Genesis
Manuscrits – Recherche – Invention

51 | 2020
Intertextualité - Exogenèse

Détection automatique de phénomènes intertextuels

Jean-Gabriel Ganascia

Édition électronique
URL : https://journals.openedition.org/genesis/5671
DOI : 10.4000/genesis.5671
ISSN : 2268-1590

Éditeur :
Presses universitaires de Paris Sorbonne (PUPS), Société internationale de génétique artistique littéraire et scientifique (SIGALES)

Édition imprimée
Date de publication : 15 décembre 2020
Pagination : 63-77
ISBN : 979-10-231-0704-3
ISSN : 1167-5101

Référence électronique
Jean-Gabriel Ganascia, « Détection automatique de phénomènes intertextuels », *Genesis* [En ligne], 51 | 2020, mis en ligne le 20 décembre 2021, consulté le 07 février 2022. URL : http://journals.openedition.org/genesis/5671 ; DOI : https://doi.org/10.4000/genesis.5671

Tous droits réservés

Détection automatique de phénomènes intertextuels

Jean-Gabriel Ganascia

L'étude des phénomènes intertextuels et exogénétiques dans les arts en général, et dans la littérature en particulier, est certainement l'une des plus délicates, des plus difficiles et des plus stimulantes que les disciplines d'érudition aient à affronter. Comme l'exprime Harold Bloom dans son essai intitulé, de façon évocatrice, *The Anxiety of Influence*[1], il existe une tension entre l'aspiration à l'originalité de l'artiste et la fidélité à une tradition ou à un auteur admiré dans laquelle il ancre son œuvre. Depuis plus de quarante ans, les théories modernes de la littérature[2] insistent sur le rôle des paraphrases, réécritures, citations, emprunts et réutilisations de toute sorte. Les notions d'*intertextualité*[3], de *transtextualité*, d'*hypertextualité* vs. d'*hypotextualité*[4] ont été introduites dans les années soixante-dix et quatre-vingt du XX^e^ siècle pour approcher ces phénomènes. L'analyse attentive des traces ou des marqueurs est particulièrement intéressante pour évaluer la distance que les créateurs introduisent volontairement ou involontairement par rapport à leurs maîtres. Cela aide non seulement à comprendre le contexte intellectuel environnant les auteurs et les facteurs qui les ont motivés à créer, mais aussi à approcher leurs objectifs et leurs stratégies. Il ne s'agit évidemment pas de plagiat, à savoir de l'utilisation frauduleuse et masquée de matériel existant, mais soit d'un processus de différenciation et de distinction intentionnelles, soit d'influences inconscientes que la détection de références, plus ou moins saillantes, permettrait de rendre explicite.

Nous souhaitons montrer ici comment la mise en évidence de marqueurs de ressemblances, qu'il s'agisse de fragments textuels faits de séquences de mots ou de topos ou encore de motifs rythmiques et/ou syntaxiques, pourrait aider à repérer des influences. Sans doute, certaines ressemblances adviennent-elles de façon fortuite, d'autres relèvent de formules ou de conventions ou encore de traits d'expression propres à une époque. À cela s'ajoute l'existence d'influences si travesties qu'elles ne se laissent pas dévoiler. Nulle équivalence donc entre ressemblances et influences, mais mise à profit de ressemblances comme autant d'indices dans l'enquête sur les influences. La mise en lumière de marqueurs peut aider à éclairer, moyennant des informations complémentaires, un passage obscur en suggérant une référence implicite à un topos, soit dans le même texte, soit dans d'autres du même auteur, soit encore dans des écrits complètement différents, comme des articles de journaux ou des ouvrages scientifiques, où des expressions très semblables ont été utilisées. Nous distinguerons alors, comme la thématique de ce numéro nous y invite, deux volets : celui qui relève de l'exogenèse, à savoir de ce qui, dans le processus de création, porte « sur des informations

1. Harold Bloom, *The Anxiety of Influence: a Theory of Poetry*, Oxford University Press, 1997.
2. Antoine Compagnon, *La Seconde Main ou le travail de la citation*, Paris, Seuil, 1979.
3. Julia Kristeva, *Sèméiotikè. Recherches pour une sémanalyse*, Paris, Seuil, 1969, et *La Révolution du langage poétique*, Paris, Seuil, 1974.
4. Gérard Genette, *Introduction à l'architexte*, Paris, Seuil, 1979, et *Palimpsestes. La littérature au second degré*, Paris, Seuil, 1982.

(documentaires, référentielles, autobiographiques, littéraires, stylistiques, etc.) émanant d'une source extérieure à la conception et à la réalisation proprement dites de l'œuvre », et celui qui est propre à l'endogenèse, qui situe l'œuvre d'un auteur à l'intérieur de cette œuvre même, ou pour reprendre la définition originaire du terme en biologie, au processus de naissance de cellules dans l'intérieur d'autres cellules.

Nous proposons ici une méthode de détection automatique des ressemblances de fragments textuels sur des corpus très conséquents de centaines de milliers, voire de millions d'ouvrages. Indépendamment des influences et des emprunts, celles-ci mettent en évidence des formes langagières propres à une époque, comme les formules de remerciement, des références à des autorités ou des citations communes qui aident à interpréter les textes en établissant des liens entre différents passages, et en mettant en évidence d'autres contextes où les mêmes locutions ont été utilisées.

L'approche adoptée repose sur la détection d'homologies textuelles ou lexicales, c'est-à-dire sur la découverte de fragments qui paraissent, eu égard à une mesure de ressemblance donnée, assez proches pour que l'on puisse faire l'hypothèse qu'il s'agit d'une réutilisation ou d'une référence commune ou encore d'un même topos. C'est ce que certains ont essayé de faire avec succès ces dernières années dans le secteur des humanités numériques[5]. Dans une première partie, nous verrons comment les techniques de détection de plagiats ont été modifiées pour permettre la détection de similitudes textuelles sur de gros corpus. Nous montrerons alors que cela a permis de mettre en évidence des phénomènes d'intertextualité, sachant, faut-il le préciser, que le repérage d'homologies n'agit que comme un catalyseur, en ce sens qu'il facilite et accélère la confrontation avec des corpus extérieurs à l'œuvre, mais qu'il ne détecte pas toutes les influences et que, parmi les ressemblances trouvées, beaucoup s'avèrent non pertinentes. Nous verrons ensuite que le nombre de réutilisations détectées, de l'ordre de plusieurs centaines de milliers, voire du million ou même de la dizaine de millions, étant prohibitif, il a fallu développer des techniques fondées sur la théorie mathématique des graphes pour les appréhender. Nous verrons enfin en quoi une telle démarche sert tant à la détection de phénomènes d'exogenèse que d'intertextualité externe ou interne à l'œuvre.

Détection de similitudes textuelles

Technique de détection de plagiats

Le principe sur lequel repose la détection des similitudes textuelles ressemble fortement à celui qui est en œuvre dans les techniques de détection de plagiats. Sans entrer dans les détails, parmi ces techniques, certaines sont inappropriées pour aborder notre problème,

5. Parmi ces travaux, nous pouvons citer : Timothy Allen, Charles Cooney, Stéphane Douard, Russell Horton, Robert Morrisse, Mark Olsen, Robert Voer, « Plundering Philosophers : Identifying Sources of the Encyclopédie », *Journal of the Association for History and Computing*, vol. XIII, nº 1, printemps 2010 (en ligne sur quod.lib.umich.edu) ; Andrew Kane et Frank Tompa, « Janus: the intertextuality search engine for the electronic *Manipulus florum* project », *Digital Scholarship in the Humanities,* vol. XXVI, nº 4, mai 2011, p. 407-415 (en ligne sur academic.oup.com) ; Marco Büchler, Gregory Crane, Maria Moritz et Alison Babeu, « Increasing recall for text re-use in historical documents to support research in the humanities », dans P. Zaphriris, G. Buchanan, E. Rasmussen et F. Loizides, *Theory and Practice of Digital Libraries*, Berlin – Heidelberg, Springer, 2012, p. 95-100 ; Jean-Gabriel Ganascia, Pierre Glaudes et Andrea Del Lungo, « Automatic Detection of Reuses and Citations », dans *Literary Texts. Literary and Linguistic Computing*, nº 29 (3), 2014, p. 412-421 (en ligne sur hal.archives-ouvertes.fr).

