Humanités numériques et science ouverte
*Les collectifs*

Clarisse Bardiot, Esther Dehoux, Émilien Ruiz (dir.)

# La fabrique numérique des corpus en sciences humaines et sociales

Centrales pour toutes les disciplines relevant des arts, des lettres et des sciences humaines et sociales, les questions relatives à l'identification, la sélection, le classement, les modalités d'exploitation et de diffusion des matériaux nécessaires à la production de connaissances ne sont pas nées avec l'ère dite « numérique ». En histoire, pour prendre l'exemple qui nous est le plus familier, le rapport à la documentation fut ainsi d'emblée au cœur des réflexions méthodologiques qui ont accompagné la construction des savoirs historiques en discipline et l'émergence du métier d'historien. Dès 1898, dans leur *Introduction aux études historiques*, Charles-Victor Langlois et Charles Seignobos formalisent les opérations qui composent ——

Septentrion
PRESSES UNIVERSITAIRES

# Le projet *eBalzac* : construire une bibliothèque hypertextuelle des sources intellectuelles

Andrea Del Lungo et Karolina Suchecka

## Introduction

1. Le projet *Phœbus-eBalzac*[1] consiste à mettre en résonance l'ensemble de l'œuvre balzacienne (*La Comédie humaine*, les romans de jeunesse, les contes drolatiques, le théâtre et les œuvres diverses) avec un vaste corpus d'écrits contemporains, à aires culturelles multiples et de nature variée, qui ont pu la nourrir : œuvres romanesques d'autres auteurs de l'époque ; recueils collectifs de littérature panoramique (auxquels Balzac apporta des contributions) ; ouvrages scientifiques susceptibles d'avoir influencé la création balzacienne, notamment dans les domaines de la médecine, de la physiologie et des sciences naturelles. L'objectif du projet est de permettre des recherches et des comparaisons intertextuelles élaborées, à l'intérieur de l'œuvre de Balzac, et dans le corpus plus vaste de textes littéraires et scientifiques de l'époque, entre 1800 et 1850 afin de faire émerger des correspondances, de repérer des emprunts, des citations, des reprises, des plagiats éventuels, et de constituer ainsi une cartographie de l'univers intellectuel de Balzac à partir des traces que d'autres textes ont laissées dans l'œuvre.

2. Ce projet vise à fournir aux chercheurs en littérature et sciences humaines de nouveaux outils d'interrogation de vastes corpus textuels, susceptibles de permettre une connaissance approfondie de phénomènes génétiques, poétiques, stylistiques, ainsi que leurs implications en termes idéologiques et de mieux comprendre les processus qui régissent l'apparition d'une réutilisation, qu'elle soit avérée ou issue inconsciemment des traces de lecture. Du point de vue de la méthodologie littéraire, il se situe dans le domaine de l'intertextualité, de l'inter-discursivité et de la génétique de l'imprimé, auquel le corpus balzacien offre un champ d'application particulièrement fécond : Balzac a la spécificité de multiplier les supports d'écriture (livres, volumes collectifs, feuilletons, articles de journal), en réutilisant ses textes antérieurs ; en même temps, son œuvre, qui définit par son ampleur l'état socio-historique contemporain, se nourrit de l'apport d'autres textes (notamment ceux de la littérature panoramique), et intègre diverses formes de savoir qui dépassent le champ littéraire. Au terme de ce projet, on sera en mesure de cartographier les sources que Balzac a

1. Le projet a été financé par l'Agence nationale de la recherche pour la période 2015-2019 et porté par Andrea Del Lungo (ALITHILA, université de Lille), Pierre Glaudes (CELLF, Centre d'études de la langue et de la littérature françaises, Sorbonne Université) et par Jean-Gabriel Ganascia (LIP6, Laboratoire d'informatique de Paris 6).

utilisées au cours de l'écriture des différents romans de *La Comédie humaine* et de livrer un développement permettant d'effectuer le même type de recherche sur tout autre corpus textuel.

3. Dans le cadre de ce chapitre, nous présenterons le projet ANR *Phœbus-eBalzac* à partir de sa première réalisation, le site ebalzac.com[2], ouvert en avril 2017, pour se focaliser ensuite sur son dernier axe qui consiste en l'édition hypertextuelle de l'œuvre de Balzac, encore en phase de préparation. Il s'agira alors d'exposer la chaîne de traitement avec les logiciels TextPAIR et Galaxies, de souligner les problèmes que nous avons dû affronter notamment en ce qui concerne le tri et la visualisation des résultats, et de montrer enfin le prototype de visualisation des homologies détectées entre les textes de notre corpus, en commentant quelques résultats.

## *ebalzac.com* : édition numérique, génétique et hypertextuelle

4. Pour réaliser les objectifs du projet, il fallait naturellement disposer d'une édition numérisée fiable de *La Comédie humaine*, ce qui n'était pas une mince affaire : 95 textes, 25 millions de signes ! C'est donc par là que nous avons commencé, en rendant accessible cette édition, jusqu'alors inédite en ligne, grâce à l'ouverture du site ebalzac.com. Sa création a constitué l'occasion de définir un élargissement considérable du périmètre du projet *Phœbus*. En effet, dans le but de créer une édition exhaustive de l'œuvre, il a été décidé d'intégrer à ce site un volet sur l'histoire du texte balzacien, qui consiste à numériser et à rendre accessibles en ligne tous les états imprimés des textes, publiés du vivant de l'auteur, afin de permettre leur comparaison avec le texte de référence que constitue l'édition dite « Furne corrigé ». Le site *eBalzac*, outre l'accueil et la description du projet, comporte quatre grandes rubriques.

5. La première propose une édition électronique de l'œuvre de Balzac, à commencer par *La Comédie humaine*, dans une version inédite en ligne et philologiquement exacte, qui intègre les corrections apportées par Balzac sur son exemplaire personnel et qui corrige de nombreuses éditions antérieures se basant sur ce dernier état du texte. L'entrée dans les textes se fait suivant trois critères au choix de l'utilisateur, via des menus déroulants : plan de l'œuvre, ordre chronologique, ordre alphabétique. L'ensemble du site a été conçu graphiquement avec une séparation verticale au centre qui mime la page du livre et qui permet surtout d'articuler deux espaces en vis-à-vis : dans l'édition, la colonne de droite est consacrée au texte numérisé, et la colonne de gauche à l'ouverture (en cliquant sur le numéro de la page, figure 1) en mode image de la page du support d'origine (dans ce cas, l'exemplaire personnel de l'édition Furne, présentant les corrections de la main de Balzac). L'édition est multi-format, et donne notamment la possibilité de télécharger les textes en EPUB.

