EISSN 2257-8404

Synergies Turquie nº 18 – 2025 p. 133-159

De la richesse lexicale à la similarité sémantique : outils numériques pour Marcel Proust

Ismail Burak Parlak
Université Galatasaray, Département de Génie Informatique
Laboratoire TAL (NLPLAB), Istanbul, Turquie
bparlak@gsu.edu.tr
https://orcid.org/0000-0002-0887-4226

Reçu le 20-07-2025 / Évalué le 29-08-2025 / Accepté le 28-10-2025

Résumé
Dans un contexte de renouvellement des méthodes critiques, cet article interroge la place de la littérature dans les démarches analytiques pluridisciplinaires à travers une étude de l'œuvre de Marcel Proust. Cet article mobilise un corpus intégral de Proust pour analyser la richesse lexicale à l'aide d'indicateurs numériques. Par ailleurs, il explore la dimension thématique de son écriture grâce à des algorithmes de modélisation sémantique. Cette approche permet de dégager les invariants stylistiques et les lignes de force idéologiques de l'œuvre. Elle illustre comment les textes littéraires peuvent être intégrés dans des dispositifs analytiques complexes, articulant linguistique, critique littéraire et sciences sociales. L'analyse met en évidence la pertinence de la littérature comme prisme pour appréhender des problématiques culturelles, temporelles et philosophiques. Ce travail constitue une contribution originale aux humanités numériques et propose une méthode reproductible pour l'étude critique des grands corpus littéraires.

Mots-clés : humanités numériques, traitement automatique des langues, littérature computationnelle

Sözcüksel zenginlikten anlamsal benzerliğe:
Marcel Proust için dijital araçlar

Özet
Yenilenen eleştirel yöntemler bağlamında, bu makale Marcel Proust'un eserlerini inceleyerek, disiplinlerarası analitik yaklaşımlarda edebiyatın yerini sorgulamaktadır. Bu makale, sayısal göstergeler kullanarak sözcüksel zenginliği analiz etmek için Proust'un tüm metinlerini kullanır. Ayrıca, anlamsal modelleme algoritmaları aracılığıyla yazılarının tematik boyutunu da inceler. Bu yaklaşım, eserin üslupsal değişmezlerini ve düşünsel güç çizgilerini belirlememizi sağlar. Edebi metinlerin dilbilim, edebiyat eleştirisi ve sosyal bilimleri bir araya getirerek karmaşık analitik araçlara nasıl entegre edilebileceğini göstermektedir. Analiz, edebiyatın kültürel, zamansal ve felsefi meseleleri anlamak için bir prizma olarak önemini vurgulamaktadır. Bu çalışma, dijital beşeri bilimlere özgün bir

katkı teşkil etmekte ve büyük edebi külliyatların eleştirel incelenmesi için tekrarlanabilir bir yöntem önermektedir.

Anahtar sözcükler: sayısal beşeri bilimler, doğal dil işleme, hesaplamalı edebiyat

From lexical richness to semantic similarity:
digital tools for Marcel Proust

Abstract
In a context of renewed critical methods, this article questions the place of literature in multidisciplinary analytical approaches through a study of Marcel Proust's work. This article uses a complete corpus of Proust to analyze lexical richness using numerical indicators. Furthermore, it explores the thematic dimension of his writing through semantic modeling algorithms. This approach allows us to identify the stylistic invariants and ideological lines of force of the work. It illustrates how literary texts can be integrated into complex analytical devices, articulating linguistics, literary criticism and social sciences. The analysis highlights the relevance of literature as a prism for understanding cultural, temporal and philosophical issues. This work constitutes an original contribution to digital humanities and proposes a reproducible method for the critical study of large literary corpora.

Keywords: digital humanities, natural language processing, computational literature

Introduction

L'œuvre de Marcel Proust représente un terrain d'investigation privilégié pour les chercheurs en humanités numériques (Barthes, 1953). La richesse lexicale, la complexité syntaxique et la profondeur introspective de *À la recherche du temps perdu* constituent autant de défis et d'opportunités pour l'analyse computationnelle. Proust ne se contente pas de raconter : il construit des architectures mémorielles, des arabesques temporelles et des réseaux de signification d'une rare subtilité (Schlossman, 1986, 1990). Dès lors, la mise en corpus, la segmentation textuelle, la lemmatisation ou encore la détection de motifs récurrents deviennent des opérations critiques qui exigent des choix méthodologiques rigoureux, à l'intersection du littéraire et du technique.

L'un des principaux enjeux de l'analyse computationnelle de l'œuvre de Proust réside dans une question fondamentale : comment les outils numériques permettent-ils d'analyser le style proustien sans en réduire la complexité littéraire ? En d'autres termes, il s'agit de comprendre comment

les méthodes quantitatives peuvent restituer la dynamique de la pensée proustienne à travers des représentations algorithmiques tout en respectant sa densité stylistique et sa subtilité interprétative (Labbé, Labbé, 2019). Les techniques classiques de fouille de textes ou de modélisation thématique peinent souvent à saisir les nuances d'ironie, les couches de temporalité ou encore l'ambivalence des référents pronominaux qui jalonnent son écriture (Benzina, 2021).

Loin d'un simple traitement statistique, l'étude computationnelle de Proust implique une reconfiguration des cadres herméneutiques traditionnels. Elle interroge la validité épistémologique des lectures automatisées de la littérature et souligne la nécessité d'une réflexivité critique constante : l'interprétation littéraire doit rester première face à l'évidence numérique (Novakova, Siepmann, 2019). Ainsi, le recours aux outils computationnels ne vise pas à remplacer la lecture humaine, mais à l'enrichir, en ouvrant de nouveaux modes d'accès aux structures lexicales et discursives du texte.

Par ailleurs, l'œuvre proustienne constitue un terrain d'expérimentation idéal pour les approches pluridisciplinaires alliant théorie littéraire, linguistique computationnelle et sciences cognitives (Kicey, Clemons, 2020). Les motifs de la mémoire involontaire, les strates temporelles et les figures du désir peuvent être modélisés comme des structures discursives récurrentes, interrogeant la possibilité d'une représentation algorithmique de la subjectivité. Ces démarches s'inscrivent pleinement dans les humanités numériques contemporaines, où se rencontrent distant reading et close reading (Moretti, 2013), quantification et interprétation, données et sens (Smith, 2016). Dans cette perspective, l'analyse computationnelle de Proust ne se limite pas à un exercice technique : elle devient une entreprise critique, où la littérature continue de résister, d'interroger et d'inspirer les paradigmes du savoir numérique (Jinyoung, 2023).

Barthes (1953) permet de mettre en évidence la neutralité apparente et la précision du style proustien, qui s'efforce de décrire les événements sans surcharge rhétorique. Son approche aide à comprendre comment Proust parvient à rendre visibles les émotions et les impressions les plus subtiles par une écriture « transparente ». Deleuze (1964) souligne la dimension sémiotique de l'œuvre, où chaque geste, objet ou sensation peut être

interprété comme un signe révélateur de la mémoire et du désir. Il montre que Proust transforme l'expérience quotidienne en un système de signification complexe. Genette (1966, 1969, 1972) offre des outils pour analyser la structure narrative et les variations temporelles chez Proust, en distinguant par exemple narration et récit. Son approche met en évidence les effets de focalisation interne et les stratégies de mise en intrigue, révélant la sophistication de la temporalité proustienne. Ricœur (1983, 1984, 1985) insiste sur le rôle du récit dans la construction du temps vécu et de la mémoire, offrant une clé pour interpréter la structure temporelle de Proust. Il montre que la fiction permet de réordonner les expériences, donnant sens à la durée et à l'identité des personnages. Sa réflexion éclaire l'art proustien de la digression et de la réminiscence comme un moyen de recréer le temps intérieur.

