parfaitement : la présence de doublons, problème typique de corpus construits en combinant des collections différentes, a émergé avec force. Dans ce type de projet, il s'agit d'un problème central. Pour réaliser une modélisation en graphe des échanges intertextuels, il est fondamental que chaque texte soit unique, car deux textes identiques auraient tendance à générer un nombre très élevé de cooccurrences entre eux, ce qui aurait un impact significatif sur les mesures les concernant, ainsi que sur celles des nœuds qui leur sont reliés.

34 Or, l'identification de ces doublons n'est pas si simple : la présence d'un même titre s'est révélée fréquente, mais demeure quand même un cas très particulier. En effet, les façons de désigner le titre d'un texte (avec ou sans sous-titres, avec ou sans ce qui suit les signes de ponctuation forts) peuvent varier considérablement. Il est nécessaire de prendre en compte les écarts découlant des pratiques éditoriales (publication anonyme ou pseudonyme, sous un faux-titre ou un titre modifié lors d'une nouvelle édition), qui introduisent des variantes que l'éditeur numérique peut choisir de respecter ou d'ignorer. Ainsi, une simple comparaison des métadonnées se révèle insuffisante.

35 Comme cela a déjà été remarqué, les systèmes d'extraction de réemplois textuels peuvent alors constituer une aide utile et servir à détecter les textes dupliqués et les rééditions au sein d'un corpus hétérogène [24]. En général, il s'agit de repérer les cooccurrences les plus longues, soit en dimension absolue (un réemploi de plusieurs milliers de mots est difficilement considérable en tant que tel) et relative (lorsque la dimension d'un réemploi dépasse un certain seuil, en pourcentage, de la longueur totale du document source ou cible). Encore une fois, les suggestions générées par les systèmes automatiques ont fait l'objet d'une analyse supervisée capable de trancher dans les cas ambigus. Par exemple, les anthologies ou les recueils encyclopédiques peuvent inclure de longs extraits sans pour autant être considérés comme des doublons.

36 Malgré l'harmonisation des métadonnées et la suppression des doublons, nos mesures de PageRank semblaient ne pas évoluer de manière significative, à l'exception des positions gagnées par un Carmontelle enfin unifié. Cela nous a conduits à approfondir notre réflexion et à rentrer dans le vif du sujet : qu'est-ce qu'un réemploi ?

Filtrage des résultats

37 En règle générale, les outils de repérage des réemplois textuels ne sont pas conçus pour évaluer le contenu de ces cooccurrences mais se contentent de détecter la présence simultanée de chaînes de mots similaires. Bien sûr, avec des outils comme Text-PAIR, il est parfois possible de spécifier à l'avance des listes de mots ou de phrases à filtrer car on sait déjà qu'ils seront peu utiles (par exemple, les formules épistolaires). Cependant, cela nécessite de prévoir en amont ce qu'il faut exclure, ce qui comporte le risque d'éliminer également des cas intéressants. Si l'on filtre par mots, on peut supprimer des occurrences utiles ; dans la phrase « Je suis, Seigneur, votre humble serviteur », par exemple, aucun mot ne peut être exclu sans risque. En revanche, les locutions récurrentes changent souvent de forme (par exemple, « Seigneur, je suis votre humble serviteur [...] » ou « Je suis, Monseigneur, [...] » ou « Je suis, Madame, [...] », etc.), ce qui les rend plus complexes à gérer.

38 Et pourtant, le filtrage de ce « bruit documentaire » est essentiel dans des projets comme le nôtre : si, dans des études ponctuelles, le chercheur peut se limiter à ignorer les résultats qui ne l'intéressent pas, une modélisation en réseau des échanges intertextuels à grande échelle perd une bonne partie de sa valeur si l'on tolère l'introduction incontrôlée et en grande quantité de connexions fautives. Si toutes les lettres de notre corpus apparaissent connectées par le seul biais de leurs formules de politesse, l'idée même d'une analyse de réseau perd son sens.

39 Alors, comment procéder ? En réalité, l'une des caractéristiques majeures de ce genre de « bruit » est sa tendance à être très « bruyant ». En d'autres termes, sa structure relativement fixe entraîne un effet d'échos significatif dans les résultats. Par exemple, si plusieurs missives A, B, C, etc., présentent la même formule de clôture, le nombre de connexions intertextuelles augmente de manière exponentielle (A est lié à B, A à C, B à C, et ainsi de suite). Pour gérer efficacement ce problème, il est nécessaire d'appliquer des techniques avancées de filtrage et d'analyse de texte. Cela peut inclure l'utilisation de méthodes de détection de motifs récurrents, l'identification de schémas de répétition, ou même l'application d'algorithmes d'intelligence artificielle pour reconnaître des structures textuelles spécifiques à exclure. En outre, l'expertise humaine demeure essentielle pour superviser et ajuster les processus de filtrage, en décidant quels motifs doivent être éliminés en fonction du contexte et des objectifs de l'analyse.

40 Dans nos résultats, les indications paratextuelles présentes dans les documents numériques se sont révélées particulièrement significatives, et pour cause. Au XVIII^e^ siècle, par exemple, toute édition légale d'une œuvre comportait la mention du « privilège royal », qui accordait à l'éditeur des droits exclusifs de publication, ou l'approbation de la censure, nécessaire pour l'impression [25]. À l'exception de la référence à l'édition spécifique, ces formules sont standardisées et sont signalées par Text-PAIR en tant que cooccurrences. Leur impact ne doit pas être sous-estimé : après avoir ajouté dans nos systèmes de filtrage d'autres structures répétitives identifiées empiriquement, comme les formules paratextuelles qui introduisent

les pièces de théâtre [26], des 76 761 cooccurrences initialement détectées, seules 14 520 se sont avérées des réemplois non affectés par le « bruit », ce qui équivaut à seulement 19 % des résultats initiaux.