2. Cf. http://ebalzac.com/

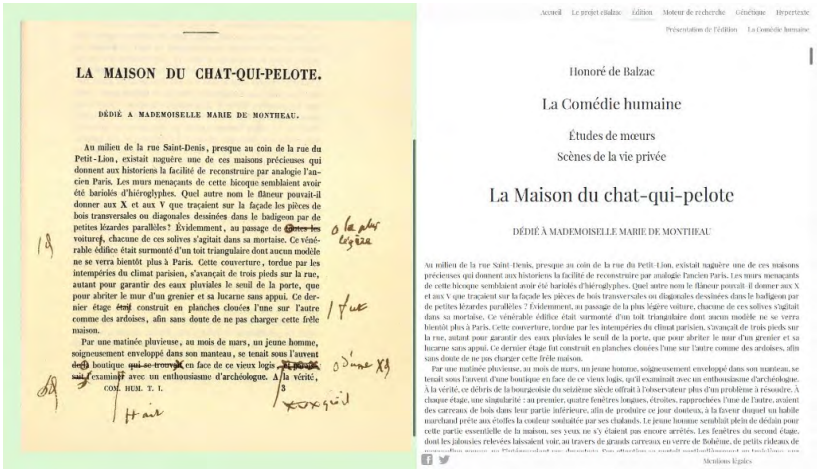



Figure 1. *La Maison du chat-qui-pelote* : édition Furne corrigé

Crédit : Andrea Del Lungo et Karolina Suchecka

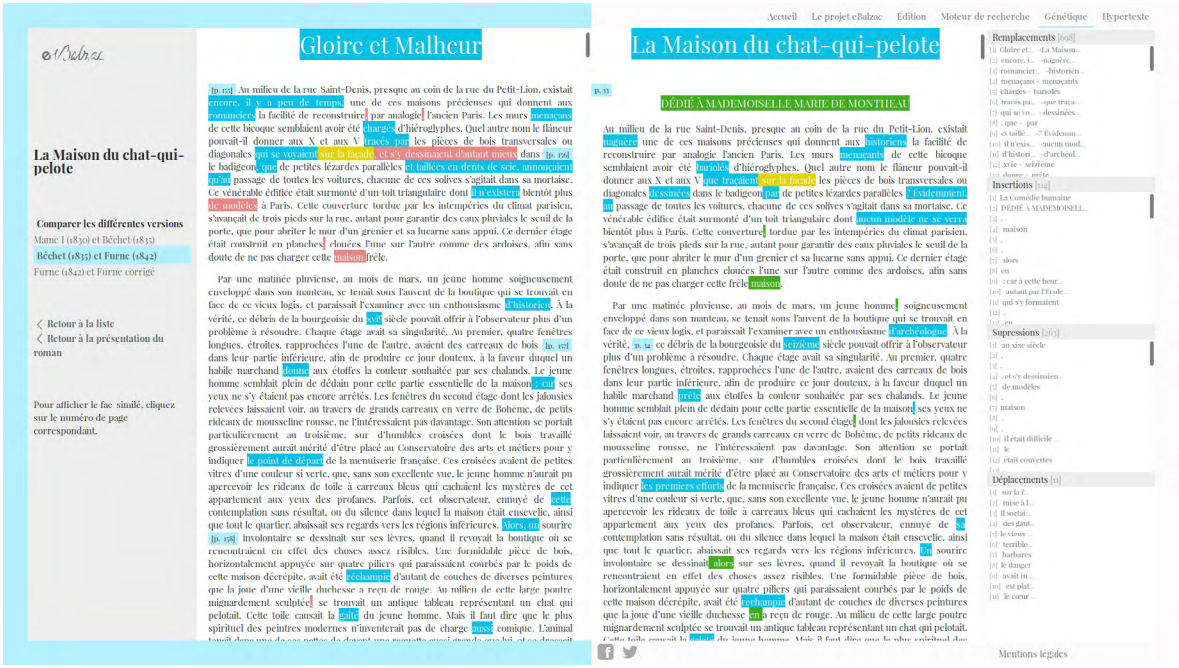



Figure 2. Comparaison des éditions Béchet (1835) et Furne (1842) de *La Maison du chat-qui-pelote*

Crédit : Andrea Del Lungo et Karolina Suchecka

6. Le deuxième axe propose l'édition des multiples états imprimés des textes de *La Comédie humaine* et offre la possibilité d'une comparaison informatique des différents états du texte (exploitant toujours le système graphique de la double colonne pour mettre les textes en vis-à-vis) via le logiciel Medite, permettant ainsi une étude génétique de l'œuvre de Balzac[3].

7. Medite est un outil de comparaison des versions d'une œuvre qui puise entre autres dans l'algorithme d'alignement par fragments grâce à la détection des homologies, une méthode de détection des séquences de macromolécules (ADN ou protéines) afin de faire ressortir leurs régions homologues (Ganascia et Bourdaillet 2006 ; Fenoglio et Ganascia 2008 ; Ganascia 2011). Actuellement, il propose une comparaison de deux textes en format brut : les blocs communs sont analysés et les différentes variantes sont signalées grâce à des codes-couleur. Les remplacements sont marqués en bleu, les insertions en vert, les suppressions en rouge et les déplacements en jaune.

8. Pour cet affichage génétique, nous avons développé une chaîne de traitement permettant d'exploiter la structure XML* afin d'introduire la mise en page éditoriale (alors que l'affichage initial n'admettait que le format de texte brut). Cela implique la facilitation de lecture par la prise en compte des éléments structurels (styles de paragraphe et de caractères, mais aussi images, etc.), l'affichage des fac-similés pour les deux textes comparés et l'alignement automatique au niveau des variantes et des blocs communs (figure 2).

9. Actuellement, la comparaison entre Furne et Furne corrigé est possible pour tous les textes ; la comparaison avec des états antérieurs, déjà disponible pour quelques textes (dont *Le Cousin Pons*), sera à court terme généralisée à l'ensemble, afin de montrer l'intégralité du parcours génétique balzacien à partir de la publication en feuilleton.

10. Le troisième espace du site est consacré à un moteur de recherche lexical, grâce au partenariat avec le projet *ARTFL* de l'université de Chicago, qui met à disposition les fonctionnalités d'un logiciel déjà développé (Philologic4[4]). Trois modes d'exploration sont possibles :

- la concordance (c'est-à-dire, l'extrait du texte avec le mot recherche)
- le KWIC (*Key word in context**) (qui permet un triage par le contexte droit et gauche)
- et enfin, la collocation qui montre le nuage des mots en cooccurrence

11. Enfin, le dernier axe du projet *eBalzac* est orienté vers une édition hypertextuelle afin de recomposer une bibliothèque virtuelle qui comprend l'ensemble des textes lit-

3. Sur la réécriture éditoriale chez Balzac et les enjeux de sa présentation numérique, cf. (Del Lungo 2017).

4. Cf. https://artfl-project.uchicago.edu/philologic4 et (Allen, Gladstone et Whaling 2013 ; Whaling 2010 ; Olsen 2008).

téraires et non littéraires dont on a repéré la trace dans l'œuvre de Balzac. Par son ampleur, mais aussi par le caractère hétérogène de ses sources, *La Comédie humaine* constitue un objet idéal pour ce type d'édition expérimentale qui pourra prendre valeur de paradigme. En effet, le modèle ainsi constitué vise à être opératoire pour d'autres auteurs chez qui l'usage d'une intertextualité abondante et éclectique est avéré.

## Détecter et visualiser les correspondances : prototype de l'édition hypertextuelle

12. Cette partie du projet, qui est la plus expérimentale et aussi la plus ambitieuse, reste encore à développer, mais des premiers résultats probants ont pu être obtenus grâce à la collaboration interdisciplinaire et internationale menée dès le début du projet.

13. Les travaux des ingénieurs et de chercheurs en informatique du laboratoire ACASA de LIP6 (Sorbonne Université, sous la direction de Jean-Gabriel Ganascia) et du projet *ARTFL* de l'université de Chicago (sous la direction de Clovis Gladstone) ont abouti au développement de deux prototypes de logiciels qui, ensemble, permettent de détecter les reprises assez subtiles entre les différents textes d'un grand corpus textuel et de visualiser les résultats à l'aide de graphes.

14. Notre rôle principal a été d'enrichir et d'adapter ces prototypes pour qu'ils répondent le mieux aux besoins des utilisateurs visés, notamment des chercheurs en littérature et en linguistique. Cela inclut, principalement, le développement des visualisations modulables, la préparation du corpus XML des textes analysés (optimisation des métadonnées*, calcul des fréquences, etc.). La détection des communautés pour les graphes de taille supérieure à 300 nœuds a été prise en charge par Fleur Gaudfernau (master d'analyse de données et d'intelligence artificielle à AgroParis Tech). Communément, nous avons également mis en place des fonctionnalités permettant de cibler les résultats les plus pertinents (scores, listes des lemmes communs, requêtes de filtrage, statistiques générales, etc.).