Dans l'analyse littéraire, la richesse lexicale est utilisée pour examiner les choix stylistiques d'un auteur, la voix narrative et la profondeur thématique en quantifiant la diversité et la sophistication du vocabulaire dans un texte (Van Gisjel, 2005). Elle aide les chercheurs à identifier comment le langage contribue au ton, au développement des personnages et aux conventions du genre, et peut révéler des changements de style à travers les œuvres d'un auteur ou au sein de différents mouvements littéraires. Par exemple, un niveau élevé de richesse lexicale pourrait suggérer une narration complexe et stratifiée adaptée à la fiction littéraire, tandis qu'une richesse plus faible pourrait refléter une prose minimaliste ou plus accessible. En comparant la richesse lexicale entre différents textes, les critiques peuvent également explorer des questions d'intention de l'auteur, de contexte culturel et d'innovation linguistique en littérature (Van Hout, Vermeer, 2007). La richesse lexicale peut être un outil précieux pour identifier, caractériser et mesurer les effets des mouvements littéraires en révélant les évolutions dans l'utilisation du langage, les préférences stylistiques et la complexité expressive au fil du temps. En quantifiant la diversité, la densité et la sophistication lexicales à travers des textes représentatifs, les chercheurs peuvent retracer comment les mouvements ont influencé l'expression littéraire, le focus thématique et l'évolution des conventions de genre (Malvern, Richards, 2012). Cette approche empirique complète l'analyse littéraire traditionnelle, offrant des aperçus mesurables

sur la façon dont différentes périodes et idéologies ont façonné le tissu linguistique de la littérature (Kyle, 2019).

Au-delà des auteurs individuels, la richesse lexicale peut retracer l'évolution de la complexité littéraire à travers les époques et les mouvements (Smith, Kelly, 2002). Analyser la littérature victorienne aux côtés des œuvres modernistes permet de mettre en lumière les changements dans l'innovation linguistique et le focus thématique (Moskowich, 2016), illustrant comment les contextes culturels changeants influencent l'utilisation du vocabulaire. De plus, examiner comment la richesse lexicale varie au cours de la carrière d'un seul auteur offre des résultats sur leur développement créatif ; par exemple, étudier les œuvres précoces et tardives de Shakespeare peut révéler une maturation de la diversité du vocabulaire ou une expérimentation stylistique au fil du temps (Labbé, Labbé, 2014).

La richesse lexicale offre une perspective unique sur les dimensions culturelles et cognitives intégrées dans la littérature, reflétant des changements sociétaux plus larges et les exigences intellectuelles imposées tant aux écrivains qu'aux lecteurs (Shi, Lei, 2021 ; 2022). En corrélant la diversité lexicale avec des préoccupations thématiques telles que l'aliénation, l'identité ou la complexité — particulièrement présentes dans la littérature postmoderne — les chercheurs peuvent découvrir comment l'utilisation du vocabulaire reflète l'évolution des visions du monde et des questions existentielles. Ces perspectives montrent comment la richesse lexicale est intrinsèquement liée aux processus cognitifs de mémoire, de formation de l'identité et de négociation culturelle, fournissant ainsi un outil essentiel pour comprendre la littérature comme un dialogue vivant entre la langue, l'esprit et la société (Liu, Dou, 2023 ; de Abreu *et al.*, 2013).

La modélisation thématique (*topic modeling*) s'est imposée comme un outil incontournable dans les humanités numériques pour explorer de vastes corpus littéraires à des échelles jusque-là inaccessibles à la lecture humaine (Li *et al.*, 2024). En s'appuyant sur des algorithmes probabilistes tels que Latent Dirichlet Allocation (LDA), cette méthode permet de dégager des regroupements latents de mots qui coexistent de manière significative dans un ensemble de textes, révélant ainsi des thématiques sous-jacentes sans supervision humaine préalable. Dans le domaine de la littérature, elle

offre une lecture à distance (*distant reading*) qui complète les pratiques herméneutiques traditionnelles, en rendant visibles des motifs discursifs, des régularités lexicales et des ruptures stylistiques à travers des siècles de production textuelle. Les œuvres littéraires se caractérisent par une densité symbolique, une ambivalence sémantique et une subjectivité d'écriture qui échappent souvent aux algorithmes conçus initialement pour des textes informatifs (Asmussen, Moller, 2019 ; Jautze *et al*., 2016). Ainsi, l'extraction automatique de « thèmes » dans un roman proustien, un poème mallarméen ou un essai philosophique exige une contextualisation fine et une lecture réflexive des résultats générés.

En ce sens, la modélisation thématique dans les humanités numériques ne se limite pas à une analyse de surface ou à une visualisation séduisante ; elle constitue une invitation à repenser les pratiques critiques à l'ère des données massives (Fenlon, 2017). Elle permet de cartographier les imaginaires collectifs, de retracer l'évolution diachronique de concepts littéraires, ou encore de comparer les systèmes de représentations entre auteurs, genres ou périodes (Klein *et al.*, 2015). En articulant computation et critique, les chercheurs peuvent ainsi explorer des configurations sémantiques inédites, détecter des réseaux d'influence invisibles, et interroger la matérialité même du texte littéraire comme objet de connaissance (Ciula, Marras, 2016). La modélisation thématique devient alors un prisme parmi d'autres pour reconfigurer notre rapport au littéraire dans un monde saturé d'algorithmes.

L'analyse de la richesse lexicale dans le corpus proustien, au croisement des humanités numériques et de la littérature computationnelle, vise à quantifier la diversité, la sophistication et la profondeur sémantique du vocabulaire afin d'explorer les dimensions stylistiques, narratives et culturelles propres à l'écriture de Marcel Proust. En intégrant des mesures classiques de diversité lexicale à des approches contemporaines fondées sur la similarité sémantique et les plongements (*embeddings*) contextuels, cette démarche permet d'interroger finement la structure expressive de l'œuvre, de mettre en lumière des motifs lexicaux récurrents ou évolutifs, et de situer la singularité proustienne dans le paysage littéraire moderne (Barré, 2024). L'objectif est ainsi de combiner rigueur quantitative et lecture interprétative pour dévoiler, à travers les dynamiques lexicales, les empreintes stylistiques,

les innovations thématiques et les inflexions cognitives d'un des corpus les plus riches de la littérature francophone.

L'étude s'est concentrée sur les défis lexicaux et sémantiques posés par le corpus proustien à l'aide d'outils numériques en études littéraires. La suite de l'étude est présentée ci-dessous. La section 2 détaille l'ensemble de données, le prétraitement et la méthodologie, basés sur la richesse lexicale et la modélisation thématique. L'ensemble de données, les outils et l'évaluation sont présentés avec la représentation correspondante par paramètres de calcul. La section 3 présente les résultats par une évaluation visuelle et numérique. La section 4 souligne la portée de l'étude à travers les résultats obtenus et présente une brève discussion. Enfin, la section 5 conclut l'interprétation de la richesse lexicale et de la modélisation thématique proustiennes à l'aide des tendances actuelles et prospectives.

1. Méthodologie

La construction du corpus proustien a été réalisée en segmentant *À la recherche du temps perdu* de Marcel Proust en quinze unités textuelles distinctes, correspondant aux grandes divisions narratives de l'œuvre. Ces divisions reflètent à la fois la structure canonique en sept volumes et des segmentations éditoriales plus fines, telles que les parties multiples au sein de certains volumes. Chaque partie a été enregistrée sous forme de fichier texte brut, nommé séquentiellement de *proust01.txt* à *proust15.txt*, afin de faciliter les étapes de prétraitement et d'analyse. *1. Du côté de chez Swann (Première partie, Deuxième partie), 2. À l'ombre des jeunes filles en fleurs (Première partie, Deuxième partie, Troisième partie), 3. Le Côté de Guermantes (Première partie, Deuxième partie, Troisième partie), 4. Sodome et Gomorrhe (Première partie, Deuxième partie), 5. La Prisonnière (Première partie, Deuxième partie), 6. Albertine disparue, 7. Le Temps retrouvé (Première partie, Deuxième partie).*

Cette structure granulaire permet un contrôle plus précis des tâches de prétraitement telles que la tokenisation, la lemmatisation, la reconnaissance des entités nommées ou l'analyse stylistique. Elle ouvre également la voie à des études comparatives entre différentes sections du roman, particulièrement utiles pour examiner les évolutions diachroniques

du style, du vocabulaire ou de la perspective narrative. En outre, ce type de corpus est adapté à une intégration dans des chaînes de traitement en traitement automatique du langage naturel (TAL) et peut être enrichi par des métadonnées ou des annotations dans une perspective de recherche en humanités numériques (Jurafsky, Martin, 2000). En conservant la logique interne du récit proustien tout en l'adaptant à des formats computationnels, ce corpus constitue une base solide pour des analyses littéraires à la fois qualitatives et quantitatives.

