Une expérience d'attribution d'auteur. Le corpus Saint-Jean.

Dominique Labbé

PACTE
CNRS – Université de Grenoble-Alpes

Une expérience d'attribution d'auteur

Le corpus Saint-Jean

Rapport de recherche
octobre 2017

Dominique Labbé
dominique.labbe@umrpacte.fr

Résumé :

Avec la collaboration de J. Savoy, dans le but de tester les méthodes d'attribution d'auteur, il a été constitué un corpus de 200 extraits tirés de 68 romans par 31 auteurs différents. Les différences de vocabulaire entre les textes sont mesurées grâce à la distance intertextuelle. Tous les extraits sont correctement attribués avec la technique du plus proche voisin mais cette méthode exige que les auteurs aient au moins deux textes dans le corpus. En l'absence de cette condition, on utilise les plus petites distances, en définissant un intervalle de confiance. Cette méthode permet d'attribuer, sans erreur, 8 extraits sur 10. Deux classifications (hiérarchique et arborée) aboutissent aux mêmes résultats. Une échelle standardisée de la distance intertextuelle permet d'attribuer un texte de manière simple et sûre sans avoir à reprendre l'ensemble de la procédure.

Abstract

With the collaboration of J. Savoy, a corpus has been compiled in order to test the methods of authorship attribution (200 excerpts drawn out of 68 novels by 31 authors). The vocabulary differences between texts are measured by the intertextual distance. With the help of the nearest neighbour method, all excerpts are correctly attributed but this attribution requires that every author should have at least two texts in the corpus. In the absence of this condition, the smallest distances are used (associated with a confidence interval). This method attributes, without error, 8 excerpts out of 10. Two classifications (hierarchical and tree-classification) lead to the same results. A standardized scale of the intertextual distance makes it possible to attribute a text in a simple and safe way without having to repeat the whole procedure.

Reconnaissance

Jacques Savoy (Université de Neuchatel) a eu l'idée de cette expérience et a participé à la réalisation du corpus mis en ligne sur le site du Centre de Linguistique de Corpus (www.unine.ch/clc) à la disposition des chercheurs qui souhaiteraient l'utiliser pour leurs propres expériences.

Cyril Labbé (Laboratoire d'Informatique de Grenoble) a collaboré à la mise au point de la distance intertextuelle et de son application à l'attribution d'auteur.

Cyril Labbé, Denis Monière et Jacques Savoy ont relu une première version de ce rapport.

A la lecture d'un texte, il est impossible de dire "qui a écrit ça ?" Bien sûr, les lecteurs érudits font des rapprochements, parfois pertinents, mais ils sont incapables de fonder leurs intuitions sur des caractéristiques objectives et mesurables intrinsèques aux textes. De ce fait, chez les universitaires de lettres, il n'y a aucun accord sur les éléments de vocabulaire ou de style qui pourraient caractériser un auteur par rapport à tous les autres.

Naturellement, l'attribution d'un texte douteux à un auteur connu reste possible mais en recourant à des éléments extrinsèques au texte (Love 2002). Ces éléments de contexte ne sont pas toujours disponibles ou sont parfois peu convaincants.

Cela a conduit à un certain nombre de conventions implicites : l'auteur est celui dont le nom figure sur la couverture du livre ou celui que les éditeurs et la tradition désignent comme tel. Parfois ces conventions sont bien fragiles mais l'habitude, voire la routine, empêchent d'en prendre conscience.

Cette impuissance a facilité des pratiques bien connues : pseudonymes, "plumes de l'ombre", "collaborations", etc. Pour peu que les associés et leur éditeur gardent le silence, ces associations restent indécelables. Pour un exemple politique : Monière et Labbé 2006. En littérature, les cas sont innombrables et beaucoup restent à découvrir.

Comment surmonter cet aveuglement ?

De même que le microscope ou le télescope aident à voir l'infiniment petit ou l'infiniment lointain hors de portée de notre vision, de même l'ordinateur peut suppléer les limites du cerveau humain et aider à l'identification de l'auteur d'un texte. Un certain nombre de méthodes ont été proposées (Koppel & Al. 2009 ; Stamatatos 2009 ; Savoy 2012a, 2012b, 2013).

En principe, toutes ces méthodes consistent à confronter des textes d'origine douteuse ou inconnue avec d'autres considérés comme "sûrs" (sous les réserves indiquées ci-dessus quant à la fragilité de ces certitudes) et à rechercher des proximités qui permettront de rattacher les premiers aux seconds.

Jacques Savoy et nous-même pensons que cette perspective mérite d'être creusée. Pour cela, il est souhaitable de mettre à l'épreuve les différents algorithmes en organisant des expériences sur de grands ensembles de textes (corpus) et en y conviant les chercheurs intéressés. Dans ce but, nous avons constitué un vaste corpus en langue française que nous mettons à la disposition de la recherche (en ligne sur le Centre de Linguistique de Corpus). L'expérience proposée a été baptisée Saint-Jean – non à cause du jugement dernier – mais parce que le travail a été réalisé par Jacques Savoy et nous-même dans un village de ce nom.

Cette note de recherche présente le corpus Saint-Jean et propose une procédure d'attribution d'auteur reposant sur la distance entre les textes (section I). Elle permet d'identifier le plus proche voisin de chaque texte et, ainsi, de reconnaître sans erreur – sous certaines conditions – les auteurs des 200 extraits constituant le corpus Saint-Jean (sections II et III). Diverses classifications donnent une représentation synthétique de ce corpus et offrent des outils intéressants pour l'histoire littéraire (section IV). Enfin, cette expérience valide l'échelle de la distance intertextuelle qui permet d'attribuer, sous certaines conditions, des textes d'origine inconnue ou douteuse (section V).

I. Sélection et préparation des textes, calcul des distances.

Notre attention s'est portée sur le roman français du XIXe siècle parce qu'il répond aux deux caractéristiques nécessaires à ce type d'expérience : un grand nombre d'auteurs et de textes dans le domaine public (pour pouvoir les communiquer à tous les chercheurs qui le souhaiteront sans violer les copyrights).

Le corpus

L'annexe 1 présente les textes retenus.

1. Une vaste population : 31 auteurs dont aucun ne pèse d'un poids tel que sa présence ou son absence puissent influer sur les résultats…

2. Deux cents extraits d'oeuvres écrites dans un même genre (le roman) et à des époques pas trop éloignées (ici un siècle) ;

3. Des extraits tirés au sort dans les oeuvres – ce qui explique que la plupart des extraits ne se suivent pas - avec cependant quelques contraintes dans la sélection :

- chaque extrait doit contenir 10 000 mots sans recouvrement ou collage avec d'autres extraits, ce qui élimine par exemple les textes courts comme les contes de Maupassant ;

- pas de proportion significative (>1%) de mots étrangers, d'argot ou de jargon. Comme on le verra plus bas, ces mots augmentent d'autant la distance entre ce texte et les autres, donnant alors au facteur thème un poids plus important que celui de l'auteur. Ceci a conduit, par exemple, à retirer de l'expérience certains romans comme *l'Ami Fritz* (Erckmann-Chatrian), *Une vieille maîtresse* (J. Barbey d'Aurévilly), *Le cousin Pons* (H. de Balzac) – qui contiennent respectivement beaucoup d'alsacien, de bas-normand et de jargon supposé restituer l'accent allemand de l'ami du cousin Pons - ou certains passages comme ceux contenant une proportion significative d'argot parisien dans E. Sue (premier Livre des *Mystères de Paris*) et dans V. Hugo (*Les Misérables*). On verra plus bas qu'il aurait fallu faire de même avec *De la terre à la lune* de J. Verne dont quelques passages comportent une proportion notable d'anglais (américain) et dont l'un de ces passages se trouve dans le corpus…

- pas de proportion significative du texte dans d'autres genres (pour les mêmes raisons). Ainsi sont éliminés les examens critiques, les avant-propos, les préfaces souvent assez longues ; les notes de bas de page, spécialement les renvois bibliographiques et, plus généralement, tous les « corps étrangers » comme les longues digressions topographiques, historiques et politiques de V. Hugo (septième livre des *Misérables*, troisième et cinquième livres de *Notre-Dame de Paris*)…

- tout auteur aura au moins deux extraits mais pas forcément tirés d'un même livre. Nous avons introduit des extraits uniques de certaines œuvres afin de juger l'aptitude des algorithmes à reconnaître l'auteur quand les deux textes portent sur les thèmes éloignés et/ou ont été rédigés à des époques différentes ;

- anonymat des extraits. Naturellement, il est facile de reconnaître l'auteur de textes contenant "Bovary", "Fabrice" ou "Goriot" ! Le remplacement des titres par des numéros signifie que l'identification doit se faire en "aveugle" - c'est-à-dire sans que les algorithmes utilisent aucune information sur les écrivains ni sur les titres des œuvres – et que l'objectif premier est d'identifier les textes écrits par une même personne et ceux écrits par des auteurs différents.

Enfin, la numérotation évite de former des blocs d'auteurs. Quand plusieurs extraits sont tirés d'un même ouvrage, cette numérotation respecte leur ordre dans le livre.

Traitements préalables des textes.

Les opérations préalables effectuées sur les textes sont décrites dans : Labbé & Labbé 2013b, Labbé 2002, Labbé 1990. Elles comportent :

- Le balisage. En tête du texte, figurent les références bibliographiques, la source électronique, la date des traitements. Puis, dans le cœur du texte, des balises isolent tout ce qui n'est pas le texte proprement dit. Par exemple, pour le théâtre, les "didascalies" : noms des acteurs, numéro des actes et des scènes, indications scéniques…

- Correction orthographique et standardisation des graphies. Par exemple : M., Mr., Monsieur, monsieur… voire M'sieu (J. Vallès). La première est la plus fréquente mais ne peut être reconnue automatiquement : monsieur, Marcel, Maurice, Marie, mètre(s), million(s)…? Quant à la dernière, tous les étiqueteurs automatiques (et les correcteurs orthographiques) la décomposent en "me" (pronom) et "sieu" (forme inconnue.) La question n'est pas anecdotique : dans le corpus Saint-Jean, le substantif le plus employé est "monsieur" avec une fréquence de 2,78 pour mille mots. Il est écrit, la plupart du temps- mais pas toujours – sous la forme "M.". Si l'on n'avait pas attaché à chacun de ces "M." une étiquette indiquant sa véritable identité, on risquait une cascade d'erreurs. D'une part, on conclurait à tort que le corpus contient très peu de « monsieur ». D'autre part, du, fait que ces points risquent d'être comptés comme des ponctuations, le découpage et la longueur des phrases seront erronés, etc.

Dans tout texte français, les variantes graphiques de ce genre concernent plus d'un mot sur dix, sans compter les fautes d'orthographe et les noms communs affublés d'une majuscule qui sont très courants chez certains auteurs. Que dirait-on de recensements de population affectés d'une incertitude supérieure à 10% et dont les incohérences dans les relevés interdiraient toute comparaison d'un recensement à l'autre ? C'est pourtant ce que font les logiciels dits d'"analyse des données textuelles" dont aucun n'opère de standardisation des graphies.

- Etiquetage : chaque mot du texte se voit doter d'une étiquette comportant sa graphie standard, son entrée de dictionnaire et sa catégorie grammaticale. Quand il s'agit bien de "monsieur", "M." est doté de l'étiquette "monsieur, nom masculin". Ou encore "Est" et "est" peuvent recevoir deux étiquettes : "être, verbe indicatif présent" ou "est, nom masculin". Ces étiquettes ne se substituent pas au texte original, elles s'y ajoutent et servent à établir le vocabulaire de ce texte, d'une œuvre, d'un auteur, d'une époque, d'un genre… et à identifier l'auteur en cas d'origine douteuse ou inconnue.

En effet, grâce à cette standardisation des unités d'observation, il devient possible de mesurer la proximité ou l'éloignement des textes d'un corpus entre eux mais aussi de comparer les résultats obtenus sur différents corpus (permettant de rapporter les résultats à un étalon unique ou "échelle").

Calcul des distances

Le calcul utilisé pour l'attribution d'auteur est présenté dans l'annexe 2.

Soit deux textes (A et B). L'ordinateur compte les mots différents au sein de ce couple. Ce nombre de mots différents constitue la distance qui varie uniformément entre 0 (tous les mots sont communs aux deux textes) et 1 (aucun mot commun). Par exemple, une valeur de 0,20

signifie qu'un mot sur cinq est différent ou encore que 80% des mots sont communs aux deux textes.

Cette distance est une réalité physique, comme le nombre de kilomètres séparant deux villes. Elle présente les propriétés d'une distance dans un espace euclidien : identité, symétrie, inégalité triangulaire.

Les expériences effectuées pour mettre au point la méthode ont permis d'identifier et de mesurer l'importance des principaux facteurs qui déterminent la distance entre textes. Par importance décroissante, il s'agit de :

- une distance minimale : un auteur n'écrit jamais deux fois le même texte avec les mêmes mots ;

- le genre : oral et écrit, prose, vers, comédie et tragédie, etc.

- l'écrivain. Chaque écrivain est caractérisé par la propension à utiliser un certain vocabulaire et à donner à certains mots ou catégories grammaticales un poids spécifique qui peut être mesuré grâce à leur "contribution" à la distance (sur cette notion : Labbé & Labbé 2003) ;

- l'époque où a été rédigé le texte car chaque époque possède un vocabulaire particulier et le lexique de chaque écrivain évolue avec le temps ;

- le thème (vocabulaire propre à ce thème, nom de personnages, de lieux, principaux motifs).

Les deux équations ci-dessous systématisent ces constats :

(1) $D_{(A,B)}=f\{Dmin_{(A,B)}, (Genre_A,Genre_B), (Auteur_A,Auteur_B), (Epoque_A,Epoque_B), (Thème_A,Thème_B)\}$

(2) $(Genre_A,Genre_B) > (Auteur_A,Auteur_B) > (Epoque_A,Epoque_B) > (Thème_A,Thème_B)$

Dans l'équation (1), le symbole *f()* signifie que la distance est **fonction** des termes indiqués entre crochets. Ces termes sont nommés "variables", car elles peuvent être chiffrées. Pour donner un **poids** à chacune de ces variables, on recherche des cas où toutes les autres sont nulles ou négligeables (raisonnement "toutes choses égales par ailleurs" que nous appliquerons dans la dernière section de cette note). Ces expériences ont conduit à l'inégalité (2) qui signifie que le poids du genre l'emporte sur celui de l'auteur qui est supérieur à celui des différences temporelles – à l'échelle d'une vie de création - et aux différences de thèmes (sous réserve que celles-ci ne s'accompagnent pas d'une proportion significative de mots étrangers, argots, jargons…).

On tire de (1) et (2) que **l'attribution à un auteur donné de certains textes anonymes ou d'origine douteuse est possible si l'on dispose de textes contemporains dont l'attribution est certaine et qui sont écrits dans le même genre que le ou les textes à attribuer**. Cette attribution sera facilitée si ces textes portent sur des thèmes pas trop éloignés et que leurs compositions sont séparées par un court laps de temps.

Sachant que les textes du corpus Saint-Jean sont tous des romans, l'inégalité (2) indique que le facteur auteur est celui qui pèse le plus lourd dans les distances entre chacun d'eux, c'est-à-dire que les textes les plus proches seront du même auteur, même s'ils ont été écrits à des époques différentes et/ou sur des thèmes éloignés.

On tire également de (1) et (2) une certitude et une question. Premièrement, la proximité sera d'autant plus forte qu'époque d'écriture et thèmes sont plus proches, ce qui est normalement

le cas entre extraits tirés d'un même ouvrage (hormis les "recueils"). Deuxièmement, une question reste en suspens : le cumul des facteurs thème et temps peut-il être supérieur au facteur auteur ?

On en déduit la loi suivante que l'expérience devrait permettre de tester : dans le corpus Saint-Jean – où tous les auteurs ont au moins deux textes -, pour tout texte A, son voisin le plus proche (B) doit être de la même plume. Ou encore : tous les plus proches voisins doivent avoir le même père.

II. Attribution d'auteur par les plus proches voisins

Remarque : pour aider le lecteur, la suite de cette note donne les titres et les auteurs correspondant aux numéros des textes mais il faut se souvenir que toute la procédure se déroule en aveugle (l'ordinateur ignore tout des textes et des auteurs) et qu'elle est entièrement automatique (aucune intervention de l'opérateur). Naturellement, le corpus étant du domaine public, tout le monde pourra refaire ces expériences et en contrôler les résultats.

La notion de plus proche voisin et son emploi dans les procédures de partition des grandes populations ont été présentés dans Cover & Hart 1967 et 1991 et son application aux textes et aux grandes bases de données dans : Hall, Park & Samworth 2008 ; Han, Karypis & Kumar 1999.

Procédure de classification par les plus proches voisins

Le tableau en annexe 3 donne les plus proches voisins de chacun des 200 textes (classés par distance croissante).

La procédure consiste à agréger ensemble chaque texte et son plus proche voisin, en descendant dans le tableau, ligne à ligne, jusqu'au 200e texte. Ainsi, les deux premières lignes indiquent que les textes 030 et 116 sont séparés par une distance de 0,1777 (soit 1 777 mots différents pour 10 000 mots) et qu'il s'agit d'un voisinage réciproque (030 et 116 sont plus proches voisins l'un de l'autre). Au passage, cela permet de vérifier l'une des propriétés des distances euclidiennes (la symétrie) : la distance est la même qu'on la mesure d'un point à un autre ou réciproquement ($d_{(A,B)} = d_{(B,A)}$).

L'automate crée un groupe n° 1 dans lequel il range ces deux voisins. La huitième ligne permet d'ajouter à ce groupe le texte 150 grâce à son voisinage avec 116. Ici le voisinage n'est pas réciproque mais, du fait d'une autre propriété de toute distance euclidienne (l'inégalité triangulaire), il est possible d'affirmer que la distance entre 030 et 150 est la plus faible qui puisse exister entre l'extrait n° 150 et un autre quelconque des 196 textes restant à classer. De fait, le texte 150 est le second plus proche voisin de 030, avec une distance de 0,2012 (onzième ligne du tableau en annexe 5). Du fait de cette *transitivité*, les trois extraits sont classés ensemble dans le groupe n°1. Lorsqu'on parviendra à la 17e ligne, le texte n° 013 vient s'agréger à ce même groupe grâce à cette même propriété (plus proche voisin n°150). Le tableau de correspondance en annexe 1 permet de savoir que G. Sand est l'auteur de ces quatre extraits et qu'ils sont tous tirés de *La petite Fadette*.

Puis viennent, dans les lignes 3 à 7 de l'annexe 3, cinq extraits de *Delphine* (A.-L. de Staël-Holstein), classés en deux groupes (2 et 3) car {157, 183 et 191} ne partagent aucun plus proches voisins avec {123, 141 et 105}. La même absence de plus proche voisin commun explique que les extraits des *Mystères de Paris* (E. Sue) sont classés en deux groupes distincts (groupes n° 27 et 47 dans le tableau ci-dessous).

Etc.

Remarque : cette classification est entièrement automatique. Même en procédant manuellement, tout le monde trouvera les mêmes résultats, mais le recours à un programme informatique permet de parer aux erreurs toujours possibles dans les opérations manuelles quand elles portent sur de grandes masses de données.

La classification complète est donnée dans le tableau 1 ci-dessous. Les groupes sont numérotés en fonction de l'ordre de la première agrégation qui marque leur création. De ce fait, plus on s'élève dans le tableau plus les groupes sont hétérogènes (distance élevée entre plus proches voisins).

Tableau 1 Classification de textes par leur plus proche voisin dans l'ordre de leur première agrégation à partir du tableau des plus proches voisins (annexe 3).