41 Cette différence est tellement significative que l'incohérence de nos premiers résultats est devenue plus explicable. La mesure surprenante du PageRank de certains auteurs n'était pas due à un biais de l'historiographie littéraire mais plutôt à un biais de notre corpus : les textes des auteurs qui apparaissaient si haut dans le classement présentaient tous ces éléments paratextuels courants (par exemple, Corneille avec la mention du « privilège du roi » et les indications scéniques initiales). Une fois le filtrage effectué, un nouveau classement est apparu (tableau 3).

Tableau 3. Classement des auteurs par PageRank après filtrage

Auteur	PageRank
Latin Vulgate Bible	0,077028
Voltaire (1694-1778)	0,034277
Michel de Montaigne (1533-1592)	0,030131
Nouveau Testament	0,0294
La Sainte Bible	0,029175
La passion de Jésus-Christ	0,021726
Pierre Corneille (1606-1684)	0,021608
Jean Racine (1639-1699)	0,017759
Molière (1622-1673)	0,016921
Jean Bodin (1530-1596)	0,015486
Thomas Corneille (1625-1709)	0,015269
Clément Marot (1496-1544)	0,014557
Jacques Bénigne Bossuet (1627-1704)	0,012189

42 Il ne faut pas être expert de littérature du XVIII^e^ siècle pour comprendre que ces résultats sont beaucoup plus cohérents que les précédents. Nous constatons l'apparition de textes liturgiques ou d'inspiration religieuse (comme Marot ou Bossuet), dont les formules récurrentes génèrent un nombre exceptionnel de réemplois. Nous voyons également émerger les grands écrivains du XVI^e^ et du XVII^e^ siècle, constamment cités et commentés (il suffit de penser aux *Commentaires sur Corneille* de Voltaire ou au *Lycée* de La Harpe dont il a été déjà question). Le filtrage s'est avéré non pas une précaution excessive mais une étape essentielle pour assurer la fiabilité de l'analyse. Il est important de comprendre sur quelles bases et à l'aide de quelles méthodologies et quels outils numériques il a été effectué.

43 Afin de choisir la méthode de filtrage la plus efficace, il est nécessaire de déterminer les principaux problèmes du corpus et les caractéristiques générales de ses réemplois. Dans le nôtre, les considérations principales à prendre en compte sont les suivantes :

1. En raison de la dimension cosmopolite des Lumières, nos textes en français comportent de nombreuses citations ou expressions en d'autres langues, notamment en anglais, en italien et en latin. Par conséquent, le système de filtrage ne doit pas être spécifique à la langue du texte analysé.

2. La typologie des réemplois textuels varie à la fois du point de vue de la forme et du contenu. Leur longueur observée varie en effet de 3 mots minimum à 140 993 mots maximum (dimension du doublon le plus long repéré). En sus, la variabilité des significations potentielles est très étendue, ce qui rend impossible toute présélection préalable basée sur des critères sémantiques sans recourir à des techniques d'apprentissage automatique.

3. En même temps, afin de distinguer les réemplois significatifs de ceux qui relèvent d'une nature paratextuelle, il n'est pas pertinent de se baser uniquement sur des caractéristiques formelles telles que la longueur ou la morphologie. Il devient nécessaire de réintroduire une distinction typologique basée sur des critères sémantiques.

44 Parmi les différentes approches disponibles, nous avons mis en œuvre une procédure permettant d'obtenir des résultats satisfaisants, sans pour autant nécessiter de calculs excessivement coûteux ou complexes. Nous avons opté pour la création d'un classifieur binaire visant à filtrer les réemplois authentiques, en nous basant sur les plongements lexicaux extraits du modèle BERT (Bidirectional Encoder Representations from Transformers) multilingue [27]. Les modèles de langage comme BERT contiennent des vecteurs contextuels qui sont efficaces pour capturer la ressemblance sémantique. Selon nos évaluations, cette capacité était cruciale pour distinguer les réemplois authentiques des faux positifs [28]. Parmi les modèles spécifiques basés sur la structure de BERT, nous avons opté pour sa version *BertForSequenceClassification*, que nous avons affinée pour la tâche de classification de texte [29]. Nous avons effectué un peaufinage (*fine tuning*) du modèle de base en utilisant des données d'entraînement annotées pour une tâche de classification binaire spécifique, en ajustant les poids du modèle préentraîné pour l'adapter à la nouvelle tâche.

45 Le modèle prend en entrée des réemplois, extraits par Text-PAIR à partir du corpus d'échantillonnage, les prétraite, puis les encode à travers plusieurs couches de l'architecture du transformeur. Il génère ainsi une représentation vectorielle de la séquence d'entrée. Après la phase de peaufinage, cette représentation finale du réemploi est utilisée pour prédire une distribution de probabilité sur des catégories prédéfinies (« vrai réemploi » ou « faux réemploi »). La catégorie qui présente la probabilité la plus élevée est sélectionnée comme étiquette prédite pour le texte d'entrée. L'étape de peaufinage permet au modèle de mieux comprendre les caractéristiques distinctives des réemplois par rapport aux exemples faux positifs [30].