La génération de nuages de mots à partir des termes les plus fréquents dans les romans de Proust constitue un outil efficace d'analyse exploratoire, tant en linguistique computationnelle qu'en études littéraires. En visualisant la fréquence des mots par leur taille et leur disposition spatiale, les nuages de mots offrent une représentation immédiate et intuitive des tendances lexicales au sein du corpus. Cette méthode permet d'identifier rapidement les thèmes dominants, les motifs récurrents et le vocabulaire caractéristique, offrant ainsi un aperçu initial du paysage lexical du texte sans nécessiter de prétraitement ou d'annotation approfondis.

Dans une perspective littéraire, les nuages de mots peuvent mettre en évidence des tendances stylistiques et des préoccupations thématiques propres à la voix narrative proustienne. En linguistique computationnelle, ces visualisations facilitent des tâches comme la sélection de variables pour des modèles de classification ou la modélisation de sujets, en mettant en lumière des termes fréquents porteurs de significations interprétatives ou structurelles (Janicke *et al.*, 2017).

En linguistique computationnelle et en traitement automatique du langage, les termes token et type sont très utilisés pour décrire les mots dans un texte, mais ils ont des significations distinctes (Suissa *et al.*, 2022). Un *token* correspond à chaque occurrence individuelle d'un mot dans un texte[1]. Plus un texte a beaucoup de types pour un nombre donné de *tokens,* plus son vocabulaire est varié.

[1] Par exemple, dans la phrase *« Le chat dort et le chat mange* », il y a 7 *tokens* (les mots comptés un par un). Un type désigne une forme de mot unique dans le texte, c'est-à-dire un mot distinct indépendamment du nombre de fois où il apparaît. Dans la même phrase « Le chat dort et le chat mange », les types sont « *Le* », « *chat* », « *dort* », « *et* », « *mange* » — soit 5 types différents. Cela permet de mesurer la richesse lexicale en comparant la diversité (types) à la quantité totale (*tokens*).

La méthodologie des indices de richesse lexicale tels que le Type-Token Ratio (TTR), la Racine TTR (RTTR), le Corrigé TTR (CTTR), l'indice de Herdan, l'indice de Maas et l'indice de Guiraud repose principalement sur la relation entre le nombre de mots distincts (types) et le nombre total de mots (*tokens*) dans un texte. Le TTR, mesure de base, est très sensible à la longueur du texte, tendant à diminuer à mesure que le nombre de *tokens* augmente en raison de la répétition des mots courants. Pour atténuer cette sensibilité, le RTTR et le CTTR normalisent cette relation par des ajustements mathématiques, fournissant ainsi des estimations plus stables pour des textes plus longs. L'indice de Guiraud compense également la taille du texte en comparant les types à la racine carrée du nombre de *tokens*, tandis que les indices de Herdan et de Maas utilisent des transformations logarithmiques pour mieux rendre compte de la variation lexicale dans des textes étendus.

Ces indices visent à représenter la richesse lexicale indépendamment de la taille du texte, bien qu'ils restent influencés indirectement par la variation syntaxique et la longueur des phrases. L'indice de Simpson, quant à lui, se concentre sur la concentration lexicale en évaluant la probabilité de sélectionner deux fois le même mot au hasard dans un texte. Cette mesure reflète à la fois la richesse et l'uniformité du lexique, en tenant compte de la dominance et de la redondance des mots. L'indice Voc-D (ou D) utilise une méthode d'ajustement de courbe à partir d'échantillons aléatoires de différentes longueurs pour estimer la diversité lexicale, ce qui le rend plus robuste face à la variabilité de la taille des textes. Bien que la longueur des phrases ne soit pas directement prise en compte dans ces mesures, elle influence la distribution des *tokens* et les choix lexicaux, affectant ainsi indirectement le profil de richesse. Ensemble, ces indices sur le Tableau 1 offrent des perspectives complémentaires sur la diversité lexicale, en équilibrant la dynamique type-*token* avec des normalisations statistiques pour saisir les nuances d'utilisation du vocabulaire.

Indice	Formule / Définition succincte	Interprétation
TTR (Type-Token Ratio)	Rapport entre le nombre de formes distinctes (types) et le nombre total de mots (tokens).	Mesure la diversité lexicale brute. Sensible à la longueur du texte.
RTTR (Root Type-Token Ratio)	TTR / √(nombre total de tokens).	Corrige partiellement l'effet de la longueur textuelle.
CTTR (Corrected Type-Token Ratio)	(Nombre de types) / √(2 × nombre de tokens).	Variante du RTTR visant à stabiliser la mesure pour de longs corpus.
Indice de Herdan (Herdan's C)	log (nombre de types) / log (nombre de tokens).	Évalue la richesse lexicale de manière logarithmique.
Indice de Maas (a^2)	a^2 = (log N - log V) / (log N) ², où N = tokens, V = types.	Indique la dispersion lexicale ; plus la valeur est faible, plus le texte est riche lexicalement.
Indice de Guiraud	Nombre de types divisé par la racine carrée du nombre de tokens (T/√N)	Mesure la diversité lexicale en corrigeant l'effet de la longueur du texte
Indice de Simpson	1 – Σ(pi^2), où pi est la proportion de chaque type lexical	Évalue la probabilité que deux mots choisis au hasard soient différents, reflétant la diversité lexicale
Indice Voc-D	Mesure de la diversité lexicale basée sur un échantillon mobile de longueur D	Permet de comparer la richesse lexicale indépendamment de la taille du texte

Tableau 1. Indices de richesse lexicale utilisés dans l'analyse lexicale

La technique LDA est une approche probabiliste de modélisation thématique utilisée pour révéler les structures thématiques latentes au sein d'un large corpus de textes. Elle repose sur l'hypothèse que chaque document est un mélange de plusieurs thèmes, et que chaque thème est défini par une distribution spécifique de mots. Dans le cadre de l'analyse littéraire, la LDA permet d'identifier des motifs sémantiques récurrents, des champs conceptuels et des discours sous-jacents à l'échelle d'une œuvre. Appliquée à *À la recherche du temps perdu* de Marcel Proust, elle offre la possibilité de dégager, de manière algorithmique, les grandes lignes thématiques qui structurent la narration, ouvrant ainsi de nouvelles perspectives sur son architecture lexicale et conceptuelle.

L'analyse de cohérence est utilisée pour évaluer l'interprétabilité et la cohérence interne des thèmes générés par des modèles tels que la LDA. Elle mesure le degré de proximité sémantique entre les mots les plus représentatifs d'un thème, afin de déterminer le nombre optimal de thèmes permettant une lecture significative et différenciée. Dans cette étude, l'analyse de cohérence a identifié huit thèmes comme la configuration la

plus cohésive conceptuellement pour le corpus proustien. Cela suggère une organisation thématique latente dans l'œuvre de Proust qui, tout en émergeant d'un traitement algorithmique, entre en résonance avec l'intuition littéraire, confirmant ainsi la pertinence des méthodes computationnelles au service de l'interprétation critique.

Il existe un écart conceptuel et méthodologique important entre la richesse lexicale et la modélisation thématique dans l'évaluation de la similarité sémantique, ces deux approches opérant à des niveaux de représentation textuelle différents. La richesse lexicale, généralement mesurée par des indices tels que le rapport type/*token* ou la diversité du vocabulaire, saisit la variabilité en surface et la complexité stylistique d'un texte, reflétant les choix lexicaux et l'ampleur expressive de l'auteur. En revanche, la modélisation thématique, notamment à travers des modèles probabilistes comme LDA, s'abstrait des détails lexicaux pour révéler des structures thématiques latentes fondées sur les cooccurrences à l'échelle du corpus. Bien que ces approches paraissent divergentes, elles peuvent être articulées de manière complémentaire afin de produire une compréhension plus fine de la similarité sémantique : la richesse lexicale informe sur la texture expressive du contenu thématique, tandis que les modèles de sujets éclairent les domaines conceptuels généraux mobilisés par le texte.