à savoir pour étudier l'intertextualité en général, ou plus particulièrement l'intertextualité comme indice de l'exogenèse. Ainsi en va-t-il de la comparaison des bibliographies d'un article[6] ; en effet, si l'on conçoit que cette approche puisse fonctionner pour repérer des articles scientifiques grossièrement plagiés, on voit mal comment elle pourrait fonctionner pour des œuvres littéraires comme les romans, ou pour des essais philosophiques. Il en va de même pour la simple comparaison de chaînes de caractères : elle est utilisée avec succès dans des situations spécifiques, lorsque les textes sont très similaires, comme en génétique textuelle[7] où l'on confronte différentes versions du même texte. De même, les modèles vectoriels qui réduisent les textes à des sacs de mots[8] ne sont pas appropriés, et ce, pour au moins deux raisons. D'une part, portant sur la globalité du texte, et non sur des petits fragments, ces approches détectent des plagiats massifs, mais ne repèrent pas les marqueurs d'influence locaux. D'autre part, en réduisant l'ensemble du texte à un sac de mots dont ils évaluent les fréquences globales, ils négligent deux aspects essentiels de la textualité qui tiennent l'un à la syntaxe, l'autre à la composition, à savoir à l'organisation du texte.

Lorsqu'on limite ces modèles vectoriels aux mots outils comme les déterminants, les pronoms ou les prépositions, ces approches sont couramment utilisées pour caractériser le style ou les pastiches d'un auteur[9] ; ce sont là des pistes fort intéressantes, même si elles n'identifient pas à proprement parler les réemplois au sens où nous l'entendons ici, à savoir les réutilisations ou les reprises locales. Sans doute, la distinction entre pastiches et réemplois demanderait-elle de plus longs développements qui déborderaient le cadre de cet article ; notons simplement que notre travail se centre sur des réemplois locaux, qu'il s'agisse de fragments textuels ou d'expressions, voire de tournures ou de motifs syntaxiques, alors que dans le cas des modèles vectoriels, la similitude porte sur la globalité du texte.

L'approche la plus efficace pour la détection des plagiats, la plus largement utilisée et la plus pertinente pour l'étude des réemplois repose sur ce que l'on appelle les « empreintes digitales[10] », c'est-à-dire sur la réduction de documents à des ensembles de *n-grammes*, définis comme des séquences de *n* mots consécutifs, où *n* vaut entre 5 et 10. Une fois extraits, ces *n-grammes* sont indexés dans une immense base de données où il arrive assez rarement que

6. Bela Gipp et Joeran Beel, « Citation based plagiarism detection – a new approach to identify plagiarized work language independently », 21th ACM Conference on Hypertext and Hypermedia, 2010.

7. Jean-Gabriel Ganascia, « Medite, a unilingual text aligner for humanities. Application to textual genetics and to the edition of text variants », *Supporting Digital Humanities (SDH)*, 2011, en ligne sur hal.archives-ouvertes.fr.

8. Voir Amit Singhal et Gerard Salton, « Automatic Text Browsing Using Vector Space Model », *Proceedings of Dual-Use Technologies and Applications Conference*, 1995, p. 318-324 (en ligne sur citeseerx.ist.psu.edu) ; ou Antonio Si, Hong Va Leong et Rynson W. H. Lau, « A document plagiarism detection system », *ACM Symposium on Applied Computing*, New-York, ACM, 1997, p. 70-77 (en ligne sur dl.acm.org).

9. Liviu P. Dinu, Vlad Niculae, Octavia-Maria Sulea, « Pastiche detection based on stopword rankings. Exposing impersonators of a romanian writer », dans EACL, *Workshop on Computational Approaches to Deception Detection*, Avignon, Association for Computational Linguistic, 2012, p. 72-77 (en ligne sur aclweb.org).

10. Voir Tymothy C. Hoad et Justin Zobel, « Methods for identifying versioned and plagiarized documents », *Journal of the American Society for Information Science and Technology*, no 54-3, janvier 2003, p. 203-215 (en ligne sur asistdl.onlinelibrary.wiley.com) ; Steven Burrows, S. M. M. Tahaghoghi et Justin Zobel, « Efficient plagiarism detection for large code repositories », *Software – Practice and Experience*, no 37, 2006, p. 151-175 (en ligne sur onlinelibrary.wiley.com) ; ou Martin Potthast, Beno Stein, Alberto Barron-Cedeno et Paolo Rosso, « An evaluation framework for plagiarism detection », 23rd International Conference on Computational Linguistics (COLING 10), 2010, Stroudburg (Pennsylvanie, USA), p. 997-1005 (en ligne sur aclweb.org).

la même séquence apparaisse dans deux textes différents et où il arrive plus rarement encore que quelques séquences consécutives appartiennent aux deux mêmes textes. Dans de tels cas, que l'on repère très facilement, il y a toutes les chances que cela corresponde à un plagiat.

Adaptation des techniques de détection de plagiats

Quoique la technique des empreintes digitales paraisse bien adaptée à la détection de petits fragments de texte réutilisés dans d'énormes corpus, elle n'est pas immédiatement transposable à la mise en évidence de phénomènes d'intertextualité. En effet, lorsque les écrivains réutilisent les œuvres d'autrui, surtout lorsqu'ils réemploient de petits fragments, ce qui est le cas le plus fréquent, ils transforment leurs emprunts pour les adapter à leur contexte d'insertion. À titre d'illustration, dans son roman *Beatrix*, Honoré de Balzac reprend des expressions tirées du portrait de Mademoiselle Georges par son ami Théophile Gautier publié dans *Le Figaro*. Cependant, la plupart des fragments repris ont été légèrement modifiés : « mains royales frappées de fossettes » est devenu « main mignonne frappée de fossettes ». Au sens mathématique, en définissant les mots comme des chaînes de caractères alphabétiques délimitées par des séparateurs (blancs, apostrophes ou ponctuations), les seuls mots communs à ces deux locutions sont « de » et « fossettes ». Pour pouvoir détecter l'homologie, nous avons besoin de techniques insensibles à la flexion des mots, et à l'insertion, suppression ou remplacement de quelques mots, par exemple ici « royales » remplacé par « mignonne ».

Pour surmonter ces difficultés, certains chercheurs[11] ont développé ces dernières années un ensemble de techniques spécifiques tirées en partie du traitement automatique du langage naturel (TAL) et en partie de la « stringologie », c'est-à-dire de la discipline qui étudie la manipulation des chaînes de caractères. Ces techniques passent par l'élimination de la ponctuation et de certaines catégories syntaxiques de mots supposés peu importants, par exemple des mots qualifiés de vides tels les articles ou les pronoms, par la lemmatisation (c'est-à-dire par la réduction des formes fléchies de mots à une forme « canonique » : singulier pour les noms, masculin singulier pour les adjectifs, infinitif pour les verbes, etc.) ou la racinisation – *stemming* en anglais – (c'est-à-dire la réduction au radical de mots), par l'introduction de trous dans les *n*-grammes, qui correspondent alors à ce que l'on appelle des *k-skip-n-grams*, c'est-à-dire à des *n*-grammes contenant *k* trous, et enfin par le raboutage des *k*-skip-*n*-grams détectés.