N°	Numéros des extraits	Auteurs et ouvrages
1	013, 030, 116, 150	George Sand. *La petite Fadette*
2	157, 183, 191	Anne-Louise de Staël-Holstein. *Delphine*
3	105, 123, 141	Anne-Louise de Staël-Holstein. *Delphine*
4	068, 078	George Sand. *La mare au diable*
5	011, 028, 044	Alfred de Musset. *La Confession d'un enfant du siècle*
6	115, 133, 149, 163	Henri de Régnier. *Les Rencontres de Monsieur de Bréot*
7	106, 124, 142, 158	Emile Erkmann & Alexandre Chatrian. *Histoire d'un conscrit de 1813*
8	111, 129, 145, 161	Alphonse de Lamartine. *Geneviève*
9	108,126, 144, 160, 171	Eugène Fromentin. *Dominique*
10	045, 057, 134, 164	George Sand. *Indiana*
11	083, 087, 091, 096	Guy de Maupassant. *Mont Oriol / Une vie*
12	049, 060	Émile Zola. *L'Assommoir*
13	084, 088	Émile Zola. *La Fortune des Rougon*
14	185, 193, 197, 199	Jules Vallès. *L'Enfant*
15	058, 069, 099	Henri Beyle Stendhal. *Le Rouge et le Noir*
16	085, 095	Honoré de Balzac. *Le père Goriot.*
17	112, 130, 146, 162, 172	Pierre Loti. *Pêcheur d'Islande*
18	102, 120, 138, 154, 168, 176	Paul Bourget. *Une idylle tragique*
19	051, 062, 073	Alexandre Dumas. *Les trois mousquetaires*
20	038, 052, 063	Gustave Flaubert. *Madame Bovary*
21	155, 169, 177	Alphonse Daudet. *Le Petit Chos*e
22	040, 054, 065	Edmond et Jules de Goncourt. *Germinie Lacerteux*
23	070, 079	Alfred de Vigny. *Servitude et grandeur militaires*
24	003, 020, 037	Alexandre Dumas. *Le comte de Monte Cristo*
25	010, 027	Guy de Maupassant. *Bel-Ami*
26	014, 031, 046	Henri Beyle Stendhal. *La Chartreuse de Parme*
27	118, 136, 152, 166, 174, 182, 195, 198, 200	Eugène Sue. *Les Mystères de Paris*
28	067, 077	Guy de Maupassant. *Fort comme la mort*
29	104, 122, 140, 156	Alexandre Dumas. *Le Vicomte de Bragelonne*
30	103, 121, 139	Alphonse Daudet. *Le Petit Chose*
31	086, 090, 094	Gustave Flaubert. *Salammbô – Hérodias*
32	170, 178, 186, 196	Anatole France. *La Rôtisserie de la reine Pédauque*
33	071, 080	Emile Zola. *La bête humaine*
34	002, 019, 036	François-René de Chateaubriand. *Atala / René*
35	107, 125, 143, 159	Anatole France. *Le crime de Sylvestre Bonnard*
36	016, 033, 048	Alfred de Vigny. *Cinq-Mars*
37	117, 135, 151, 165, 181, 189	Charles-Augustin Sainte-Beuve. *Volupté*
38	092, 098	Emile Zola. *Germinal*
39	017, 034	Zola Emile. *L'Argent*
40	089, 093	Honoré de Balzac. *Eugénie Grandet*
41	001, 018, 035	Honoré de Balzac. *La cousine Bette*
42	101, 119, 137, 153	Jules Barbey d'Aurevilly. *Le chevalier des Touches*
43	043, 056	Guy de Maupassant - *Notre cœur*
44	012, 029	Gérard de Nerval. *Aurélia*
45	152, 190	Eugène Sue - *Les Mystères de Paris*

46	074, 076, 082	Gustave Flaubert. *Un cœur simple /l'Education sentimentale*
47	173, 180, 188, 194	Henri de Régnier. *La Double Maîtresse*
48	167, 175, 184, 192	Jules Barbey d'Aurevilly. *Les Diaboliques*
49	006, 023	Edmond et Jules de Goncourt. *Madame Gervaisais*
50	179, 187	Pierre Loti. *Madame Chrysanthème*
51	039, 053	Théophile Gautier. *Jettatura*
52	050, 061	Honoré de Balzac. *Grandeur et décadence de César Birotteau*
53	114, 132, 148	Marcel Proust. *Les plaisirs et les jours*
54	009, 026, 042	Alphonse de Lamartine. *Graziella*
55	097, 100	Honoré de Balzac. *La Maison Nucingen*
56	007, 024, 041	Victor Hugo. *Les Misérables*
57	110, 128	Joris-Karl Huysmans. *Marthe histoire d'une fille*
58	072, 081	Honoré de Balzac. *Le Colonel Chabert*
59	004, 021	Gustave Flaubert. *Bouvard et Pécuchet*
60	109, 127	Victor Hugo. *Les Misérables*
61	113, 131,147	Gérard de Nerval. *Les Illuminés*
62	055, 066, 075	Victor Hugo. *Notre Dame de Paris*
63	005, 022, 064	Théophile Gautier. *Avatar / Spirite*
64	047, 059	Jules Verne. *Le tour du monde en quatre-vingt jours*
65	015, 032	Jules Verne. *De la terre à la lune*
66	008, 025	Joris-Karl Huysmans. *A rebours*

Tous les textes sont classés sans erreur. Comme le laissaient attendre les inégalités présentées ci-dessus, la plupart des groupes correspondent aux extraits d'un même ouvrage (dans ce cas, les facteurs temps et thèmes sont minimisés sinon annulés).

Les extraits uniques sont correctement rattachés à leurs auteurs : A. de Châteaubriand (*René* à *Atala*) ; G. Flaubert, *Un cœur simple* à *l'Education sentimentale* et *Hérodias* à *Salammbô* ; T. Gautier (*Spirite* à *Avatar*) ; G. de Maupassant (*Une vie* à *Mont Oriol*). De même pour les recueils de contes qui sont en réalité autant d'ouvrages différents sur des thèmes proches : J. Barbey d'Aurévilly (*les Diaboliques*). On peut rattacher à ce cas, G. de Nerval (*les Illuminés*) et H. de Régnier (*Les rencontres de M. Bréot*). Autrement dit, comme le laissait prévoir l'équation n°1, la procédure reconnaît, sans erreur, l'auteur d'un texte isolé, quand un ou plusieurs autres textes écrits dans le même genre par le même auteur sont présents dans le corpus. Mais, pour ces extraits uniques, la distance est plus grande puisque les facteurs thème et temps jouent plus nettement.

Cependant, certains auteurs sont dispersés dans plusieurs groupes (comme H. de Balzac, A. Dumas, G. Flaubert, A. France, V. Hugo, G. de Maupassant, A. de Vigny, E. Sue, E. Zola), la plupart du temps pour des livres différents.

Malgré cette limite, les résultats valident l'inégalité (2) énoncée ci-dessus : l'auteur est le principal facteur déterminant la distance entre deux textes écrits dans un même genre. Ils confirment également que l'identification de cet auteur est d'autant plus facile que les thèmes sont proches et/ou la composition des textes n'est pas séparée par un laps de temps trop important.

Naturellement, si l'on souhaite poursuivre la classification, on prendra en compte d'une part les seconds, voire les troisièmes voisins – afin de fondre ensemble les groupes les plus proches - et, d'autre part, les plus fortes distances afin d'opérer des disjonctions entre les groupes (annexe 4).

Ces opérations apportent une réponse à la question : quels sont les textes qui n'ont pas été écrits par les mêmes auteurs ?

Procédure de disjonction des groupes par les plus lointains

La réponse est donnée par les couples séparés par les distances les plus fortes (annexe 4), en utilisant là encore la transitivité de la distance intertextuelle. Soit un texte A appartenant à un groupe où figurent également les textes B, C et D. Si X est le texte le plus lointain de A, alors l'auteur de X ne peut avoir écrit A mais aussi les textes B, C, et D. En considérant les 200 couples de textes les plus lointains (annexe 4), on aboutit à une nouvelle classification dont des extraits sont donnés dans le tableau 2.

Tableau 2. Détermination des groupes ayant des auteurs différents (entre parenthèses, le nom des auteurs – pour les œuvres, voir tableau 1).

1	Sand	34 (Chateaubriand), 41 (Balzac), 44 (Nerval), 49 (Goncourt), 51 (Gautier), 52 (Balzac), 54 (Lamartine), 61 (Nerval), 63 (Gautier), 64 et 65 (Verne), 66 (Huysmans)
2	de Stael	7 (Erkmann et Chatrian), 13 (Zola), 17 (Loti), 20 (Flaubert), 28 (Maupassant), 31 (Flaubert), 38 (Zola), 42 (Barbey d'Aurevilly), 46 (Flaubert), 47 (Régnier), 49 (Goncourt), 50 (Loti), 54 (Lamartine), 57 (Huysmans), 59 (Flaubert), 60 (Hugo), 62 (Hugo), 63 (Gautier), 65 (Verne)
(…)	(…)	(…)
30	Vigny	65 (Verne)
(…)	(…)	(…)
65	Verne.	1 (Sand), 2 et 3 (de Stael), 4 et 10 (Sand), 5 (Musset), 6 (de Régnier), 8 et 54 (Lamartine), 9 (Fromentin), 11 12 33 et 39 (Zola), 14 (Vallès), 15 (Stendhal), 16 et 58 (Balzac), 17 (Loti), 18 (Bourget), 20 (Flaubert), 21 (Daudet), 22 (Goncourt), 23 (Vigny), 24 (Dumas), 25 28 et 43 (Maupassant), 27 (Sue), 32 et 35 (France), 36 (Vigny), 37 (Sainte-Beuve), 42 (Barbey d'Aurevilly), 48 (Barbey d'Aurevilly), 51 (Gautier), 53 (Proust), 56 (Hugo), 61 (Nerval).
66	Huysmans	1 (Sand), 7 (Erkmann et Chatrian), 13 (Zola), 16 (Balzac), 19 24 (Dumas), 21 et 30 (Daudet), 24 (Dumas), 26 (Stendhal), 27 (Sue), 29 (Dumas), 30 (Daudet), 40 (Balzac), 45 (Sue), 52 (Balzac)

Certains groupes – comme le numéro 1 (G. Sand) –sont disjoints d'une minorité des autres (ici 11 auteurs sur 30). C'est le résultat moyen obtenu sur le corpus. Le résultat le plus médiocre est atteint pour A. de Vigny (*Cinq-Mars*). Ce résultat s'explique par le fait que les trois extraits de cette oeuvre sont les plus "centraux" du corpus (leur distance moyenne à tous les autres est la plus faible). Pour d'autres, la liste est presque complète. Par exemple, pour le 65e groupe (J. Verne), il ne manque à la liste que 64 (J. Verne) et 7 (Erkmann et Chatrian). Cela s'explique par le fait que les extraits de *De la terre à la lune* sont – avec ceux de *A rebours* (K.-J. Huysmans) -, les textes les plus "décalés" (leurs distances moyennes à tous les autres extraits du corpus sont les plus élevées).

Dans une attribution d'auteur, il peut être important de déterminer qui n'a pas pu écrire le texte sous revue : lorsque, dans un corpus, aucune distance n'est assez faible pour désigner l'auteur de ce texte, on pourra au moins en éliminer un certain nombre, réduisant d'autant le champ de la recherche.

Rappelons que l'attribution d'auteur consiste ici à répondre à la question : quels sont les textes écrits par un même auteur ? Puisque tous les textes ont été mariés au moins à un autre, l'expérience pourrait donc s'arrêter là. Cependant, dans le cadre de la recherche d'une méthode générale d'attribution d'auteur, une question reste pendante.

La méthode proposée n'est possible qu'à la condition de savoir que tout auteur a deux ou plusieurs textes, ce qui revient à dire : l'auteur du texte douteux est certainement présent dans le corpus. Que faire si cette caractéristique n'est pas certaine ? C'est-à-dire, dans le cas présent : on ignore qui sont les auteurs et l'on ne peut exclure la possibilité que certains de ces auteurs n'aient qu'un texte ?

III. Attribution d'auteur par les plus petites distances

D'après l'inégalité (2) ci-dessus, dans un corpus de textes écrits dans un genre unique, les distances les plus faibles désignent un même auteur pour les deux textes concernés. Mais si l'on ne sait pas combien il y a d'auteurs différents (et que l'on ne peut écarter la possibilité que certains auteurs n'aient qu'un texte), la question devient : jusqu'à quel niveau peut-on considérer que la distance entre deux textes est suffisamment petite pour conclure sans risque que l'auteur est le même ? Autrement dit, quelle est la probabilité associée à cette décision ? (Ce problème est également discuté dans Savoy 2016).

Le problème dépend du nombre possible d'auteurs et du nombre de textes de chacun d'eux. Par exemple, avec une centaine d'auteurs ayant chacun deux textes, il n'y aurait qu'une centaine de distances intra-auteurs, soit 0,5% des distances générées par la confrontation deux à deux des 200 textes ce corpus. Avec une cinquantaine d'auteurs ayant chacun 4 textes, on aurait 300 distances intra-auteurs, soit 1,5% du total, etc. Dans ces conditions, retenir les 5% des individus les plus petits – comme on le fait habituellement dans les études sur échantillons - aboutirait nécessairement à un certain nombre d'erreurs qu'il faut absolument éviter.

En définitive, la solidité des résultats dépendra du nombre de distances considérées, c'est-à-dire des seuils de décision retenus.

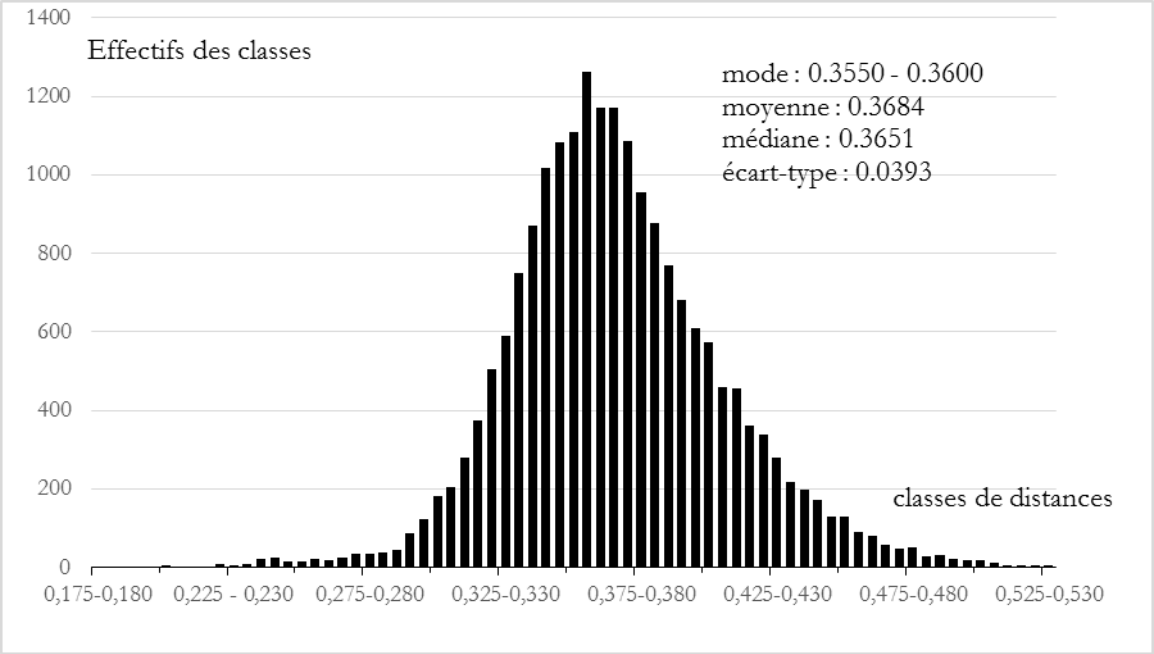
Les seuils de décision

Considérons qu'on ne sait rien de ces textes, à part leur appartenance à un genre unique, condition sans laquelle, d'après l'inégalité (2) aucune attribution d'auteur n'est possible, tant que l'on n'a pas procédé à une partition de l'ensemble pour dégager des clusters à genre unique dans lesquels l'étude des auteurs devient possible (Labbé & Labbé 2013a ; et Labbé 2014b).

Pour l'ensemble du corpus Saint-Jean, la comparaison deux à deux des 200 textes génère 19 900 distances différentes : 1/2(200*199). La division par deux se déduit de la propriété de symétrie de la distance intertextuelle ($D_{(A,B)} = D_{(B,A)}$). La multiplication par 199 (et non 200) vient de la propriété d'identité de la distance intertextuelle - $D_{(A,A)} = 0$ (le tableau des distances est une matrice carrée à diagonale nulle).

Rangeons ces 19 900 distances par ordre croissant dans des classes d'intervalles égaux dont les effectifs sont représentés dans le diagramme ci-dessous (figure 1). Les valeurs centrales de cette population sont indiquées sur la figure.

Figure 1. Distribution des distances rangées par valeur croissantes dans des intervalles de 0.5 (effectifs absolus)



Le profil en cloche de la figure 1 indique une population homogène ou encore une distribution "normale" des valeurs de la variable. On le vérifie en constatant l'identité des valeurs centrales : mode (ou valeur la plus fréquente), moyenne (ci-dessous notée $\overline{D}$) et médiane (valeur que prend la variable pour l'individu situé au milieu de la distribution). Sous réserve d'une légère asymétrie à droite, les valeurs centrales sont donc représentatives de la population.

Cette population est soumise à des fluctuations s'expliquant par plusieurs facteurs (auteurs, temps, thèmes) évoqués ci-dessus mais qui peuvent être assimilées aux fluctuations aléatoires existant dans toute vaste population normale. L'ampleur de ces variations est mesurée par l'écart-type (moyenne géométrique des écarts des valeurs de la série à sa moyenne arithmétique) qui est noté σ (sigma).

L'intérêt d'un tel profil est de pouvoir prévoir certaines des caractéristiques de la population. Par exemple, pour une variable donnée dans une population distribuée normalement, 95% des observations seront comprises dans un intervalle de $\pm 1{,}96\ \sigma$ autour de la moyenne. Il y a 2,5 % de ces valeurs qui sont "anormalement" faibles (en dessous de $-1{,}96\sigma$, soit 0,2898) et 2,5% "anormalement" élevées (supérieures à 0,4468).

Vérifications :

- avec le seuil à 5%, il est attendu 19900*0,025 = 497 distances anormalement faibles et 497 distances anormalement élevées. En classant les distances par ordre croissant, la 497^{e} distance est égale à 0,2940 ; la 19 403^{e} à 0,4550. Ces deux valeurs approchent de près les seuils théoriques présentés ci-dessus.

- avec le seuil de 1‰ on attend 50 distances inférieures à 0,2366. Le tableau en annexe 5 indique qu'il y en a 53, etc.

Ces concordances confirment que la population présente un profil typiquement "gaussien".

Il s'agit alors de choisir un seuil d'incertitude α (dit "risque d'erreur" dans les études sur échantillon[1]), pour pondérer l'écart type par un coefficient u_α, délimitant ainsi un intervalle de fluctuation normale (entre crochet dans la formule 3) autour de la moyenne ($\overline{D}$) :

(3) $(\overline{D} - u_\alpha \sigma)$ […] $(\overline{D} + u_\alpha \sigma)$

Le tableau 3 ci-dessous donne les bornes inférieures et supérieures de cet intervalle de fluctuation normale en fonction d'un certain nombre de valeurs d'α, avec $\overline{D} = 0{,}3684$ et un écart-type $\sigma = 0{,}0399$.

Tableau 3 Intervalle de fluctuation normale en fonction du seuil d'incertitude alpha

α	0,05	0.01	0.001	0.000 1	0.000 01	0.000 001
$U\alpha$	1,96	2,58	3,29	3,89	4,42	4,89
$\overline{D} - u_\alpha \sigma$	0,2898	0,2659	0,2366	0,2127	0,1915	0,1727
$\overline{D} + u_\alpha \sigma$	0,4467	0,4714	0,4999	0,5239	0,5451	0,5639

Lecture : avec une incertitude $\alpha = 0{,}05$ (ou 5%), on peut considérer que toute valeur inférieure à 0,2898 est significativement faible (ou "anormalement faible") et toute valeur supérieure à 0,4467 est significativement forte, etc.

Les distances sont classées par ordre croissant (annexe 5) en interrompant le recensement au seuil de 1% (borne inférieure : 0,2659). Ce choix est justifié par les considérations mentionnées au début de cette section. Les autres bornes plus basses sont indiquées sur le tableau.

Pour tout texte non attribué, *son ou ses frères seront le ou les textes dont il est séparé par une ou des distances inférieures à la borne basse de l'intervalle de confiance choisi*. Par exemple, en choisissant $\alpha = 0.01$, on attribue les deux textes au même auteur avec une incertitude inférieure à 0,5% à la borne inférieure de l'intervalle (0,2659). Lorsque les distances passent au-dessus de cette borne, l'attribution n'est plus possible.

Attribution d'auteur par les plus petites distances

La classification se déroule selon une procédure semblable à celle présentée ci-dessus mais s'interrompt dès qu'on atteint la borne basse de l'intervalle choisi. Elle aboutit aux résultats présentés dans le tableau 4 ci-dessous. Seuls 156 textes (78% du corpus) sont classés. Les groupes sont très proches de ceux identifiés dans le tableau 1, mais en nombre plus réduits (50 groupes différents au lieu de 66), essentiellement du fait du résidu non attribué.

[1] Nous avons renoncé à employer le terme "risque d'erreur" car d'une part, il n'est applicable qu'aux études sur échantillons – ce qui n'est pas le cas ici - et parce que, d'autre part, chez les personnes non familières avec la statistique, ce terme est compris ainsi : "les calculs peuvent être faux", ce qui conduit à une suspicion générale et à un rejet à chaque fois que les résultats sont contraires à l'intuition ou à la doxa. Dans le cas des "auteurs", cette doxa est très puissante, spécialement chez les littéraires.

Tableau 4. Classification de textes par les plus petites distances dans l'ordre de leur première agrégation à partir de l'annexe 5.