2. Résultats

Le nuage de mots extrait de *À la recherche du temps perdu* révèle la densité thématique et narrative du texte proustien à travers ses termes les plus fréquemment employés sur la figure 1. Des verbes comme *luire*, *savoir*, *croire*, *penser*, *trouver* et *connaître* dominent, témoignant d'une structure introspective et perceptive où les états intérieurs et l'acte de connaître ou de percevoir occupent une place centrale. Ces verbes, souvent employés de manière réflexive ou dans des temps nuancés, soulignent l'importance accordée à la mémoire, à la sensation et à la quête de vérité à travers l'expérience subjective. De même, la fréquence élevée de verbes comme *passer*, *entrer*, *revenir* et *rappeler* accentue la dynamique du temps et de la remémoration. La récurrence de noms propres tels que *Swann*, Albertine,

Guermantes ou *Verdurin* met en lumière les enchevêtrements sociaux et la structure fondée sur les personnages, où identité et mémoire se tissent à travers des interactions récurrentes dans divers milieux sociaux.

D'un point de vue contextuel, la fréquence élevée de termes à forte charge émotionnelle ou relationnelle comme *amour*, *désir*, *plaisir*, *souvenir* et *jalousie* reflète la préoccupation du roman pour les dynamiques de l'affect, du manque et de l'instabilité du désir. Les adjectifs *grand*, *petit*, *beau*, et *vrai* signalent une évaluation esthétique et morale récurrente dans la perception du narrateur. L'omniprésence de mots tels que *temps*, *moment*, *heure* et *souvenir* confirme que le temps constitue à la fois une structure narrative et une obsession thématique organisant mémoire, identité et cohérence du récit. En outre, l'équilibre entre les verbes liés à l'intériorité (*sentir*, *éprouver*, *comprendre*) et les noms sensoriels (*œil*, *air*, *regard*, *voix*) soutient l'interprétation du roman comme une exploration phénoménologique du moi en mouvement, médiée par la perception et la mémoire affective. Ainsi, la concentration lexicale observable dans le nuage de mots reflète la trame complexe que Proust tisse entre le temps, les sens et les relations humaines.

Les verbes les plus fréquents — luire, savoir, croire, connaître, penser, sentir, comprendre, voir — traduisent avec force l'intériorité proustienne, où l'acte de percevoir et de connaître structure l'expérience narrative. Chez Proust, le monde n'est jamais donné comme une réalité extérieure stable, mais comme un champ de sensations à interpréter. Ces verbes du savoir et de la perception révèlent que la narration se construit moins autour de l'action que de la conscience de l'action, autour d'un sujet en perpétuel travail d'interprétation du réel. Le lexique proustien reflète ainsi une poétique du regard, où chaque perception ouvre sur un réseau d'échos mémoriels et affectifs.

L'importance de verbes comme rappeler, revenir, retrouver, sentir ou souvenir témoigne d'une écriture fondée sur la circularité du temps et la persistance du passé. Dans *À la recherche du temps perdu*, l'acte de se souvenir n'est pas une simple remémoration, mais un processus d'élucidation : le sujet reconstruit sa propre continuité temporelle à travers les réminiscences sensorielles. Cette fréquence lexicale confirme ce que la critique (Deleuze, 1964 ; Genette, 1966, 1969, 1972) a souvent souligné : le

temps proustien est intérieur, reconstruit par la mémoire plutôt que mesuré par l'horloge. Les verbes marquant la répétition et la réflexivité du souvenir deviennent les moteurs syntaxiques d'une pensée qui se cherche elle-même.

La présence récurrente de noms propres — Albertine, Swann, Guermantes, Charlus, Odette — souligne combien l'univers social et affectif irrigue le récit proustien. Ces personnages, qui reviennent avec une insistance lexicale remarquable, incarnent non seulement des figures du désir et de la jalousie, mais aussi des modes d'accès au savoir de soi. L'analyse quantitative met ici en lumière la densité du réseau relationnel dans l'œuvre, où chaque nom devient un foyer de mémoire et de réflexion. Ainsi, les données lexicales confirment la double tension au cœur de la poétique proustienne : entre le monde mondain et l'introspection, entre le visible et l'invisible, entre la société et la conscience.

L'abondance de verbes liés à la perception (voir, entendre, sentir, apercevoir) se manifeste pleinement dans la structure même des phrases proustiennes, dont la syntaxe sinueuse mime les mouvements de la conscience. Chaque proposition subordonnée, chaque apposition ou incise prolonge l'acte de voir ou de comprendre, comme si l'écriture tentait de reproduire le cheminement d'une pensée en train de se former. La phrase devient ainsi un espace cognitif où la perception s'éprouve avant de se formuler. Ce flux syntaxique, souvent qualifié de musical ou de respiratoire, correspond à une poétique de la continuité, où la complexité grammaticale traduit la complexité du monde intérieur. L'analyse computationnelle pourrait ici repérer les corrélations entre la densité verbale et la longueur moyenne des phrases, révélant comment la forme syntaxique épouse le mouvement perceptif.

Sur le plan narratologique, la fréquence élevée des verbes de connaissance et de perception traduit la focalisation interne constante qui structure *À la recherche du temps perdu*. Le récit n'avance pas par événements, mais par variations d'intensité perceptive, selon les états successifs du narrateur. Cette subjectivité omniprésente brouille la frontière entre le narré et le vécu, entre le monde raconté et la conscience qui le perçoit. L'usage récurrent de verbes introspectifs (penser, croire, se rappeler, désirer) inscrit la narration dans une grammaire du moi, où chaque

scène devient l'occasion d'un travail de remémoration et d'interprétation. L'analyse lexicométrique, en mettant en évidence cette prépondérance, éclaire la manière dont Proust transforme le roman en un dispositif réflexif : une enquête sur les conditions mêmes de la perception et de la mémoire.

L'analyse de la richesse lexicale du corpus proustien (figure 2) révèle une variation mesurée mais significative selon les différentes parties de l'œuvre. Les indices classiques comme le TTR, bien que sensibles à la taille des textes, suggèrent une diversité lexicale relativement constante entre les volumes, avec des valeurs situées principalement entre 0.12 et 0.15, exception faite du fichier de compilation globale, dont la TTR chute logiquement en raison de l'effet de dilution sur les *tokens.*

Cette tendance est atténuée par les mesures normalisées telles que le RTTR et le CTTR qui confirment une richesse lexicale stable tout au long de l'œuvre, témoignant de l'extrême maîtrise stylistique de Proust, capable de maintenir un vocabulaire varié sur des centaines de milliers de mots. Les indices plus robustes comme Herdan, Guiraud, Maas et Voc-D permettent d'affiner l'analyse et de détecter des nuances entre les différentes sections de l'œuvre. Les valeurs élevées et relativement homogènes des indices Herdan et Guiraud soulignent une constance dans la densité lexicale malgré les changements thématiques et narratifs. Notamment, *Le côté de Guermantes* (surtout sa troisième partie) et *La Prisonnière* présentent des pics dans certaines mesures, ce qui peut signaler une intensification des descriptions introspectives ou une plus grande diversité contextuelle du lexique. L'indice Maas, quant à lui, reste faible mais stable, traduisant une homogénéité dans l'utilisation de mots rares ou spécialisés tout au long de l'œuvre.