À titre d'illustration, voici, lorsqu'on élimine les déterminants, les différents *1-skip-2-grams* de lemmes, c'est-à-dire les bi-grammes avec un trou, accompagnés de leur ordonnée, c'est-à-dire du rang du premier élément de chaque bi-gramme, de « mains royales frappées de fossettes » : « main, royal » (1), « main, frapper » (1), « royal, frapper » (2), « royal, fossette » (2), « frapper, fossette » (3). De même, les 1-skip-2-grams de lemmes de « main mignonne frappée de fossettes » sont : « main, mignonne » (1), « main, frapper » (1), « mignonne, frapper » (2), « mignonne, fossette » (2), « frapper, fossette » (3). On constate qu'il existe deux 1-skip-2-grams communs : « main, frapper » et « frapper, fossette », que l'on peut aisément rabouter,

11. Voir par exemple Marco Büchler, Gregory Crane, Maria Moritz et Alison Babeu, « Increasing recall for text re-use in historical documents to support research in the humanities », art. cit. ; Andrew Kane et Frank Tompa, « Janus: the intertextuality search engine for the electronic Manipulus florum project », art. cit. ; ou Jean-Gabriel Ganascia, Pierre Glaudes et Andrea Del Lungo, « Automatic Detection of Reuses and Citations », art. cit.

puisqu'ils sont jointifs dans les deux cas : « main, frapper » se trouve au rang 1 dans les deux cas et « frapper, fossette » au rang 3. Comme nous l'avons dit, cette séquence de deux 1-skip-2grams met en évidence une ressemblance assez forte qui pourrait éventuellement, si elle était confirmée par d'autres indices, attester d'une réutilisation et donc d'une influence.

Mise en évidence de traces de l'exogenèse

Équipé de cet algorythme, nous avons essayé de retrouver des traces du travail d'exogenèse en détectant des ressemblances à des éléments émanant de sources extérieures à l'œuvre et incorporées à elle.

Nous avons commencé par appliquer notre méthode à des reprises qui avaient déjà été mises en évidence, en reprenant le travail de thèse de Tania Duclos[12]. À titre d'illustration, nous avons confronté le roman *Béatrix* d'Honoré de Balzac aux portraits de femmes de Théophile Gautier parus dans *Le Figaro* et nous avons, sans surprise, détecté des similitudes qui, d'après Tania Duclos, correspondent à des réutilisations. À titre d'illustration, voici quelques lignes écrites par Théophile Gauthier sur Mademoiselle Jenny Colon dont nous surlignons les reprises identifiées par notre logiciel sur un fragment de *Béatrix* de Balzac mentionné ci-après :

> Les costumes romanesques de Piquillo conviennent beaucoup au type de beauté de Mlle Colon ; les grandes robes de lampas ou de brocatelle aux plis soutenus et puissants, les hautes fraises godronnées et frappées à l'emporte-pièce, comme on en voit dans les dessins de Romain de Hooge ; les manches à crevés et à sabots de dentelles, dont la main sort comme le pistil du calice d'une fleur, les feutres à ganse de perles, à plumes crespelées, les chaînes et les rivières de diamants écaillant d'étincelles papillotantes la blancheur mate de la poitrine, les corsets pointus à échelles de rubans s'élançant minces et frêles de l'ampleur étoffée des jupes : – toute la toilette abondante et fantasque du seizième siècle s'adapte merveilleusement à la physionomie de Mlle Colon, que l'on prendrait, dans un de ses costumes capricieux, pour une de ces belles dames des gravures d'Abraham Bosse, qui marchent gravement une tulipe à la main, suivies du petit page nègre qui porte leur queue, leur chien et leur manchon, dans les allées bordées de buis d'un parterre du temps de Louis XIII[13].

> Si elle pouvait par un artifice quelconque porter le costume du~~es anci~~ temps où les femmes avaient des corsets pointus à échelles de rubans s'élançant minces et frêles de l'ampleur étoffée des jupes en brocatelle à plis soutenus et puissants, s'entouraient de fraises godronnées cachaient leurs bras dans des manches à crevés et à sabots de dentelles d'où la main sortait comme une fleur de sa capsule, et qui rejetaient ~~leurs~~ les mille boucles de leur chevelure ~~sur leurs épaules~~ au delà d'un chignon ficelé de pierreries, elle lutterait avec avantage ~~ont~~ avec les beautés les plus célèbres que vous voyez vêtues ainsi dit-elle en montrant un tableau à Calyste, ### debout, ~~devant un~~ tenant une m un papier et chantant avec un seigneur brabançon, pendant qu'un nègre verse dans un verre à patte du vieux vin d'Espagne et qu'~~une~~ la vieille femme de charge arrange des biscuits[14].

12. Tania Duclos, *L'Intertextualité dans* Une fille d'Ève *et* Béatrix *d'Honoré de Balzac*, thèse, Paris, Université Paris-Sorbonne, 2013.
13. Théophile Gautier, *Œuvres complètes*, *Critique théâtrale*, t. I, Paris, Honoré Champion, 2007, p. 782-783.
14. Honoré de Balzac, *Manuscrit de la première partie de Béatrix*, Bibliothèque de l'Institut de France, cote A6, collection Lovenioul, f° 42.

Nous avons aussi comparé les poésies de Lautréamont[15] aux moralistes français du XVIIe siècle comme Pascal, La Rochefoucauld ou La Bruyère, et nous avons retrouvé de très nombreuses réutilisations correspondant à celles mentionnées dans les apparats savants. Parmi celles-ci, certaines présentent d'intéressantes distorsions que le logiciel a surmontées. Ainsi, le logiciel a repéré l'aphorisme de Pascal « Nous naissons injustes, car chacun tend à soi : cela est contre tout ordre » dans sa réécriture en « Nous naissons justes. Chacun tend à soi. C'est envers l'ordre ».

Enfin, il nous est apparu intéressant de confronter les romans de Balzac aux écrits de personnalités considérées à l'époque par Balzac comme des scientifiques dignes du plus grand intérêt. Ce faisant, le logiciel rapproche le fragment surligné en vert du phrénologiste François-Joseph Gall (« Toutes les fois que le front est bas et rétréci, les circonvolutions qu'il recouvre sont petites ; ce qui emporte des facultés intellectuelles médiocres. Le contraire a lieu lorsque le front est haut, large et bombé ») du passage lui aussi surligné en vert de la Grenadière de Balzac (« Tout en lui dénotait une santé robuste, de même que son front large et haut, heureusement bombé, semblait trahir un caractère énergique »).

Quoique ces similitudes ne prouvent rien à elles seules, elles apportent néanmoins des éléments suffisamment intrigants pour susciter une investigation plus poussée.

Extension à de gros corpus

En indexant les *k*-skip-*n*-grams détectés, la méthode peut être étendue à la détection efficace des réutilisations sur des corpus très volumineux. Deux logiciels développés indépendamment, Phœbus, que j'ai conçu et que j'ai implémenté dans le cadre du Labex OBVIL[16], et Philomine, développé par le projet ARTFL de l'Université de Chicago, ont été fusionnés pour donner Textpair qui permet de faire varier de nombreux paramètres, par exemple le nombre *k* de sauts, la taille *n* des n-grams, l'étiquette – nom, adjectif, verbe – et la nature des termes – lemmes, racines, étiquettes syntaxiques, etc. – que l'on retient dans les *k*-skip-*n*-grams.

Toutes ces techniques se sont avérées très efficaces dans le sens où il est maintenant possible de détecter de minuscules fragments de textes réutilisés dans de très grands corpus comprenant des centaines de milliers de livres. Comme le nombre de livres et de volumes imprimés dans les principales bibliothèques nationales, par exemple à la Bibliothèque nationale de France (BnF) ou à la Bibliothèque du Congrès des États-Unis, atteint des dizaines de millions de volumes (14 millions à la BnF et 33 millions à la Bibliothèque du Congrès), nous pensons qu'il existe un horizon quantitatif ultime qui correspond au traitement de dizaines de millions de livres et qu'il sera possible de l'atteindre lorsque ceux-ci seront tous numérisés.