46 Pour affiner le classifieur pour la tâche de filtrage des réemplois, nous avons d'abord construit l'ensemble d'entraînement [31]. Pour constituer l'ensemble des réemplois faux positifs, nous avons utilisé l'extraction d'informations à l'aide de mots-clés tels que « Par la grâce et Privilège du Roy », « lu par ordre de Monseigneur le Chancelier », etc. Nous avons adopté une approche pragmatique en alimentant progressivement l'ensemble d'entraînement. Après chaque cycle d'entraînement du modèle, la sortie du classifieur était vérifiée manuellement, en particulier pour les réemplois pour lesquels le modèle donnait des probabilités basses d'appartenance à l'une des deux classes (faux ou vrai réemplois). Les réemplois avec des probabilités élevées étaient ajoutés à l'ensemble d'entraînement.

47 Après le troisième réentraînement du modèle, les résultats liés au paratexte ont été efficacement filtrés, comme nous l'avons déjà décrit. Cependant, de nouveaux cas problématiques ont émergé, cette fois-ci liés au contenu des textes. Le langage humain, voire littéraire, se caractérise par un certain degré de répétitivité, avec l'utilisation de formules stylistiques ou de périphrases récurrentes que la machine ne peut que détecter comme des cooccurrences. Il est essentiel d'identifier ces éléments et de tenter de les traiter, car leur influence sur les résultats peut être très significative.

Tout texte est-il un intertexte ?

48 En soi, la nature répétitive d'un fragment textuel n'est pas un critère décisif pour l'écarter de nos résultats. Citer le début du *Pater Noster*, qui tire justement de son caractère immuable tout son sens, est logiquement différent d'utiliser l'expression « Le théâtre représente [nom de la ville/région] [...] La scène est à [lieu de l'action] », où la cooccurrence dérive de l'adoption d'une structure éditoriale fixe. Mais que peut-on dire de cas très fréquents tels que « baptisa au nom du Père, et du Fils, et du Saint Esprit », ou du triplet canonique de louanges « bon citoyen, bon époux et bon père » ? Il s'agit d'expressions pour lesquelles le nombre élevé d'occurrences nous suggère leur nature « figée », alors qu'*a priori* il serait très compliqué de la leur attribuer.

49 Il est évident que la machine, toute seule, ne peut pas suffire, et que le chercheur doit faire des choix explicites et spécifiques en fonction de ses intérêts de recherche. Dans un projet de ce type, il semble impossible de prévoir tous les cas potentiels et la construction d'une typologie de réemplois est d'autant plus nécessaire qu'elle est soumise aux aléas des résultats obtenus. Une véritable recherche *data driven*, telle que celle entreprise dans ce projet, est orientée par les données repérées, avec une marge de prédiction limitée. Cependant, il est possible de faire quelques observations d'ensemble, afin que l'on puisse, sans pouvoir donner une règle générale ou déterminer une taxonomie définitive, avancer quelques observations théoriques pour servir comme point de départ pour la recherche.

50 Ainsi, nous pouvons creuser davantage la différence entre l'occurrence d'une prière et l'incipit traditionnel d'une édition d'un texte dramaturgique. Tout d'abord, ce qui semble distinguer les deux exemples est « l'univocité » du processus de coréférence. Dans le premier cas, nous pouvons présumer une volonté précise de l'auteur qui décide de citer spécifiquement cette prière et pas une autre. Le choix du sujet « remployant » est clair et net, et il est impossible de douter de sa volonté de se référer à ces mots en particulier. En revanche, l'annonce du lieu de l'action théâtrale n'est qu'un élément formel parmi d'autres et son sens spécifique réside justement dans l'indication du lieu, qui peut varier indéfiniment, remplaçant ainsi les crochets.

51 Cependant, baser notre raisonnement sur la catégorie de l'intention de l'auteur peut être trompeur. Lorsqu'un auteur cite le *Pater Noster*, nous sommes certains de son intention de faire référence à une proposition spécifique. Mais avec qui établit-il un rapport intertextuel ? Est-ce directement avec les Évangiles ou plus généralement avec une formule tellement répandue qu'il semble impossible d'établir une dépendance textuelle ? Il

est évidemment possible de se limiter aux cas clairement intentionnels, c'est-à-dire aux citations explicites où l'auteur cité est indiqué et qui seront identifiables dans de vastes ensembles de données en utilisant des outils de repérage d'entités nommées, par exemple. Cependant, ce type d'occurrence est très rare, plusieurs facteurs pouvant expliquer l'absence de ces caractéristiques formelles (par exemple, lorsqu'on s'adresse à un auteur en utilisant des périphrases). Une telle limitation réduirait de manière significative le nombre de cooccurrences analysables.

52 En général, l'intention d'un auteur de se référer à un contenu verbal singulier n'est pas toujours accompagnée d'une « volonté intertextuelle » certaine, c'est-à-dire du dessein de se référer à un texte particulier. Si l'échange se produit indéniablement, l'absence de l'un de ses deux pôles (celui du texte remployé) compliquera nos catégories. Prenons un autre exemple, cette fois-ci séculaire (tableau 4).