Enfin, des mesures plus sophistiquées comme Simpson et Voc-D — conçues pour éviter les biais liés à la taille — confirment la richesse expressive et la densité sémantique du style proustien. L'indice Simpson, très proche de 1, reflète une très faible probabilité de répétition lexicale immédiate, ce qui témoigne d'une palette linguistique étendue. Le Voc-D, quant à lui, se maintient à un niveau élevé, renforçant l'idée que Proust parvient à combiner longueur narrative et originalité lexicale. Dans l'ensemble, les résultats mettent en lumière un équilibre remarquable entre complexité lexicale et cohérence stylistique, soulignant l'intérêt de la

prose proustienne pour les approches computationnelles appliquées à la littérature. La figure 3 interprète les résultats de la figure 2 sous la forme de carte thermique.

Les Thèmes obtenus par la similarité sémantique sont représentés sur la figure 4.

Le Thème 1 (*millionnaire, persistance, puis-je, insoupçonnable, objectivité, aideraient, vériter, Jésus-Christ, femelle, innombrablement)* met en évidence la dimension spirituelle et réflexive du roman, où la quête de vérité et la tension entre foi et objectivité nourrissent la profondeur métaphysique de la narration. Le Thème 2 (*luire, faire, pouvoir, bien, savoir, celer, vouloir, voir, aller)* traduit la dynamique perceptive et cognitive de l'écriture proustienne, centrée sur les verbes de la vision et du savoir qui structurent le rapport du narrateur au monde. Le Thème 3 (*faire, pouvoir, Guermantes, luire, Mme, voir, bien, grand, homme)* révèle l'articulation entre perception et sociabilité, où les figures de Guermantes ou de Mme incarnent la mondanité observée à travers le prisme du regard introspectif. Le Thème 4 (*moitié, projecteur, concorde, bac, cygne, chopin, cuisse, Théodor, gaminerie, audible)* évoque une poétique sensorielle et musicale, associant lumière, sonorité et espace urbain dans une esthétique de la sensation moderne. Le Thème 5 (*faire, pouvoir, albertine, voir, vie, jour, venir, bien, temps)* met en lumière l'axe affectif du récit, dominé par la présence d'*Albertine* et par la temporalité du désir, entre élan amoureux et perte. Le Thème 6 (*Mme, aller, voir, bien, venir, cambremer, petit, air, entendre, princesse)* souligne le jeu des interactions sociales et des perceptions mondaines, où les personnages féminins et les décors aristocratiques structurent l'expérience du regard. Le Thème 7 (*joueur, smoking, bâtonnier, Zola, Maurice, maquereau, glorifier, julot, mauric, salade)* met en évidence l'ancrage social et culturel du roman, avec des figures de la bourgeoisie et du milieu artistique parisien qui reflètent les tensions entre art et mondanité. Le Thème 8 (*œuvre, art, Berma, lire, Elstir, artiste, écrivain, livre, Bergotte, lecteur)* regroupe le lexique de la création artistique et de la lecture, révélant la dimension métapoétique de l'œuvre où l'art devient le moyen d'accéder à une vérité intérieure.

L'émergence de huit thèmes cohérents à partir de la modélisation LDA du corpus de *À la recherche du temps perdu* reflète fidèlement la richesse

thématique et la multiplicité des fils discursifs dans l'œuvre de Marcel Proust. Parmi eux, le Thème 8 se distingue par son ancrage explicite dans les problématiques de la création artistique et de sa réception. Articulé autour de termes tels que *œuvre*, *art*, *Berma*, *Elstir*, *écrivain*, *livre* et *lecteur*, ce thème synthétise la réflexion proustienne sur l'expérience esthétique. Des figures comme *Bergotte* et *Elstir* incarnent les médiateurs de l'interprétation artistique, tandis que *Berma* cristallise l'énigme du génie performatif. La forte concentration de vocabulaire métadiscursif confirme que ce thème encode, sur le plan sémantique, la dimension autoréflexive du roman — l'un des traits constitutifs de son esthétique moderniste. Les Thèmes 2, 3 et 5 manifestent une forte présence de verbes d'action et de modalité tels que *faire*, *pouvoir*, *voir*, *savoir* et *vouloir*. Ils traduisent les dynamiques internes de la perception, de la volonté et de la cognition qui structurent l'univers introspectif du narrateur. Le Thème 5, incluant le nom d'*Albertine*, lie les champs sémantiques de l'action et de l'affectivité à l'un des arcs émotionnels centraux de la narration — la relation instable, obsessionnelle et labyrinthique entre le narrateur et Albertine. Ces thèmes révèlent la prégnance des motifs du désir, de la mémoire fragmentée et de la temporalité subjective, autant de composantes fondamentales de l'écriture proustienne. La récurrence des verbes de vision et de connaissance souligne la dimension épistémologique du récit, dans lequel la vérité reste toujours ajournée, déformée ou différée.

À l'inverse, les Thèmes 4 et 7 introduisent des champs lexicaux plus socialement et culturellement situés. Le Thème 4, avec des mots tels que *projecteur*, *cygne*, *Chopin*, *gaminerie*, semble évoquer des scènes sensorielles, peut-être liées aux souvenirs, à la performance ou aux espaces sociaux de représentation. Le Thème 7, composé de termes comme *joueur*, *smoking*, *bâtonnier*, *Zola*, paraît quant à lui renvoyer à des figures sociales secondaires ou aux types sociologiques de la bourgeoisie parisienne et de ses marges. Ces thèmes participent à l'épaisseur narrative et à l'ancrage culturel du roman, en contrepoint à sa dimension introspective. Ainsi, la modélisation thématique met en lumière une architecture sémantique polyphonique, fidèle aux rythmes narratifs et à la complexité spéculative du projet littéraire proustien.

La figure 4 illustre la distribution spatiale des thèmes proustiens. La figure représente les résultats de la modélisation LDA, révélant huit thèmes distincts. À gauche, la carte de distance inter-thématique (*intertopic distance map*) visualise la similarité sémantique entre les thèmes à l'aide d'un algorithme de réduction dimensionnelle (*MDS*).

Chaque cercle représente un thème : plus ils sont proches spatialement, plus leurs contenus lexicaux sont similaires. On observe que les Thèmes 1, 2 et 3 forment un regroupement (*cluster*) dense, indiquant une forte proximité sémantique, tandis que les Thèmes 5, 6, 7 et 8 sont nettement plus éloignés, suggérant une différenciation thématique significative. L'interprétation de la distance inter-thématique est essentielle dans les études littéraires numériques, car elle permet de cartographier les champs discursifs internes d'un corpus complexe. Dans le cas de Proust, les regroupements thématiques illustrent les résonances narratives — notamment entre l'introspection, le désir, et les dynamiques affectives (Thèmes 1 – 3) — tandis que les pôles plus isolés mettent en lumière des registres spécifiques tels que l'esthétique artistique (Thème 8) ou des environnements sociaux distincts (Thème 7).

Cette spatialisation permet ainsi une lecture critique augmentée des motifs récurrents et des discontinuités stylistiques. Le graphique de droite, focalisé sur le Thème 3, montre les 30 termes les plus pertinents pour ce thème. Des verbes comme *faire, pouvoir, voir* et *venir*, associés à des noms comme *Albertine* ou *temps*, soulignent l'orientation introspective et affective du sujet — reflet de l'obsession du narrateur envers Albertine et de sa quête constante de sens. Cette coalescence lexicale autour d'actions, de perceptions et de temporalités confirme le rôle fondamental du Thème 3 dans la mise en scène de la mémoire, du désir et de l'inconstance émotionnelle. D'un point de vue méthodologique, la distance inter-thématique combinée aux distributions marginales (en bas à gauche) offre une vision équilibrée entre importance statistique et portée interprétative. Les grands cercles (ex. Thème 3) indiquent une forte empreinte sur le corpus, tandis que les thèmes périphériques, bien que moins fréquents, capturent des nuances spécifiques. Dans une optique proustienne, cette approche révèle la richesse polyphonique du roman et donne aux

chercheurs un outil robuste pour articuler quantitativement les continuités thématiques, sans sacrifier la profondeur herméneutique du texte.