De notre côté, nous avons testé avec succès ces approches sur de très grands corpus, en particulier sur une collection de 130 000 livres mise gracieusement à disposition du Labex OBVIL par la BnF. Cependant, le nombre de réemplois est si important que la plupart du temps le lecteur se perd dans leur masse. À titre d'illustration, notre logiciel a détecté 309 474 fragments de textes de l'Encyclopédie de Diderot et d'Alembert dans le corpus des 130 000 livres fournis par la BnF. De même, il a trouvé 874 606 similitudes textuelles de la base

15. Lautréamont, *Œuvres complètes*, Paris, Gallimard, coll. « Bibliothèque de la Pléiade », 2009.
16. Jean-Gabriel Ganascia, Pierre Glaudes et Andrea Del Lungo, « Automatic Detection of Reuses and Citations », art. cit.

Frantext sur ce même corpus. Une autre expérimentation conduite par le projet ARTFL de Chicago sur le corpus ECCO (*Eighteen Century Corpus Online*) qui contient 250 000 livres généra 17 millions de fragments similaires.

Or, cette quantité énorme de similitudes textuelles est en partie due à une multiplication artificielle lorsqu'un même fragment est souvent repris. Pour comprendre ce point, supposons qu'un fragment *P* ait été dupliqué *n* fois dans un corpus, il s'ensuivra que le nombre de réutilisations, c'est-à-dire le nombre de couples (α, β), où α et β sont deux emplacements distincts de *P* dans le corpus sera : $n \times (n - 1)/2$. Ainsi, si un fragment comme l'ouverture très célèbre du monologue du prince Hamlet « être ou ne pas être, telle est la question » avait été citée 1 000 fois – et il n'est pas absurde qu'il en aille ainsi sur 130 000 ouvrages – cela générerait environ un million de réutilisations. Pour surmonter cette multiplication artificielle de réutilisations qui brouille notre appréhension, la solution proposée est de regrouper les réutilisations de fragments identiques ou approximativement similaires, ce qui réduit considérablement leur nombre et évite également de se perdre dans de multiples duplications de réemplois provenant tous de la réutilisation du même fragment. En d'autres termes, nous n'étudierons pas les similitudes par elles-mêmes, mais leurs regroupements en ensembles de fragments similaires. Ces regroupements se feront en représentant les réutilisations sous forme de graphes, le terme étant entendu au sens de la théorie mathématique des graphes dont nous allons d'abord rappeler les fondements.

Représentation des similitudes sur graphes

Un graphe[17] se définit par des entités appelées *sommets* et par des relations entre ces entités appelées *arêtes* auxquelles il est possible d'affecter un poids ; dans cette dernière éventualité, le graphe résultant est alors qualifié de *pondéré*.

Une fois les similitudes extraites, lorsqu'un même fragment de texte a plusieurs occurrences dans le corpus, nous représentons chacune d'entre elles par un sommet et chaque ressemblance par une arête dans un graphe que nous appelons le *graphe des similitudes*. Si le même passage est repris plusieurs fois, plusieurs arêtes seront émises depuis le sommet correspondant à ce passage. Cependant, il peut aussi arriver que les multiples fragments réutilisés ne soient pas parfaitement identiques, tout en se chevauchant partiellement. Dans ce cas, une étape importante consiste à agréger les différents passages qui se recouvrent, en les prolongeant.

Pour faciliter la compréhension, supposons que nous ayons détecté :

– une similarité R_1 entre le fragment *Adeo ista toto mundo consensere, quanquam discordi sibi et* du texte T_1 et le fragment *adeo ista toto mundo consensere, quanquam discordi et sibi* du texte T_2 ;

– une similarité R_2 entre le fragment *ista toto mundo consensere, quanquam discordi sibi et ignoto* du texte T_1 et le fragment *Ista toto mundo consensere quanquam discordi et sibi ignoto* du texte T_3 ;

– et enfin, une similarité R_3 du fragment *mundo consensere, quanquam discordi sibi et* du texte T_1 et du fragment *mundo li consensere, quanquam discordi et sibi* du texte T_4.

17. Claude Berge, *Graphs and Hypergraphs*, North-Holland Publishing Company, coll. « Mathematical Library », 1973.

Comme la figure 1 le met en évidence, les trois fragments du texte T_1, *Adeo ista toto mundo consensere, quanquam discordi sibi et* dans R_1, *ista toto mundo consensere, quanquam discordi sibi et ignoto* dans R_2 et *mundo consensere, quanquam discordi sibi et* dans R_3 se recouvrent mais ne sont pas identiques. Sachant qu'ils correspondent au même passage, ils seront agrégés sur un seul sommet associé au fragment étendu qui commence par le premier mot *Adeo* et se termine par le dernier *ignoto*, autrement dit au fragment *Adeo ista toto mundo consensere, quanquam discordi sibi et ignoto*.

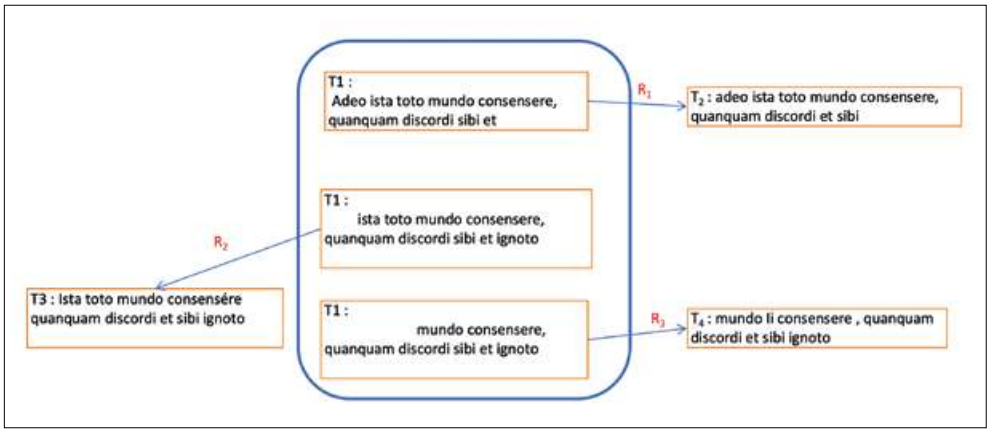

Fig. 1 : Agrégation des trois passages du texte T1 qui se chevauchent et qui sont présents dans R_1, R_2 et R_3.

Par ailleurs, la taille du graphe des similitudes étant généralement très conséquente et nombre de petites ressemblances n'étant vraiment pas intéressantes, il est apparu crucial de mettre en évidence les sommets et arêtes les plus précieux en les quantifiant. Ceci peut être fait en pondérant les arêtes, en fonction du nombre de mots ou plus exactement de lemmes communs aux deux fragments de la réutilisation comme dans la figure 2, ce qui transforme le graphe de similitudes en un graphe pondéré. Il peut également être utile de quantifier les sommets à l'aide d'au moins deux indicateurs : un indicateur local qui correspond à la taille du fragment de texte couvert par le sommet, et un indicateur global qui est la centralité du sommet. Ce dernier indicateur est classique en théorie des graphes ; il peut être calculé de différentes manières, par exemple avec le degré du sommet qui correspond au nombre d'arêtes émises depuis ce sommet ou à la somme des similitudes avec tous les autres sommets.