Tableau 4. Exemple d'une cooccurrence textuelle

G. Daniel, Voyage du monde de Descartes, 1690	J.-J. Barthélemy, Voyage du jeune Anacharsis, 1788
L'étendue de chacun est aussi proportionnée à l'excellence de sa nature : ils ont partagé, comme frères, les quatre qualités : ils en ont chacun deux, dont ils en possèdent une dans le souverain degré. **La terre est froide et sèche ; l'eau est froide et humide ; l'air est chaud et humide ; et le feu est** chaud et sec.	Aux quatre éléments sont attachées quatre propriétés essentielles : froideur, chaleur, sécheresse et humidité. Les deux premières sont actives, les deux secondes passives ; chaque élément en possède deux : **la terre est froide et sèche ; l'eau, froide et humide ; l'air, chaud et humide ; le feu**, sec et chaud.

53 Si l'on exclut l'inversion des deux derniers adjectifs associés à l'élément « feu », les deux phrases forment une cooccurrence. Pourtant, il serait très difficile d'argumenter que l'abbé Barthélemy est ici en train de « citer » Gabriel Daniel. La description des caractères des quatre éléments de la matière, héritée de la physique antique, est tellement commune que les deux écrivains la rapportent machinalement. Il serait impossible, voire erroné, d'en déduire un rapport univoque de filiation directe entre les deux.

54 Nous touchons aux limites ultimes de la relation intertextuelle, selon la définition désormais classique de Roland Barthes (1973) :

> Tout texte est un intertexte ; d'autres textes sont présents en lui, à des niveaux variables, sous des formes plus ou moins reconnaissables : les textes de la culture antérieure et ceux de la culture environnante ; tout texte est un tissu nouveau de citations révolues. Passent dans le texte, redistribués en lui, des morceaux de codes, des formules, des modèles rythmiques, des fragments de langages sociaux, etc., car il y a toujours du langage avant le texte et autour de lui. L'intertextualité, condition de tout texte, quel qu'il soit, ne se réduit évidemment pas à un problème de sources ou d'influences ; l'intertexte est un champ général de formules anonymes, dont l'origine est rarement repérable, de citations inconscientes ou automatiques, données sans guillemets.

55 Faut-il donc conclure que cette typologie de recherche est, en soi, destinée à l'échec ? Pas nécessairement. Au contraire, cette prise de conscience peut s'avérer productive et stimuler des astuces techniques ou des réflexions théoriques. Il existe différentes stratégies pour affiner la détection des réemplois textuels. Par exemple, on peut augmenter le nombre de 3-grammes requis pour identifier un réemploi, en partant du principe qu'une cooccurrence plus longue a moins de chances d'être due au hasard

ou à une réminiscence inconsciente. Cependant, il convient de noter que ce dernier exemple met en évidence les limites de cette approche. Une autre possibilité consiste à entraîner davantage le modèle de filtrage en lui apprenant à classer les résultats en fonction de ressemblances sémantiques plus subtiles. Par exemple, les réemplois liés à l'univers religieux pourraient être regroupés dans un cluster spécifique et exclus du jeu de données s'ils sont considérés comme peu significatifs. Enfin, si des ressources humaines importantes sont disponibles, il est envisageable de lancer une campagne d'annotation manuelle visant à établir une taxonomie commune permettant de distinguer et de trier les différentes typologies de cooccurrences.

56 Cependant, en adoptant une perspective plus générale, il est possible de réfléchir au sens profond de notre modélisation et de prendre conscience de ses implications. Tout d'abord, la distinction des réemplois en fonction de leur « référentialité » (ce qui nous a conduits à préférer le *Pater Noster* aux indications scéniques) semble être un point de départ solide. Cette approche nous libère de la contrainte de l'intentionnalité de l'auteur : que la source soit consciente ou qu'il s'agisse d'un élément appartenant à l'héritage culturel de l'auteur, la coréférence d'une même formule liturgique ou d'une tournure linguistique très particulière demeure un fait objectif qui relève d'un horizon de connaissances et de propositions verbales définies. De même, si la décision de représenter graphiquement les échanges intertextuels en tant que liens orientés entre deux nœuds ouvre la voie à l'exploitation et à l'amélioration des potentialités analytiques de l'analyse de réseau, il faut avoir conscience des erreurs de perspective qu'elle peut engendrer.

57 En effet, est-ce qu'un lien entre deux nœuds représente toujours une interconnexion explicite, un fait historique du type « B a lu A », ou « B cite A » ? Nous avons vu que, sauf indication explicite, un réemploi ne peut jamais se dire une citation directe d'un texte précédent. Mais il ne s'agit pas de la seule contrainte envisageable. Considérons les textes A et B, respectivement précédent et successif d'un point de vue chronologique, partageant un même contenu textuel et donc graphiquement connectés : mais se peut-il que B ait lu A, ou qu'il se soit plutôt appuyé sur un autre texte, Y, agissant comme médiateur, mais qui, étant absent du corpus, demeure invisible à l'analyse ? Et même si l'on dispose du texte Y, est-il possible de choisir entre une reconstruction de l'échange intertextuel du type « B a lu Y qui a lu A », ou y a-t-il d'autres combinaisons qui seront logiquement envisageables, comme par exemple « B et Y ont lu, de manière indépendante, A » ? Et que faire lorsque la source originelle d'un réemploi est absente ? Si A cite une phrase en italien de la *Jérusalem délivrée* du Tasse – poème épique du XVI^e^ siècle, très apprécié au XVIII^e^ siècle en France, mais exclu par définition de notre corpus – tout réemploi ultérieur de la même expression ne pourra que se traduire, dans notre représentation graphique, par des liens qui pointent vers A en tant que source principale, même si cela n'est pas historiquement fondé.