Numéro	Saint-Jean	Auteur et oeuvre
1	013, 030, 068, 078, 116, 150	George Sand. *La petite Fadette / La mare au diable*
2	105, 123, 141, 157, 183, 191	Anne-Louise de Staël-Holstein. *Delphine*
3	011, 028, 044	Alfred de Musset. *La Confession d'un enfant du siècle*
4	115, 133, 149, 163	Henri de Régnier. *Les Rencontres de Monsieur de Bréot*
5	106, 124, 142, 158	Emile Erkmann & Alexandre Chatrian. *Histoire d'un conscrit de 1813*
6	111, 129, 145, 161	Alphonse de Lamartine. *Geneviève*
7	108, 126, 144, 160, 171	Fromentin Eugène. *Dominique*
8	045, 057, 134, 164	George Sand. *Indiana*
9	083, 087, 091	Guy de Maupassant. *Mont Oriol*
10	049, 060	Émile Zola. *L'Assommoir*
11	084, 088	Émile Zola. *La Fortune des Rougon*
12	185, 193, 197, 199	Jules Vallès. L'Enfant
13	058, 069, 099	Henri Beyle Stendhal. *Le Rouge et le Noir*
14	085, 095	Honoré de Balzac. *Le père Goriot*.
15	112, 130, 146, 162, 172	Pierre Loti. *Pêcheur d'Islande*
16	102, 120, 138, 154, 168, 176	Paul Bourget. *Une idylle tragique*
17	051, 062, 073	Alexandre Dumas. *Les trois mousquetaires*
18	038, 052, 063	Gustave Flaubert. *Madame Bovary*
19	155, 169, 177	Alphonse Daudet. *Le Petit Chos*e
20	040, 054, 065	Edmond et Jules de Goncourt. *Germinie Lacerteux*
21	070, 079	Alfred de Vigny. *Servitude et grandeur militaires*
22	003, 020	Alexandre Dumas. *Le comte de Monte Cristo*
23	010, 027	Guy de Maupassant. *Bel-Ami*
24	014, 031, 046	Henri Beyle Stendhal. *La Chartreuse de Parme*
25	118, 136, 152, 174, 182, 190, 198, 200	Eugène Sue. *Les Mystères de Paris*
26	043, 056, 067, 077	Guy de Maupassant. *Notre cœur / Fort comme la mort*
27	104, 122, 140, 156	Alexandre Dumas. *Le Vicomte de Bragelonne*
28	103, 121, 139	Alphonse Daudet. *Le Petit Chose*
29	090, 094	Gustave Flaubert. *Salammbô*
30	170, 178, 186, 196	Anatole France. *La Rôtisserie de la reine Pédauque*
31	71, 80	Emile Zola. *La bête humaine*
32	19, 36	François-René de Chateaubriand. *Atala / René*
33	125, 143, 159	Anatole France. *Le crime de Sylvestre Bonnard*
34	16, 33, 48	Alfred de Vigny. *Cinq-Mars*
35	117, 135, 151, 165, 181, 189	Charles-Augustin Sainte-Beuve. *Volupté*
36	092, 098	Emile Zola. *Germinal*
37	017, 034	Emile Zola. *L'Argent*
38	089, 093	Honoré de Balzac. *Eugénie Grandet*
39	001, 035	Honoré de Balzac. *La cousine Bette*
40	119, 153	Jules Barbey d'Aurevilly. *Le chevallier des Touches*
41	012, 029	Gérard de Nerval. *Aurélia*
42	076, 082	Gustave Flaubert. *L'Education Sentimentale*
43	180, 188	Henri de Régnier. *La Double Maîtresse*
44	167, 175, 192	Jules Barbey d'Aurevilly. *Les Diaboliques*
45	006, 023	Edmond et Jules de Goncourt. *Madame Gervaisais*
46	179, 187	Pierre Loti. *Madame Chrysanthème*
47	039, 053	Théophile Gautier. *Jettatura*
48	170, 186	Anatole France. La Rôtisserie de la reine Pédauque
49	117, 181	Charles-Augustin Sainte-Beuve. Volupté
50	050, 061	Honoré de Balzac. *Grandeur et décadence de César Birotteau*

L'attribution d'auteur proprement dite s'interrompt juste avant les trois extraits tirés des *Plaisirs et des jours* (Marcel Proust : n° 114, 132 et 148) dont les distances sont respectivement de 0,2666, 0,2952 et 0,2676.

Si l'on retient le seuil de 0.001 pour une attribution sans erreur : 111 textes seulement sont attribués. Parmi les 89 textes non attribués figurent notamment tous les extraits uniques et tous les textes de plusieurs auteurs, au premier rang desquels : V. Hugo.

De plus, du point de vue statistique, cette procédure par les plus petites distances et la détermination de seuils de décision pourrait se voir opposer une objection.

La procédure consiste, pour chaque couple de textes, à confronter deux hypothèses : H_0 (un seul écrivain pour les deux textes) et H_1 (deux écrivains différents). L'acceptation de H_1 (deux écrivains différents) n'a pas le même statut que celle de H_0 (un seul écrivain). Dans le premier cas, cette décision se voit associer une certaine incertitude alors que le second (non rejet de H_0 : un seul écrivain) est une décision par défaut qui ne signifie pas pour autant que H_0 est vraie (au seuil choisi). En effet, un élargissement du corpus ou une modification substantielle de celui-ci risquent de changer la valeur de l'écart-type. Cela peut suffire pour faire passer une valeur donnée de la zone d'acceptation à celle du rejet de H_0. D'où, dans les études comme celles-ci, l'acceptation prudente de l'hypothèse nulle *en l'état des observations disponibles*.

C'est pour cette raison que l'un des critères qui ont guidé la constitution du corpus était que, non seulement il devait y avoir un grand nombre d'auteurs mais que, de plus, aucun de ces auteurs ne devait peser d'un poids tel que son retrait puisse influer sur les résultats.

Il n'en reste pas moins que, en l'absence de l'indication selon laquelle tout auteur a au moins deux textes, il faut renoncer à classer tous les textes et associer aux résultats restant une marge d'incertitude, du moins lorsque la distance est supérieure à certains seuils.

Ces résultats sont-ils décevants ?

Il faut se souvenir qu'à ce stade, **il ne s'agit pas d'attribuer le maximum de textes mais de le faire sans erreur** quitte à ne pas répondre dans plus de 2 cas sur 10.

En résumé, l'attribution de la totalité des textes est possible – grâce à la méthode des plus proches voisins – lorsqu'on est certain que tous les auteurs ont au moins deux textes. Sans cette condition, une attribution solide est possible grâce aux plus petites distances, mais seulement dans moins de 8 cas sur 10 et à condition qu'il y ait un grand nombre d'auteurs différents (au moins une trentaine).

Trois remarques.

- En utilisant la borne supérieure de l'intervalle, on peut également répondre à la question : qui n'a pas écrit ce texte. La procédure n'est pas présentée ici car elle est proche de celle déjà évoquée à propos des plus lointains, avec des résultats très comparables.

- Les opérations manuelles sont toujours fastidieuses et source d'erreur. Il est préférable de confier ces opérations à des automates et de les faire précéder d'une phase exploratoire qui donne des "ordres de grandeur" permettant de détecter d'éventuelles erreurs de classification ou cas "aberrants" à éliminer de la recherche.

- On peut souhaiter aller plus loin. Par exemple, en admettant que le corpus offre un panorama des romans du XIXe siècle, quels sont les auteurs (ou les œuvres) les plus proches, les plus centraux ou les plus originaux, les plus "décalés". Ou encore, peut-on apercevoir quelques grands groupes ou "courants" littéraires ?

Les classifications automatiques permettent de répondre à ces deux souhaits.

III. Classifications automatiques

La classification automatique est un outil indispensable pour traiter les grandes collections de textes. Elle vise à opérer des regroupements dans ces vastes populations. La meilleure classification possible est celle qui minimise les distances entre les individus classés dans le même groupe et qui maximise ces distances entre les différents groupes. Pour une présentation de la taxinomie numérique, voir Sneath & Sokal 1973 et Roux 1985.

Toute classification repose sur le calcul d'une distance séparant les objets à classer. Dans les limites indiquées en annexe 2, la distance intertextuelle présente les caractéristiques d'une distance dans un espace euclidien, ce qui permet de réaliser des classifications représentant, sans déformation, la population étudiée.

Deux méthodes sont utilisées : la classification hiérarchique ascendante et la classification arborée.

Classification hiérarchique ascendante

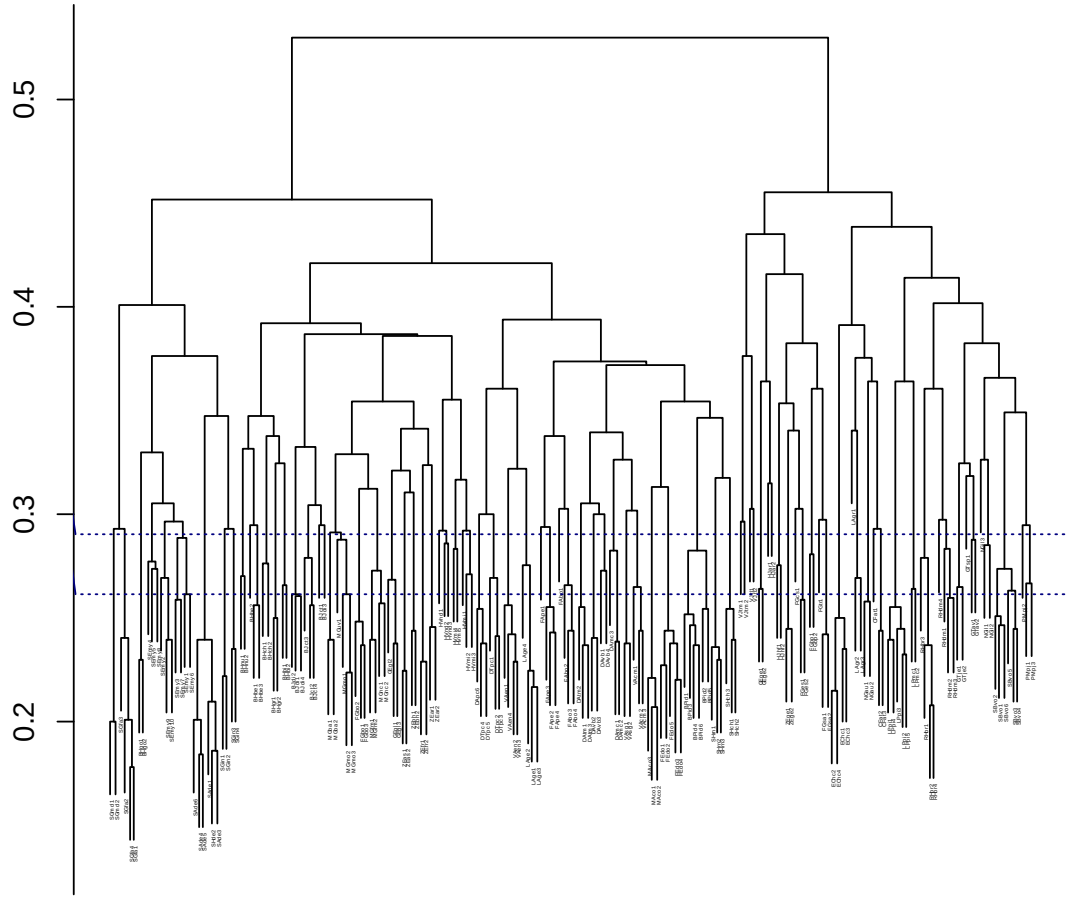
L'algorithme commence par regrouper les deux textes les plus proches (ici les deux extraits de la *Petite Fadette* et il calcule les distances entre ce groupe et tous les autres textes en effectuant la moyenne arithmétique simple des distances originelles et ainsi de suite jusqu'à la formation d'un ensemble unique. Ces groupements successifs sont représentés dans un "dendrogramme" avec en ordonnées les distances relatives correspondants aux différents niveaux d'agrégation (Figure 2).

Avec des populations nombreuses, ces figures comportent beaucoup de feuilles terminales et les étiquettes se superposent, sauf à les réduire drastiquement. Ici elles comportent les initiales de l'auteur (en majuscules) et deux lettres du titre en minuscules (il est possible de lire ces étiquettes grâce au zoom : FApe = France Anatole. *La Rôtisserie de la reine Pédauque*.). La légende sous le graphique énumère ces étiquettes de gauche à droite, avec les groupes formés par l'algorithme.

L'ordre dans lequel sont présentés les groupes sur l'axe horizontal n'a pas d'importance. Seule compte la hauteur du trait horizontal qui marque la distance entre les différents textes groupés sous ce trait. Il suffit donc de couper le graphique aux seuils choisis pour identifier les textes d'un même auteur au seuil de confiance choisi. Sur la figure ci-dessous, sont portés les deux traits (0,267 et 0,291) correspondant aux seuils discutés plus haut concernant l'attribution d'auteur.

Par exemple, pour Anatole France, les quatre extraits de la *Rôtisserie de la reine Pédauque* sont séparés des quatre extraits du *Crime de Sylvestre Bonnard* par une distance moyenne de 0,34, ce qui ne permet pas trancher entre les deux hypothèses : (1) un même auteur écrivant sur des thèmes différents et/ou à une époque différente ; (2) deux auteurs contemporains travaillant sur des thèmes proches.

Figure 2. Classification hiérarchique ascendante Corpus Saint-Jean (méthode de la moyenne)



Réalisé avec R (août 2017)

De gauche à droite :

Sand George - *La mare au diable* (2 extraits)
Sand Georges - *La petite Fadette* (4 extraits)
Balzac (Honoré de) - *Le père Goriot* (2 extraits)
Sue Eugène - *Les Mystères de Paris* (10 extraits)
Staël-Holstein (Anne-Louise de) – *Delphine* (6 extraits)
Sand George - *Indiana* (4 extraits)
Balzac Honoré de – *La Maison Nucingen* (2 extraits)
Balzac Honoré de - *La cousine Bette* (3 extraits)
Balzac Honoré de - *Le Colonel Chabert* (2 extraits)
Balzac Honoré de - *Eugénie Grandet* (2 extraits)
Balzac Honoré de - *César Birotteau* (2 extraits)
Barbey d'Aurevilly Jules - *Les Diaboliques* (3 extrait)
Barbey d'Aurevilly Jules - *Le chevalier des Touches* (4 extraits)
Maupassant Guy de - *Bel-Ami* (2 extraits)
Maupassant Guy de - *Une vie* (1 extrait)
Maupassant Guy de - *Mont Oriol* (3 extraits)
Flaubert Gustave - *Madame Bovary* (3 extraits)
Maupassant Guy de - *Fort comme la mort* (2 extraits)
Maupassant (Guy de) - *Notre coeur* (2 extraits)
Goncourt Edmond et Jules de - *Germinie Lacerteux* (3 extraits)
Zola Émile - *L'Assommoir* (2 extraits)
Zola Emile - *La bête humaine* (2 extraits)
Zola Émile - *La Fortune des Rougon* (2 extraits)
Zola Emile - *L'Argent* (2 extraits)
Hugo Victor - *Notre Dame de Paris* (3 extraits)
Hugo Victor - *Les Misérables* (5 extraits)
Daudet Alphonse – *Le Petit Chose* (6 extraits)
Vallès Jules – *L'enfant* (4 extraits)
Lamartine Alphonse de – *Geneviève* (4 extraits)

France Anatole - *La Rôtisserie de la reine Pédauque* (4 extraits)
France Anatole - *Le crime de Sylvestre Bonnard* (4 extraits)
Dumas Alexandre – *Les trois mousquetaires* (3 extraits)
Dumas Alexandre - *Le Vicomte de Bragelonne* (4 extraits)
Dumas Alexandre - *Le comte de Monte Cristo* (3 extraits)
Vigny Alfred de - *Servitude et grandeur militaires* (2 extraits)
Vigny Alfred de - *Cinq-Mars* (3 extraits)
Musset Alfred de - *La Confession d'un enfant du siè*de (3 extraits)
Fromentin Eugène – *Dominique* (5 extraits)
Bourget Paul - *Une idylle tragique* (6 extraits)
Stendhal(Henri Beyle) - *Le Rouge et le Noir* (3 extraits)
Stendhal (Henri Beyle) *La Chartreuse de Parme* (3 extraits)
Verne Jules - *Le tour du monde en quatre-vingt jours* (2 extraits)
Verne Jules - *De la terre à la lune* (2 extraits)
Goncourt Edmond et Jules de - *Madame Gervaisais* (2 extraits)
Huysmans Joris-Karl - *A rebours* (2 extraits)
Huysmans Joris-Karl – *Marthe histoire d'une fille* (2 extraits)
Zola Emile - *Germinal* (2 extraits)
Flaubert Gustave - *Un coeur simple* (1 extrait)
Flaubert Gustave - *L'Education Sentimentale* (2 extraits)
Flaubert Gustave - *Bouvard et Pécuchet* (2 extraits)
Flaubert Gustave - *Hérodias* (1 extrait)
Flaubert Gustave - *Salammbô* (2 extraits)
Erckmann Emile & Chatrian Alexandre - *Histoire d'un conscrit* (4 extraits)
Lamartine (Alphonse de) - *Graziella* (3 extraits)
Nerval Gérard de - *Aurélia* (2 extraits)
Chateaubriand François-René de - *Atala* (3 extraits)
Loti Pierre - *Pêcheur d'Islande* (4 extraits)
Loti Pierre - *Madame Chrysanthème* (2 extraits)
Régnier Henri de - *Les Rencontres de Monsieur de Bréot* (4 extraits)
Régnier Henri de - *La Double Maîtresse* (4 extraits)
Gautier Théophile – *Jettatura* (2 extraits)
Gautier Théophile – *Spirite* (1 extrait)
Gautier Théophile - *Avatar* (2 extraits)
Nerval Gérard de - *Les Illuminés* (3 extraits)
Sainte-Beuve Charles-Augustin – *Volupté* (6 extraits)

Proust Marcel – *Des plaisirs et des jours* (3 extraits)

Tous les extraits tirés d'une même œuvre sont classés ensemble. L*es Diaboliques* de J. Barbey d'Aurevilly sont un recueil de nouvelles dont nous avons extraits celles de 10 000 mots et plus. Il s'agit donc en fait de livres différents. En plus de J. Barbey d'Aurévilly, quelques auteurs sont dans deux groupes pour des ouvrages différents. Par exemple, *Madame Bovary* (G. Flaubert) est groupée avec G. de Maupassant, alors que le reste des œuvres de G. Flaubert sont bien groupées (à droite du graphe) mais apparemment plus proches d'E. Zola, de K.-J. Huysmans ou des frères Goncourt que de *Mme Bovary*. Est-ce que cela correspond réellement à une évolution significative dans l'œuvre de G. Flaubert ? De même, le *Père Goriot* de H. de Balzac (classé entre E. Sue et A. de Musset) est-il singulier par rapport aux autres œuvres de H. de Balzac (classées entre G. Sand et G. de Maupassant) ou encore *Germinal* est-il à part des autres romans d'E. Zola ?

Avec la classification hiérarchique ascendante, il n'est pas possible de répondre à ces questions car la méthode produit des "effets de chaîne" qui effacent les liens individuels entre les textes groupés et ceux qui restent en dehors de ce groupe.

Ces inconvénients sont moins sensibles dans la classification arborée.

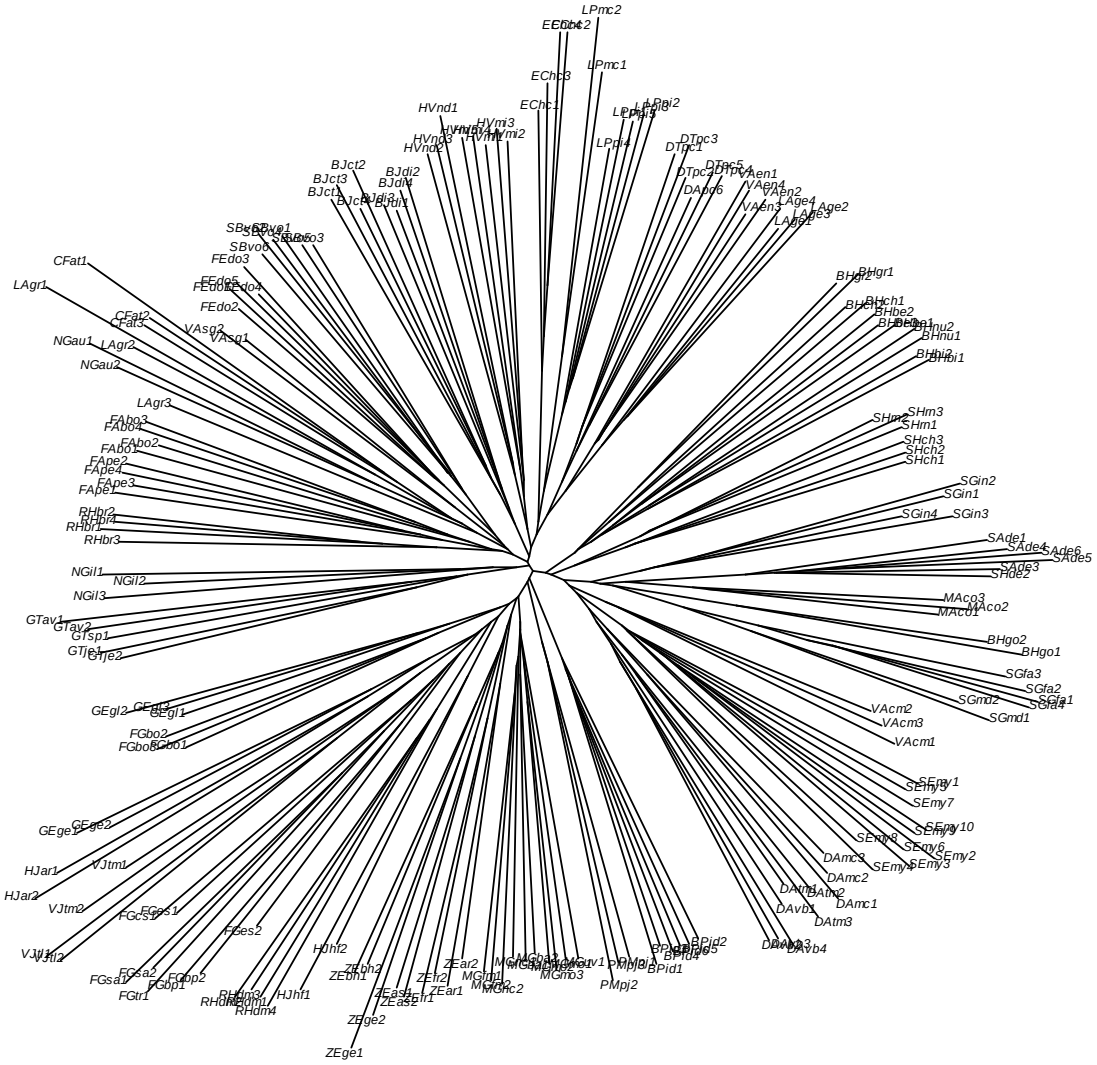
Classification arborée

La méthode dite de la "classification arborée" est classique en génétique (Felsenstein 2004) ou en linguistique historique (Embleton 1986, Holm 2007). Elle repose sur la propriété suivante : si toutes les distances séparant les individus étudiés présentent bien les propriétés requises d'une

distance dans un espace euclidien, il existe un "arbre" qui représente le mieux possible les positions respectives de ces individus les uns par rapport aux autres et les meilleurs groupements possibles entre eux. L'algorithme est présenté dans X. Luong (1988 et 1994). Voir également, Barthélémy & Guénoche 1988 ; Rulhman 2003 ; Labbé & Labbé 2006.