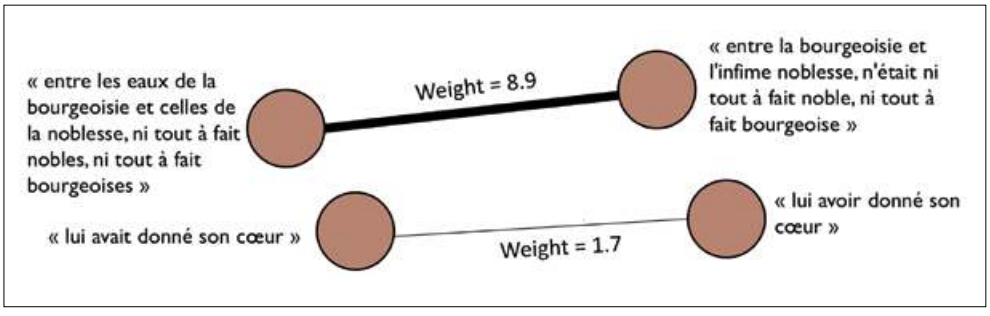

Fig. 2 : Deux similitudes avec des pondérations différentes.

Enfin, selon la motivation qui anime le chercheur, il y a au moins deux façons d'envisager les similitudes : soit en comparant un corpus avec un autre, soit en le comparant avec lui-même. Par exemple, si nous étudions l'influence de Jean-Jacques Rousseau sur sa postérité littéraire et philosophique, il suffit d'extraire les citations ou les emprunts de Rousseau par les auteurs qui vinrent après lui, alors que si nous sommes intéressés par le phénomène des réutilisations mutuelles ou par les sujets et clichés communs à la littérature du XIXe siècle, il est préférable de comparer un corpus de littérature du XIXe siècle à lui-même.

Dans le premier cas, notre approche contribue à l'étude de l'exogenèse en mettant en évidence des similitudes avec des textes extérieurs à l'œuvre qui seraient susceptibles, sous réserve d'analyse, soit de traduire l'influence qu'un auteur a subie au cours du processus d'écriture, soit l'influence qu'il a exercée sur d'autres (voir fig. 3) ; dans le second, elle pourrait contribuer à mettre en évidence des récurrences ou des particularités stylistiques ou encore des reprises à l'intérieur d'une même œuvre (voir fig. 4 et 5).

Composantes connexes et communautés

Les sommets du graphe de similitudes représentent des passages, c'est-à-dire des occurrences de fragments de textes, tandis que les liens entre ces sommets, à savoir les arêtes du graphe de similitudes, attestent d'une ressemblance entre eux. En termes mathématiques, un groupe de sommets entièrement connectés, c'est-à-dire dans lequel chacun est connecté à tous les autres, est appelé une *clique*. Dans notre cas, cela correspond à un fragment présent à l'identique, ou presque, dans de nombreux textes différents. La détection automatique de

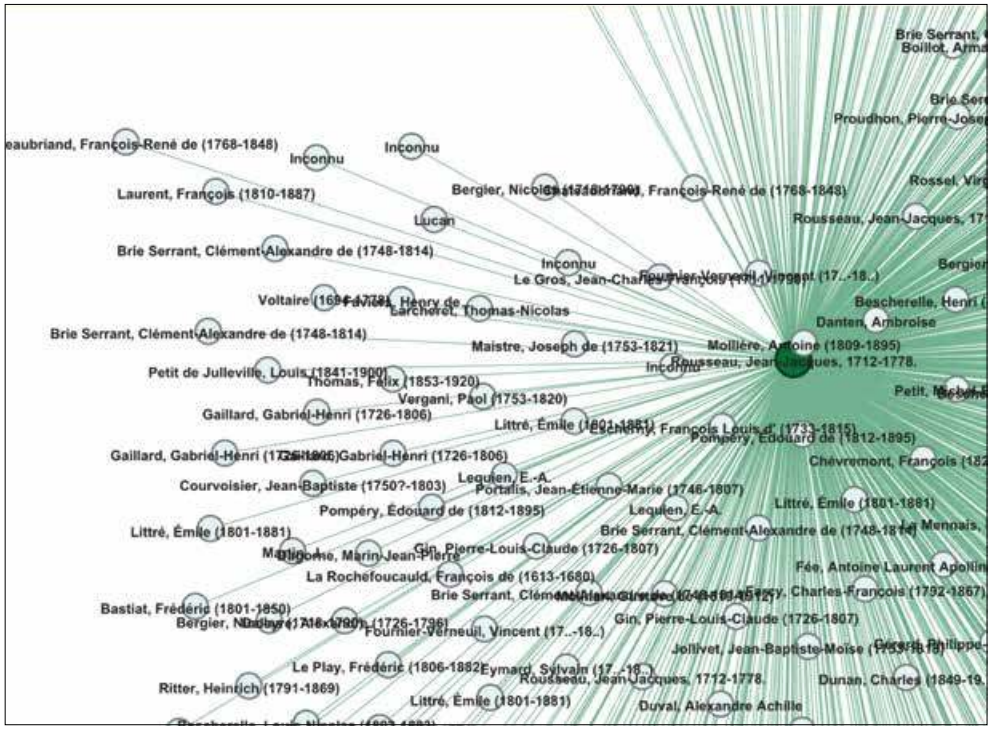

Fig. 3 : Fragment d'une composante connexe du graphe des réutilisations issu de la confrontation de la base Frantext et d'un corpus de 130 000 ouvrages mis gracieusement à la disposition du Labex OBVIL par la BnF.

cliques serait idéale. Cependant, ce n'est pas possible pour au moins deux raisons : d'une part, parce que cette opération est trop coûteuse pour être réalisée effectivement sur de grands graphes ; d'autre part, parce que, très souvent, les fragments associés aux différents sommets sont si différents que même si le fragment F_1 est similaire au fragment F_2 et si le fragment F_2 est similaire au fragment F_3, il n'est pas certain que le fragment F_1 soit semblable au fragment F_3.

À défaut de détecter les cliques, on identifie les ensembles de nœuds *E* entre lesquels il existe au moins une chaîne de similitudes. De tels groupes de nœuds sont appelés des *composantes connexes*. Il existe de nombreux algorithmes efficaces capables de les détecter sur de grands graphes[18]. Cela atteste d'un « air de famille » entre des passages. En termes mathématiques, cela signifie que pour toutes les paires de sommets E_1 et E_2 appartenant à *E*, il existe un chemin reliant E_1 à E_2 dans le graphe de similarité, autrement dit une séquence d'arêtes qui permet de passer de E_1 à E_2.

L'extraction des composantes connexes est sans doute très utile, mais parfois ces sous-graphes sont si grands qu'ils sont difficiles à visualiser et que la ressemblance familiale entre deux nœuds de ces composantes est trop éloignée pour être significative. Il apparaît alors utile d'identifier des sous-graphes « denses » dont les sommets sont fortement liés, même si ce ne sont pas des cliques. Ces dernières années, de nombreux travaux ont porté sur la détection de ce que l'on appelle des *communautés*[19] en vue d'étudier les réseaux sociaux sur le web[20] ou les réseaux biologiques[21]. Les communautés sont des graphes « denses », à savoir entre les sommets desquels il y a beaucoup de liens. Ainsi, dans le cas de l'analyse des réseaux sociaux, en considérant le web comme un immense graphe, les communautés correspondent à des groupes de personnes qui échangent régulièrement.

Nous utilisons les mêmes algorithmes pour détecter des communautés dans le graphe de similitudes ; ces communautés peuvent éventuellement se chevaucher ; ce qui importe, c'est que les sommets qu'elles comprennent et qui correspondent à des fragments textuels partagent tous beaucoup de liens de similitudes entre eux. Cela répond à notre objectif qui est de détecter les sous-graphes denses de ressemblances textuelles.

À titre d'illustration, la figure 4 montre une communauté du graphe de similitudes obtenu en confrontant *La Comédie humaine* de Balzac à elle-même, puis en cherchant des fragments qui contiennent le mot « boucle ».

---

18. John Hopcroft et Robert Tarjan, « Algorithm447: efficient algorithms for graph manipulation », *Communications of the ACM*, vol. XVI, nº 6, juin 1973, p. 372-378 (en ligne sur dl.acm.org).