### La chaîne de traitement : collecter et exploiter le corpus des sources

15. La chaîne du traitement comporte actuellement trois étapes, dont la première est l'établissement d'un corpus des textes structurés à l'aide du standard XML-TEI*[5]. Ce corpus a été progressivement constitué dès le début du projet. À terme, il regroupera au total presque 500 œuvres de 56 auteurs différents, dont la totalité de *La Comédie humaine* de Balzac, qui constitue notre corpus principal (le corpus cible). Le corpus associé (source) est partagé en trois sous-ensembles :

5. Le lecteur pourra également se référer à l'entrée « OCR : *Optionnal character recognition* » pour plus d'informations sur le procédé de numérisation.

- le corpus romanesque d'autres auteurs contemporains ou antérieurs (George Sand, Chateaubriand, Gautier, Sue, etc.), qui est pour le moment le plus riche[6]
- les ouvrages de la littérature panoramique (depuis Mercier jusqu'aux recueils collectifs des années 1840)
- les ouvrages scientifiques contemporains, notamment dans les domaines des sciences naturelles, de la médecine et de la physiologie[7]

16. En ce qui concerne la détection des correspondances, un prototype de logiciel, nommé TextPAIR, a été conçu par le projet *ARTFL*[8]. Il permet de procéder, à partir d'un corpus XML, à la détection des correspondances selon des paramètres assez modulables. On peut, entre autres, choisir le nombre minimal des mots communs détectés, soumettre une liste des mots à ne pas prendre en compte et définir si l'on veut effectuer la recherche sur les mots pleins, les lemmes ou les stemmes (mots racinisés). C'est un logiciel d'alignement de séquences conçu pour identifier des passages similaires dans de grands corpus de textes en s'appuyant sur les techniques d'analyse de séquences employées dans les sciences dures, comme la bio-informatique, avec des applications allant du séquençage du génome à la détection du plagiat. Dans un premier temps, il génère un ensemble des séquences de mots qui se chevauchent pour chaque texte du corpus, puis stocke et indexe les informations à analyser par rapport aux séquences des autres textes. Par exemple, la déclaration liminaire du *Contrat social* de Rousseau, « L'homme est né libre, et partout il est dans les fers. Tel se croit le maître des autres, qui ne laisse pas d'être plus esclave qu'eux », sera traduite en séquences tri-grammes (avec lemmatisation, accents aplatis et mots faibles supprimés), comme : homme_naitre_libre, naitre_libre_partout, libre_partout_fer, partout_fer_croire, fer_croire_maitre, croire_maitre_laisser et maitre_laisser_esclave. Les séquences communes entre les textes indiquent de nombreux types d'emprunts textuels, des citations directes aux utilisations les plus ambiguës et non attribuées.

17. Deux problèmes principaux liés à TextPAIR ont mené à la création d'un logiciel de visualisation nommé Galaxies. Premièrement, le nombre important des résultats, présentés sous forme d'une base de données, où chaque couple de correspondances est inscrit sur une ligne, est très

6. 175 textes du sous-corpus romanesque ont été numérisés et structurés en XML-TEI de manière semi-automatique. Pour ce faire, des transformations XSL ont été conçues à partir du format adaptable DAISY DTBook, disponible sur Gallica pour certaines œuvres les plus connues, et à partir du format EPUB pour les œuvres numérisées par les bibliothèques numériques, dont une partie nous a été mise à disposition par le projet ANR *Chapitre*. Malgré quelques erreurs d'OCR qui persistent, ces numérisations sont à taux d'erreur beaucoup plus faible que par exemple celles disponibles sur Gallica en format texte, ce qui allège de manière significative la question de la gestion des bruits au sein du logiciel.

7. Ce corpus regroupe pour le moment six ouvrages : Nacquart, J-B. 1808. *Traité sur la nouvelle physiologie du cerveau* ; Gall, F. J. et J. G. Spurzheim. 1810. *Anatomie et physiologie du système nerveux en général et du cerveau en particulier*. 4 vol. ; Lavater, J. G. 1820. *L'art de connaître les hommes par la physionomie*. Vol. 1 ; Gall, F. J. 1832. *Sur les fonctions du cerveau et sur celles de chacune de ses parties*. 6 vol. ; Spurzheim, J. G. 1832. *Manuel de phrénologie* ; et Bourdon, l. 1842. *La physiognomonie et la phrénologie*.

8. Cf. http://artfl-project.uchicago.edu/text-pair et, par exemple, (Abdul-Rahman *et al.* 2017 ; Horton, Olsen et Roe 2011). Initialement, les expérimentations sur la détection des homologies ont été menées, dans le cadre du projet *eBalzac*, avec le logiciel Phœbus (http://obvil-dev.paris-sorbonne.fr/phoebus/). Cf. par exemple (Boukhaled, Sellami et Ganascia 2015 ; Ganascia, Glaudes et Del Lungo 2014).

difficile à explorer. Le format de sortie n'est pas adapté aux utilisateurs non-spécialisés et reste très peu lisible. Ensuite, cette détection binaire rend difficile la détection des correspondances croisées (où le même extrait d'un texte correspond à plusieurs autres extraits). Enfin, le nombre des banalités détectées est resté trop important pour que les résultats puissent être présentés au public des chercheurs. Un des buts de Galaxies a donc été de trouver une manière, en combinant les deux logiciels, de limiter le nombre des homologies non-pertinentes et de cibler les recherches selon les besoins spécifiques des chercheurs. Cela inclut, d'un côté, le développement des visualisations modulables à l'aide des graphes et des fonctionnalités comme le calcul de scores, la liste des mots communs, les requêtes de filtrage et les statistiques générales et, de l'autre, l'optimisation du traitement TextPAIR, notamment en écartant les mots les plus fréquents du corpus traité et en conformant la structuration du corpus à l'arborescence des métadonnées prise en compte par le logiciel.

18. Les résultats du logiciel sont ensuite analysés et enrichis par Galaxies afin de construire des graphes de correspondances (figure 3). Les couples de correspondances sont aussi comparés pour retrouver les éléments communs, que nous présentons sous forme de lemmes (pour rendre compte également des correspondances établies sur les différentes formes d'un même mot) et pour calculer le score de chaque couple. Ce score a été conçu au sein du projet, en se basant sur les calculs déjà existants et en expérimentant plusieurs méthodes afin de trouver la plus performante. Les résultats ont ensuite été soumis aux chercheurs littéraires afin de juger de leur pertinence. Le calcul retenu pour le moment prend en compte les proportions de la longueur de deux correspondances (le nombre des chaînes de caractères qui les composent), la fréquence inversée de chaque mot commun par rapport au corpus traité (moins le mot est fréquent, plus il aura de poids pour le score) et le ratio des chaînes de caractères appartenant aux mots communs par rapport à la totalité des chaînes de l'extrait (sur dix caractères de l'extrait, combien, en moyenne, appartiennent à des mots communs ?). Les recherches continuent afin d'améliorer ces résultats, mais la première version du logiciel permet déjà d'écarter la plupart des banalités et d'identifier immédiatement les correspondances les plus proches.

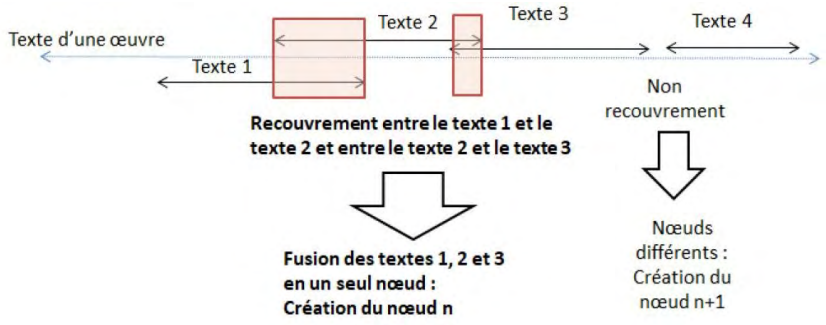



Figure 3. Schéma du traitement des résultats TextPAIR par Galaxies
© Fleur Gaudfernau, AgroParis Tech

19. La liste des galaxies affichée dans l'interface peut être triée selon plusieurs facteurs (score moyen, nombre de nœuds, etc.). L'utilisateur peut ensuite choisir la galaxie qui l'intéresse et l'afficher dans le navigateur, formuler

une requête de filtrage ou accéder aux statistiques générales concernant la totalité du corpus.