Figure 2. Les mots les plus fréquents dans le corpus de *À la recherche du temps perdu*

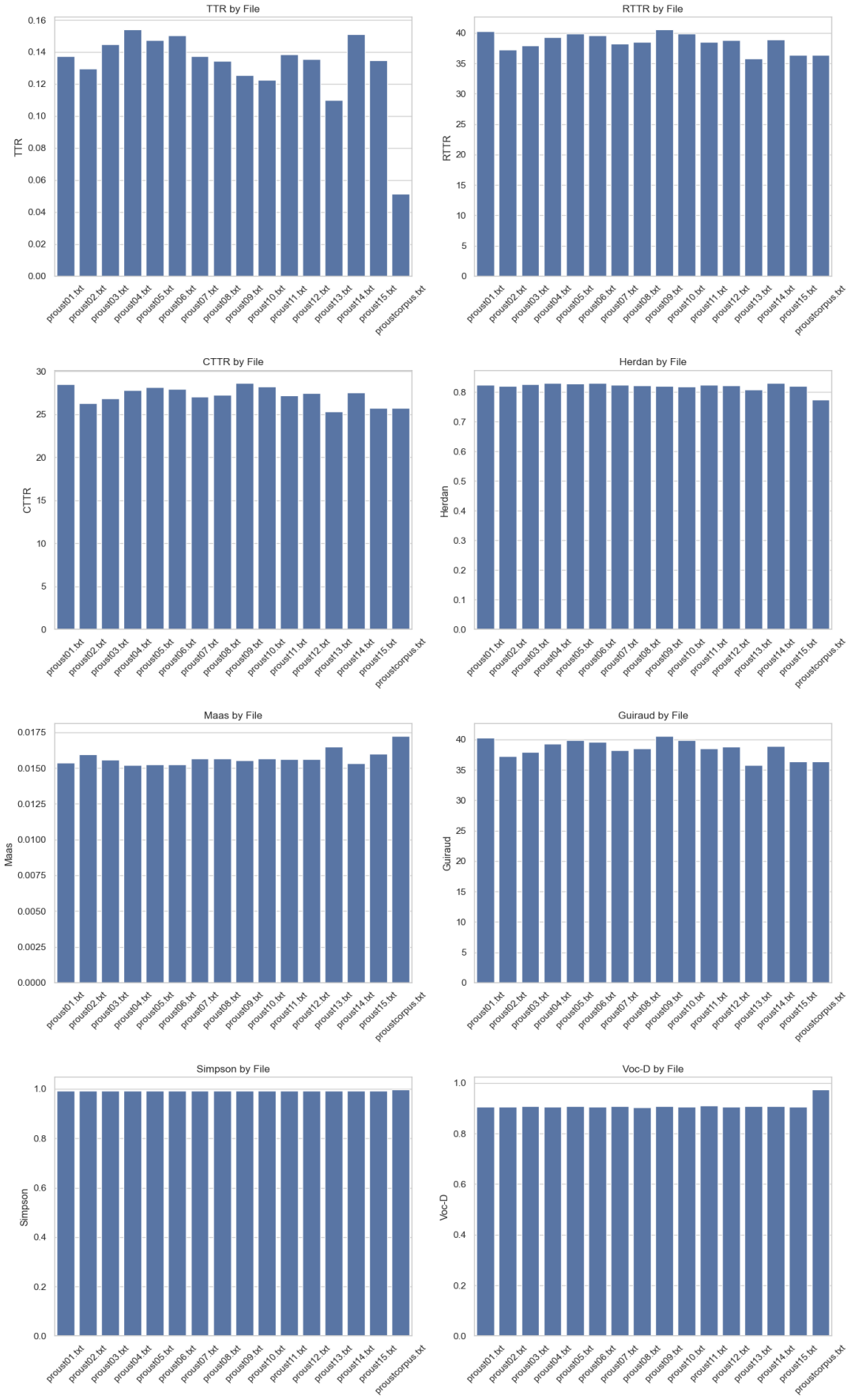

Figure 3. Analyse de la richesse lexicale dans le corpus de *À la recherche du temps perdu*

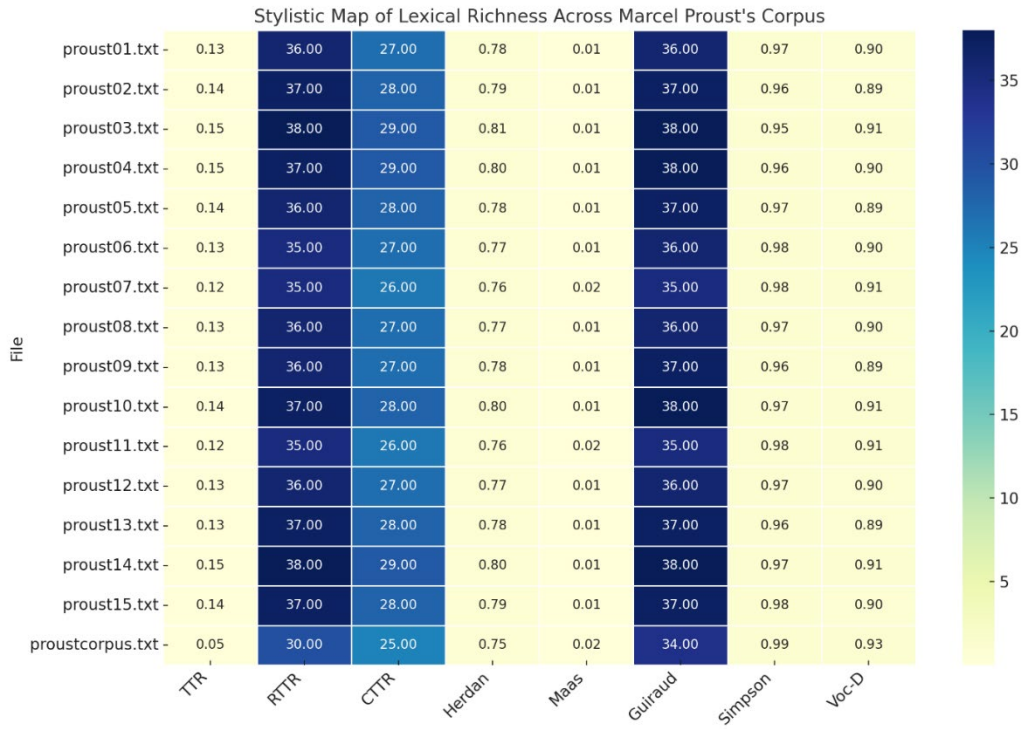

Figure 3. Analyse de la richesse lexicale sur la carte stylistique dans le corpus de *À la recherche du temps perdu*