19. Le lecteur soucieux de détails techniques se reportera à Zhao Yang, René Algesheimer et Claudio J. Tessone, « A comparative analysis of community detection algorithms on artificial networks », *Scientific Reports*, 1er août 2016 (en ligne sur nature.com) ; M. E. J. Newman, « Modularity and community structure in networks », *Proceedings of the National Academy of Sciences*, 6 juin 2002, p. 8577-8582 (en ligne sur pnas.org) ; ou Andrea Lancichinetti et Santo Fortunato, « Community detection algorithms: A comparative analysis », *Physical Review*, 30 novembre 2009 (en ligne sur journals.aps.org).

20. Vincent Labatut et Jean-Michel Balasque, « Detection and Interpretation of Communities in Complex Networks: Practical Methods and Application », dans A. Abraham, A.-E. Hassanien, *Computational Social Networks*, Londres, Springer, 2012, p. 81-113 (en ligne sur link.springer.com).

21. Georgios A. Pavlopoulos, Panagiota I. Kontou, Athanasia Pavlopoulou, Costas Bouyioukos, Evripides Markou et Pantelis G. Bagos, « Bipartite graphs in systems biology and medicine: a survey of methods and applications », *GigaScience*, vol. VII, nº 4, avril 2018 (en ligne sur academic.oup.com).

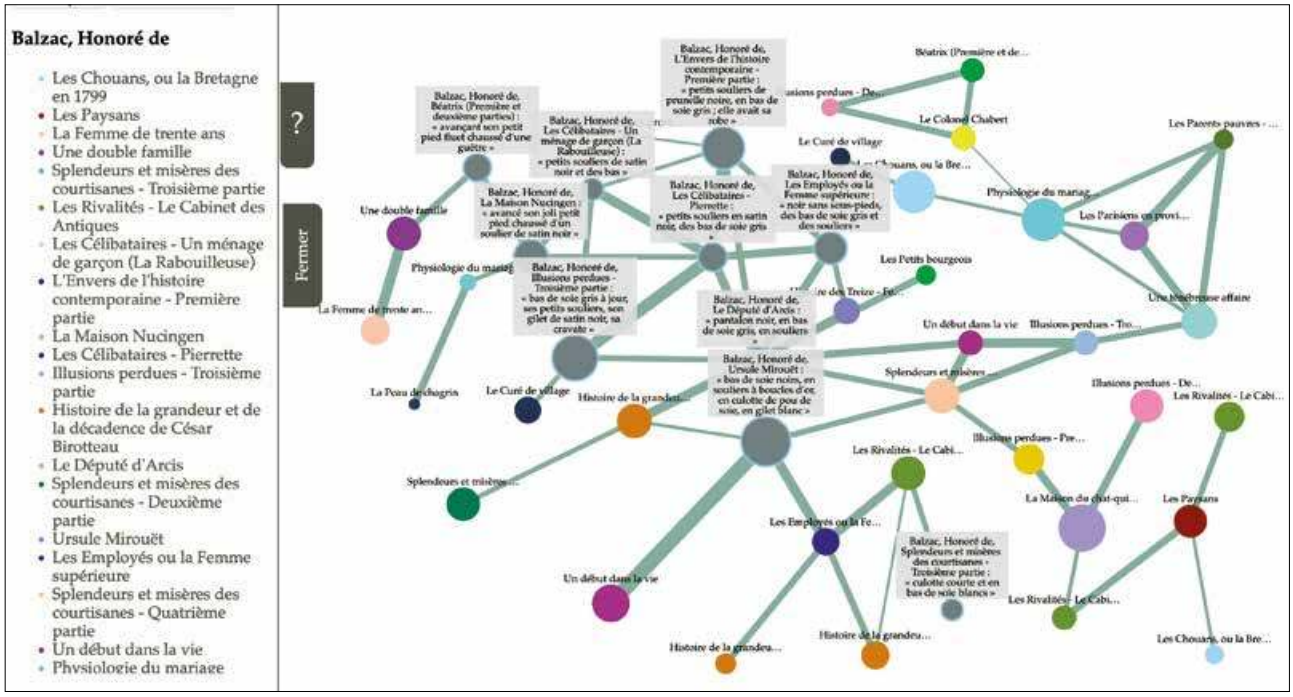

Fig. 4 : Communauté de fragments de *La Comédie humaine* de Balzac dans laquelle le mot « boucle » apparaît.

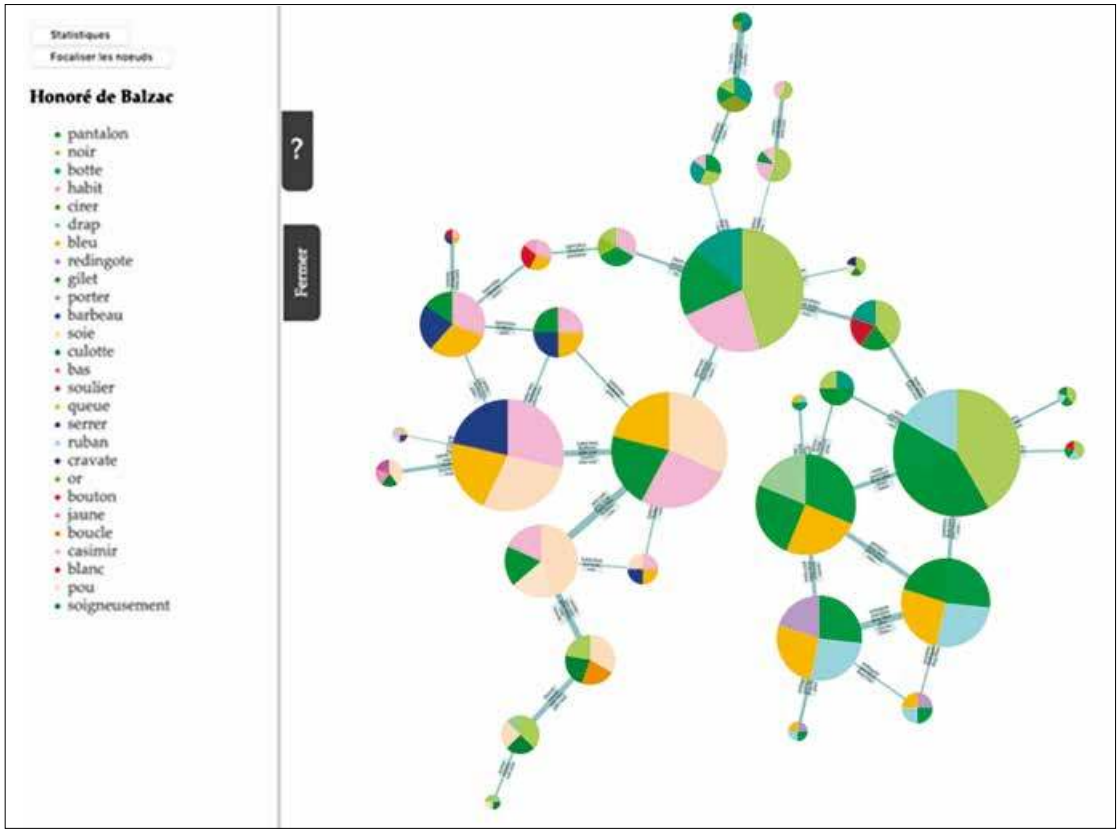

Fig. 5 : « Garde-robes de Balzac » obtenue en retenant les lemmes les plus fréquents des fragments de la communauté de la figure 4.

Si maintenant, on extrait les lemmes les plus fréquents dans les sommets de ce graphe, on accède automatiquement à la « garde-robes de Balzac » (voir fig. 5), c'est-à-dire aux habits dont il affuble un certain nombre de ses personnages.