La figure 3 présente l'arbre du corpus Saint-Jean.

Figure 3. Classification arborée du corpus Saint-Jean



Réalisé avec R (ape - août 2017)

Les principaux clusters (Dans le sens des aiguilles d'une montre en partant de midi (haut) :

A
EChc 1, 3, 4, 2 : Emile Erckmann, Alexandre Chatrian. *Histoire d'un conscrit de 1813*
LPmc 2, 1 : Pierre Loti. *Madame Chrysanthème*
LPpi 4, 1, 5, 3, 2 : Pierre Loti. *Pêcheur d'Islande*
DTpc 1, 3, 2, 6, 5, 4 : Alphonse Daudet. *Le petit Chose*
VAen 1, 4, 2, 2 Jules Vallès. *L'enfant*
LAge 4, 1, 3, 2 : Alphonse de Lamartine. *Geneviève*

B
BHgr 2, 1 : Honoré de Balzac. *Eugénie Grandet*
BHch 2, 1 : Honoré de Balzac. *Le Colonel Chabert*

BHbe 2, 3, 1 : Honoré de Balzac. *La cousine Bette*
BHnu 2, 1 : Honoré de Balzac. *La maison Nucingen*
BHbi 2, 1 : Honoré de Balzac. *Histoire de la grandeur et de la décadence de César Birotteau*

C
SHrn 2, 3, 1 : Henri Beyle Stendhal. *Le Rouge et le Noir*
SHch 3, 2, 1 : Henri Beyle Stendhal. *La chartreuse de Parme*

D
SGin 2, 1, 4, 3 : George Sand. *Indiana*
SAde 1, 4, 6, 5, 3, 2 Anne-Louise de Staël-Holstein. *Delphine*
MAco 3, 2, 1 : Alfred de Musset. *La Confession d'un enfant du siècle*
BHgo 2, 1 : Honoré de Balzac. *Le père Goriot*
SGfa 3, 2, 1, 4 : George Sand. *La petite fadette*
SGmd 2, 1 : George Sand. *La mare au diable*
VAcm 2, 3, 1 : Alfred de Vigny. *Cinq-Mars*
SEmy 1, 5, 7, 9, 2, 10, 6, 3, 8, 4 : Eugène Sue. *Les mystères de Paris*.
DAmc 3, 2, 1 : Alexandre Dumas. *Monte Cristo*
DAtm 2, 3, 1 : Alexandre Dumas. *Trois mousquetaires*
DAvb 1, 4, 3, 2 : Alexandre Dumas. *Le vicomte de Bragelonne*

E
BPid 2, 5, 6, 4, 3, 1 : Paul Bourget. *Une idylle tragique*

F
PM 1, 3, 2 : Marcel Proust. *Des plaisirs et des jours*

G
MGuv : Guy de Maupassant. *Une vie* (extrait unique)
MGmo 1, 3, 2 : Guy de Maupassant. *Mont Oriol*
MGba 2, 1 : Guy de Maupassant. *Bel ami*
MGnc 1, 2 : Guy de Maupassant. *Notre coeur*
MGfm 2, 1 : Guy de Maupassant. *Fort comme la mort*
ZEar 2, 1 : Emile Zola. *L'argent*
ZEfr 1, 2 : Emile Zola. *La fortune des Rougon*
ZEas 2, 1 : Emile Zola. *L'assommoir*
ZEge 2, 1 : Emile Zola. *Germinal*
ZEbh 2, 1 : Emile Zola. *La bête humaine*
HJhf 2, 1 : Huysmans Karl-Joris. *Histoire d'une fille*
RHdm 4, 1, 3, 2 : Henri de Régnier. *La Double Maîtresse*
FGes 2 : Gustave Flaubert. *Education sentimentale*
FGbp 2, 1 : Gustave Flaubert. *Gustave et Pécuchet*
FGtr 1 : Gustave Flaubert. *Hérodias*
FGsa 2, 1 : Gustave Flaubert. *Salammbô*
FGes 2 : Gustave Flaubert. *Education sentimentale*
FGcs : Gustave Flaubert. *Un cœur simple*
VJtl 2, 1 : Jules Verne. *De la terre à la lune*
VJtm 2, 1 : Jules Verne. *Le tour du monde en 80 jours*
HJar 2, 1 : Huysmans Karl-Joris. *A rebours*
GEge 2, 1 : Edmond et Jules Goncourt. *Madame Gervaisais*
FGbo 1, 3, 2 : Gustave Flaubert. *Madame Bovary*
GEgl 1, 3, 2 Edmond et Jules de Goncourt. *Germinie Lecerteux*

H
GTje 2, 1 : Théophile Gautier. *Jettatura*
GTsp 1 : Théophile Gautier. *Spirite*
GTav 2, 1 : Théophile Gautier. *Avatar*
NGil 3, 2, 1 : Gérard de Nerval. *Les Illuminés*

I
RHbr 3, 4, 1, 2 : Henri de Régnier. *Les Rencontres de Monsieur de Bréot*
FApe 1, 3, 4, 2 : Anatole France. *La rôtisserie de la reine Pédauque*
FAbo 4, 2, 3 ,1 : Anatole France. *Le crime de Sylvestre Bonnard*
LAgr 3 : Alphonse de Lamartine. *Graziella*
NGau 2, 1 : Gérard de Nerval. *Aurélia*
LAgr 2, 1 : Alphonse de Lamartine. *Graziella*
CFat 3, 2, 1 : François-René de Chateaubriand. *Atala et René*
VAsg 2, 1 : Alfred de Vigny. *Servitude et grandeur militaires*
FEdo 2, 1, 4, 3 : Eugène Fromentin. *Dominique*
SBvo 6, 4, 1, 2, 5, 3 : Charles-Augustin Sainte-Beuve. *Volupté*

J
BJct 1, 3, 4, 2 : Jules Barbey d'Aurevilly. *Le chevallier des Touches*
BJdi 3, 1, 4, 2 : Jules Barbey d'Aurevilly. *Les Diaboliques*

K
HVnd 2, 3, 1 : Victor Hugo. *Notre-Dame de Paris*
HVmi 4, 1, 3, 2 : Victor Hugo. *Les Misérables*

Dans cette figure, les feuilles terminales figurent chacune un texte ; les nœuds intermédiaires donnent les meilleurs groupements possibles, c'est-à-dire ceux pour lesquels les distances entre les éléments qui composent le groupe sont les plus faibles et les distances les séparant des autres les plus grandes possibles. Les segments de droite, ou arêtes, sont des branches quand elles relient des feuilles à des nœuds et des troncs quand elles relient des nœuds entre eux. La distance entre deux points quelconques est figurée par le chemin unissant ces points et la longueur de ce chemin est proportionnelle à la distance originelle correspondante. Il ne faut

pas attacher d'importance au placement des branches rattachées à un même nœud, spécialement pour les branches terminales. Par exemple, tout en haut de l'arbre, les 4 extraits EChc (*l'Histoire d'un conscrit de 1813* d'E. Erckmann et A. Chatrian) sont dans un ordre indifférent (1, 3, 4, 2). En revanche, la longueur des branches terminales indiquent que l'extrait n°1 est le plus central et le n° 4 le plus décalé.

Au plus près de la racine de l'arbre, la classification a isolé 11 clusters principaux (notés de A à K dans le sens des aiguilles d'une montre). Plus le nœud est proche de la racine de l'arbre, plus les distances internes sont grandes entre les différents individus formant ce groupe, et plus le regroupement est hétérogène. C'est tout particulièrement le cas du groupe I dont les différentes branches se rejoignent quasiment au nœud de l'arbre. Il ne faut donc pas attacher trop d'importance à ce regroupement. En revanche, d'autres semblent plus caractéristiques, tout particulièrement celui qui rassemble autour de G. Flaubert : G. de Maupassant, E. Zola, les frères Goncourt et H. de Régnier.

Au sein de ces 11 groupes, tous les extraits des mêmes ouvrages sont classés ensemble. La plupart des œuvres d'un même auteur sont également regroupées. Quelques exceptions apparaissent, dont certaines ont déjà été notées lors de la classification hiérarchique, notamment :

- *le père Goriot* (dans le groupe D) semble singulier par rapport au reste de l'œuvre d'H. de Balzac (dans le groupe B). Du moins les extraits sélectionnés de ce roman (la conclusion revient sur cet aspect).

- A. de Lamartine : *Graziella* (groupe I) et *Geneviève* (groupe A) sont presque à l'opposé l'une de l'autre (alors qu'il s'agit d'œuvres contemporaines) :

Certains auteurs semblent connaître des ruptures non seulement thématiques mais aussi stylistiques telles qu'elles peuvent mettre en échec l'attribution d'auteur ou au moins en limiter la portée.

En revanche, *Germinal* est bien classé avec les autres œuvres d'E. Zola ; tous les contes tirés de *Diaboliques* (J. Barbey d'Aurevilly) sont également groupés. On peut donc penser que la disjonction de ces textes par la classification hiérarchique était un artefact de la méthode.

Naturellement, il n'est pas possible de conclure quant aux courants littéraires et aux proximités entre auteurs, car on a utilisé des échantillons et non les œuvres complètes. De plus, il n'y a pas toutes les œuvres des principaux auteurs comme H. de Balzac, V. Hugo, E. Zola, et.

Ce graphique traduit-il fidèlement l'organisation du corpus ? Peut-on se fier à un graphique de ce genre ?

Qualité de la classification

Avec quelle fidélité l'arbre représente-t-il les distances originales ?

Pour chaque couple de textes, la réponse est donnée par le rapport entre leur distance intertextuelle et sa représentation sur l'arbre, c'est-à-dire la longueur du chemin reliant les deux feuilles correspondantes (Labbé & Labbé 2008).

Soit $L_{(A,B)}$ la longueur du chemin reliant les feuilles correspondant aux textes A et B et $Qual_{(A,B)}$ l'indice de qualité (ou de confiance) de ce chemin :

$$Qual_{(A,B)} = 1 - \left| \frac{L_{(A,B)} - D_{(A,B)}}{D_{(A,B)}} \right|$$

De même, la qualité d'un nœud est la moyenne des indices de tous les chemins qui passent par ce nœud. Enfin la qualité de l'arbre est la moyenne des indices de tous les chemins.

L'indice varie uniformément entre :

- 1 si $L_{(A,B)} = D_{(A,B)}$

- 0 si $L_{(A,B)} = 0$ (ou $L_{(A,B)} = 2\ D_{(A,B)}$)

Une valeur de 0,5 indique que la longueur du chemin s'écarte (par excès ou défaut) de 50% de la distance qu'elle est censée représenter.

En ce qui concerne l'étalonnage de cet indice, on dispose d'une indication - donnée par X. Luong dans le code source reproduit dans sa thèse de 1988 -, selon laquelle, un écart au plus égal à 10% entre la longueur des chemins et les distances correspondantes ne remet pas en cause la classification. Nos propres expériences confirment qu'un indice supérieur à 90% (chemin, nœud ou arbre) indique une classification fiable et qu'à l'inverse tout indice inférieur doit être soigneusement examiné avant d'accepter la classification proposée.

Pour l'arbre présenté ci-dessus, les principaux indices de qualité sont les suivants.

- 95,5% pour l'arbre entier. L'information initiale contenue dans la matrice des distances est donc restituée avec une incertitude moyenne inférieure à 5%, ce qui indique une classification tout à fait fiable.

- Les nœuds dont la qualité est la plus faible ont des indices de 92,9% et 93,2% respectivement pour les 269^{e} et 388^{e} dans l'ordre d'agrégation. Le premier concerne l'agrégation de *Graziella* d'A. de Lamartine avec *Atala* d'A. de Châteaubriand. Le second concerne, dans le groupe A situé en haut de l'arbre, la classification du *Père Goriot* avec la *Petite Fadette de G. Sand*. Le groupement de ces œuvres ensemble est donc à considérer avec prudence.

- 93% des 19 900 chemins ont un indice supérieur à 90%. 34 (1,7‰) ont un indice inférieur à 80%. Les plus médiocres sont ceux qui relient le *Père Goriot* avec la *Petite Fadette* de G. Sand. L'arbre exagère la distance moyenne qui sépare ces deux œuvres de près de 25%. Parmi les œuvres posant également problème, on trouve *Les Rencontres de M. Bréot* d'H. de Régnier dans le même cluster en haut du graphe.

Avec ces réserves, aucune autre méthode de classification – sur une grande population de textes - ne parvient à un tel degré de fiabilité.

Pour ce qui concerne l'attribution d'auteur proprement dite, tous les nœuds rassemblant des extraits d'une même œuvre sont d'une excellente qualité, mais, comme indiqué plus haut, il est possible de conclure parce que l'on sait que tout auteur a au moins deux textes.

Au total, les diverses procédures présentées montrent la capacité de la distance intertextuelle à identifier les textes écrits par un même auteur à une époque pas trop éloignée. Cela s'explique par la prépondérance du facteur "auteur" sur le temps et les thèmes. La dernière section de ce rapport en donne la mesure, ce qui permet de présenter une échelle étalonnée de la distance intertextuelle.

V. Echelle de la distance intertextuelle

A l'aide des renseignements de l'annexe 1, les 19 900 distances différentes peuvent être classées en trois familles :

- les distances "intra-livre", séparant les extraits d'un même ouvrage ;
- les distances "intra-auteur et inter-livres", séparant les extraits d'un même auteur mais d'ouvrages différents ;
- les distances "inter-auteurs" entre extraits d'auteurs différents.

Naturellement, ce classement repose sur le postulat selon lequel les 31 auteurs désignés comme tels par la tradition ont bien composé les textes parus sous leur nom…

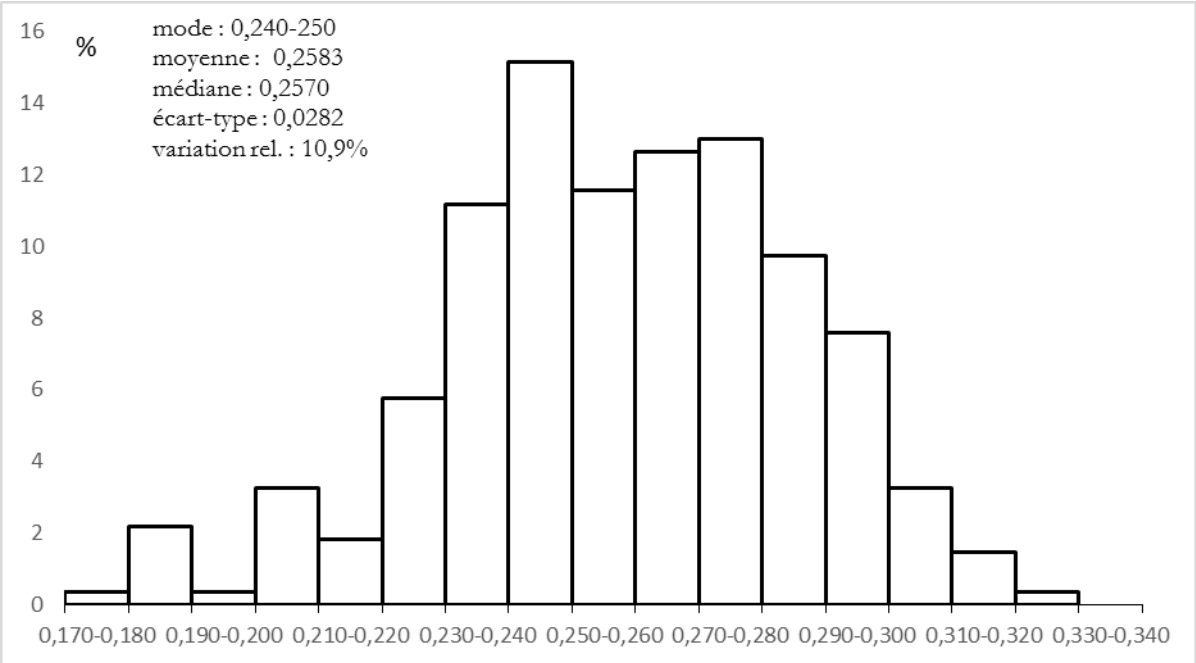
Les distances intra-livres

Ici non seulement le facteur auteur est neutralisé mais thème et temps sont minimisés (avec quelques réserves indiquées ci-dessous).

Il y a 269 distances entre extraits tirés d'un même livre, soit moins de 1,4% de toutes les distances. Cette faible proportion est la conséquence normale des contraintes imposées au départ : grand nombre d'auteurs et faible poids de chacun.

Ces distances sont rangées par ordre croissant dans des classes d'intervalle égal. Leur distribution est décrite dans l'histogramme ci-dessous (Figure 4).

Figure 4. Histogramme des distances entre extraits tirés d'un même livre, classées par ordre croissant dans des classes d'intervalles égaux (% du total, avec les valeurs centrales et indices de dispersion)



Contrairement à la Figure 1 ci-dessus l'échelle verticale (ordonnées) ne représente plus les effectifs absolus des classes mais leurs effectifs relatifs (en % du total). Par exemple, 15% des 269 distances "intra" sont comprises dans la classe 0,240-0,250 (classe la plus peuplée ou "mode"). Cette échelle relative permettra de comparer la distribution des trois sous-populations (intra, intra-inter et inter) malgré la grande différence d'effectifs.

La distribution n'a pas tout à fait la forme en cloche qui est attendue dans une population homogène. La présence d'un mode secondaire (0,260-0,280) laisse supposer une hétérogénéité de la population (ou plutôt le mélange d'au moins deux populations proches mais différentes). Le premier (mode 240-0,250) correspond aux extraits composés à la même période et portant sur des thèmes voisins, alors que la seconde ne présente pas ces deux caractéristiques pour les raisons suivantes :

- la composition de certains ouvrages s'étend sur une période assez longue : pour eux, le facteur temps n'est pas neutralisé ;

- certains ouvrages, surtout les plus longs comme *Monte Cristo*, *les Misérables* ou *les Mystères de Paris* – qui comptent plus d'un demi-million de mots – sont nécessairement multi-thématiques,

- certains auteurs – comme V. Hugo, K.-J. Huysmans, E. Sue ou J. Verne – abordent plusieurs thèmes dans un même roman et n'hésitent pas y insérer des digressions parfois très longues,

- le tirage au sort a pu amener des extraits assez éloignés ;

- pour J. Barbey d'Aurévilly (*Diaboliques*), G. de Nerval (*les illuminés*), de H. de Régnier (*Monsieur Bréot*), il s'agit de recueils de textes sur des thèmes souvent éloignés qui auraient dû être traités comme des livres différents ainsi que cela a été fait pour G. Flaubert (*Un cœur simple* et *Hérodias*) ou T. Gautier…

- la plus grande distance (0.3402) sépare deux extraits de *Graziella* (Lamartine) dont nous dirons plus bas la singularité.