## Galaxies des relations : visualisation modulable des résultats

20. L'entrée dans les statistiques générales s'effectue par un graphe des auteurs dans lequel le nœud représentant Balzac est situé au centre. En cliquant sur un nœud représentant un des auteurs du corpus romanesque, l'utilisateur peut accéder aux informations relatives au nombre de correspondances détectées avec Balzac (figure 4). Le clic sur l'arête qui lie cet auteur avec Balzac transfère l'utilisateur à la page relative, qui présente à gauche le tableau des correspondances et, à droite, les cinq résultats les plus importants (où les textes sont jugés les plus proches).

21. Nous proposons également un graphique statistique qui illustre le nombre des correspondances détectées dans toutes les œuvres de Balzac (figure 5). Il est partagé en trois diagrammes, selon les différentes sous-sections de *La Comédie humaine*. En pointant la souris sur un des bâtons, les informations détaillées concernant les auteurs avec lesquels les correspondances ont été détectées s'affichent en infobulle.

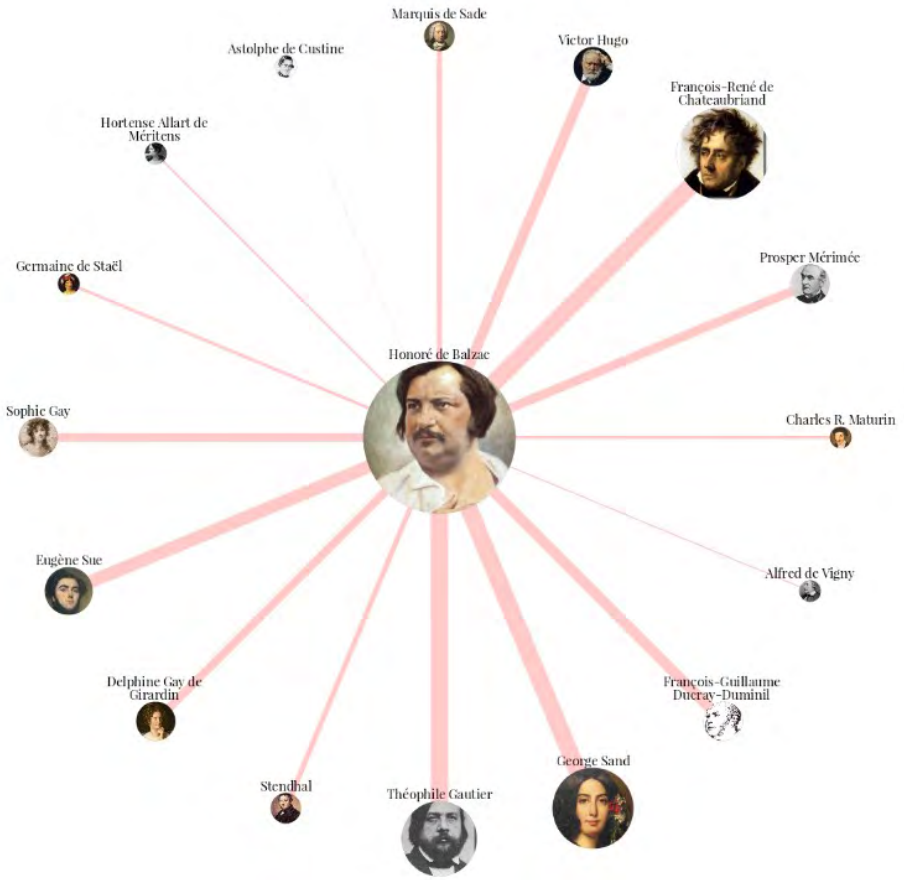



Figure 4. Statistiques générales : graphe des auteurs
Crédit : Andrea Del Lungo et Karolina Suchecka

22. En ce qui concerne les différents graphes qui ont été constitués pour les correspondances croisées, deux modes de visualisation sont disponibles : une concentrée sur les auteurs et les titres (figure 6) et une qui focalise les mots qui ont permis d'établir une correspondance

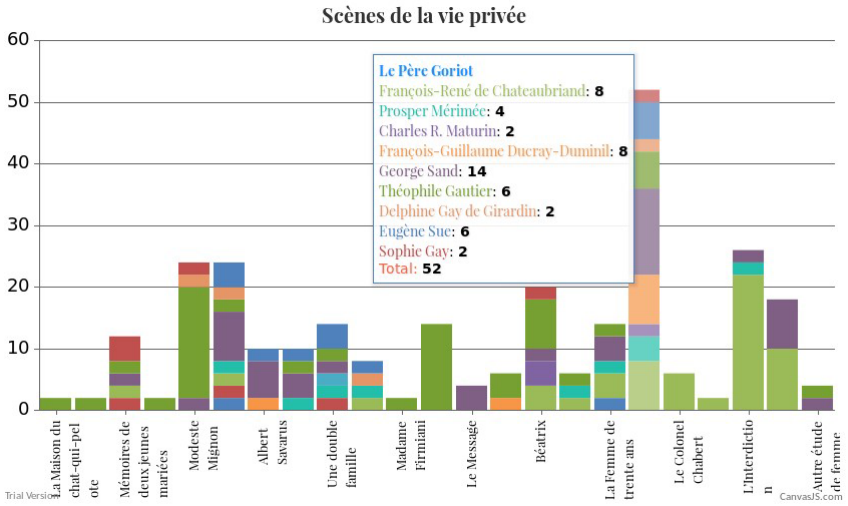



Figure 5. Statistiques générales : graphique des correspondances
Crédit : Andrea Del Lungo et Karolina Suchecka

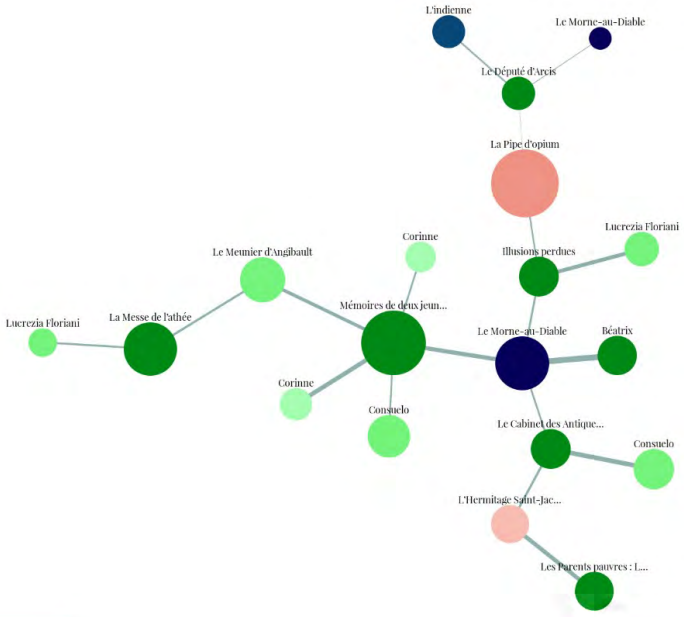



Figure 6. Visualisation focalisée sur les auteurs
Crédit : Andrea Del Lungo et Karolina Suchecka