Perspectives

Cette « garde-robes » de Balzac ne donne qu'un avant-goût de toutes les perspectives ouvertes par l'utilisation des graphes de similitudes dans le domaine de l'intertextualité, en particulier de l'exogenèse et de l'endogenèse. Bien d'autres se dégagent aujourd'hui. Esquissons-en, ci-dessous, quelques-unes.

Influences mutuelles

Il est possible de déterminer des influences mutuelles, ou tout au moins des ressemblances entre des auteurs en construisant un graphe dont les sommets sont des auteurs et les arêtes résument, par une pondération, le nombre des fragments similaires que l'on repère entre les œuvres de ces mêmes auteurs. Établie à partir d'un corpus romanesque d'œuvres du XIXe siècle, la figure 6 visualise les influences mutuelles de leurs auteurs.

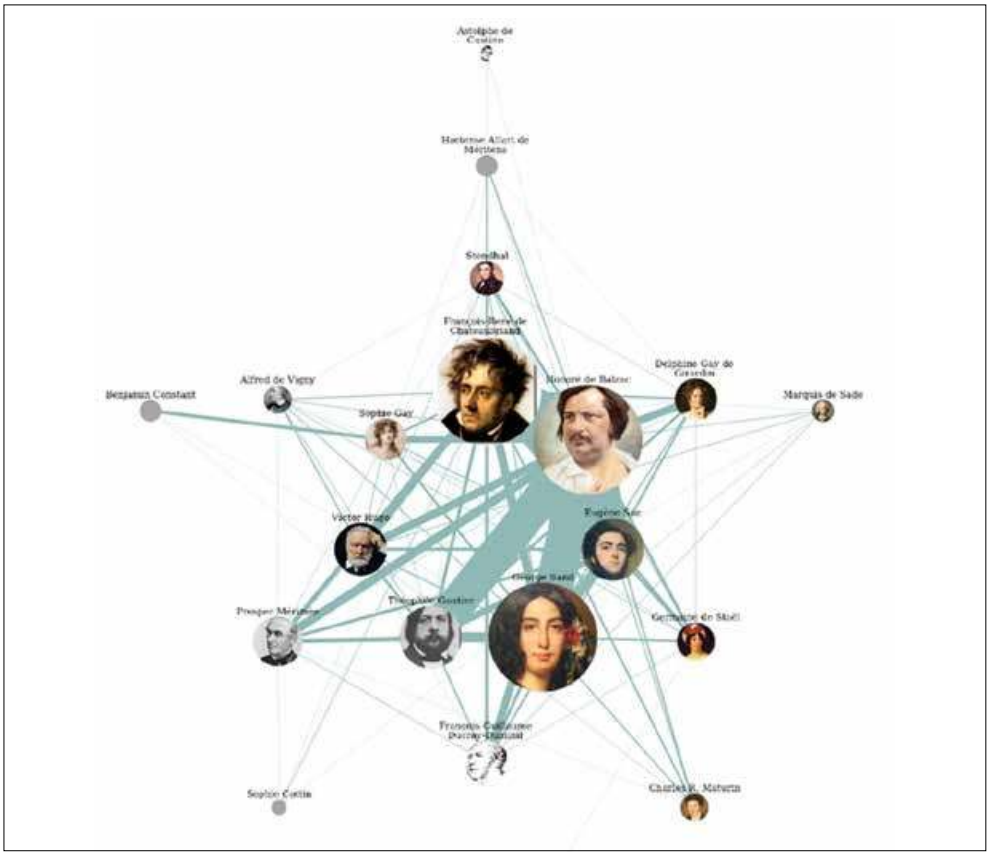

Fig. 6 : Influences mutuelles de différents romanciers du XIXe siècle.

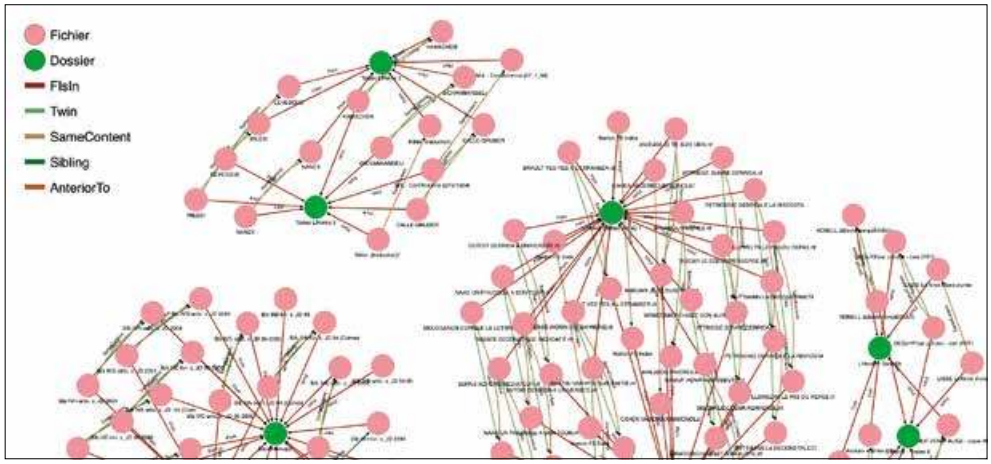

Fig. 7 : Visualisation du graphe des fichiers et dossiers inclus dans le dossier nommé HERNE du disque dur de Jacques Derrida.

Étude diachronique des influences

Étant donné que les œuvres sont datées, il est également possible d'analyser, dynamiquement, l'impact d'un auteur dans le temps, par exemple la façon dont les références explicites et/ou implicites à Jean-Jacques Rousseau ont augmenté ou diminué entre le XVIIIe et le XXe siècle.

Ce type d'approche semble particulièrement utile pour les disciplines d'érudition, par exemple pour les spécialistes d'histoire de la littérature qui pourraient ainsi quantifier l'évolution de la valeur littéraire accordée à un auteur en évaluant le nombre de références à son œuvre. Plus généralement, cette approche permet de mesurer avec précision l'évolution de l'influence dans le temps. Pour ce faire, le plus simple est d'utiliser des fenêtres temporelles glissantes et de voir comment le graphe de réutilisations évolue. D'un point de vue technique, nous nous proposons de recourir à ce que l'on appelle des « graphes de flux » (*stream graphs* en anglais) et des « flux de liens » (*link stream* en anglais) qui ont été récemment introduits pour traiter des graphes évolutifs[22].

Perspectives en génétique textuelle

Enfin, les graphes de similitudes peuvent aussi rendre de grands services à la génétique textuelle de l'âge numérique. Les auteurs classiques ont souvent laissé différentes versions de leurs textes sur papier. Leur collation permet de saisir les étapes du processus créatif en examinant les différences entre versions et en caractérisant les transformations avec les opérateurs génétiques spécifiques (addition, suppression, remplacement, déplacement) mis en évidence par les spécialistes de génétique textuelle.

Aujourd'hui, comme les écrivains utilisent des ordinateurs, sans toujours imprimer toutes les versions intermédiaires de leurs travaux, les généticiens sont confrontés à de nouveaux problèmes : ils doivent extraire directement les différentes versions des œuvres des disques durs des auteurs et ensuite reconstruire la chronologie entre les fichiers correspondant aux différents états des textes.

En raison de la facilité avec laquelle il est possible de dupliquer des fichiers, différentes copies prolifèrent, parmi lesquelles il est très difficile de s'orienter. Pour donner une idée de cette prolifération, la figure 7 montre quelques-uns des fichiers présents sur l'un des disques durs de Jacques Derrida dans un seul dossier nommé HERNE qui renvoie à un numéro de ce journal.

22. Matthieu Latapy, Tiphaine Viard, Clémence Magnien, « Stream graphs and link streams for the modeling of interactions over time », *Social Network Analysis and Mining*, 3 octobre 2018 (en ligne sur link.springer.com).