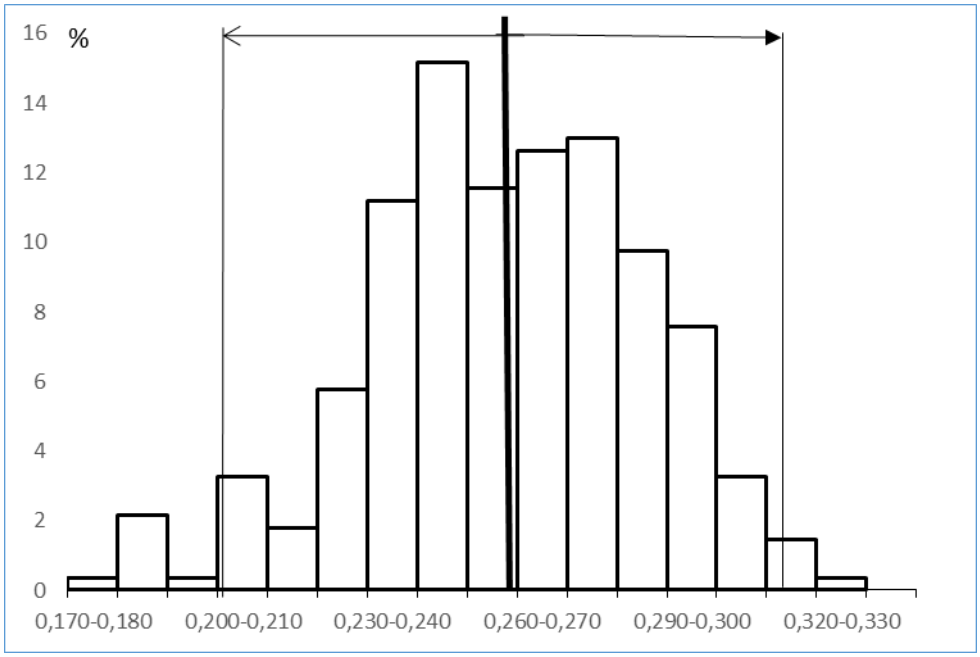
Ces réserves admises, la proximité des valeurs centrales et la dispersion modérée des observations autour de la moyenne autorisent à raisonner comme s'il s'agissait d'une population homogène. On peut alors associer à la moyenne un "intervalle de confiance" avec $\alpha = 5\%$ autour de la moyenne :

- borne inférieure : 0,2583 - (0,0282 * 1,96) = 0,2030

- borne supérieure : 0,2583 + (0,0282 * 1,96) = 0,3137

Ces bornes sont reportées sur la figure 5.

Figure 5. Intervalle de confiance (α = 5%) de la distribution des distances intra-livres.



Toute distance inférieure à la borne supérieure de cet intervalle peut-elle laisser supposer que les deux textes concernés sont d'une même plume ? Pour le savoir, il faut examiner les deux autres familles de distances.

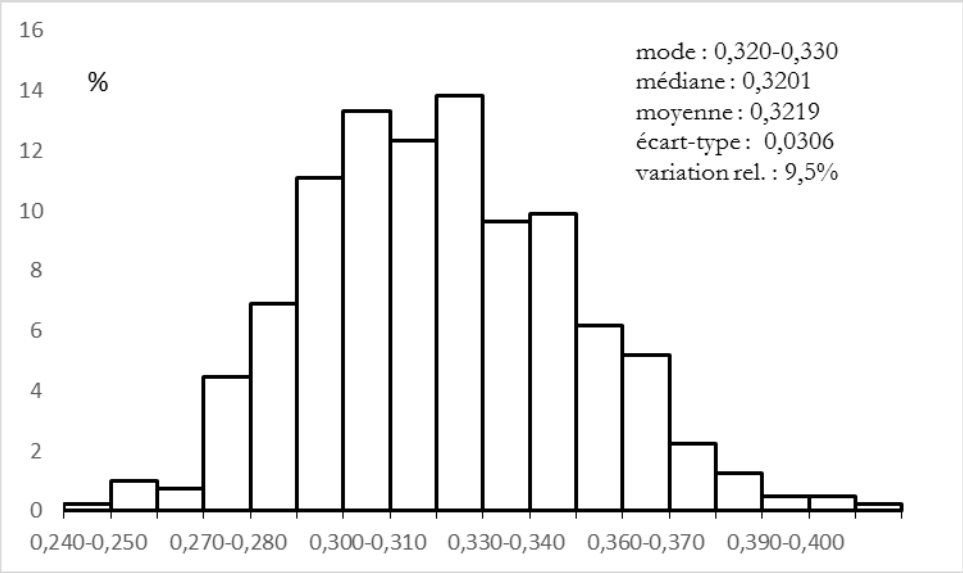
Les distances inter-livres et intra-auteurs

Dans ce cas, le facteur auteur est toujours neutralisé mais les facteurs temps et thèmes ne le sont plus et cumulent souvent leurs effets. Par exemple, 31 ans séparent la parution de *Notre-Dame de Paris* de celle des *Misérables* (V. Hugo), et si leur scène est toujours Paris, c'est à 4 siècles de distance…

Il y a 393 distances inter-livres et intra-auteurs, soit 2% du total. Là encore, cette faible proportion s'explique par le grand nombre d'auteurs et le faible poids de chacun.

La distribution de ces distances est décrite dans la figure 6 (dont l'ordonnée est l'effectif relatif des classes).

Figure 6. Histogramme des distances entre extraits tirés de livres différents d'un même auteur, classées par ordre croissant dans des classes d'intervalles égaux (% du total avec les valeurs centrales et indices de dispersion)



Ce profil est plus proche de la cloche attendue en cas de distribution aléatoire mais présente encore une distribution multimodale et un étalement des observations. Quelques livres sont assez proches dans le temps et/ou les thèmes et sont logiquement séparés par les plus petites distances. Par exemple : *Atala* et *René* de F.-R. de Chateaubriand (0.2434) ; *La petite Fadette* et *la Mare au diable* de G. Sand (0.2545 et 0.2558). La plupart des autres sont plus franchement éloignés car, pour eux, temps et thèmes cumulent leurs effets.

Là encore, un intervalle de confiance ($\alpha = 5\%$) est associé à la moyenne : [0,2620 – 0,3818].

La borne supérieure permettrait – avec une incertitude de 5% - d'exclure que deux textes séparés par une distance supérieure à cette borne soient d'un même auteur. Effectivement, l'annexe 4 indique bien que tous les couples de textes les plus lointains sont d'écrivains différents (avec des distances supérieures à 0.382).

Cependant, huit distances "intra-inter" se trouvent au-dessus de ce seuil :

- Cinq concernent *Graziella* et *Geneviève* (A. de Lamartine). Ce sont les plus fortes (0,463 0,447 ; 0,412 ; 0,401 et 0,394). Elles sont d'autant plus remarquables que les deux œuvres sont presque contemporaines (la première est parue en feuilleton dans *la Presse* en 1849 et la seconde en 1851 pourvue d'une longue préface de l'auteur). Versatilité exceptionnelle de cet écrivain ou deux plumes différentes, c'est-à-dire une collaboration inconnue ? La conclusion revient sur ce point.

- Une concerne J.-K. Huysmans (*A rebours – Marthe* : 0,401), mais l'explicationt tient ici à la singularité de l'extrait déjà signalée ci-dessus et à la coupure considérable que représente *A rebours* par rapport aux œuvres précédentes de cet auteur ;

- Deux concernent un extrait du *Père Goriot* (Balzac) avec la *Cousine Bette* et la *Maison Nuncingem* mais pour des valeurs proches du seuil (0,399 et 0,388). Ce dernier résultat suggère une singularité du *Père Goriot* dans l'œuvre de Balzac.

La conclusion reviendra sur ces cas.

Enfin, ces données permettent d'estimer le poids moyen combiné des variables temps et thèmes en comparant la moyenne de cette série avec la précédente :

$\delta = (0{,}3219 - 0{,}2583)/ 0{,}2583 = + 25{,}8\ \%$

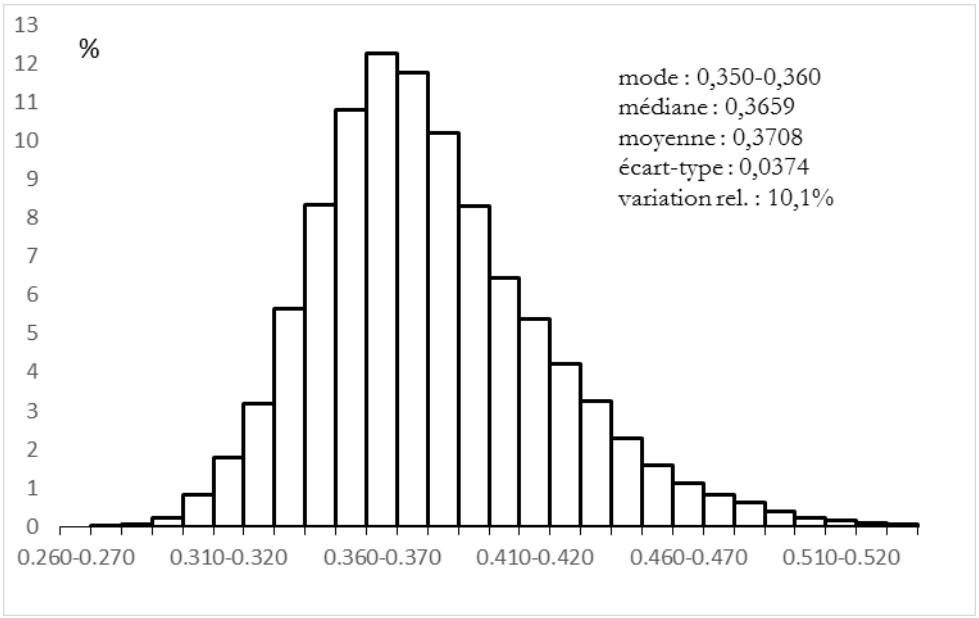
En moyenne, les facteurs temps et thèmes combinés augmentent de plus d'un quart la distance intertextuelle entre textes d'un même auteur.

Distances inter-auteurs

Enfin, il y a 19 230 distances inter-auteurs, soit 96,6% du total. Cette proportion indique combien cette attribution d'auteur n'est pas facile et ne peut être réussie par chance.

La distribution de ces distances inter-auteurs est décrite dans la figure 7.

Figure 7. Histogramme des distances entre extraits tirés de livres par des auteurs différents, classées par ordre croissant dans des classes d'intervalles égaux (% du total avec les valeurs centrales et indices de dispersion)



Ici la courbe en cloche est presque parfaite, principalement parce que la population est très vaste. Les trois valeurs centrales sont pratiquement confondues (légère asymétrie à droite).

Deux conclusions sont possibles.

Premièrement, l'écart entre les séries donne une mesure du poids du facteur auteur. Etant donné que les distances de ce troisième groupe sont l'effet de trois facteurs (auteur, temps et thèmes), le calcul doit logiquement être fait avec les distances intra-inter où seuls temps et thèmes interviennent :

$\delta = (0{,}3708 - 0{,}3219)/ 0{,}3219 = + 15{,}2\%$

En moyenne, le facteur auteur seul pèse moins lourd que la combinaison du temps et du thème (+25,8 %).

D'autres expériences ont montré que le temps – quand il dépasse une décennie - est en général supérieur au thème. La solution pour l'attribution d'auteur réside donc dans la comparaison de textes contemporains et sur des thèmes pas trop éloignés. Cette conclusion rejoint d'ailleurs la simple logique : pour que deux auteurs puissent collaborer, il faut bien qu'ils soient contemporains et qu'ils partagent une certaine vision (sinon, il ne reste qu'une explication le plagiat…)

Deuxièmement, on associe un intervalle de confiance à cette troisième distribution :

- borne inférieure : 0,3708 - (1,96 *0,0374) = 0,2974

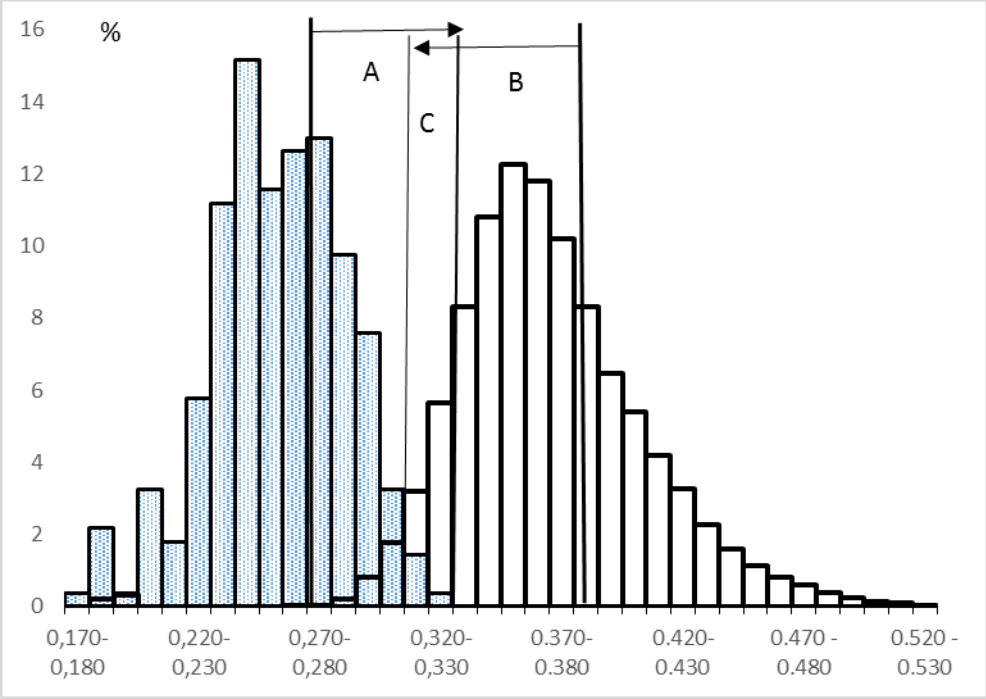
- borne supérieure : 0,3708 + (1,96 *0,0374) = 0,4427

La comparaison des trois distributions conduit à l'échelle de la distance intertextuelle pour l'attribution d'auteur

Echelle étalonnée pour l'attribution d'auteur

Pour comprendre le raisonnement qui conduit à cette échelle, il faut d'abord considérer les intervalles de confiance définis ci-dessus. La combinaison du premier intervalle intra-livre : même auteur, même thème et même époque - et du troisième (inter- auteurs) est représentée dans la figure 8. Les intervalles sont symbolisés par les deux flèches partant des deux moyennes et orientées respectivement vers la borne supérieure de l'intervalle intra-livre et de la borne basse de l'intervalle inter-auteur. NB : en ordonnées, figurent les effectifs relatifs des classes.

Figure 8. Association des séries intra-livres et inter-auteurs avec leurs intervalles de confiance respectifs.



Les deux moyennes sont situées en dehors de la borne extrême de l'intervalle de l'autre : les deux populations sont certainement différentes et les distances situées entre ces deux moyennes et les bornes peuvent être classées en trois situations :

- en dessous de 0,2974 (zone A) : les deux textes sont du même auteur avec une incertitude de 2,5% à la borne. Mais cela fait encore 19 230 * 0,025 ≈ 480 distances. Dans les faits, il y a 159 distances entre auteurs différents inférieures à ce seuil. Il est donc nécessaire de choisir des seuils de décision plus bas (comme nous l'avons fait dans la section consacrée à l'attribution d'auteur par les plus petites distances)

- en dessus de 0,3137 (zone B) : si les deux textes sont contemporains et écrits dans le même genre sur des thèmes proches, l'hypothèse de deux auteurs différents est privilégiée mais son acceptation dépend là encore des seuils choisis. Par exemple, avec α = 5%, il y a théoriquement 269*0,25 ≈ 7 distances intra-livres supérieures à ce seuil. Dans les faits, il y en a 9 (voir annexe 3) dont les extraits déjà signalés de *Graziella* (Lamartine), de *A rebours* (J.-K. Huysmans) et des *Illuminés* de G. de Nerval qu'il aurait fallu compter comme des livres différents.

- entre 0,2974 et 0,3137 (zone C) : chevauchement des deux intervalles de confiance. Si les deux textes sont contemporains et écrits dans un même genre sur des thèmes proches, aucune décision n'est possible. Cela concerne un peu moins d'un millier de distances, soit 5% de celles-ci et une douzaine de textes.

Pour diminuer l'incertitude, il faut agrandir la zone C (aucune conclusion possible). Avec :

Avec α < 1 % : [0.2766 – 0.3319] :

Il y a 4 distances inter-auteurs inférieures à la borne basse : *Les Confessions d'un enfant du siècle* (Musset) avec *Dominique* (Fromentin) et avec Vigny (*Grandeur et servitude militaires*) ; Cinq-Mars (Vigny) avec *les Trois mousquetaires* (Dumas) ; *Volupté* (Sainte-Beuve) avec *Dominique* (Fromentin). Les proximités des dates, des thèmes et parfois des auteurs, expliqueraient ces distances remarquablement faibles. A l'inverse, pour les distances intra-auteurs, l'annexe 3 indique que 27 plus proches voisins (intra-auteurs) sont comprises dans cet intervalle ;

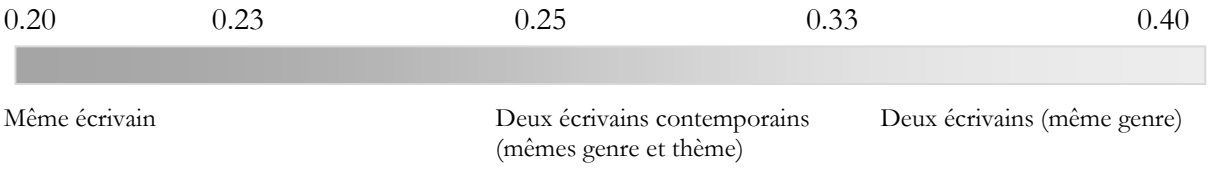
Avec α < 0.1% : [0,2498 – 0,3520] :

Il n'y a plus aucune distance inter-auteurs inférieure à la borne basse de cet intervalle. En contrepartie, seulement 117 des distances entre plus proches voisins sont inférieures à cette borne basse, soit 58,5% des plus proches voisins.

Le seuil de 0,25 (ou 25% des mots différents) correspond donc à une décision sérieuse (à la borne, moins d'une chance sur mille de se tromper en considérant que les deux textes sont de la même plume).

De multiples expériences - comme celle qui vient d'être présentée - ont été réalisées et ont abouti à la calibration d'une échelle de la distance (présentée pour la première fois dans Labbé & Labbé 2001). Soit deux textes A et B, écrits dans un même genre à la même époque dont les longueurs sont supérieures à 5 000 mots et inférieures à 25 000 mots, l'échelle présentée dans la figure 9 permet de prendre une décision sans se reporter à un corpus de référence.

Figure 9. Schéma de principe de l'acceptation ou du rejet de l'hypothèse de l'écrivain unique pour deux textes en fonction de leur distance.



- une distance inférieure ou égale à 0.20 indique un auteur unique avec une incertitude infinitésimale;

- une distance inférieure à 0.23 indique un auteur unique avec moins une incertitude de 1 sur 1000 (à la borne supérieure) ;

- une distance comprise entre 0.23 et 025 indique un auteur unique avec une incertitude inférieure à 1% (à la borne supérieure) ;

- une distance comprise entre 0.25 et 0.40 peut indiquer soit (1) un auteur unique écrivant à des époques et/ou sur des thèmes plus ou moins éloignés, soit (2) deux auteurs contemporains travaillant sur des thèmes plus ou moins proches. A partir de 0.33 (2) devient l'hypothèse la plus probable ;

- une distance supérieure à 0,40 indique soit (1) un changement de genre chez un même auteur, soit (2) deux auteurs différents dans un genre unique sur des thèmes éloignés.

Ces seuils d'attribution ont été calibrés, dans les années 1990, à l'aide d'un très grand nombre de textes français de toute nature allant du XVIIe au XXe siècle. Depuis 20 ans, cette procédure a été mise à l'épreuve à de nombreuses reprises sans être prise en défaut.

Conclusions

Quatre questions étaient posées à propos du corpus Saint-Jean.

Premièrement, quels sont les textes écrits par le même auteur ?

Tous les textes ont été "mariés" sans erreur à leur plus proche voisin. Dans la mise en œuvre d'une attribution réelle, cette méthode implique que l'on soit certain que tout texte douteux a bien au moins un "frère" parmi les textes non douteux utilisés comme références.

En revanche, si l'on n'est pas certain que l'auteur recherché ait au moins un texte dans le corpus – cas normal dans une attribution d'auteur - il faut renoncer à une classification exhaustive en se concentrant sur les plus petites distances. En contrepartie, certains textes ne peuvent être attribués (22 % des textes du corpus Saint-Jean). Comme indiqué, lors de la présentation de cette méthode en 2001, c'est le prix à payer pour une décision rapide, solide et transparente.

Deuxièmement, déterminer les textes écrits par des auteurs différents ?

Tous les textes du corpus Saint-Jean ont été associés avec leurs plus lointains. Même si cette méthode ne permet pas de classer l'ensemble du corpus, elle peut être utile pour écarter certains auteurs et restreindre le champ de la recherche. Là encore l'échelle de la distance apporte une réponse rapide et sûre, à condition de renoncer à l'exhaustivité.