58 Il est évident que contrairement à d'autres réseaux, tels que les réseaux d'échanges épistolaires liés à une pratique concrète de communication, tout réseau intertextuel n'est qu'une modélisation abstraite. Il ne prétend pas décrire un réseau « historiquement donné » d'emprunts ou de citations, mais se limite à détecter des cooccurrences textuelles dans le but de décrire les parcours de circulation de certains concepts ou idées. Circula-

tion qui s'est faite plus tortueusement que ce qu'un lien direct entre deux nœuds peut suggérer et dont nous ne gardons finalement que quelques points singuliers, mais qui profite de la formalisation graphique en réseau pour être enfin analysée dans son ensemble, selon des critères quantifiables et à grande échelle.

Conclusions et prochaines étapes

59 Compte tenu des réserves exprimées ci-dessus, il est peut-être nécessaire, pour les étapes à venir du projet, de mettre en valeur de manière plus approfondie nos bases, tant méthodologiques que théoriques. Cela n'est qu'un début. Afin de pallier les lacunes un peu trop quantitatives de l'analyse traditionnelle des réseaux, nous pourrions nous orienter par exemple vers la théorie de l'acteur-réseau, la *actor-network theory* (ANT) notamment promue par Bruno Latour (2005). Développée au cours des trente dernières années par des chercheurs dans le domaine des études des sciences et des techniques, l'ANT est une approche des réseaux qui redéfinit les acteurs (ou actants) non pas en tant qu'agents volontaires ou intentionnels, mais plutôt en tant qu'entités – humaines ou non humaines – qui influencent de quelque manière l'activité d'un système sociotechnique. Les implications de l'ANT pour notre projet nous permettront de mieux retracer des constellations hybrides et hétérogènes de textes (et d'auteurs) tels qu'ils apparaissent dans nos données. Ces constellations, ou réseaux, seront considérées comme des ontologies relationnelles qui établissent des liens entre les acteurs. Les intertextes qui émergeront de ces réseaux seront qualifiés d'« influents » précisément parce qu'ils créent, ou cocréent, des liens puissants et durables à travers l'espace et le temps. La nature de ces liens (sociale, affective, esthétique, politique) et la manière dont ils sont établis (techniques d'invention, d'emprunt, de contestation, de réfutation et de diffusion) deviendront les éléments qualitatifs clés de la méthodologie de recherche du projet *Modern*.

60 D'un point de vue quantitatif, l'approche *data driven* de *Modern* sera peut-être en adéquation avec l'« empirisme radical » de l'ANT, dont l'accent sur la contingence et la variabilité des associations et des assemblages laisse place à des résultats inattendus (Latour 1993). En combinant des mesures quantitatives de graphes avec l'apport qualitatif des spécialistes de la littérature, nous nous efforcerons d'identifier, de reconstituer et d'examiner les réseaux tels qu'ils se dessinent à partir des données, sans *a priori* historique ou théorique. Cette approche nous permettra de progresser de manière itérative, à travers plusieurs étapes de raffinement et d'ajustement des modèles, afin d'acquérir une meilleure compréhension des forces d'influence à l'œuvre au sein de nos réseaux. De ce fait, l'influence se révèle un concept heuristique central pour ce projet, en résonance avec l'usage récent de l'ANT par Rita Felski en littérature comparée (2016). Selon elle, certains textes et auteurs connaissent un succès empirique supérieur à d'autres, non seulement sur le plan esthétique ou institutionnel, mais en grande partie grâce à des réseaux d'acteurs hybrides qui perdurent au fil du temps grâce à des alliances, de la diplomatie, des compromis, des persuasions, des réfutations, des négociations, etc. Pour détecter et analyser ces indicateurs d'influence, *Modern* se consacrera dans les années à venir à cartographier les réseaux d'acteurs au sein de ses

vastes corpus, ainsi qu'à étudier les mécanismes par lesquels ses intertextes ont été médiatisés, réutilisés, transformés et enrichis lors de leur circulation de nœud en nœud, ou d'acteur en acteur.

CONTRIBUTIONS

Valentina Fedchenko : curation des données, analyse formelle, méthodologie, développement informatique, rédaction – version originelle

Dario Maria Nicolosi : conceptualisation, recherche, méthodologie, rédaction – version originelle

Glenn Roe : conceptualisation, obtention du financement, recherche, méthodologie, supervision, rédaction – révision et correction

(Contributor Roles Taxonomy, Credit)

Bibliographie

Ahnert, Ruth, Sebastian E. Ahnert, Catherine Nicole Coleman et Scott Weingart. 2020. *The Network Turn. Changing Perspectives in the Humanities*. Cambridge : Cambridge University Press.

Armstrong, Elizabeth. 1990. *Before Copyright. The French Book-Privilege System 1498-1526*. Cambridge : Cambridge University Press.

Barthes, Roland. 1973. « Texte (théorie du) ». Dans *Encyclopædia Universalis* 15 : 1013-1020.

Barthes, Roland. 1984. *Le Bruissement de la langue*. Paris : Le Seuil.

Büchler, Marco, Philip R. Burns, Martin Müller, Emily Franzini et Greta Franzini. 2014. « Towards a Historical Text Re-Use Detection ». Dans *Text Mining. From Ontology Learning to Automated Text Processing Applications*, édité par Chris Biemann et Alexander Mehler, 221-238. Cham : Springer. https://doi.org/10.1007/978-3-319-12655-5_11.