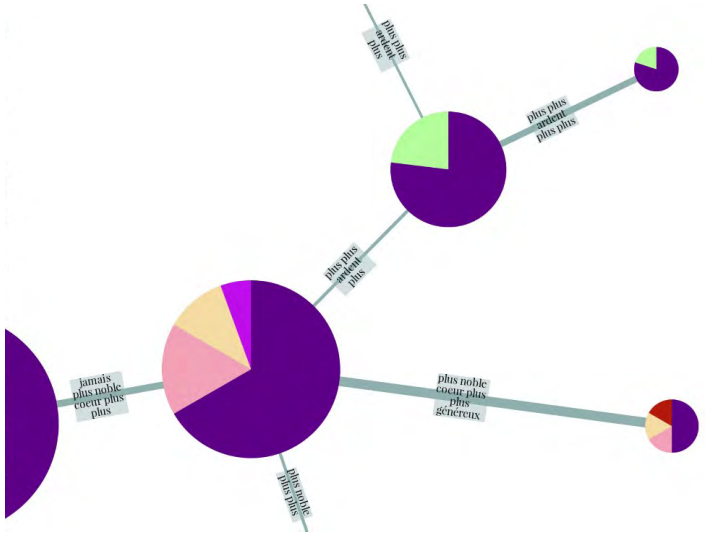



Figure 7. Visualisation concentrée sur les mots communs
Crédit : Andrea Del Lungo et Karolina Suchecka

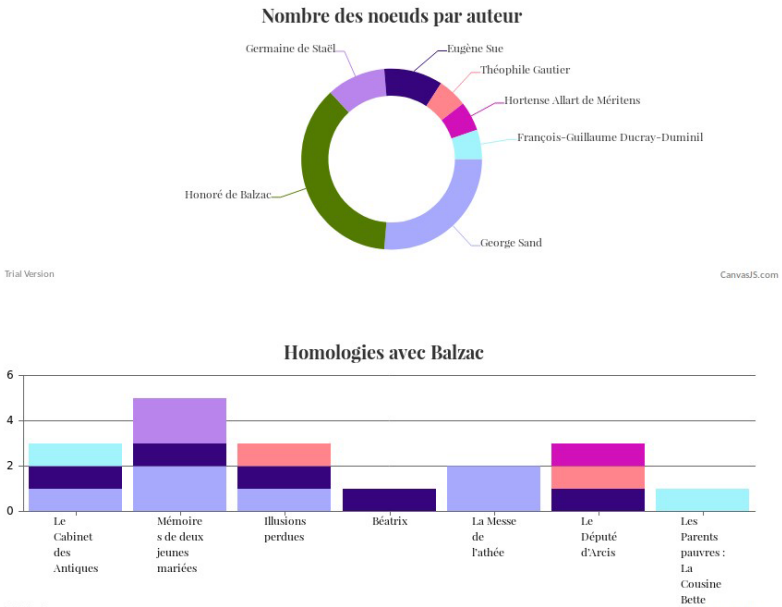
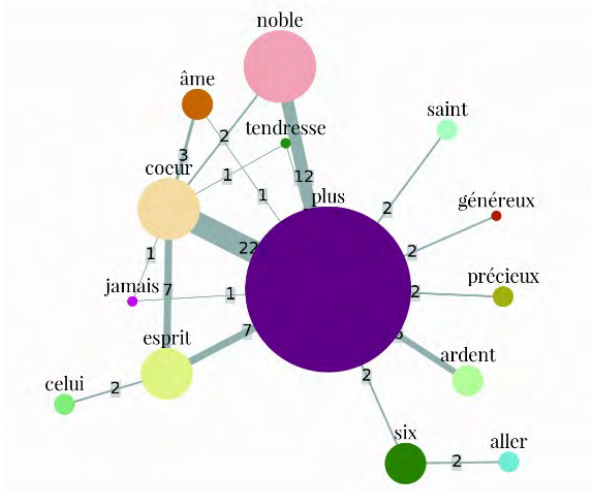



Figure 8. Graphiques statistiques
Crédit : Andrea Del Lungo et Karolina Suchecka

(figure 7). Il est également possible d'accéder aux diverses statistiques concernant la galaxie affichée : deux graphiques qui visualisent les proportions de la présence des différents auteurs et la répartition des correspondances dans les différentes œuvres de Balzac (figure 8), le graphe illustrant les cooccurrences des mots communs détectés, la liste des lemmes communs et leur nombre d'occurrences et enfin, le score maximal, minimal et moyen de la galaxie (figure 9).




**Lemmes communs**

| Lemme | Occ. | % |
|---|---|---|
| plus | 16 | 29.09 |
| noble | 7 | 12.73 |
| coeur | 6 | 10.91 |

**Scores de la galaxie**

| | |
|---|---|
| Score maximal : | 10.945 |
| Score moyen : | 5.07 |
| Score minimal : | 0.648 |

Figure 9. Graphe de cooccurrences, liste des mots communs et scores
Crédit : Andrea Del Lungo et Karolina Suchecka

## Exemple de l'exploration des résultats : Chateaubriand et *Sur Catherine de Médicis*

23. Plusieurs types de traitements ont été opérés : nous avons, entre autres, confronté le corpus balzacien aux autres textes du corpus romanesque en écartant les mille mots les plus fréquents ou en ignorant uniquement les mots faibles (les articles, les auxiliaires, etc.). En comparant ces deux approches, nous avons constaté que cette première restriction permet d'écarter un grand nombre de banalités, au risque d'omettre les correspondances pertinentes. Elle résulte également en un nombre restreint de galaxies complexes. Pour Balzac *vs* corpus romanesque sans les mots faibles, la plus grande galaxie compte 3 835 nœuds, dont les mots les plus courants sont « vingt », « sept », « an », « trois » et « bien ». En écartant les mille mots les plus fréquents, la galaxie la plus grande n'en compte que 18, mais les mots communs sont indubitablement plus pertinents : « Henri », « Charles », « François », « Marguerite », « reine », « roi » (figure 10 et figure 11).

| Identifiant | nombre de noeuds | score moyen | termes réutilisés les plus courants |
|---|---|---|---|
| 1 | 3835 | 6.6 | vingt sept an trois bien |
| 1-124 | 270 | 7.4 | somme neuf cent mille franc |
| 1-84 | 137 | 7.8 | quatre vingt mille livre rente |
| 1-91 | 121 | 6.6 | cent mille franc cinq cinquante |
| 1-69 | 100 | 5.6 | vingt cinq an lieu aller |
| 1-18 | 100 | 4.7 | eh bien partir dire oui |
| 1-11 | 100 | 4.7 | bien reprendre eh oui monsieur |

Figure 10. Balzac *vs* corpus romanesque sans les mots faibles
Crédit : Andrea Del Lungo et Karolina Suchecka

| Identifiant | nombre de noeuds | score moyen | termes réutilisés les plus courants | galaxie marquée |
|---|---|---|---|---|
| 59 | 18 | 12.3 | François II Charles IX Henri | |
| 25 | 16 | 7.5 | pièce servir cuisine salle manger | |
| 8 | 12 | 2.5 | salle manger rez donner | |
| 77 | 10 | 11.3 | cent rente viager foi payer | |
| 35 | 9 | 8.2 | habit bleu bouton ciseler porter | |
| 85 | 8 | 12.4 | tribunal premier instance département Seine | |
| 13 | 8 | 5.4 | tel | |

Figure 11. Balzac *vs* corpus romanesque sans les mille mots les plus fréquents
Crédit : Andrea Del Lungo et Karolina Suchecka