Troisièmement, comment ça marche ?

La méthode présentée dans cette note est transparente et reproductible. Elle prend en compte la totalité du texte sans aucune sélection (qui peut toujours être soupçonnée) ; chaque mot pèse dans le calcul exactement son poids dans les textes comparés ; il n'y a aucune intervention de l'opérateur, pas d'apprentissage et aucune supervision. La distance intertextuelle présente les propriétés d'une distance euclidienne. Ses résultats sont faciles à interpréter. Depuis 2001, l'ensemble des données et des algorithmes sont mis à la disposition des chercheurs qui souhaitent refaire ces expériences.

Quatrièmement, comment classer de manière optimale une grande collection de textes ?

La classification arborée est, en l'état de l'art, la meilleure technique, puis la classification hiérarchique ascendante. Il faut associer à la classification arborée un indice de confiance – pourcentage de l'information contenue dans le tableau original qui est restituée par la figure – et un indice de qualité pour chaque nœud et feuille de l'arbre. Soulignons que la quasi-totalité des méthodes de classification ne donnent pas ces éléments de jugement.

Une difficulté particulière réside dans les notions de « genre » et de « thème ». En effet, il est difficile d'isoler ces catégories à l'état pur et l'on peut distinguer un grand nombre de sous-genres et de sous-thèmes. Par exemple, dans le corpus Saint-Jean, on aurait pu distinguer les romans historiques (*Cinq-Mars*, *les Trois mousquetaires*, *Notre-Dame de Paris*, *Histoire d'un conscrit de 1813* ou les *Illuminés*), la science-fiction (*De la terre à la lune*, *le Tour du monde en 80 jours*, *Jettatura*…), le roman naturaliste, etc. De même, la plupart des romans du corpus Saint-Jean sont multi-thématiques, de telle sorte que ces facteurs ne sont pas totalement neutralisés dans le calcul portant sur des extraits d'une même œuvre.

Sous cette réserve, à condition que les textes aient été préparés (correction et standardisation orthographiques, lemmatisation) et que la même procédure de calcul soit utilisée, les résultats d'expériences successives deviennent comparables et leurs enseignements peuvent se cumuler. En respectant les règles précises des sciences expérimentales, il est possible de tirer des conclusions solides et de jeter un nouveau regard sur l'auteur.

Un nouveau regard sur l'auteur

Face à un texte d'origine inconnue ou douteuse, le lecteur – ne disposant d'aucun indice "extra-textuel" - doit faire des milliers de comparaisons entre ce texte et tous ceux dont il se souvient, plus ou moins vaguement, pour tenter de détecter une même plume ou d'éventuelles collaborations ou des "influences". Une telle recherche dépasse la capacité du cerveau humain.

Du fait de cette impuissance, on considère habituellement que l'auteur est celui dont le nom figure sur la couverture du livre ou celui que les éditeurs, la critique et la tradition désignent comme tel. De multiples supercheries comme celle de Gary-Ajar – ou, plus récemment celle de Elena Ferrante (Savoy 2017) – montrent combien ces conventions sont fragiles.

La méthode présentée dans cette note permet donc de jeter un regard neuf sur l'auteur en dissipant le brouillard qui l'entourait jusqu'à maintenant. Naturellemet, pour acquérir des certitude, il faut traiter l'ensemble des œuvres concernées et approfondir les indices historiques disponibles. En effet, une attribution doit résulter non seulement d'une proximité remarquable entre textes mais aussi d'un faisceau d'indices convergents. Outre les éléments historiques, on examine d'autres indices de vocabulaire - comme les combinaisons de mots les plus fréquentes, le sens des mots les plus usuels (Labbé & Labbé 2005) - ou des indices stylistiques comme les longueurs des phrases (Labbé & Labbé 2010 ; Labbé 2014b).

Cette rapide discussion aura néanmoins suffi à suggérer combien ces méthodes ouvrent des perspectives nouvelles à l'histoire littéraire : vocabulaire et le style d'un auteur, proximité ou éloignement relatifs par rapport aux autres créateurs contemporains, "influences" et "collaborations", etc.

Signalons enfin que les méthodes présentées dans cette note ont bien d'autres utilités comme la détection des fraudes scientifiques (Byrne & Labbé 2016 ; Van Noorden 2014 ; Labbé & Labbé 2012a) ou des plagiats (Labbé & Labbé 2012b, 2014).

Bibliographie

Nos publications citées ci-dessous sont consultables sur HAL et sur researchgate

Barthélémy Jean-Pierre et Guénoche Alain (1988). *Les arbres et les représentations de proximité*. Paris, Dunod.

Baudelaire Charles (1845). Comment on paie ses dettes quand on a du génie. Article paru anonymement dans *Le Corsaire-Satan*, 24 novembre 1845.

Byrne Jennifer A. & Labbé Cyril (2016). Striking similarities between publications from China describing single gene knockdown experiments in human cancer cell lines. *Scientometrics*. Published on line: December 2016.

Cover Thomas M. & Hart Peter E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*. 13, 1967, 21–27.

Cover Thomas M. & Thomas J. A. (1991). *Elements of Information Theory*. John Wiley & Sons, 1991.

Croisille Christian (1997). Répertoire de la correspondance de Lamartine (1807-1829) et lettres inédites. *Cahiers d'étude sur les correspondances du XIXe siècle.* Clermont-Ferrand : Université Blaise-Pascal.

Embleton Sheila M. (1986). *Statistics in Historical Linguistics*. Bochum : Brokmeyer.

Felsenstein Joseph (2004). *Inferring Phylogenies*. Sunderland : Sinauer Ass.

Hall P., Park B. U. & Samworth R. J. (2008). "Choice of neighbor order in nearest-neighbor classification". *Annals of Statistics* 36 (5): 2135–2152.

Han Eui-Hong, Karypis George & Kumar Vipin (1999). Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification. *Lecture Notes in Computer Science*. Volume 2035, 21, p. 53-65.

Koppel Moshe, Schler Jonathan & Argamon Shlomo (2009). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology*. 60-1, 9-26.

Labbé Cyril & Labbé Dominique (2001). "Inter-Textual Distance and Authorship Attribution Corneille and Molière". *Journal of Quantitative Linguistics*. 8-3, December 21, p. 213-231.

Labbé Cyril & Labbé Dominique (2003). La distance intertextuelle. *Corpus*, 3, p. 95-118.

Labbé Labbé Cyril & Labbé Dominique (2005). How to measure the meanings of words? Amour in Corneille's work. *Language Resources Evaluation*. 25, 39, p. 335-351.

Labbé Cyril & Labbé Dominique (2006). A Tool for Literary Studies. Intertextual Distance and Tree Classification. *Literary and Linguistic Computing*. 21-3, p. 311-326.

Labbé Cyril & Labbé Dominique (2010). Ce que disent leurs phrases. In Bolasco Sergio, Chiari Isabella, Giuliano Luca (Eds). *Proceedings of 10th International Conference Statistical Analysis of Textual Data*. Rome : Edizioni Universitarie di Lettere Economia Diritto, 2010, Vol 1, p. 297-307.

Labbé Cyril & Labbé Dominique (2011a). La classification des textes. Comment trouver le meilleur classement possible au sein d'une collection de textes ? *Images des mathématiques. La recherche mathématique en mots et en images*. (http://images.math.cnrs.fr/La-classification-des-textes.html). 28 mars 2011.

Labbé Cyril & Labbé Dominique (2012a). Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science? *Scientometrics*. 22 june 2012.

Labbé Cyril & Labbé Dominique (2012b). Detection of Hidden Intertextuality in the Scientific Publications. In Dister Anne, Longrée Dominique, Purnelle Gérald (éds). *Proceedings of the 11th International Conference on Textual Data Statistical Analysis*. Liège : LASLA - SESLA, p.537-551.

Labbé Cyril & Labbé Dominique (2013a). Existe-t-il un genre épistolaire ? Hugo, Flaubert et Maupassant. In Banks David. *Le texte épistolaire du XVIIe siècle à nos jours*. Paris : L'Harmattan, p. 53-85.

Labbé Cyril & Labbé Dominique (2013b). Lexicométrie : quels outils pour les sciences humaines et sociales ? Communication aux journées d'étude *Usages de la lexicométrie en sociologie*. Versailles (12-13 juin).

Labbé Cyril & Labbé Dominique (2014). Who wrote this scientific text? Technical report. Grenoble : Laboratoire d'Informatique de Grenoble (LIG). September 2014.

Labbé Dominique (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble : Cahiers du CERAT.

Labbé Dominique (2002). La lemmatisation des grandes bases de textes. Un exemple : Corneille, Molière et Racine. Communication au colloque *L'édition électronique en littérature et dictionnairique, évaluation et bilan*. Rouen : 17-21 juin 2002.

Labbé Dominique (2007). Experiments on Authorship Attribution by Intertextual Distance in English. *Journal of Quantitative Linguistics*. 14-1, 1, April 27, p. 33-80.

Labbé Dominique (2014a). Identification de l'auteur d'un texte (Hugo, Lamartine, Musset et Vigny). Séminaire *L'œuvre et son auteur : problèmes d'attribution*. Université de Lille-Nord de la France, 21 mai 2014.

Labbé Dominique (2014b). *Qui a écrit Juba, Aétius et Tachmas ? Une attribution d'auteur par ordinateur.* Rapport technique. Grenoble : Pacte, décembre 2014.

Love Harold (2002). *Attributing Authorship: An Introduction*. Cambridge : Cambridge University Press.

Luong Xuan (1988). *Méthodes d'analyse arborée. Algorithmes, applications*. Thèse pour le doctorat ès sciences. Paris, Université de Paris V.

Luong Xuan (1994). L'analyse arborée des données textuelles : mode d'emploi. *Travaux du cercle linguistique de Nice*. 16, p. 25-42.

Monière Denis & Labbé Dominique (2006). L'influence des plumes de l'ombre sur les discours des politiciens. In Condé Claude et Viprey Jean-Marie. *Actes des 8e Journées internationales d'Analyse des données textuelles*. Besançon, II, p. 687-696.

Roux Maurice (1985). *Algorithmes de classification*. Paris : Masson. Ouvrage disponible en ligne : http://www.imep-cnrs.com/docu/mroux/algoclas.pdf.

Ruhlman Mathieu (2003). *Analyse arborée. Représentation par la méthode des groupements*. Grenoble : Polytech' – CERAT.

Savoy Jacques (2012a). Authorship Attribution: A Comparative Study of Three Text Corpora and Three Languages. *Journal of Quantitative Linguistics*. 19(2): 132-161.

Savoy Jacques (2012b). Authorship Attribution Based on Specific Vocabulary. *ACM Transactions on Information Systems*. 30, 2, September 2012.

Savoy Jacques (2013). Authorship attribution based on a probabilistic topic model. *Information Processing and Management*. 49, 1, p. 341-354.

Savoy Jacques (2016). Estimating the probability of an authorship attribution. *Journal of the Association for Information Science and Technology*. 67, 6, June 2016, p. 1462–1472.

Savoy Jacques (2017). *Elena Ferrante Unmasked*. Université de Neuchatel. Article consultable à : https://www.researchgate.net/publication/320131096

Sneath Peter & Sokal Robert (1973). *Numerical Taxonomy*. San Francisco: Freeman, 1973.

Stamatatos Efstathios (2008). Author Identification: Using Text Sampling to Handle the Class Imbalance Problem. *Information Processing & Management*. Volume 44-2. March 28, Pages 790–799.

Stamatatos Efstathios (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*. 60-3, p. 538-556.

Van Noorden Richard (2014). Publishers withdraw more than 120 gibberish papers. *Nature*. 24 February 2014.

Annexe 1.

Le corpus Saint-Jean

N°	Etiquette	Auteur et titre	Année
001	BHbe1	Balzac Honoré de - La cousine Bette	1846
002	CFat1	Chateaubriand François-René de - Atala	1801
003	DAmc1	Dumas Alexandre - Le comte de Monte Cristo	1845
004	FGbp1	Flaubert Gustave - Bouvard et Pécuchet	1881
005	GTav1	Gautier Théophile – Avatar	1856
006	GEge1	Goncourt Edmond et Jules de - Madame Gervaisais	1869
007	HVmi1	Hugo Victor - Les Misérables	1862
008	HJar1	Huysmans Joris-Karl - A rebours	1887
009	LAgr1	Lamartine (Alphonse de) - Graziella	1852
010	MGba1	Maupassant Guy de - Bel-Ami	1885
011	MAco1	Musset Alfred de - La Confession d'un enfant du siècle	1836
012	NGau1	Nerval Gérard de - Aurélia	1855
013	SGfa3	Sand George - La Petite Fadette	1851
014	SHch1	Stendhal (Henri Beyle) La Chartreuse de Parme	1839
015	VJtl1	Verne Jules - De la terre à la lune	1865
016	VAcm1	Vigny Alfred de - Cinq-Mars	1826
017	ZEar1	Zola Emile - L'Argent	1891
018	BHbe2	Balzac Honoré de - La cousine Bette	1846
019	CFat2	Chateaubriand François-René de - Atala	1801
020	DAmc2	Dumas Alexandre - Le comte de Monte Cristo	1845
021	FGbp2	Flaubert Gustave - Bouvard et Pécuchet	1881
022	GTav2	Gautier Théophile - Avatar	1856
023	GEge2	Goncourt Edmond et Jules de - Madame Gervaisais	1869
024	HVmi2	Hugo Victor - Les Misérables.	1862
025	HJar2	Huysmans Joris-Karl - A rebours	1887
026	LAgr2	Lamartine (Alphonse de) - Graziella	
027	MGba2	Maupassant Guy de - Bel-Ami	1885
028	MAco2	Musset Alfred de - La Confession d'un enfant du siècle	1836
029	NGau2	Nerval Gérard de - Aurélia	1855
030	SGfa4	Sand George - La Petite Fadette	1851
031	SHch2	Stendhal (Henri Beyle) La Chartreuse de Parme	1839
032	VJtl2	Verne Jules - De la terre à la lune	1865
033	VAcm2	Vigny Alfred de- Cinq-Mars	1826
034	ZEar2	Zola Emile - L'Argent	1891
035	BHbe3	Balzac Honoré de - La cousine Bette	1846
036	CFat3	Chateaubriand François-René de - René	1802
037	DAmc3	Dumas Alexandre - Le comte de Monte Cristo	1845
038	FGbo1	Flaubert Gustave - Madame Bovary	1857
039	GTje1	Gautier Théophile – Jettatura	1856
040	GEgl1	Goncourt Edmond et Jules de - Germinie Lacerteux	1864
041	HVmi3	Hugo Victor - Les Misérables.	1862

042	LAgr3	Lamartine (Alphonse de) - Graziella	1852
043	MGnc1	Maupassant (Guy de) - Notre coeur	1890
044	MAco3	Musset Alfred de - La Confession d'un enfant du siècle	1836
045	SGin1	Sand George - Indiana	1832
046	SHch3	Stendhal (Henri Beyle) La Chartreuse de Parme	1839
047	VJtm1	Verne Jules - Le tour du monde en quatre-vingt jours	1872
048	VAcm3	Vigny Alfred de - Cinq-Mars	1826
049	ZEas1	Zola Émile - L'Assommoir	1879
050	BHbi1	Balzac Honoré de - Histoire de la grandeur et de la décadence de César Birotteau	1837
051	DAtm1	Dumas Alexandre – Les trois mousquetaires	1844
052	FGbo2	Flaubert Gustave - Madame Bovary : moeurs de province	1857
053	GTje2	Gautier Théophile – Jettatura	1856
054	GEgl2	Goncourt Edmond et Jules de - Germinie Lacerteux	1864
055	HVnd1	Hugo Victor - Notre Dame de Paris	1831
056	MGnc2	Maupassant (Guy de) - Notre coeur	1890
057	SGin2	Sand George - Indiana	1832
058	SHrn1	Stendhal (Henri Beyle) - Le Rouge et le Noir	1830
059	VJtm2	Verne Jules - Le tour du monde en quatre-vingt jours	1872
060	ZEas2	Zola Émile - L'Assommoir	1879
061	BHbi2	Balzac Honoré de - Histoire de la grandeur et de la décadence de César Birotteau	1837
062	DAtm2	Dumas Alexandre – Les trois mousquetaires	1844
063	FGbo3	Flaubert Gustave - Madame Bovary	1857
064	GTsp1	Gautier Théophile – Spirite	1865
065	GEgl3	Goncourt Edmond et Jules de - Germinie Lacerteux	1864
066	HVnd2	Hugo Victor - Notre Dame de Paris	1831
067	MGfm1	Maupassant Guy de - Fort comme la mort	1889
068	SGmd1	Sand George - La mare au diable	1846
069	SHrn2	Stendhal (Henri Beyle) - Le Rouge et le Noir -	1830
070	VAsg1	Vigny Alfred de - Servitude et grandeur militaires	1835
071	ZEbh1	Zola Emile - La bête humaine	1890
072	BHch1	Balzac Honoré de - Le Colonel Chabert	1832
073	DAtm3	Dumas Alexandre – Les trois mousquetaires	1844
074	FGcs1	Flaubert Gustave - Un coeur simple	1877
075	HVnd3	Hugo Victor - Notre Dame de Paris	1831
076	FGes1	Flaubert Gustave - L'Education Sentimentale	1869
077	MGfm2	Maupassant Guy de - Fort comme la mort	1889
078	SGmd2	Sand George - La mare au diable	1846
079	VAsg2	Vigny Alfred de - Servitude et grandeur militaires	1835
080	ZEbh2	Zola Emile - La bête humaine	1890
081	BHch2	Balzac Honoré de - Le Colonel Chabert	1832
082	FGes2	Flaubert Gustave - L'Education Sentimentale	1869
083	MGmo1	Maupassant Guy de - Mont Oriol	1886
084	ZEfr1	Zola Émile - La Fortune des Rougon	1871
085	BHgo1	Balzac (Honoré de) - Le père Goriot	1835
086	FGtr1	Flaubert Gustave - Hérodias - Trois contes	1877

087	MGmo2	Maupassant Guy de - Mont Oriol	1886
088	ZEfr2	Zola Émile - La Fortune des Rougon	1871
089	BHgr1	Balzac Honoré de - Eugénie Grandet	1833
090	FGsa1	Flaubert Gustave - Salammbô	1862
091	MGmo3	Maupassant Guy de - Mont Oriol	1886
092	ZEge1	Zola Emile - Germinal	1885
093	BHgr2	Balzac Honoré de - Eugénie Grandet	1833
094	FGsa2	Flaubert Gustave - Salammbô	1862
095	BHgo2	Balzac Honoré de - Le père Goriot	1835
096	MGuv1	Maupassant Guy de - Une vie	1883
097	BHnu1	Balzac Honoré de –La Maison Nucingen	1838
098	ZEge2	Zola Emile - Germinal	1885
099	SHrn3	Stendhal (Henri Beyle) - Le Rouge et le Noir	1830
100	BHnu2	Balzac Honoré de - La Maison Nucingen	1838
101	BJct1	Barbey d'Aurevilly Jules - Le chevallier des Touches	1864
102	BPid1	Bourget Paul - Une idylle tragique	1896
103	DTpc1	Daudet Alphonse – Le Petit Chose	1868
104	DAvb1	Dumas Alexandre - Le Vicomte de Bragelonne	1847
105	SAde1	Staël-Holstein Anne-Louise (Madame de) - Delphine	1803
106	EChc1	Erckmann Emile & Chatrian Alexandre - Histoire d'un conscrit de 1813	1864
107	FAbo1	France Anatole - Le crime de Sylvestre Bonnard	1881
108	FEdo1	Fromentin Eugène – Dominique	1862
109	HVmi4	Hugo Victor - Les Misérables.	1862
110	HJhf1	Huysmans Joris-Karl – Marthe histoire d'une fille	1876
111	LAge1	Lamartine Alphonse de – Geneviève	1851
112	LPpi1	Loti Pierre - Pêcheur d'Islande	1886
113	NGil1	Nerval Gérard de - Les Illuminés	1852
114	PMpj1	Proust Marcel - Les plaisirs et les jours	1896
115	RHbr1	Régnier Henri de - Les Rencontres de Monsieur de Bréot	1901
116	SGfa1	Sand Georges - La petite Fadette	1832
117	SBvo1	Sainte-Beuve Charles-Augustin – Volupté	1834
118	SEmy1	Sue Eugène - Les Mystères de Paris	1842
119	BJct2	Barbey d'Aurevilly Jules - Le chevallier des Touches	1864
120	BPid2	Bourget Paul - Une idylle tragique	1896
121	DTpc2	Daudet Alphonse – Le Petit Chose	1868
122	DAvb2	Dumas Alexandre - Le Vicomte de Bragelonne	1847
123	SAde2	Staël-Holstein Anne-Louise (Madame de) - Delphine	1803
124	EChc2	Erckmann Emile & Chatrian Alexandre - Histoire d'un conscrit de 1813	1864
125	FAbo2	France Anatole - Le crime de Sylvestre Bonnard	1881
126	FEdo2	Fromentin Eugène – Dominique	1862
127	HVmi5	Hugo Victor - Les Misérable	1862
128	HJhf2	Huysmans Joris-Karl – Marthe histoire d'une fille	1876
129	LAge2	Lamartine Alphonse de – Geneviève	1851
130	LPpi2	Loti Pierre - Pêcheur d'Islande	1886
131	NGil2	Nerval Gérard de - Les Illuminés	1852