Buscaldi, Davide, Ghazi Felhi, Dhaou Ghoul, Joseph Le Roux, Gaël Lejeune et Xudong Zhang. 2020. « Calcul de similarité entre phrases : quelles mesures et quels descripteurs ? ». Dans *Actes de la 6e conférence conjointe Journées d'études sur la parole (JEP, 33e édition), Traitement automatique des langues naturelles (TALN, 27e édition), Rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues (RÉCITAL, 22e édition). Atelier Défi fouille de textes*, édité par Rémi Cardon, Natalia Grabar, Cyril Grouin et Thierry Hamon, 14-25. ATALA et AFCP. https://aclanthology.org/2020.jeptalnrecital-deft.2.

Coffee, Neil, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde et Sarah L. Jacobson. 2013. « The Tesserae Project : Intertextual Analysis of Latin Poetry ». *Literary and Linguistic Computing* 28 (2) : 221-228. https://doi.org/10.1093/llc/fqs033.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee et Kristina Toutanova. 2019. « BERT : Pre-Training of Deep Bidirectional Transformers for Language Understanding ». arXiv. https://doi.org/10.48550/arXiv.1810.04805.

Edelstein, Dan, Robert Morrissey et Glenn Roe. 2013. « To Quote or not to Quote : Citation Strategies in the *Encyclopédie* ». *Journal of the History of Ideas* 74 (2) : 213-236. https://www.jstor.org/stable/43291299.

Edmondson, Chloe et Dan Edelstein, éd. 2019. *Networks of Enlightenment. Digital Approaches to the Republic of Letters*. Oxford : Voltaire Foundation.

Felski, Rita. 2016. « Comparison and Translation : a Perspective from Actor-Network Theory ». *Comparative Literature Studies* 53 (4) : 747-765. https://doi.org/10.5325/complitstudies.53.4.0747.

Franzini, Greta, Marco Carlo Passarotti, Maria Moritz et Marco Büchler. 2019. « Using and Evaluating TRACER for an Index Fontium Computatus of the Summa Contra Gentiles of Thomas Aquinas ». Zenodo. https://doi.org/10.5281/zenodo.3362130.

Genette, Gérard. 1992. *Palimpsestes. La littérature au second degré*. Paris : Le Seuil.

Kristeva, Julia. 1969. *Sēmeiōtikē. Recherches pour une sémanalyse*. Paris : Le Seuil.

Labatut, Vincent et Xavier Bost. 2019. « Extraction and Analysis of Fictional Character Networks : a Survey ». *ACM Computing Surveys* 52 (5) : 1-40. https://doi.org/10.1145/3344454.

Latour, Bruno. 1993. *Aramis ou l'amour des techniques*. Paris : La Découverte.

Latour, Bruno. 2005. *Reassembling the Social. An Introduction to Actor-Network-Theory*. Oxford : Oxford University Press.

Liu, Qi, Biao Xiang, Nicholas Jing Yuan, Enhong Chen, Hui Xiong, Yi Zheng et Yu Yang. 2017. « An Influence Propagation View of PageRank ». *ACM Transactions on Knowledge Discovery from Data* 11 (3) : 1-30. https://doi.org/10.1145/3046941.

Mahadevan, Ananth, Michael Mathioudakis, Eetu Mäkelä et Mikko Tolonen. 2024. « Optimizing a Data Science System for Text Reuse Analysis ». arXiv. https://doi.org/10.48550/arXiv.2401.07290.

Mullen, Lincoln. 2020. « Detect Text Reuse and Document Similarity ». Textreuse. https://docs.ropensci.org/textreuse/.

Olsen, Mark, Russell Horton et Glenn Roe. 2011. « Something Borrowed : Sequence Alignment and the Identification of Similar Passages in Large Text Collections ». *Digital Studies/Le champ numérique* 2 (1). https://doi.org/10.16995/dscn.258.

Romanello, Matteo et Simon Hengchen. 2021. « Detecting Text Reuse with Passim ». *Programming Historian*, mai. https://programminghistorian.org/en/lessons/detecting-text-reuse-with-passim.

Rosson, David, Eetu Mäkelä, Ville Vaara, Ananth Mahadevan, Yann Ryan et Mikko Tolonen. 2023. « Reception Reader : Exploring Text Reuse in Early Modern British Publications ». *Journal of Open Humanities Data* 9 : 5. https://doi.org/10.5334/johd.101.

Salmi, Hannu, Petri Paju, Heli Rantala, Asko Nivala, Aleksi Vesanto et Filip Ginter. 2020. « The Reuse of Texts in Finnish Newspapers and Journals, 1771-1920 : a Digital Humanities Perspective ». *Historical Methods : a Journal of Quantitative and Interdisciplinary History* 54 (1) : 14-28. https://doi.org/10.1080/01615440.2020.1803166.

Samoyault, Tiphaine. 2005. *L'Intertextualité. Mémoire de la littérature*. Paris : Armand Colin.

Smith, David, Ryan Cordell et Elizabeth Dillon. 2013. « Infectious Texts : Modeling Text Reuse in Nineteenth-Century Newspapers ». Dans *Proceedings of the 2013 IEEE International Conference on Big Data*, édité par Xiaohua Hu, Tsau Young Lin, Vijay Raghavan, Benjamin Wah, Ricardo Baeza-Yates, Geoffrey Fox, Cyrus Shahabi, Matthew Smith, Qiang Yang, Rayid Ghani, Wei Fan, Ronny Lempel et Raghunath Nambiar, 86-94. New York : IEEE. https://doi.org/10.1109/BigData.2013.6691675.