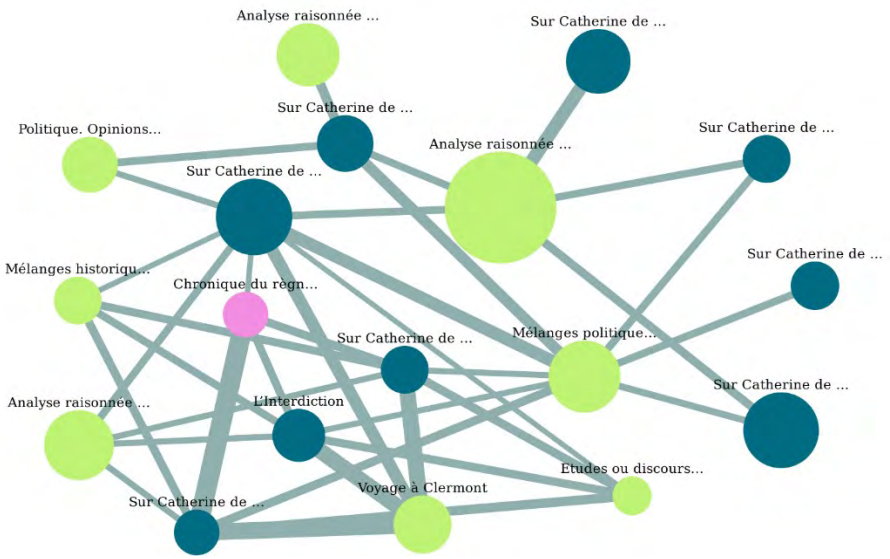



Figure 12. Galaxie n° 59 (18 nœuds)
Crédit : Andrea Del Lungo et Karolina Suchecka

24. Parmi les 18 nœuds textuels de cette galaxie (figure 12), neuf sont issus des textes de Balzac (en bleu), dont huit de *Sur Catherine de Médicis* et un de *L'Interdiction*, huit viennent des textes divers de Chateaubriand (en vert). Nous recensons également une occurrence de *Chronique du règne de Charles* IX de Prosper Mérimée (en rose). Si nous regardons les différents extraits (figure 13), nous pouvons nous rendre compte qu'il ne s'agit pas ici de plagiats ou réécritures, mais plutôt de proximités sémantiques, principalement des énumérations de rois et de reines de France.

25. Grâce à la visualisation focalisée sur les mots communs (figure 14), on peut se rendre compte des glissements thématiques au sein de la galaxie. Pour toute la partie gauche, les extraits traitent principalement d'Henri IV et d'Henri III (nœuds violets, roses et jaunes), alors qu'à droite, les entités nommées* communes sont plutôt Charles IX et François II.

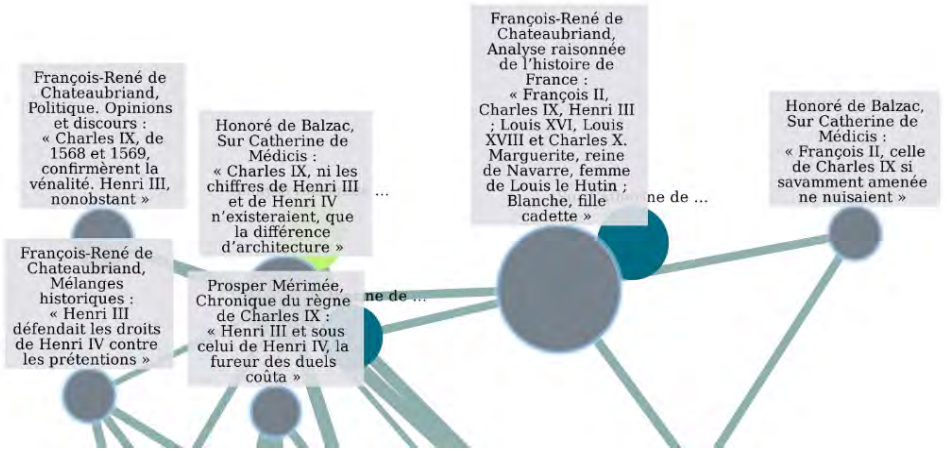



Figure 13. Galaxie n° 59 (18 nœuds) : extraits textuels
Crédit : Andrea Del Lungo et Karolina Suchecka

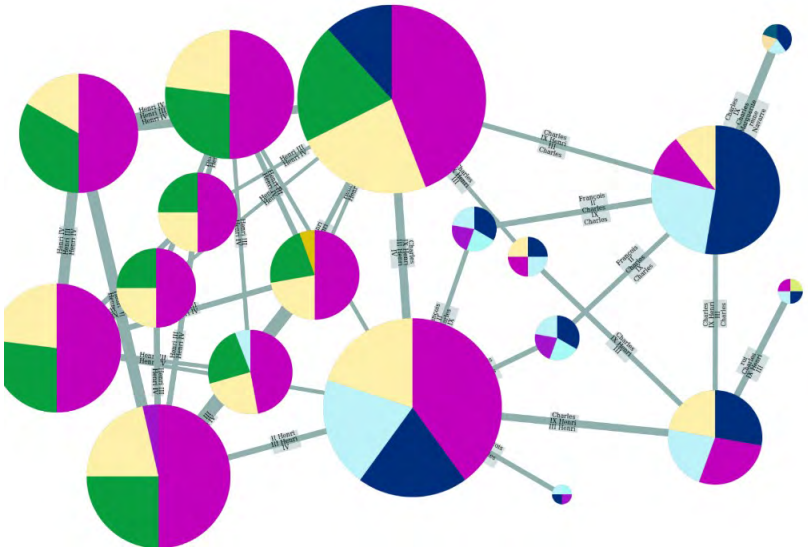


Figure 14. Galaxie n° 59 (18 nœuds) : glissements thématiques

Crédit : Andrea Del Lungo et Karolina Suchecka

Si nous filtrons tous les résultats pour ne garder que les galaxies contenant les textes de Chateaubriand et *Sur Catherine de Médicis*, nous obtenons 18 galaxies au total (dont le n° 59). Presque toutes les correspondances ont été constituées sur les noms propres de personnages historiques (Henri II, Diane de Poitiers, Marie Stuart, Catherine de Médicis, etc.) ou sur les noms de lieux constitutifs des noms de rois et reines (Poitiers, Navarre, etc.). Mais quelques résultats s'écartent de ce schéma.

26. La galaxie n° 53 (figure 15) constituée de trois nœuds, présente des extraits textuels qui sont quasiment les mêmes entre *Analyse raisonnée de l'histoire de France* de Chateaubriand, *Sur Catherine de Médicis* de Balzac et *L'Hermitage Saint-Jacques* de Ducray-Duminil : « Reine, n'ayant de femme que le sexe, l'âme entière aux choses viriles, l'esprit puissant aux affaires, le cœur invincible aux adversités ». Grâce au référencement direct de Chateaubriand qui fait partie des résultats, nous pouvons donc identifier automatiquement cette citation qu'Agrippa d'Aubigné formule à propos de Jeanne d'Albret, et qui reste inavoué chez Balzac, faisant partie d'une réplique du chancelier de Navarre.

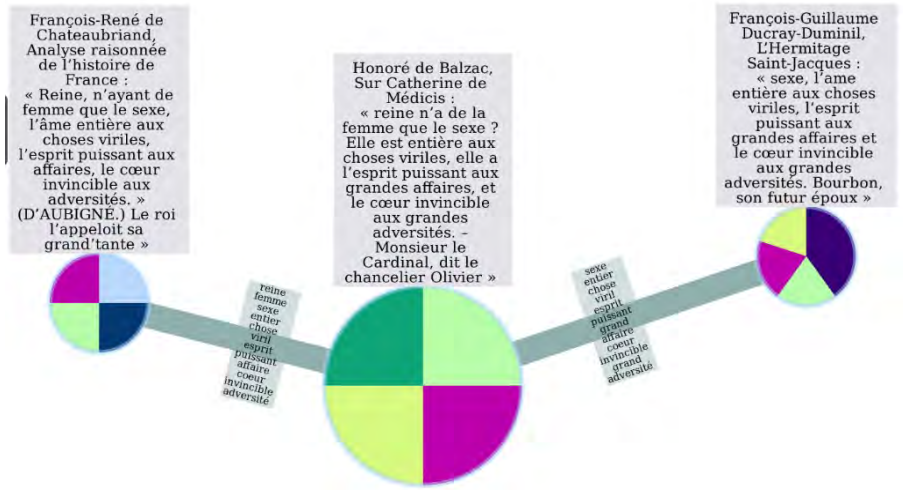



Figure 15. Galaxie n° 53 (trois nœuds), score : 30,9 – citation d'Agrippa d'Aubigné

Crédit : Andrea Del Lungo et Karolina Suchecka

27. Les deux correspondances suivantes (figure 16 et figure 17) restent dans l'ordre de l'énumération, mais la spécificité des entités nommées cooccurrentes (« Inde », « Perse », « Égypte », « Grèce »), les mots savants employés communément (« connétable ») et les adjectifs qualificatifs

(« fameux ») nous permettent, nous semble-t-il, de formuler la thèse suivante : sans que nous puissions parler de réécriture, Balzac documente les aspects historiques de ses œuvres, et notamment *Sur Catherine de Médicis*, en s'appuyant sur les œuvres de Chateaubriand.

