Identification de l'auteur d'un texte (Hugo, Lamartine, Musset et Vigny)

Dominique Labbé

Université de Lille-Nord de la France

Ecole doctorale – Science de l'homme et de la société

Séminaire

L'œuvre et son auteur : problèmes d'attribution

21 mai 2014

Dominique Labbé

Laboratoire PACTE (CNRS - Institut d'Etudes Politiques de Grenoble)
(dominique.labbe@umrpacte.fr)
http://www.pacte-grenoble.fr/blog/membres/labbe-dominique/

Identification de l'auteur d'un texte

Hugo, Lamartine, Musset et Vigny

Abstract
Can lexical statistics identify the author of a text? In 1988, E. Brunet had answered negatively by using plays, novels and poems by Hugo, Lamartine and Musset. We propose to revisit this trial: cleared of its bias and assumptions, it shows that these three authors are clearly identifiable both in their vocabularies as their styles. This helps answering the question "What is an author?" establishing a kind of "identity card". The introduction, in this trial, of Vigny illustrates also how this method can highlight similarities between contemporary authors.

Résumé
La statistique lexicale permet-elle d'identifier l'auteur d'un texte ? En 1988, E. Brunet avait répondu par la négative en utilisant des pièces de théâtre, des romans et des poésies d'Hugo, Lamartine et Musset. Nous proposons de revisiter cette expérience : débarrassée de ses biais et de ses présupposés, elle montre que ces trois auteurs sont clairement identifiables tant au niveau de leurs vocabulaires que de leurs styles. Cela permet de répondre à la question « qu'est-ce qu'un auteur ? » en identifiant les caractéristiques particulières de son vocabulaire et de son style par rapport à ses contemporains. L'introduction de Vigny permet en outre de mettre en valeur des proximités et des influences entre auteurs.

> Avec quelle force une idée s'empare de nous, comme elle nous fait sa dupe, et combien il faut de temps pour l'user !
> (Alfred de Vigny. *Servitude et grandeur militaires*. Chapitre 1. Pourquoi j'ai rassemblé ces souvenirs).

Répondre à la question « Qu'est-ce qu'un auteur ? » c'est d'abord l'identifier. Notre cerveau n'a pas cette capacité de même que notre œil n'est pas équipé pour voir les bactéries ou les galaxies lointaines… Faute de l'équivalent du télescope ou du microscope, la théorie littéraire a été jusqu'à maintenant impuissante à dire ce qu'est un auteur et à proposer des critères susceptibles de l'identifier. Quelques cas célèbres comme celui de R. Gary – évoqué par V. Chepiga lors de cette journée d'étude (voir aussi Chepiga 2009) - montrent qu'il s'agit d'une cécité générale[1]. De telle sorte qu'une convention s'est imposée : de même que le père d'un enfant est l'époux de la mère, le père d'un texte est celui qui a son nom sur la couverture et, en cas de doute, celui que désignent les éditeurs, les critiques, les universitaires, l'opinion commune…

Certains ont déduit de cette impuissance que l'auteur d'un texte anonyme ou d'origine douteuse est définitivement impossible à connaître. Dans les années 1960-70, on a conclu que, pour la critique, l'auteur est mort (Barthes 1968), ou que c'est un lieu vide (Foucault 1969). Depuis, cette idée – de l'auteur impossible à identifier - s'est emparée de la critique littéraire et des universitaires et - comme le dit A. de Vigny à propos du militarisme - elle les a dupés à tel point qu'elle ne semble pas prête d'être dissipée. D'autant plus que, en 1988, cette position a reçu un renfort inattendu de la part de deux spécialistes de statistique lexicale (Brunet & Muller 1988), grâce à une expérience présentée comme décisive. Nous allons reprendre cette expérience et montrer que, non seulement, elle permet d'identifier les auteurs mais aussi de les mieux connaître.