Smith, David, Ryan Cordell et Abby Mullen. 2015. « Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers ». *American Literary History* 27 (3) : E1-E15. https://doi.org/10.1093/alh/ajv029.

Tharsen, Jeffrey et Clovis Gladstone. 2020. « Using Philologic for Digital Textual and Intertextual Analyses of the Twenty-Four Chinese Histories 二十四史 ». *Journal of Chinese History 中國歷史學刊* 4 (2) : 558-563. https://doi.org/10.1017/jch.2020.27.

Vesanto, Aleksi, Asko Nivala, Heli Rantala, Tapio Salakoski, Hannu Salmi et Filip Ginter. 2017. « Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771-1910 ». Dans *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, édité par Gerlof Bouma et Yvonne Adesam, 54-58. Gothenburg : Linköping University Electronic Press. https://aclanthology.org/W17-0510.

Wu, Zhengxuan et Desmond C. Ong. 2021. « On Explaining Your Explanations of BERT : an Empirical Study With Sequence Classification ». arXiv. https://doi.org/10.48550/arXiv.2101.00196.

Notes

1 Projet de recherche financé par l'Union européenne (ERC Consolidator Grant 101043369). Les points de vue et opinions exprimés sont toutefois ceux des auteurs et ne reflètent pas nécessairement ceux de l'Union européenne ou de l'Agence exécutive du Conseil européen de la recherche. Ni l'Union européenne ni l'autorité subventionnaire ne peuvent en être tenues pour responsables [*Research funded by the European Union (ERC Consolidator Grant 101043369). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them*].

2 Plusieurs projets similaires, se servant de la détection automatique des réemplois (*text reuse detection*) pour fouiller dans des corpus massifs ont vu le jour au niveau international. On mentionne, entre autres, le projet *Viral texts* sur la presse américaine du XIX^e^ siècle (Smith, Cordell et Dillon 2013 ; Smith, Cordell et Mullen 2015) ainsi que le travail tout récent du Computational History Group à l'université d'Helsinki sur le corpus ECCO (Eighteenth-Century Collections Online) (Mahadevan *et al.* 2024).

3 La littérature sur l'intertextualité est immense : pour une belle synthèse, voir Samoyault 2005.

4 Sur ce « tournant » des réseaux dans les sciences humaines, voir Ahnert *et al.* 2020 et Labatut et Bost 2019. Sur l'usage de la SNA aux études dix-huitiémistes, voir Edmondson et Edelstein 2019.

5 Grâce à des accords institutionnels conclus avec l'université d'Oxford, l'université de Chicago et la Bibliothèque nationale de France (BNF), le projet *Modern* dispose d'un accès à divers corpus de textes francophones, à savoir : la base de données ARTFL-Frantext et les collections ARTFL du XVIII^e^ siècle (American and French Research on the Treasury of the French Language, université de Chicago, 5 000 volumes) ; les ressources numériques du Voltaire Lab de la Voltaire Foundation (Oxford, 2 000 documents) ; Electronic Enlightenment (Oxford, 60 000 documents) ; Gallica (BNF, 250 000 volumes) ; la collection des pamphlets français (Newberry Library, 40 000 documents) ; Eighteenth-Century Collections Online (ECCO) (Gale, 5 000 volumes) ; Goldsmiths'-Kress Collection (Gale, 3 000 volumes). Nous exprimons notre profonde gratitude envers nos partenaires pour cette précieuse collaboration. Il faut admettre, par contre, que plusieurs de ces collections relèvent des produits commerciaux, ou sont réservées en accès limité aux abonnés (ARTFL-Frantext, ECCO, Goldsmiths'-Kress Collection, RetroNews, etc.) : il n'est donc pas possible de partager ces données, ni le corpus d'échantillonnage. Le projet compte partager tous les résultats de ses analyses, en format ouvert, une fois effectuées.

6 En gros, ce corpus restreint contient des textes « canoniques » tirés des bases ARTFL-Frantext (https://artfl-project.uchicago.edu/content/artfl-frantext/), Théâtre classique de Paul Fièvre (https://theatre-classique.fr) et TOUT Voltaire (https://artfl-project.uchicago.edu/tout-voltaire/).

7 En collaboration avec Frédéric Glorieux, nous avons développé un outil de visualisation performant pour les données du catalogue de la BNF, offrant la possibilité d'appliquer divers filtres tels que le nom de l'auteur, le titre, la période de publication, la date, etc. Cet outil, nommé Cataviz, est accessible à l'adresse suivante : http://modern.huma-num.fr/cataviz/.

8 https://github.com/ropensci/textreuse/. Voir aussi Mullen 2020.

9 https://www.etrap.eu/research/tracer/. Voir aussi Büchler *et al.* 2014 ; Franzini *et al.* 2019.

10 https://github.com/tesserae/tesserae/. Voir aussi Coffee *et al.* 2013.

11 https://github.com/dasmiq/passim/. Voir aussi Romanello et Hengchen 2021.

12 https://blast.ncbi.nlm.nih.gov/Blast.cgi. Voir aussi Vesanto *et al.* 2017 ; Salmi *et al.* 2020.

13 https://github.com/ARTFL-Project/text-pair/. Voir aussi Olsen, Horton et Roe 2011.