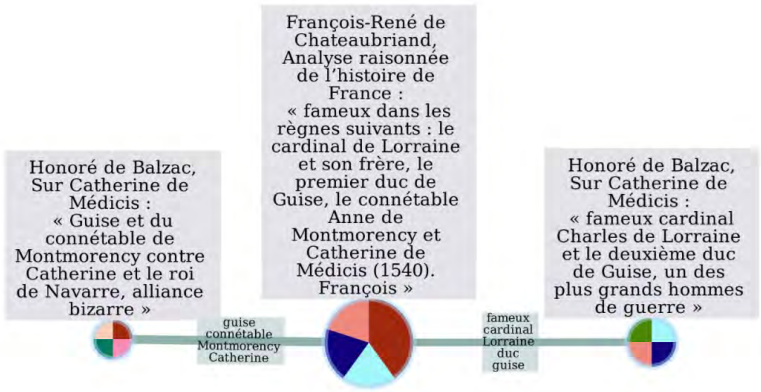



Figure 16. Galaxie n° 123 (trois nœuds), score : 12,1

Crédit : Andrea Del Lungo et Karolina Suchecka


Figure 17. Galaxie n° 65 (trois nœuds), score : 8,4

Crédit : Andrea Del Lungo et Karolina Suchecka

## Quelques autres découvertes : des écritures à quatre mains aux proximités sémantiques

28. L'expérimentation que nous avons présentée ici n'a montré qu'une parmi les nombreuses découvertes qui ont été permises par le logiciel et qui sont observables dans les premiers résultats.

29. Nous retrouvons également des réécritures à quatre mains (figure 18), notamment entre Balzac (*Une fille d'Ève*, *Madame de Firmiani*) et Théophile Gautier (*Mademoiselle de Maupin*) qui ont déjà été identifiées de manière « traditionnelle » par (Duclos 2013), et dont les extraits communs sont quasiment identiques. Toutefois, elles ne pourraient pas être retrouvées dans leur totalité par un logiciel de détection de plagiat classique, notamment à cause de quelques ajouts de Gautier (« avec leurs mille têtes chevelues ») et quelques reformulations (« sans faire saigner le cœur, sans que de ta tige brisée suintent des gouttes rouges » ou « sans faire saigner les cœurs à tous ses recoins, et de la tige brisée suintent des gouttes rouges »).

30. Des proximités sémantiques assez surprenantes ont également été relevées par exemple en ce qui concerne les descriptions des lieux et des personnages balzaciens très proches de celles d'Eugène Sue (figure 19). Si nous nous penchons sur les termes communs retrouvés dans toutes les galaxies contenant les textes de ces deux auteurs, nous nous rendons compte qu'ils appartiennent principalement à deux champs lexicaux : le vestimentaire (« redin-

gote », « bouton », « habit », « veste », « chapeau », etc.) et le mobilier (« rez-de-chaussée », « chambre », « cuisine », « salle à manger », etc.). Là encore, les résultats issus du traitement automatique permettent d'appuyer les analyses des chercheurs littéraires : les échanges entre les deux auteurs ont été, entre autres, analysés par (Lascar 2010). Mais l'approche critique que Balzac manifeste envers l'écriture de Sue, surtout à partir de 1836, rend cette proximité assez inattendue.


Figure 18. Galaxie n° 316 (deux nœuds), score : 84,3
Crédit : Andrea Del Lungo et Karolina Suchecka

31. Nous recensons également quelques expressions figées qui ne sont plus utilisées de nos jours et dont les occurrences sont assez importantes dans plusieurs œuvres du corpus (« méchant comme un âne rouge », « vouloir manger » ou « faire manger de la vache enragée », « conduite est un chef-d'œuvre de politique », etc.). Une expression empruntée au duc d'Albe « une tête de saumon vaut mieux que dix mille grenouilles » a été, par exemple, utilisée une fois par Prosper Mérimée (*Chronique du règne de Charles IX*) qui la cite mot à mot. Au contraire, deux réutilisations de Balzac, ont été légèrement reformulées : dans *La Paix du ménage*, il écrit qu'« un saumon vaut mieux que mille grenouilles », dans *Les Secrets de la princesse de Cadignan*, « une tête d'un seul saumon vaut celle de toutes les grenouilles ». Outre l'analyse littéraire, ce type de résultats a donc un certain intérêt également pour les linguistes et les historiens de la langue.

| Identifiant | nombre de noeuds | score moyen | termes réutilisés les plus courants | galaxie |
|---|---|---|---|---|
| 21 | 3 | 16.2 | rez chaussée premier étage chambre | |
| 34 | 3 | 18.3 | redingote bleu boutonner jusque cou | |
| 12 | 5 | 10.8 | chausser Antin faubourg saint Germain | |
| 41 | 2 | 11.7 | croisée dont carreau remplacer papier | |
| 25 | 16 | 7.5 | pièce servir cuisine salle manger | |
| 35 | 9 | 8.2 | habit bleu bouton ciseler porter | |
| 39 | 2 | 9.2 | expression cheveu noir ressortir oeil | |
| 30 | 2 | 9 | manoeuvre couronner plein succès | |
| 40 | 2 | 8.8 | longue perche charger linge | |
| 37 | 2 | 8.7 | environ soixante mille livre rente | |
| 31 | 2 | 8.5 | jambe droit gauche | |
| 43 | 3 | 8 | escalier conduire étage supérieur | |
| 33 | 2 | 8 | taille moyen svelte physionomie | |
| 23 | 3 | 7.3 | cravate noir nouer négligemment pantalon | |
| 36 | 3 | 6.4 | pantalon noir bas soie soulier | |
| 28 | 2 | 7.7 | partir éclat rire interdire | |
| 32 | 2 | 7.2 | chapeau forme rond bord | |
| 26 | 2 | 6.9 | profond régner troubler coup | |
| 22 | 4 | 6 | veste gros drap bleu chapeau | |
| 27 | 2 | 6.3 | nez bec oiseau proie | |
| 24 | 2 | 5.9 | paraître âgé trente quarante an | |
| 20 | 2 | 4.3 | sourire satisfaction lèvre commencer | |
| 38 | 2 | 4.1 | expression peindre trait | |
| 29 | 2 | 3.5 | perdre qualité mauvais | |
| 42 | 2 | 0 | | |

Figure 19. Liste de correspondances entre Balzac et Sue
Crédit : Andrea Del Lungo et Karolina Suchecka

32. Enfin, le logiciel permet de formuler ou de confirmer plusieurs hypothèses, et notamment celle de l'influence des théories scientifiques de l'époque – par exemple celles de la physiognomonie et la phrénologie – sur l'écriture balzacienne, surtout dans la description physique des personnages. Ainsi par exemple, une homologie significative a été retrouvée entre le traité de Gall *Anatomie et physiologie du système nerveux* (publié en français entre 1810 et 1819, en collaboration avec son disciple Johann Gaspar Spurzheim) et un passage de *La Grenadière* (figure 20). Autant Balzac s'inspire des hypothèses de la phrénologie pour décrire l'aspect extérieur du personnage comme indice de son caractère, autant il adapte ces théories et les modifie dans le contexte de la fiction. Chez Gall, le front haut et bombé est un signe d'intelligence, alors que chez Balzac il renvoie à l'énergie de la vigueur. On observe ainsi un déplacement du paradigme herméneutique du domaine intellectuel au domaine physique.