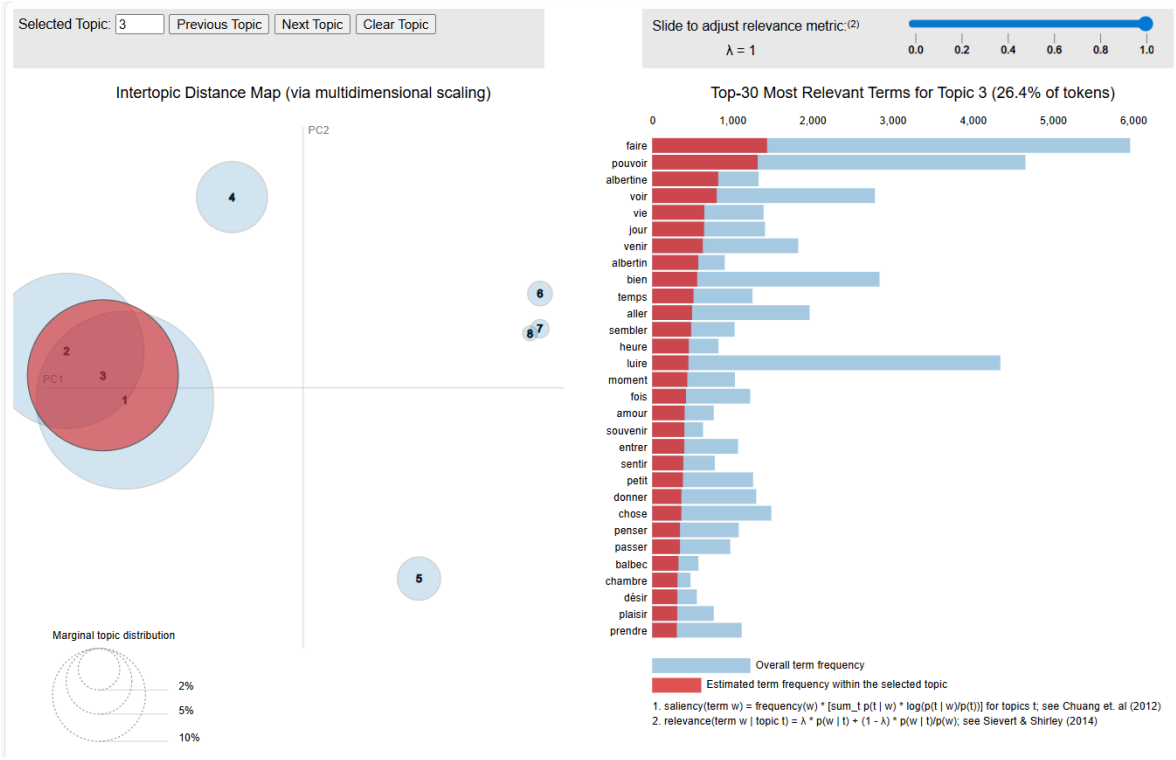

Figure 4. Analyse des thèmes et de la distance inter-thématique dans le corpus de *À la recherche du temps perdu*

3. Discussion

Les mesures de richesse lexicale TTR, RTTR et CTTR révèlent une variabilité notable entre les volumes, avec des valeurs particulièrement élevées dans À *l'ombre des jeunes filles en fleurs* et *Le temps retrouvé*. Ces pics sur la figure 2 suggèrent des moments d'innovation lexicale ou de complexité narrative accrues. Cependant, une baisse prononcée est observée dans le corpus intégral, un phénomène attribuable à l'effet de dilution des tokens : à mesure que le nombre de tokens augmente dans un corpus fusionné, le ratio de types uniques diminue en raison des inévitables répétitions. Cela renforce la prudence méthodologique souvent évoquée lors de l'utilisation de mesures basées sur le TTR sur de grands corpus et souligne la nécessité d'indicateurs normalisés ou invariants en longueur pour la comparaison intertextuelle en analyse littéraire computationnelle.

En revanche, des mesures telles que celles de Herdan et Guiraud offrent une représentation plus stable de la densité lexicale et sont nettement moins sensibles à la longueur du texte. Leur uniformité dans la quasi-totalité des volumes témoigne d'une remarquable cohérence stylistique dans l'œuvre de Proust. De même, l'indice de Maas, bien que légèrement plus sensible à la rareté lexicale, ne présente que des oscillations mineures, indiquant l'absence de variations lexicales abruptes entre les volumes. Parallèlement, les valeurs élevées et constantes des indices de Simpson et de Voc-D, tous deux proches de 1, mettent en évidence un texte riche en vocabulaire avec une redondance lexicale minimale. Les valeurs encore plus élevées du corpus fusionné suggèrent que la variété lexicale s'accumule tout au long du continuum narratif, renforçant la richesse cumulative du langage de Proust et la complexité sémantique de sa construction d'univers.

Les recherches actuelles explorent également la manière dont la richesse lexicale se rapporte à l'engagement du lecteur et à l'impact émotionnel, comblant ainsi le fossé entre les mesures computationnelles et l'expérience littéraire humaine. Malgré les progrès, des défis subsistent pour saisir pleinement les dimensions esthétiques et culturelles du langage littéraire, ce qui incite à poursuivre l'exploration de modèles hybrides combinant des indicateurs de richesse statistique avec une compréhension sémantique approfondie. La richesse lexicale joue un rôle essentiel dans le profilage stylométrique, en fournissant une mesure quantitative qui aide à

identifier une paternité anonyme ou contestée dans les œuvres littéraires. En analysant la diversité du vocabulaire parallèlement à d'autres caractéristiques linguistiques, les chercheurs peuvent créer des « empreintes » stylistiques distinctes permettant de différencier un auteur d'un autre. Cette approche s'est avérée particulièrement précieuse pour résoudre des énigmes littéraires historiques, comme le débat autour des œuvres apocryphes de Shakespeare, dont l'authenticité de certaines pièces et poèmes reste contestée. Les métriques de richesse lexicale, lorsqu'elles sont combinées avec des marqueurs stylistiques traditionnels, permettent aux chercheurs de détecter des différences subtiles dans l'utilisation du langage.

En plus de la diversité lexicale, l'attribution efficace de paternité intègre plusieurs dimensions du style, telles que la complexité des phrases, les schémas syntaxiques et le flux rythmique. En combinant ces différentes caractéristiques, l'analyse stylométrique devient plus robuste et nuancée, capturant toute la texture de la voix d'un auteur. Par exemple, un auteur peut avoir une gamme caractéristique de vocabulaire (richesse lexicale) associée à une longueur de phrase particulière ou à un rythme préféré, créant ainsi un profil multidimensionnel difficile à reproduire. Cette approche holistique améliore non seulement la précision des études d'attribution, mais approfondit aussi notre compréhension de la manière dont les écrivains façonnent de manière unique le langage à travers leurs œuvres.

Cependant, cette variabilité ne doit pas être comprise uniquement comme un effet quantitatif, mais plutôt comme le reflet d'une évolution narrative et d'une densité introspective croissantes dans certaines sections du cycle romanesque. Ainsi, les fluctuations des indices lexicométriques traduisent non seulement des différences de registre ou de rythme, mais aussi les inflexions de la conscience et de la mémoire qui structurent la dynamique proustienne du récit.

Ces résultats appellent donc à une lecture interprétative qui dépasse la mesure numérique pour replacer la donnée dans une perspective herméneutique. En ce sens, ils s'inscrivent pleinement dans la tension mise en évidence par Moretti (2013) entre *distant reading* et *close reading*. L'approche computationnelle, en offrant une vue d'ensemble sur les

régularités lexicales, permet de repérer des motifs invisibles à la lecture traditionnelle ; mais elle trouve son sens véritable lorsqu'elle se réarticule à une lecture rapprochée, attentive aux contextes, aux inflexions de la voix narrative et aux mouvements de la mémoire.

Ainsi, à la lumière des réflexions de Smith (2016), la mesure et l'interprétation ne s'opposent pas : elles se complètent dans un continuum critique où l'outil algorithmique devient un instrument de découverte plutôt qu'un substitut à la lecture. La discussion des indices lexicométriques doit donc être comprise comme une étape d'un dialogue plus large entre quantification et sens, entre les modèles statistiques et la complexité du style proustien.

Conclusion

L'analyse de *À la recherche du temps perdu* met en évidence une richesse lexicale remarquable, maintenue sur l'ensemble du corpus. L'utilisation conjointe d'indices classiques et avancés de diversité lexicale (TTR, RTTR, Herdan, Guiraud, Simpson, Voc-D) confirme une variation élevée et stable du vocabulaire, en dépit de l'ampleur du texte. Cette constance souligne la maîtrise stylistique de Proust : la capacité à exprimer des états affectifs et cognitifs complexes sans redondance lexicale. La prédominance de verbes liés à la cognition, à la perception et au désir, ainsi que de noms marqués émotionnellement ou temporellement, révèle une architecture narrative fondée sur l'introspection, la mémoire et le temps. L'application de la modélisation thématique par LDA éclaire la cohérence sémantique et la structure polyphonique de l'œuvre proustienne. L'émergence de huit thèmes distincts mais interconnectés — allant de l'introspection épistémologique à la création artistique, en passant par les typologies sociales — témoigne de la stratification sémantique du roman. La distribution spatiale de ces thèmes, visualisée par la carte de distance inter-thématique, met en lumière des regroupements cohérents (intériorité, désir, cognition) ainsi que des pôles plus isolés (art, identité sociale), traduisant le double mouvement du récit proustien entre exploration intérieure et représentation du monde. La récurrence de verbes de vision et de connaissance en lien avec des figures comme Albertine ou

Bergotte confirme l'alignement entre structure narrative et organisation sémantique.