132	PMpj2	Proust Marcel - Les plaisirs et les jours	1896
133	RHbr2	Régnier Henri de - Les Rencontres de Monsieur de Bréot	1901
134	SGin3	Sand George - Indiana	1832
135	SBvo2	Sainte-Beuve Charles-Augustin – Volupté	1834
136	SEmy2	Sue Eugène - Les Mystères de Paris	1842
137	BJct3	Barbey d'Aurevilly Jules - Le chevallier des Touches	1864
138	BPid3	Bourget Paul - Une idylle tragique	1896
139	DTpc3	Daudet Alphonse – Le Petit Chose	1868
140	DAvb3	Dumas Alexandre - Le Vicomte de Bragelonne	1847
141	SAde3	Staël-Holstein Anne-Louise (Madame de) - Delphine	1803
142	EChc3	Erckmann Emile & Chatrian Alexandre - Histoire d'un conscrit de 1813	1864
143	FAbo3	France Anatole - Le crime de Sylvestre Bonnard	1881
144	FEdo3	Fromentin Eugène – Dominique	1862
145	LAge3	Lamartine Alphonse de – Geneviève	1851
146	LPpi3	Loti Pierre - Pêcheur d'Islande	1886
147	NGil3	Nerval Gérard de - Les Illuminés	1852
148	PMpj3	Proust Marcel - Les plaisirs et les jours	1896
149	RHbr3	Régnier Henri de - Les Rencontres de Monsieur de Bréot	1901
150	SGfa2	Sand George - La Petite Fadette	1851
151	SBvo3	Sainte-Beuve Charles-Augustin – Volupté	1834
152	SEmy3	Sue Eugène - Les Mystères de Paris	1842
153	BJct4	Barbey d'Aurevilly Jules - Le chevallier des Touches	1864
154	BPid4	Bourget Paul - Une idylle tragique	1896
155	DTpc4	Daudet Alphonse – Le Petit Chose	1868
156	DAvb4	Dumas Alexandre - Le Vicomte de Bragelonne	1847
157	SAde4	Staël-Holstein Anne-Louise (Madame de) - Delphine	1803
158	EChc4	Erckmann Emile & Chatrian Alexandre - Histoire d'un conscrit de 1813	1864
159	FAbo4	France Anatole - Le crime de Sylvestre Bonnard	1881
160	FEdo4	Fromentin Eugène – Dominique	1862
161	LAge4	Lamartine Alphonse de – Geneviève	1851
162	LPpi4	Loti Pierre - Pêcheur d'Islande	1886
163	RHbr4	Régnier Henri de - Les Rencontres de Monsieur de Bréot	1901
164	SGin4	Sand George - Indiana	1832
165	SBvo4	Sainte-Beuve Charles-Augustin – Volupté	1834
166	SEmy4	Sue Eugène - Les Mystères de Paris	1842
167	BJdi1	Barbey d'Aurevilly Jules - Les Diabolique	1874
168	BPid5	Bourget Paul - Une idylle tragique	1896
169	DTpc5	Daudet Alphonse – Le Petit Chose	1868
170	FApe1	France Anatole - La Rôtisserie de la reine Pédauque	1893
171	FEdo5	Fromentin Eugène – Dominique	1862
172	LPpi5	Loti Pierre - Pêcheur d'Islande	1886
173	RHdm1	Régnier Henri de - La Double Maîtresse	1900
174	SEmy5	Sue Eugène - Les Mystères de Paris	1842
175	BJdi2	Barbey d'Aurevilly Jules - Les Diabolique	1874
176	BPid6	Bourget Paul - Une idylle tragique	1896

177	DTpc6	Daudet Alphonse – Le Petit Chose	1868
178	FApe2	France Anatole - La Rôtisserie de la reine Pédauque	1893
179	LPmc1	Loti Pierre - Madame Chrysanthème	1899
180	RHdm2	Régnier Henri de - La Double Maîtresse	1900
181	SBvo5	Sainte-Beuve Charles-Augustin – Volupté	1834
182	SEmy6	Sue Eugène - Les Mystères de Paris	1842
183	SAde5	Staël-Holstein Anne-Louise (Madame de) - Delphine	1803
184	BJdi3	Barbey d'Aurevilly Jules - Les Diaboliques	1874
185	VAen1	Vallès Jules - L'Enfant	1879
186	FApe3	France Anatole - La Rôtisserie de la reine Pédauque	1893
187	LPmc2	Loti Pierre - Madame Chrysanthème	1899
188	RHdm3	Régnier Henri de - La Double Maîtresse	1900
189	SBvo6	Sainte-Beuve Charles-Augustin – Volupté	1834
190	SEmy7	Sue Eugène - Les Mystères de Paris	1842
191	SAde6	Staël-Holstein Anne-Louise (Madame de) - Delphine	1803
192	BJdi4	Barbey d'Aurevilly Jules - Les Diaboliques	1874
193	VAen2	Vallès Jules - L'Enfant	1879
194	RHdm4	Régnier Henri de - La Double Maîtresse	1900
195	SEmy8	Sue Eugène - Les Mystères de Paris	1842
196	FApe4	France Anatole - La Rôtisserie de la reine Pédauque	1893
197	VAen2	Vallès Jules - L'Enfant	1879
198	SEmy9	Sue Eugène - Les Mystères de Paris	1842
199	VAen4	Vallès Jules - L'Enfant	1879
200	SEmy10	Sue Eugène - Les Mystères de Paris	1842

Annexe 2

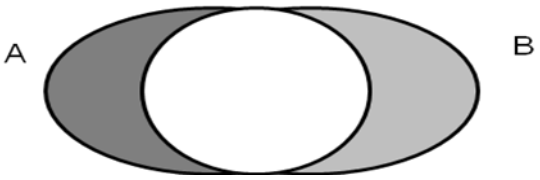
Calcul de la distance intertextuelle

Un exposé détaillé - en français et destiné aux non-mathématiciens - est disponible en ligne dans *Images des mathématiques*, revue des mathématiciens du CNRS destinée à un large public (Labbé & Labbé 2011). Voir également : Labbé 2007 ; Labbé & Labbé 2003 (également en ligne).

Principes du calcul

Soit deux textes A et B.

Ces deux textes sont superposés et on compte le nombre de mots différents (zones grisées dans le schéma ci-dessous).



On note :

- N_A et N_B : nombre de **mots** ("tokens" en anglais) dans A et respectivement B, ou **longueurs** de A et de B, ici 8 mots dans les deux cas ;

- V_A et V_B : nombre de "**vocables**" ("types" en anglais) dans A et respectivement B. C'est l'étendue de leurs vocabulaires respectifs. $V_{(A,B)}$ est le vocabulaire total de A et B ;

- F_{iA} et F_{iB} : nombre de fois qu'un vocable i est utilisé dans A et respectivement B. Ce sont les **effectifs** ou les "fréquences absolues" de ce vocable ;

- $|F_{iA} - F_{iB}|$ la différence absolue des effectifs du vocable i dans A et dans B. L'adjectif "absolue" signifie que l'on ne tient pas compte du signe dans le résultat.

- $D_{(A,B)}$: la **distance** entre A et B.

Cette distance est le **nombre de mots différents** dans A par rapport à B (ou réciproquement).

$$(1)\ D_{(A,B)} = \sum_{i \in (A,B)}^{V_{(A,B)}} \left| F_{iA} - F_{iB} \right| \ avec\ N_A = N_B$$

Dans cette formule, la lettre majuscule sigma signifie "somme" et les notations en dessous et en dessus de ce symbole signifient "effectuer le calcul, figurant à droite de ce symbole, pour les vocables de rang i appartenant à A et/ou B, avec i variant de 1 à $V_{(A,B)}$".

Pour pouvoir comparer les résultats obtenus sur des populations importantes de textes, la distance relative est calculée :

$$(2)\ D_{rel(A,B)} = \frac{\sum_{i \in (A,B)}^{V_{(A,B)}} \left| F_{iA} - F_{iB} \right|}{N_A + N_B}$$

Cet indice varie entre 0 (mêmes vocables avec les mêmes effectifs dans les deux textes) et 1 aucun mot en commun. Cette variation est uniforme (ni seuil ni saut).

$D_{(A,B)}$ est une **distance euclidienne** (longueur du segment de droite unissant deux points). L'adjectif "euclidien" signifie "conforme à la géométrie d'Euclide" (par un point il ne passe qu'une parallèle à une droite située hors de ce point). Les propriétés d'une distance euclidienne sont :

- l'**identité** (la distance d'un point à lui-même est nulle),

- la **symétrie** (le résultat est le même que l'on mesure *AB* ou *BA*), ce qui dispense d'utiliser les vecteurs,

- l'**inégalité triangulaire** (le chemin direct entre *A* et *B* est toujours plus court qu'en passant par un point *C* n'appartenant pas au segment *AB*).

Ces propriétés ont d'importantes conséquences. Par exemple, on peut réaliser une représentation graphique de toutes les distances au sein d'une vaste population de textes, comme on dresse la carte d'une ville ou d'un quartier.

Trois remarques :

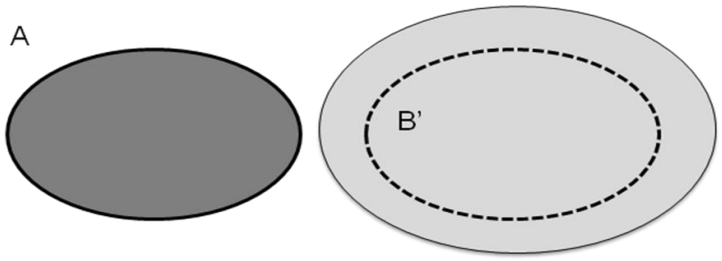
- le calcul porte sur la totalité des vocables et leurs effectifs, autrement dit sur l'ensemble du texte ("intertextuelle"),

- l'ordre des mots dans le texte importe peu, seule compte l'idée véhiculée (comme le maître de rhétorique l'explique à M. Jourdain),

- la longueur des textes est la même. Généralement ce n'est pas le cas. La formule a donc été adaptée pour s'appliquer à des textes de longueurs différentes.

Calcul sur des textes de longueurs différentes

Dans le cas de deux textes de longueurs inégales ($N_A < N_B$), la distance est estimée en "réduisant" B à la longueur de A (schéma ci-dessous) puis en superposant A et B' (comme dans le schéma précédent) et en comptant le nombre de mots différents entre A et B'.



Soit :

- *U* : le **rapport des longueurs** de *A* et *B*, c'est-à-dire la proportion dont il faut réduire *B* pour obtenir *B'* (ou "coefficient de proportionnalité") :

$$U = \frac{N_A}{N_B}$$

- $E_{iA(u)}$: **l'effectif théorique** dans un texte *B'* de la longueur de A d'un vocable *i* appartenant au vocabulaire de *B*. Cet effectif théorique est obtenu en pondérant l'effectif de *i* dans *B* par *U* (formule 3) :

$$(3)\ E_{iA(u)} = F_{iB} * U \ \ avec U = \frac{N_A}{N_B}$$

Pour chacun des vocables de *B*, la formule (3) permet de calculer le nombre de fois que ce vocable apparaîtrait si *B* avait la longueur de *A*. En remplaçant, dans la formule (1), l'effectif de chacun des vocables de *B* par cet effectif théorique, on obtient une **estimation** de la distance intertextuelle (formule 4) :

$$(4)\ D_{(A,B')} = \sum_{i \in (A,B)}^{V_{(A,B)}} \left| F_{iA} - E_{iA(u)} \right|$$

Pour le calcul de la distance relative, on remplace, dans la formule (2) N_B par la somme des effectifs théoriques, c'est-à-dire la longueur théorique de *B'* ($N_{B'}$) :

$$N_{B'} = \sum_{i \in B}^{V_B} E_{iA(u)}$$

Aux arrondis près, $N_{B'}$ est égale à N_A. La formule (2) devient :

$$(5)\ D_{rel(A,B')} = \frac{\sum_{i \in (A,B)}^{V_{(A,B)}} \left| F_{iA} - E_{iA(u)} \right|}{N_A + N_{B'}}$$

Il s'agit d'une **estimation** et ceci pour au moins deux raisons.

Les effectifs dans A sont des entiers naturels et les effectifs théoriques dans *B'* des rationnels *approchant* des entiers naturels (inconnus). Autrement dit, le résultat de la soustraction - au numérateur de (4) et (5) - comportera des décimales sans signification mais qui seront pourtant additionnées pour obtenir la distance… Ces décimales pèseront d'autant plus lourd que le vocable considéré aura des effectifs faibles - observés dans A et théoriques dans *B'*. Or, dans tout texte en langue naturelle, les vocables qui n'apparaissent qu'une fois sont toujours plus nombreux que ceux survenant deux fois, eux-mêmes plus nombreux que les effectifs trois, etc. Le fait que dans les formules (4) et (5), on cumule des différences absolues ne permet pas à ces "erreurs" de s'annuler. Au contraire, elles se cumuleront. Au passage, on remarquera que cette caractéristique est considérablement aggravée quand on élève au carré le résultat de la soustraction au numérateur de (4). C'est ce que font les "analyses en composantes principales" ou l'"analyse factorielle des correspondances", sans que les usagers aient toujours conscience des déformations massives auxquelles conduit cette élévation au carré.

Pour limiter cet inconvénient, on élimine du calcul :

- Les vocables absents de *A* et pour lesquels l'effectif théorique dans *B'* est inférieur à 1. La formule (3) devient :

$$\text{(3 bis)}\ E_{iA(u)} = \begin{cases} 0 \quad si\ \ F_{iA} = 0\ \ et\ \ F_{iB} * U < 1 \\ F_{iB} * U\ \ si\ \ F_{iA} > 0\ ou\ F_{iB} * U \geq 1 \end{cases}$$

- La différence des effectifs observés en *A* et des effectifs théoriques en *B'* lorsque cette différence est inférieure à 0.5. En effet, puisqu'il s'agit d'estimer un entier, ce résultat équivaut à zéro. La formule (4) devient :

$$\text{(4 bis)}\ D_{(A,B')} = \sum_{i \in (A,B)}^{V_{(A,B)}} \left| F_{iA} - E_{iA(u)} \right| \ \text{avec} \left| F_{iA} - E_{iA(u)} \right| = 0\ \ si\ \ \left| F_{iA} - E_{iA(u)} \right| < 0.5$$

La formule (5) est complétée pour intégrer ces deux éléments.

La seconde raison, pour laquelle la formule (5) ne donne qu'une estimation de la distance, tient aux postulats qui fondent le calcul de l'effectif théorique d'un vocable dans *B'* (formule 3 bis). Cette formule repose sur deux postulats[2].

- premier postulat : l'effectif d'un vocable augmente proportionnellement à l'allongement du texte. Ce premier postulat n'est valable que pour les mots les plus fréquents et non-spécialisés (ces derniers surviennent par paquets dans les passages où sont traités les thèmes auxquels ils appartiennent),

- deuxième postulat : l'apparition des vocables nouveaux se fait toujours au même rythme. En réalité, ce rythme est très rapide au début du texte – donc la formule (3 bis) ne peut pas s'appliquer à des textes trop courts – puis il décline ensuite lentement pour devenir presque linéaire. Dès lors la formule (5) n'est pleinement valable que lorsque les deux textes comparés ne sont pas de longueurs trop différentes et lorsque la longueur du plus court excède le point à partir duquel le rythme d'apparition des mots nouveaux devient sensiblement linéaire.

Limites du calcul

Une série d'expériences indiquent que :

- les deux textes doivent avoir plus de 1 000 mots, et que, en-dessous de 5 000 mots, le résultat de (5) peut être instable,

- le rapport (*U*) doit être inférieur à 1 : 10. En fait, plus ce rapport s'élève, plus le résultat doit être examiné avec prudence.

- dans ces limites, l'incertitude qui pèse sur la distance estimée est comprise entre ±1% (90% des valeurs sont comprises dans cet intervalle), avec des textes de longueurs supérieures à 5 000 mots et avec $U \geq 0.5$) et ±5%, avec la longueur du petit texte au moins égal à 5 000 mots et lorsque $U = 0.2$ (rapport de 1:5).

[2] Labbé Cyril, Labbé Dominique et Hubert Pierre. Automatic Segmentation of Texts and Corpora. *Journal of Quantitative Linguistics,* 11-3, 2004, p. 193-213.

Les textes du corpus Saint-Jean comptent tous 10 000 mots. Le calcul se fait avec la formule (1). Les résultats sont présentés avec quatre décimales (ordre de précision du calcul). Ils se lisent directement : un indice de 0,2000 indique que, entre les deux textes 2000 mots sont différents sur les dix mille mots.

Annexe 3

Les plus proches voisins des 200 textes du corpus Saint-Jean (classement par distance croissante)

N°	texte	1e voisin	Distance	Auteur et titres
1	030	116	0,1777	George Sand. *La petite Fadette*
2	116	030	0,1777	George Sand. *La petite Fadette*
3	157	183	0,1845	Anne-Louise Staël-Holstein. *Delphine*
4	183	157	0,1845	Anne-Louise Staël-Holstein. *Delphine*
5	191	183	0,1848	Anne-Louise Staël-Holstein. *Delphine*
6	123	141	0,1858	Anne-Louise Staël-Holstein. *Delphine*
7	141	123	0,1858	Anne-Louise Staël-Holstein. *Delphine*
8	150	116	0,1876	George Sand. *La petite Fadette*
9	105	123	0,1900	Anne-Louise Staël-Holstein. *Delphine*
10	068	078	0,2002	George Sand. *La mare au diable*
11	078	068	0,2002	George Sand. *La mare au diable*
12	011	028	0,2073	Alfred de Musset. *La Confession d'un enfant du siècle*
13	028	011	0,2073	Alfred de Musset. *La Confession d'un enfant du siècle*
14	133	163	0,2082	Henri de Régnier. *Les Rencontres de Monsieur de Bréot*
15	163	133	0,2082	Henri de Régnier. *Les Rencontres de Monsieur de Bréot*
16	044	011	0,2116	Alfred de Musset. *La Confession d'un enfant du siècle*
17	013	150	0,2147	George Sand. *La petite Fadette*
18	124	158	0,2151	Emile Erkmann & Alexandre Chatrian. *Histoire d'un conscrit de 1813*
19	158	124	0,2151	Emile Erkmann & Alexandre Chatrian. *Histoire d'un conscrit de 1813*
20	111	145	0,2162	Alphonse de Lamartine– *Geneviève*
21	145	111	0,2162	Alphonse de Lamartine– *Geneviève*
22	144	160	0,2201	Fromentin Eugène – *Dominique*
23	160	144	0,2201	Fromentin Eugène – *Dominique*
24	129	111	0,2214	Alphonse de Lamartine– *Geneviève*
25	045	057	0,2216	George Sand. *Indiana*
26	057	045	0,2216	George Sand. *Indiana*
27	087	091	0,2239	Guy de Maupassant. *Mont Oriol*
28	091	087	0,2239	Guy de Maupassant. *Mont Oriol / Une vie*
29	049	060	0,2244	Émile Zola. *L'Assommoir*
30	060	049	0,2244	Émile Zola. *L'Assommoir*
31	126	160	0,2250	Fromentin Eugène – *Dominique*
32	108	126	0,2273	Fromentin Eugène – *Dominique*
33	142	158	0,2275	Emile Erkmann & Alexandre Chatrian. *Histoire d'un conscrit de 1813*
34	084	088	0,2288	Émile Zola. *La Fortune des Rougon*
35	088	084	0,2288	Émile Zola. *La Fortune des Rougon*
36	193	197	0,2291	Jules Vallès. *L'Enfant*
37	197	193	0,2291	Jules Vallès. *L'Enfant*
38	069	099	0,2296	Henri Beyle Stendhal. *Le Rouge et le Noir*
39	099	069	0,2296	Henri Beyle Stendhal. *Le Rouge et le Noir*
40	171	160	0,2299	Fromentin Eugène – *Dominique*
41	085	095	0,2300	Honoré de Balzac. *Le père Goriot*.
42	095	085	0,2300	Honoré de Balzac. *Le père Goriot*.
43	115	163	0,2301	Henri de Régnier. *Les Rencontres de Monsieur de Bréot*
44	130	172	0,2326	Pierre Loti. *Pêcheur d'Islande*
45	172	130	0,2326	Pierre Loti. *Pêcheur d'Islande*
46	146	130	0,2338	Pierre Loti. *Pêcheur d'Islande*
47	164	045	0,2343	George Sand. *Indiana*