33. Cet exemple de Balzac s'inspirant d'un ouvrage de phrénologie de Gall ne serait plus alors à considérer comme une simple source, mais deviendrait le nœud d'un réseau conceptuel susceptible de montrer comment les modèles scientifiques sont investis, mais aussi déformés dans une œuvre de fiction qui prend la valeur d'une forme de connaissance, et qui contribue à établir de nouveaux paradigmes. Il serait alors possible de repenser la relation de l'auteur avec un ensemble de disciplines positivistes qui ont pu fonder son œuvre, en dépassant la traditionnelle étude des sources afin de situer l'auteur dans un réseau : celui d'un savoir partagé de la culture d'une époque.

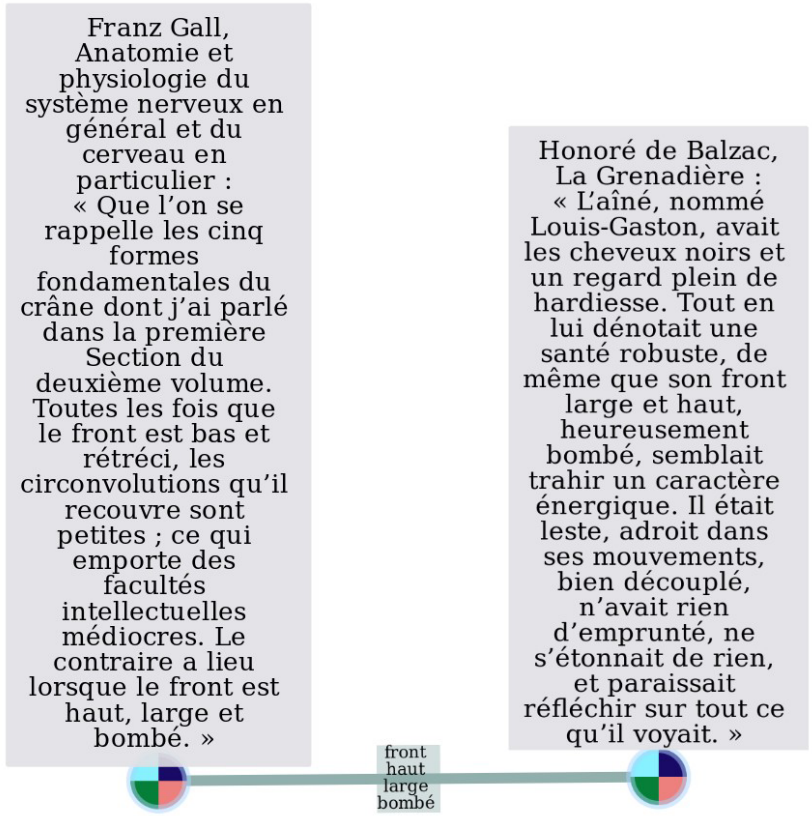



Figure 20. Galaxie n° 4 (deux nœuds), score : 7
Crédit : Andrea Del Lungo et Karolina Suchecka

34. Ces relations croisées sont par ailleurs visibles également dans les premiers résultats de la confrontation du corpus romanesque à lui-même (figure 21) : en approfondissant l'analyse de ce traitement, plus problématique du point de vue du traitement informatique puisque les mêmes

textes constituent le corpus source et le corpus cible, nous pensons qu'il est possible de prouver que les procédés de réécriture ne sont pas caractéristiques uniquement à Balzac, mais marquent l'ensemble de l'histoire littéraire, au moins au XIX[e] siècle.

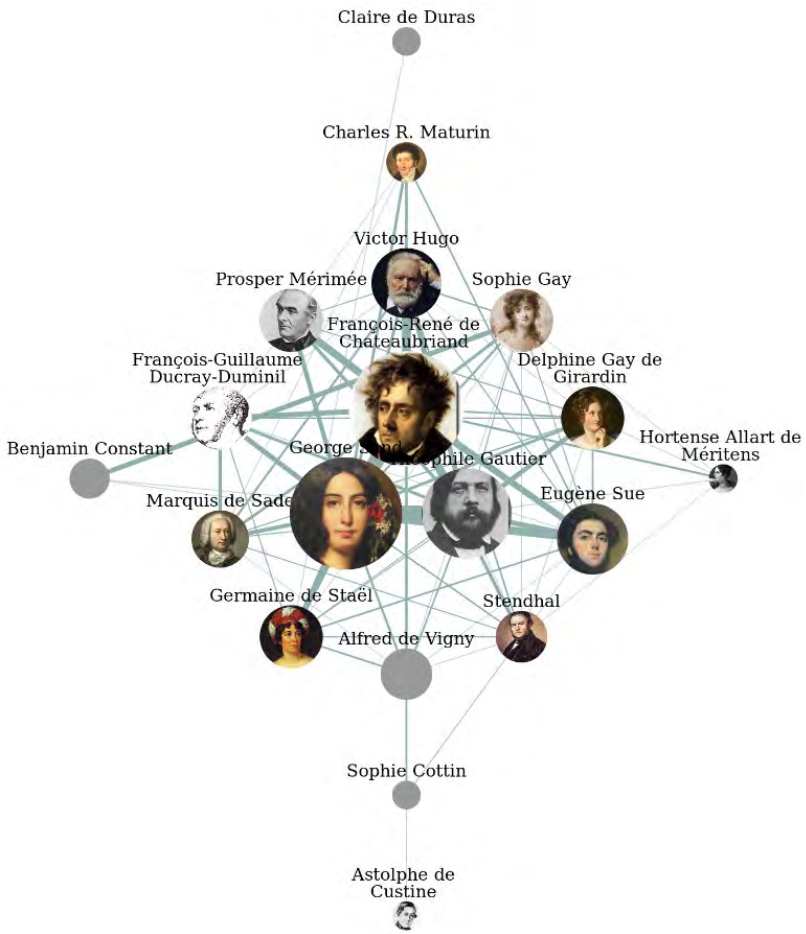



Figure 21. Graphe des auteurs pour le corpus romanesque *vs* lui-même
Crédit : Andrea Del Lungo et Karolina Suchecka

## Conclusion

35. Soucieux d'inscrire notre projet dans l'esprit des humanités numériques ouvertes*, nous tenons à ce que tous les développements et numérisations produits dans le cadre de l'édition et du projet en général soient généralisés et mis à disposition de la communauté des chercheurs. Nous souhaitons qu'ils puissent être réexploités au-delà de notre recherche particulière afin de contribuer au passage, désormais indispensable tant dans le domaine des humanités numériques littéraires que de l'édition numérique savante, du quantitatif au qualitatif. Il ne s'agit donc pas seulement d'exploiter les possibilités offertes par les outils informatiques, mais aussi de réfléchir aux modalités d'adaptation des contenus enrichis à la lecture numérique, qui diffère de la lecture papier de manière significative, notamment par sa non-linéarité.

36. Le projet *eBalzac* ne se restreint donc pas à l'exploitation, à la mise en valeur ou à la numérisation d'un corpus. L'originalité de son approche numérique tient au caractère systématique d'une investigation opérée sur des quantités considérables de textes qu'il eût été impossible d'exploiter manuellement.

Pour consulter les données mobilisées dans le chapitre, voir **https://hnso-corpus.nakala.fr/**