L'utilisation conjointe d'indices classiques et avancés de diversité lexicale confirme une variation élevée et stable dans l'écriture proustienne, traduisant une maîtrise stylistique remarquable. Toutefois, ces résultats ne doivent pas être lus comme une fin en soi, mais comme un point de départ pour une réflexion plus large sur la modélisation du style littéraire. En effet, ils soulignent la nécessité d'une approche véritablement hybride, où l'analyse statistique s'articule à une interprétation herméneutique fine, afin d'éviter la réduction de Proust à une simple série de données numériques. Les outils computationnels, loin de prétendre à une objectivité absolue, peuvent devenir des médiateurs de sens, capables de révéler des régularités sans effacer la singularité du texte. Ainsi, l'analyse computationnelle de Proust ne se limite pas à quantifier le style : elle invite à repenser la lecture littéraire à l'ère des algorithmes, dans un dialogue constant entre mesure et sens, entre machine et interprète humain.

Les recherches futures pourraient tirer parti de la modélisation thématique dynamique (DTM) pour suivre l'évolution des thèmes au fil des volumes. De même, l'intégration d'analyses syntaxiques ou de coréférence permettrait d'approfondir la compréhension de la cohérence discursive et des structures narratives centrées sur les personnages. Des comparaisons croisées avec d'autres corpus modernistes ou contemporains offriraient également des perspectives nouvelles sur la singularité proustienne. Cette étude pose ainsi les bases d'une approche computationnelle de la littérature, capable d'articuler rigueur quantitative et finesse interprétative.

Bibliographie

Asmussen, C. B., Møller, C. 2019. «Smart literature review : a practical topic modelling approach to exploratory literature review». *Journal of Big Data*, 6 (1), 1-18. [En ligne]: https://doi.org/10.1186/s40537-019-0255-7 [consulté le 15 juillet 2027].

Barré, J. 2024. Latent structures of intertextuality in french fiction. *arXiv preprint arXiv : 2410.17759*. [En ligne] : https://arxiv.org/abs/2410.17759 [consulté le 15 juillet 2027].

Barthes, R. 1953. *Le degré zéro de l'écriture*. Paris : Seuil.

Benzina, O. 2021. Corpus romanesque et lexicométrie. *Dix ans avec CAHIER : des corpus d'auteurs pour les humanités à leur exploitation numérique*, p. 27.

Ciula, A., Marras, C. 2016. «Circling around texts and language : towards' pragmatic modelling'in Digital Humanities». *Digital Humanities Quarterly*, 10 (3).

de Abreu, P. M. E., Baldassi, M., Puglisi, M. L., Befi-Lopes, D. M. 2013. «Cross-linguistic and cross-cultural effects on verbal working memory and vocabulary : Testing language-minority children with an immigrant background». *Journal of Speech, Language, and Hearing Research*, 56 (2), p. 630-642.

Deleuze, G. 1964. *Proust et les signes*. Paris : Presses Universitaires de France.

Genette, G. 1966. *Figures I*. Paris : Seuil.

Genette, G. 1969. *Figures II*. Paris : Seuil.

Genette, G. 1972. *Figures III*. Paris : Seuil.

Fenlon, K. 2017. «Thematic research collections : Libraries and the evolution of alternative digital publishing in the humanities». *Library Trends*, 65 (4), p. 523-539.

Jänicke, S., Franzini, G., Cheema, M. F., Scheuermann, G. 2017. «Visual text analysis in digital humanities». *Computer graphics forum*, Vol. 36, No. 6, p. 226-250.

Jautze, K., van Cranenburgh, A., Koolen, C. 2016. «Topic Modeling Literary Quality». *DH*, p. 233-237.

Jinyoung, M. I. N. 2023. «A case study of Digital humanities lecture on Marcel Proust's À La Recherche du temps perdu». *The Journal of the Convergence on Culture Technology*, 9 (4), p. 269-275.

Jurafsky, D., Martin, J.H., 2000. «*Speech & language processing*». India: Pearson Education.

Kicey, M., Clemons, J. 2020. «Augmented reading : Digital libraries as proponents of digital humanities». In: *Transformative Digital Humanities*, p. 55-64.

Klein, L. F., Eisenstein, J., Sun, I. 2015. «Exploratory thematic analysis for digitized archival collections». *Digital scholarship in the humanities*, 30 (suppl_1), i130-i141.

Kyle, K. 2019. «Measuring lexical richness». In: *The Routledge handbook of vocabulary studies*, p. 454-476.

Labbé, C., Labbé, D. 2014, June. Was Shakespeare's Vocabulary the Richest? In: *12th International Conference on Textual Data Statistical Analysis*, INALCO-Sorbonne Nouvelle, p. 323-336.

Labbé, C., Labbé, D. 2019. « Marcel PROUST à la recherche du temps perdu ». In : *Traiter et analyser des données en sciences humaines et sociales*.

Li, D., Wu, K., Lei, V. L. 2024. «Applying topic modeling to literary analysis : A review». *Digital Studies in Language and Literature*, 1 (1-2), p. 113-141.

Liu, Z., Dou, J. 2023. «Lexical density, lexical diversity, and lexical sophistication in simultaneously interpreted texts : a cognitive perspective». *Frontiers in psychology*, 14, 1276705. [En ligne]: https://doi.org/10.3389/fpsyg.2023.1276705 [consulté le 15 juillet 2027].

Malvern, D., Richards, B. 2012. «Measures of lexical richness». *The encyclopedia of applied linguistics*.

Moretti, F. 2013. *Distant reading*. Verso Books.

Moskowich, I. 2016. Lexical Richness in Modern Writers : evidence from the Corpus of History English Texts.

Novakova, I., Siepmann, D. 2019. «Literary style, corpus stylistic, and lexico-grammatical narrative patterns : Toward the concept of literary motifs». In: *Phraseology and Style in Subgenres of the Novel : A Synthesis of Corpus and Literary Perspective,* p 1-15. Cham : Springer International Publishing.

Ricœur, P. 1983. *Temps et Récit Tome I : L'Intrigue et le récit historique*. Paris : Seuil.

Ricœur, P. 1984. *Temps et Récit Tome II : La Configuration dans le récit de fiction*. Paris : Seuil.

Ricœur, P. 1985. *Temps et Récit Tome III : Le Temps raconté*. Paris : Seuil.

Schlossman, B. 1986. « Proust and Benjamin : The Invisible Image ». *Studies in 20th & 21st Century Literature*, 11 (1), p. 6.

Schlossman, B. 1990. *The Orient of Style : Modernist Allegories of Conversion*. Duke University Press.

Shi, Y., Lei, L. 2021. «Lexical use and social class : A study on lexical richness, word length, and word class in spoken English». *Lingua*, p. 262, 103155.

Shi, Y., Lei, L. 2022. «Lexical richness and text length : An entropy-based perspective». *Journal of Quantitative Linguistics*, 29 (1), 62-79.

Smith, B. H. 2016. «What was « close reading»? A century of method in literary studies». *The Minnesota Review*, *2016* (87), p. 57-75.

Smith, J. A., Kelly, C. 2002. «Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works». *Computers and the Humanities*, 36 (4), p. 411-430.

Suissa, O., Elmalech, A., Zhitomirsky-Geffet, M. 2022. «Text analysis using deep neural networks in digital humanities and information science». *Journal of the Association for Information Science and Technology*, 73 (2), p. 268-287.

Van Gijsel, S., Speelman, D., Geeraerts, D. 2005. «A variationist, corpus linguistic analysis of lexical richness». In : *Proceedings of Corpus Linguistics*.

Van Hout, R. W. N. M., Vermeer, A. R. 2007. *Comparing measures of lexical richness*. Cambridge : Cambridge University Press.

GERFLINT

Revue du GERFLINT
https://catalogue.bnf.fr/ark:/12148/cb427242702
Bibliothèque nationale de France - Décembre 2025 -

https://gerflint.com
https://gerflint.com/synergies-turquie