48	106	142	0,2347	Emile Erkmann & Alexandre Chatrian. *Histoire d'un conscrit de 1813*
49	134	164	0,2350	George Sand. *Indiana*
50	058	099	0,2351	Henri Beyle Stendhal. *Le Rouge et le Noir*
51	112	130	0,2363	Pierre Loti. *Pêcheur d'Islande*
52	154	176	0,2365	Paul Bourget. *Une idylle tragique*
53	176	154	0,2365	Paul Bourget. *Une idylle tragique*
54	051	073	0,2366	Alexandre Dumas. *Les trois mousquetaires*
55	073	051	0,2366	Alexandre Dumas. *Les trois mousquetaires*
56	038	063	0,2370	Gustave Flaubert. *Madame Bovary*
57	063	038	0,2370	Gustave Flaubert. *Madame Bovary*
58	161	111	0,2371	Alphonse de Lamartine– *Geneviève*
59	155	169	0,2374	Alphonse Daudet. *Le Petit Chos*e
60	169	155	0,2374	Alphonse Daudet. *Le Petit Chos*e
61	040	065	0,2379	Edmond et Jules de Goncourt. *Germinie Lacerteux*
62	065	040	0,2379	Edmond et Jules de Goncourt. *Germinie Lacerteux*
63	070	079	0,2381	Alfred de Vigny. *Servitude et grandeur militaires*
64	079	070	0,2381	Alfred de Vigny. *Servitude et grandeur militaires*
65	003	020	0,2384	Alexandre Dumas. *Le comte de Monte Cristo*
66	020	003	0,2384	Alexandre Dumas. *Le comte de Monte Cristo*
67	162	172	0,2386	Pierre Loti. *Pêcheur d'Islande*
68	010	027	0,2390	Guy de Maupassant. *Bel-Ami*
69	027	010	0,2390	Guy de Maupassant. *Bel-Ami*
70	014	031	0,2393	Henri Beyle Stendhal. *La Chartreuse de Parme*
71	031	014	0,2393	Henri Beyle Stendhal. *La Chartreuse de Parme*
72	198	200	0,2395	Eugène Sue. *Les Mystères de Paris*
73	200	198	0,2395	Eugène Sue. *Les Mystères de Paris*
74	067	077	0,2397	Guy de Maupassant. *Fort comme la mort*
75	077	067	0,2397	Guy de Maupassant. *Fort comme la mort*
76	052	063	0,2402	Gustave Flaubert. *Madame Bovary*
77	054	065	0,2405	Edmond et Jules de Goncourt. *Germinie Lacerteux*
78	122	140	0,2405	Alexandre Dumas. *Le Vicomte de Bragelonne*
79	138	154	0,2405	Paul Bourget. *Une idylle tragique*
80	140	122	0,2405	Alexandre Dumas. *Le Vicomte de Bragelonne*
81	177	169	0,2406	Alphonse Daudet. *Le Petit Chos*e
82	121	139	0,2410	Alphonse Daudet. *Le Petit Chose*
83	139	121	0,2410	Alphonse Daudet. *Le Petit Chose*
84	199	193	0,2420	Jules Vallès. *L'Enfant*
85	185	197	0,2421	Jules Vallès. *L'Enfant*
86	168	154	0,2422	Paul Bourget. *Une idylle tragique*
87	090	094	0,2423	Gustave Flaubert. *Salammbô*
88	094	090	0,2423	Gustave Flaubert. *Salammbô*
89	136	198	0,2426	Eugène Sue. *Les Mystères de Paris*
90	178	196	0,2433	Anatole France. *La Rôtisserie de la reine Pédauque*
91	196	178	0,2433	Anatole France. *La Rôtisserie de la reine Pédauque*
92	019	036	0,2434	François-René de Chateaubriand. *Atala / René*
93	036	019	0,2434	François-René de Chateaubriand. *Atala / René*
94	071	080	0,2434	Emile Zola. *La bête humaine*
95	080	071	0,2434	Emile Zola. *La bête humaine*
96	143	159	0,2438	Anatole France. *Le crime de Sylvestre Bonnard*
97	159	143	0,2438	Anatole France. *Le crime de Sylvestre Bonnard*
98	033	048	0,2440	Alfred de Vigny. *Cinq-Mars*

99	048	033	0,2440	Alfred de Vigny. *Cinq-Mars*
100	151	165	0,2444	Charles-Augustin Sainte-Beuve. *Volupté*
101	165	151	0,2444	Charles-Augustin Sainte-Beuve. *Volupté*
102	181	151	0,2444	Charles-Augustin Sainte-Beuve. Volupté
103	103	121	0,2445	Alphonse Daudet. *Le Petit Chose*
104	092	098	0,2448	Emile Zola. *Germinal*
105	098	092	0,2448	Emile Zola. *Germinal*
106	046	014	0,2452	Henri Beyle Stendhal. *La Chartreuse de Parme*
107	117	165	0,2455	Charles-Augustin Sainte-Beuve. *Volupté*
108	017	034	0,2460	Zola Emile. *L'Argent*
109	034	017	0,2460	Zola Emile. *L'Argent*
110	189	117	0,2463	Charles-Augustin Sainte-Beuve. *Volupté*
111	186	178	0,2480	Anatole France. *La Rôtisserie de la reine Pédauque*
112	174	198	0,2489	Eugène Sue. *Les Mystères de Paris*
113	083	091	0,2490	Guy de Maupassant. *Mont Oriol / Une vie*
114	102	138	0,2497	Paul Bourget. *Une idylle tragique*
115	089	093	0,2498	Honoré de Balzac. *Eugénie Grandet*
116	093	089	0,2498	Honoré de Balzac. *Eugénie Grandet*
117	135	117	0,2509	Charles-Augustin Sainte-Beuve. *Volupté*
118	016	048	0,2529	Alfred de Vigny. *Cinq-Mars*
119	062	073	0,2547	Alexandre Dumas. *Les trois mousquetaires*
120	120	168	0,2547	Paul Bourget. *Une idylle tragique*
121	156	140	0,2548	Alexandre Dumas. *Le Vicomte de Bragelonne*
122	001	035	0,2559	Honoré de Balzac. *La cousine Bette*
123	035	001	0,2559	Honoré de Balzac. *La cousine Bette*
124	119	153	0,2564	Jules Barbey d'Aurevilly. *Le chevalier des Touches*
125	153	119	0,2564	Jules Barbey d'Aurevilly. *Le chevalier des Touches*
126	182	198	0,2565	Eugène Sue. *Les Mystères de Paris*
127	043	056	0,2569	Guy de Maupassant - *Notre coeur*
128	056	043	0,2569	Guy de Maupassant - *Notre coeur*
129	012	029	0,2575	Gérard de Nerval. *Aurélia*
130	029	012	0,2575	Gérard de Nerval. *Aurélia*
131	152	190	0,2586	Eugène Sue. *Les Mystères de Paris*
132	190	152	0,2586	Eugène Sue. *Les Mystères de Paris*
133	076	082	0,2589	Gustave Flaubert. *Un cœur simple /l'Education sentimentale*
134	082	076	0,2589	Gustave Flaubert. *Un cœur simple /l'Education sentimentale*
135	180	188	0,2592	Henri de Régnier. *La Double Maîtresse*
136	188	180	0,2592	Henri de Régnier. *La Double Maîtresse*
137	167	192	0,2601	Jules Barbey d'Aurevilly. *Les Diaboliques*
138	192	167	0,2601	Jules Barbey d'Aurevilly. *Les Diaboliques*
139	175	192	0,2609	Jules Barbey d'Aurevilly. *Les Diaboliques*
140	118	182	0,2611	Eugène Sue. *Les Mystères de Paris*
141	104	140	0,2618	Alexandre Dumas. *Le Vicomte de Bragelonne*
142	006	023	0,2642	Edmond et Jules de Goncourt. *Madame Gervaisais*
143	023	006	0,2642	Edmond et Jules de Goncourt. *Madame Gervaisais*
144	179	187	0,2644	Pierre Loti. *Madame Chrysanthème*
145	187	179	0,2644	Pierre Loti. *Madame Chrysanthème*
146	125	143	0,2645	Anatole France. *Le crime de Sylvestre Bonnard*
147	039	053	0,2646	Théophile Gautier. *Jettatura*
148	053	039	0,2646	Théophile Gautier. *Jettatura*
149	149	163	0,2650	Henri de Régnier. *Les Rencontres de Monsieur de Bréot*

150	170	186	0,2651	Anatole France. *La Rôtisserie de la reine Pédauque*
151	050	061	0,2661	Honoré de Balzac. *Grandeur et décadence de César Birotteau*
152	061	050	0,2661	Honoré de Balzac. *Grandeur et décadence de César Birotteau*
153	114	148	0,2666	Marcel Proust. *Des plaisirs et des jours*
154	148	114	0,2666	Marcel Proust. *Des plaisirs et des jours*
155	184	167	0,2667	Jules Barbey d'Aurevilly. *Les Diaboliques*
156	195	198	0,2673	Eugène Sue. *Les Mystères de Paris*
157	132	148	0,2676	Marcel Proust. *Des plaisirs et des jours*
158	097	100	0,2701	Honoré de Balzac. *La Maison Nucingen*
159	100	097	0,2701	Honoré de Balzac. *La Maison Nucingen*
160	173	180	0,2706	Henri de Régnier. *La Double Maîtresse*
161	024	041	0,2708	Victor Hugo. Les Misérables
162	041	024	0,2708	Victor Hugo. *Les Misérables*
163	101	119	0,2708	Jules Barbey d'Aurevilly. *Le chevalier des Touches*
164	037	020	0,2709	Alexandre Dumas. *Le comte de Monte Cristo*
165	002	019	0,2725	François-René de Chateaubriand. *Atala*
166	110	128	0,2755	Joris-Karl Huysmans. *Marthe histoire d'une fille*
167	128	110	0,2755	Joris-Karl Huysmans. *Marthe histoire d'une fille*
168	009	026	0,2757	Alphonse de Lamartine. *Graziella*
169	096	083	0,2758	Guy de Maupassant. *Mont Oriol / Une vie*
170	166	195	0,2761	Eugène Sue. *Les Mystères de Paris*
171	072	081	0,2764	Honoré de Balzac. *Le Colonel Chabert*
172	081	072	0,2764	Honoré de Balzac. *Le Colonel Chabert*
173	018	035	0,2765	Honoré de Balzac. *La cousine Bette*
174	074	076	0,2771	Gustave Flaubert. *Un cœur simple /l'Education sentimentale*
175	107	125	0,2774	Anatole France. *Le crime de Sylvestre Bonnard*
176	004	021	0,2806	Gustave Flaubert. *Bouvard et Pécuchet*
177	021	004	0,2806	Gustave Flaubert. *Bouvard et Pécuchet*
178	109	127	0,2830	Victor Hugo. *Les Misérables*
179	127	109	0,2830	Victor Hugo. *Les Misérables*
180	113	131	0,2850	Gérard de Nerval. *Les Illuminés*
181	131	113	0,2850	Gérard de Nerval. *Les Illuminés*
182	007	041	0,2856	Victor Hugo. *Les Misérables*
183	066	075	0,2857	Victor Hugo. *Notre Dame de Paris*
184	075	066	0,2857	Victor Hugo. *Notre Dame de Paris*
185	005	022	0,2879	Théophile Gautier. *Avatar*
186	022	005	0,2879	Théophile Gautier. *Avatar*
187	055	075	0,2888	Victor Hugo. *Notre Dame de Paris*
188	183	157	0,2911	Anne-Louise Staël-Holstein. *Delphine*
189	157	183	0,2911	Anne-Louise Staël-Holstein. *Delphine*
190	086	090	0,2914	Gustave Flaubert. *Salammbô - Hérodias*
191	047	059	0,2962	Jules Verne. *Le tour du monde en quatre-vingt jours*
192	059	047	0,2962	Jules Verne. *Le tour du monde en quatre-vingt jours*
193	199	197	0,2998	Jules Vallès. *L'Enfant*
194	191	197	0,3023	Jules Vallès. *L'Enfant*
195	015	032	0,3031	Jules Verne. *De la terre à la lune*
196	032	015	0,3031	Jules Verne. *De la terre à la lune*
197	064	022	0,3065	Théophile Gautier– *Spirite / Avatar*
198	008	025	0,3153	Joris-Karl Huysmans. *A rebours*
199	025	008	0,3153	Joris-Karl Huysmans. *A rebours*
200	147	113	0,3183	Gérard de Nerval. *Les Illuminés*

Annexe 4

Les 200 couples les plus lointains (classement en fonction inverse de la distance)

Texte	Lointain	Distance
015	116	0,5299
116	015	0,5299
150	015	0,5291
030	015	0,5287
032	116	0,5259
025	116	0,5239
129	015	0,5211
068	015	0,5205
085	025	0,5196
090	183	0,5164
183	090	0,5164
008	030	0,5154
086	183	0,5145
009	183	0,5131
094	183	0,5126
078	015	0,5100
136	090	0,5094
152	090	0,5093
028	015	0,5080
191	090	0,5063
059	116	0,5037
013	015	0,5029
095	015	0,5017
187	183	0,5003
200	090	0,4979
157	032	0,4964
092	183	0,4960
006	183	0,4939
194	183	0,4937
198	015	0,4936
145	015	0,4935
182	025	0,4923
190	025	0,4922
123	032	0,4910
057	015	0,4907
141	032	0,4902
073	025	0,4894
166	025	0,4891
074	183	0,4888
158	183	0,4876
076	183	0,4875
111	015	0,4866
004	183	0,4863
003	015	0,4859
021	183	0,4859
156	025	0,4852
124	183	0,4837
105	032	0,4836
188	183	0,4832
011	015	0,4808
002	116	0,4805
045	015	0,4793
110	183	0,4783
193	015	0,4775
174	032	0,4770
044	032	0,4768
023	030	0,4758
197	015	0,4757
118	025	0,4722
140	025	0,4719
180	183	0,4712
142	183	0,4711
185	015	0,4707
196	015	0,4707
047	116	0,4704
173	183	0,4704
139	025	0,4703
179	183	0,4696
199	015	0,4693
161	015	0,4692
175	015	0,4692
130	183	0,4684
159	015	0,4681
146	183	0,4679
089	025	0,4672
155	025	0,4671
098	183	0,4670
012	116	0,4667
005	030	0,4658
143	015	0,4655
195	025	0,4644
016	032	0,4642
024	015	0,4628
169	015	0,4614
020	025	0,4612
192	015	0,4612
134	032	0,4604
154	015	0,4603
048	032	0,4593
172	183	0,4591
051	025	0,4589
144	015	0,4587
022	030	0,4585
093	025	0,4574
104	025	0,4573
178	015	0,4572
112	183	0,4570
176	032	0,4567
062	025	0,4563
082	183	0,456
060	015	0,4551
106	025	0,4551
040	015	0,4540
071	015	0,4540
054	015	0,4533
122	025	0,4528
087	015	0,4526
055	183	0,4520
165	086	0,4519
001	090	0,4518

121	025	0,4515
168	015	0,4513
177	015	0,4509
065	015	0,4507
050	025	0,4496
079	032	0,4492
164	015	0,4489
014	086	0,4485
035	090	0,4484
125	032	0,4477
133	015	0,4476
091	015	0,4469
036	116	0,4462
096	015	0,4462
148	015	0,4459
042	015	0,4458
126	015	0,4446
186	015	0,4445
031	025	0,4442
114	015	0,4441
097	090	0,4438
037	025	0,4437
108	090	0,4434
160	015	0,4433
072	090	0,4431
018	116	0,4430
007	025	0,4426
103	025	0,4425
067	015	0,4423
162	015	0,4420
102	015	0,4417
128	183	0,4415
029	116	0,4414
099	032	0,4413
064	183	0,4412
100	090	0,4412
132	015	0,4412
026	116	0,4409
010	015	0,4407
135	086	0,4404
163	015	0,4404
061	116	0,4400
081	032	0,4399
084	191	0,4395
138	032	0,4394
058	032	0,4393
113	085	0,4384
049	009	0,4380
056	015	0,4380
080	015	0,4380
046	025	0,4376
063	183	0,4374
041	009	0,4372
149	015	0,4367
033	015	0,4365
120	124	0,4359
171	015	0,4357
039	116	0,4351
137	183	0,4350
053	015	0,4348
109	191	0,4342
119	015	0,4342
101	183	0,4330
083	032	0,4328
066	183	0,4326
077	191	0,4326
107	090	0,4324
019	116	0,4322
167	032	0,4322
147	116	0,4320
070	032	0,4316
151	086	0,4314
052	183	0,4313
075	183	0,4306
027	015	0,4298
038	015	0,4295
189	086	0,4292
034	015	0,4288
127	002	0,4285
170	090	0,4281
117	090	0,4280
069	090	0,4277
153	015	0,4275
017	009	0,4269
088	025	0,4262
184	015	0,4251
115	015	0,4242
181	032	0,4237
043	015	0,4232
131	015	0,4215

Annexe 5

Les couples de textes séparés par une distance inférieure au seuil de 1%

1	030	116	0,1777
2	157	183	0,1845
3	183	191	0,1848
4	123	141	0,1858
5	123	157	0,1872
6	116	150	0,1876
7	105	123	0,1900

α = 000 01

8	141	157	0,1935
9	068	078	0,2002
10	157	191	0,2006
11	030	150	0,2012
12	011	028	0,2073
13	133	163	0,2082
14	141	183	0,2092
15	123	191	0,2095
16	123	183	0,2097
17	105	141	0,2100
18	011	044	0,2116

α = 000 1

19	013	150	0,2147
20	141	191	0,2150
21	124	158	0,2151
22	111	145	0,2162
23	144	160	0,2201
24	111	129	0,2214
25	028	044	0,2214
26	045	057	0,2216
27	087	091	0,2239
28	129	145	0,2242
29	049	060	0,2244
30	105	157	0,2248
31	126	160	0,2250
32	108	126	0,2273
33	142	158	0,2275
34	084	088	0,2288
35	193	197	0,2291
36	069	099	0,2296
37	160	171	0,2299
38	085	095	0,2300
39	115	163	0,2301
40	130	172	0,2326
41	013	116	0,2335
42	130	146	0,2338
43	045	164	0,2343
44	106	142	0,2347
45	126	144	0,2349
46	134	164	0,2350
47	058	099	0,2351
48	058	069	0,2356
49	112	130	0,2363
50	105	191	0,2364
51	154	176	0,2365
52	115	133	0,2365
53	051	073	0,2366

α = 001

54	038	063	0,2370
55	111	161	0,2371
56	155	169	0,2374
57	040	065	0,2379
58	144	171	0,2380
59	070	079	0,2381
60	003	020	0,2384
61	162	172	0,2386
62	010	027	0,2390
63	105	183	0,2390
64	014	031	0,2393
65	108	144	0,2394
66	198	200	0,2395
67	067	077	0,2397
68	112	162	0,2397
69	108	160	0,2400
70	052	063	0,2402
71	013	030	0,2404
72	122	140	0,2405
73	138	154	0,2405
74	054	065	0,2405
75	169	177	0,2406
76	121	139	0,2410
77	193	199	0,2420
78	185	197	0,2421
79	108	171	0,2422
80	154	168	0,2422
81	090	094	0,2423
82	197	199	0,2424
83	136	198	0,2426
84	178	196	0,2433
85	071	080	0,2434
86	126	171	0,2434
87	019	036	0,2434
88	143	159	0,2438

89	033	048	0,2440
90	145	161	0,2442
91	151	181	0,2444
92	151	165	0,2444
93	103	121	0,2445
94	092	098	0,2448
95	124	142	0,2448
96	146	172	0,2450
97	014	046	0,2452
98	117	165	0,2455
99	112	146	0,2457
100	017	034	0,2460
101	117	189	0,2463
102	138	176	0,2466
103	038	052	0,2478
104	178	186	0,2480
105	174	198	0,2489
106	083	091	0,2490
107	168	176	0,2494
108	106	124	0,2497
109	102	138	0,2497
110	106	158	0,2498
111	102	154	0,2498
112	089	093	0,2498
113	117	135	0,2509
114	112	172	0,2509
115	135	189	0,2510
116	130	162	0,2510
117	185	193	0,2518
118	102	176	0,2520
119	135	165	0,2520
120	165	189	0,2524
121	016	048	0,2529
122	031	046	0,2540
123	045	134	0,2542
124	030	078	0,2545
125	155	177	0,2546
126	120	168	0,2547
127	062	073	0,2547
128	140	156	0,2548
129	051	062	0,2553
130	146	162	0,2554
131	120	154	0,2554
132	186	196	0,2554
133	078	116	0,2558
134	151	189	0,2558
135	001	035	0,2559
136	185	199	0,2559
137	138	168	0,2562
138	119	153	0,2564
139	182	198	0,2565
140	181	189	0,2565
141	043	056	0,2569
142	012	029	0,2575
143	057	164	0,2578
144	078	150	0,2585
145	152	190	0,2586
146	056	067	0,2586
147	076	082	0,2589
148	180	188	0,2592
149	167	192	0,2601
150	175	192	0,2609
151	083	087	0,2611
152	118	182	0,2611
153	167	175	0,2612
154	120	138	0,2613
155	117	151	0,2614
156	104	140	0,2618
157	152	182	0,2633
158	165	181	0,2634
159	068	150	0,2640
160	006	023	0,2642
161	030	068	0,2643
162	179	187	0,2644
163	125	143	0,2645
164	039	053	0,2646
165	016	033	0,2648
166	149	163	0,2650
167	170	186	0,2651
168	117	181	0,2658
169	050	061	0.2660

$\alpha = 0.01$