14 https://github.com/ARTFL-Project/PhiloLogic4/. Voir aussi Tharsen et Gladstone 2020. Depuis 2015, Clovis Gladstone, directeur adjoint du projet ARTFL à l'université de Chicago, développe les bases de code de PhiloLogic4 et de Text-PAIR. Nous lui sommes très reconnaissants pour son soutien précieux du projet *Modern*.

15 Voir https://fr.mathworks.com/discovery/ngram.html.

16 Disponibles en libre accès sur le site *Théâtre classique* tenu par Paul Fièvre.

17 De fait, nous avons utilisé deux versions du *Lycée* : la première corrigée et rendue accessible par ARTFL (Chicago) et la deuxième en version OCR, téléchargeable gratuitement sur Gallica. Vu que la comparaison concernait seulement le théâtre de Voltaire et pour éviter de complexifier notre jeu de données avec d'autres cooccurrences aléatoires, nous avons utilisé seulement le tome X de l'édition de 1825 du *Lycée*, dédié justement au théâtre voltairien après *Mérope*.

18 https://gallica.bnf.fr/ark:/12148/bpt6k202737s/.

19 Pour cette raison, nous avons dû contourner ce défaut du programme avec deux méthodes différentes. Au cours de cette expérience, nous avons également remarqué que l'outil Text-PAIR avait davantage tendance à produire des faux positifs. Par conséquent, le post-traitement des sorties de Text-PAIR vise principalement à séparer les résultats en vrais positifs et faux positifs, même si parfois cette distinction n'est pas évidente d'un point de vue conceptuel.

20 https://networkx.org.

21 Sur PageRank, voir entre autres Liu *et al.* 2017.

22 Voir https://www.dublincore.org.

23 Sur ces deux mesures de « similarité » textuelle, voir Buscaldi *et al.* 2020.

24 Voir notamment Rosson *et al.* 2023.

25 Par exemple : « Privilège du roi Louis par la grace de Dieu, roi de France et de Navarre, à nos ames et féaux conseillers [...] », voir Armstrong 1990.

26 Par exemple : « COMÉDIE en un acte, en prose. », « M. DCC. XLI. », « [MARIVAUX] », « Représentée pour la première fois », etc.

27 https://huggingface.co/bert-base-multilingual-cased/. Voir aussi Devlin *et al.* 2019.

28 La capacité de BERT à capturer la similarité sémantique peut être démontrée empiriquement : les mots sémantiquement proches se retrouvent à proximité les uns des autres dans l'espace vectoriel. BERT est reconnu pour sa capacité à saisir des relations sémantiques complexes.

29 Les transformeurs, en tant qu'architecture spécifique de réseau neuronal, se sont révélés généralement très efficaces pour les tâches de classification de séquences, grâce à leur aptitude à capturer les dépendances à long terme et les relations contextuelles dans les données. Cela a conduit à d'importants progrès dans divers domaines d'application, allant de l'analyse des sentiments et de la traduction linguistique à la reconnaissance vocale et au sous-titrage d'images. Voir Wu et Ong 2021.

30 Le code pour le filtrage des réemplois avec BERT se trouve sur le site GitLab de notre projet : https://gitlab.huma-num.fr/groe/modern/-/blob/a1804b986b8a1a69a4e1f12c104455e9f13c772e/BERT_textual_reuses_filter.ipynb.

31 Le modèle affiné a été entraîné avec les hyperparamètres suivants : une taille de batch de 32, un taux d'apprentissage de 2e-5, l'optimiseur AdamW, et initialement, quatre époques d'entraînement sur un petit ensemble de données. Une fois que la taille de l'ensemble d'entraînement a été augmentée, trois époques ont suffi. Pour évaluer le modèle, le score MCC (Matthews Correlation Coefficient, entre – 1 et 1) a été utilisé, une mesure bien adaptée pour les classifieurs binaires et les jeux de données non équilibrés. Dans notre expérience, le modèle a atteint un score MCC de 0,95 sur les données de test.

Auteurs

Valentina Fedchenko

Projet *Modern*, Sorbonne Université, Paris, France
Valentina Fedchenko est docteure en linguistique générale, avec une spécialisation dans le domaine du traitement automatique de langues, et ingénieure de recherche au CELLF dans le cadre du projet *Modern*. Ses travaux portent sur l'étude des langues et littératures minoritaires par les méthodes du traitement automatique des langues.
ORCID 0000-0002-5119-631X
valentina.fedchenko@sorbonne-universite.fr

Dario Maria Nicolosi

Projet *Modern*, Sorbonne Université, Paris, France
Dario Maria Nicolosi est chercheur postdoctoral au CELLF dans le cadre du projet *Modern*. Ses recherches portent sur l'histoire théâtrale et littéraire de la France du XVIII^e^ siècle et se concentrent surtout sur le genre tragique et les rapports des Lumières avec l'Antiquité classique, grecque et latine.
ORCID 0000-0002-0159-2023
dario-maria.nicolosi@sorbonne-universite.fr

Glenn Roe

Projet *Modern*, Sorbonne Université, Paris, France
Glenn Roe est professeur de littérature française et humanités numériques à Sorbonne Université. Ses recherches portent sur l'histoire littéraire et intellectuelle des XVIII^e^ et XIX^e^ siècles (1750-1914) ainsi que sur l'analyse et le traitement informatiques de textes et de corpus historiques pour la recherche en lettres et sciences humaines.
ORCID 0000-0002-5611-7916
glenn.roe@sorbonne-universite.fr

Droits d'auteur