Habituellement, le nombre de fichiers est si important qu'il ne peut être traité manuellement. Pour illustrer ce point, notons que le nombre de fichiers appartenant au disque dur d'un des ordinateurs du philosophe atteint un peu plus de 20 000. Pour étudier l'œuvre de l'auteur de manière précise et scientifique, on doit être en mesure de dresser automatiquement un inventaire de tous les dossiers et ensuite d'établir des liens d'antériorité, de duplication, etc. en utilisant différents arguments philologiques, tels que le titre, la date, etc. C'est exactement ce que nous envisageons de faire dans le cadre du projet « Derrida hexadécimal[23] » auquel collaborent Aurèle Crasson, Jean-Louis Lebrave et Jérémy Pedrazzi de l'ITEM, l'IMEC et moi-même, au LIP6. Sans rentrer dans le détail, la première étape de ce projet, qui vise à poser les jalons d'une génétique textuelle à l'âge du numérique, passe par la représentation des différents états du texte sur un graphe de similitudes qui constitue l'équivalent du dossier génétique. Dans ce cas, les sommets ne sont plus des passages d'œuvres comme sur les graphes de similitudes précédents, mais des états de textes, et les arêtes, des liens d'antériorité entre ces états de texte, qu'il faut inférer automatiquement à partir d'indices multiples.

23. Ce travail est tributaire des efforts de Karolina Suchecka qui a programmé les interfaces ayant permis la visualisation des graphes des figures 2, 3, 4, 5 et 6, de Fleur Gaudferneau qui a étudié et comparé les différents algorithmes de détection des communautés sur les graphes et, enfin, des travaux de Maëlle Dagot qui a défriché le projet « Derrida hexadécimal » en représentant, sous forme d'un graphe, le contenu de l'un des disques durs de Jacques Derrida. Je les remercie toutes les trois chaleureusement.

Professeur d'informatique à Sorbonne Université, **JEAN-GABRIEL GANASCIA** poursuit ses recherches au LIP6 où il dirige l'équipe ACASA. Spécialiste d'intelligence artificielle (EurAI Fellow) et d'apprentissage machine, ses travaux actuels portent sur les humanités numériques et sur l'éthique computationnelle. Il est aussi président du comité d'éthique du CNRS et du comité d'orientation du Cycle des hautes études de la culture. Dernier ouvrage : *Le mythe de la Singularité*, Seuil (2017).

Jean-Gabriel.Ganascia@lip6.fr

Résumés

Détection automatique de phénomènes intertextuels

Il est désormais possible de détecter automatiquement, avec des techniques inspirées de la détection de plagiats, des fragments textuels évoquant, du fait de leurs ressemblances, des citations ou des réutilisations. Cependant, lorsque la taille des corpus est conséquente, le nombre de similitudes détectées est si grand qu'on s'y perd. De plus, des expressions figées ou des clichés enfouissent les reprises les plus intéressantes. De façon analogue, on peut repérer, sur les disques durs d'écrivains, des fichiers très semblables correspondant soit à des duplications, soit à des états différents d'un même écrit. Là encore, le nombre de fichiers semblables apparaît vertigineux. Pour surmonter ces difficultés, nous proposons de représenter les grandes masses de similitudes textuelles sur des graphes et de tirer parti des opérateurs mathématiques sur les graphes, en particulier de la détection de « communautés » ou d'arbres couvrant minimaux, pour les regrouper de manière significative.

With methods inspired by plagiarism detection techniques, we can now automatically detect textual fragments that resemble quotations or that appear to be other forms of borrowing. However, when the size of the corpus is large, the number of detectable similarities is so great that one gets lost. Fixed expressions or clichés bury the most interesting reuses. Similarly, we can find files on writers' hard drives that are either duplicates or that are very different stages of the same text. Here again, the number of similar files appears staggering. To overcome these difficulties, we propose to represent large numbers of textual similarities on graphs while taking advantage of mathematical graph operators, in particular those that detect "communities" or minimal spanning trees, which can be then clustered in a meaningful way.

Es ist nun möglich, mit Techniken, die von der Plagiatserkennung inspiriert sind, automatisch Textfragmente zu erkennen, die aufgrund ihrer Ähnlichkeit Zitate oder Wiederverwendung von fremden Texten hervorrufen. Wenn der Korpus jedoch groß ist, ist die Zahl der entdeckten Ähnlichkeiten so groß, dass man den Faden verliert. Außerdem begraben feststehende Ausdrücke oder Klischees die interessantesten Wiederverwendungen. In ähnlicher Weise kann man auf den Festplatten von Schreibern sehr ähnliche Dateien finden, die entweder Duplikaten oder verschiedenen Zuständen der gleichen Schrift entsprechen. Auch hier ist die Anzahl ähnlicher Dateien schwindelerregend. Um diese Schwierigkeiten zu überwinden, schlagen wir vor, große Massen von textlichen Ähnlichkeiten auf Graphen darzustellen und mathematische Operatoren auf Graphen, insbesondere die Erkennung von „Gemeinschaften" oder minimal bedeckenden Bäumen, zu nutzen, um sie auf signifikante Weise zu gruppieren.

En la actualidad es posible detectar automáticamente, con las técnicas inspiradas en la detección de plagios, fragmentos textuales que evocan, por sus semejanzas, citas o reutilizaciones. Sin embargo, cuando la extensión de los corpus es importante, la cantidad de similitudes detectadas es tan grande que desorienta. Además, las expresiones fijas y los clichés ocultan las recuperaciones más interesantes. De manera análoga, se pueden detectar, en los discos rígidos de los escritores, ficheros muy parecidos que corresponden ya sea a duplicaciones, ya sea a estadios diferentes de un mismo escrito. También en este caso, la cantidad de ficheros similares resulta vertiginoso. Para superar estas dificultades, lo que proponemos es representar las grandes masas de semejanzas a través de grafos y sacar partido de operaciones matemáticas aplicadas a los grafos, en particular, de la detección de "comunidades" o de arborescencias que vinculan rasgos mínimos, para reagruparlos de manera significativa.

É agora possível detetar automaticamente, com técnicas inspiradas na deteção de plágio, fragmentos textuais que evocam, pela sua semelhança, citações ou reutilizações. No entanto, quando o tamanho do *corpus* é considerável, o número de semelhanças detetadas é tão grande que nos perdemos. Além disso, frases feitas ou *clichés* escondem as reutilizações mais interessantes. Da mesma forma, pode-se encontrar, nos discos rígidos dos escritores, ficheiros muito semelhantes, que correspondem quer a duplicações, quer a diferentes estados da mesma escrita. Além disso, o número de ficheiros semelhantes é impressionante. Para ultrapassar estas dificuldades, propomos representar as grandes massas de similitudes textuais em gráficos e tirar partido dos operadores matemáticos em gráficos, em particular da deteção de "comunidades" ou árvores recobrindo formas mínimas, para as agrupar de forma que seja significativa.

Utilizzando le tecniche basate sul rilevamento dei plagi, è ormai possibile rintracciare automaticamente dei frammenti testuali che indicano, per la loro rassomiglianza, citazioni o riutilizzazioni. Tuttavia, nel caso di corpus d'importante entità, il numero delle similitudini rilevate è così grande che si rischia di perdersi. In più, le espressioni idiomatiche o i cliché seppelliscono le citazioni più interessanti. Allo stesso modo, si possono rintracciare negli hard disk degli scrittori, file molto simili, che corrispondono o a copie o a versioni diverse di uno stesso testo ; ma anche qui, il numero dei file simili appare vertiginoso. Per superare queste difficoltà, proponiamo di rappresentare le grandi masse di similitudini testuali con dei grafi e di utilizzare degli operatori matematici sui grafi, in particolare il rilevamento di "comunità" o di "alberi ricoprenti minimi", per raggrupparli in modo efficace.