I. Expérience décisive ou artefact ?

En 1988, Brunet et Muller se sont posé la question suivante : la statistique appliquée et l'ordinateur peuvent-ils reconnaître l'auteur d'un texte ? La première partie de l'article, écrite par C. Muller, bien que pessimiste, laissait prudemment la porte ouverte, mais E. Brunet la refermait à l'aide d'une expérience présentée comme décisive. Cette expérience portait sur des romans, poésies et pièces de théâtre de trois auteurs (Hugo, Lamartine et Musset). Selon Brunet l'ordinateur se "trompait" en attribuant systématiquement les romans à un auteur unique, les poésies à un deuxième lui aussi unique et les pièces à un troisième. Il en concluait que l'auteur est impossible à reconnaître tant le poids du genre est grand. Depuis lors, E. Brunet n'a cessé de répéter cette conclusion et, après lui, la quasi-totalité des littéraires, de telle sorte qu'aujourd'hui, la cause semble entendue : l'auteur serait impossible à reconnaître par informatique. Cette affirmation correspond si bien à la doxa ambiante que personne ne s'est rendu compte que l'expérience de Brunet n'a pas la portée que lui attribue son auteur.

D'une part, la formule utilisée par Brunet ne mesure pas seulement la distance entre les textes mais aussi et surtout leurs différences de longueur. Il en convenait d'ailleurs lui-même dans un autre article contemporain de son expérience où il avouait, à propos de sa méthode : "on ne peut pas être aveugle à l'inégalité de traitement qui frappe les textes longs et les textes courts." (1988b, p. 99). En effet, dans cet article, Brunet signalait clairement la dépendance de son indice à la longueur des textes (p. 83-84). C'est pourtant ce calcul qu'il a appliqué aux trois auteurs. Il est

[1] Par exemple, pour le théâtre français du XVIIe, on ignore les auteurs de plus de la moitié des pièces (Labbé 2014).

donc logique que l'expérience groupe ensemble les poésies, plus courtes que les pièces de théâtre, elles-mêmes plus courtes que les romans…

D'autre part, personne ne conteste que le genre s'impose aux auteurs – mêmes romantiques et contrairement à ce qu'Hugo affirme dans sa préface à *Cromwell* (1827) – et que changer de genre c'est un peu changer de langue. L'expérience de Brunet ne fait que confirmer ce lieu commun. Mais elle laisse en suspens la question cruciale : dans un genre donné, n'y aurait-il rien qui singularise un auteur par rapport aux autres écrivains contemporains ?

Nous proposons de répondre à cette seconde question en reprenant les mêmes auteurs que ceux sélectionnés par Brunet.

II. Le corpus

Pour des raisons qui seront dévoilées au cours de l'exposé, on y a ajouté Vigny qui est exactement contemporain des trois autres, qui a également donné des romans, des pièces de théâtre et de la poésie et qui appartient au même courant littéraire romantique. Nous supposons connus le climat intellectuel de l'époque, la vie et l'œuvre de ces quatre auteurs (voir brève bibliographie à la fin de cet exposé).

Le corpus

Tableau 1. Le corpus des « quatre auteurs »

Auteurs	Longueur (mots)	Vocabulaire
Hugo Victor (1802-1885)		
Poésie : *Les Contemplations* (1830-1855)	91 890	5 942
Théâtre (en vers) : *Hernani* (1830)	17 361	2 024
Ruy Blas (1838)	21 080	2 668
Romans : *Notre-Dame de Paris* (1831)	185 483	10 752
Les Misérables (1862)	564 292	17 387
Total V. Hugo	880 106	-
Lamartine Alphonse de (1790-1869)		
Poésie : *Méditations* (1815-1820 ; 1820-1848)	28 283	2 861
Théâtre (en vers) : *Saül* (1818)	16 259	1 710
Toussaint Louverture (1850)	24 738	2 807
Roman : *Graziella* (1852)	41 121	3 817
Total Lamartine	110 401	5 917
Musset Alfred de (1810-1857)		
Premières Poésies (1829-1835)	32 150	3 558
Théâtre (prose) : *Lorenzaccio* (1834)	36 201	3 163
André del Sarto (1833)	12 156	1 435
Roman : *Confession d'un enfant du siècle* (*La*, 1836)	96 516	5 031
Total Musset	177 023	7 127
Vigny Alfred de (1797-1863)		
Poésie *Livre mystique et livre antique (1826)*	15 618	2 393
Livre moderne (1826)	14 775	2 449
Théâtre (prose) : *Maréchale d'Ancre* (*La*, 1831)	21 575	2 046
Chatterton (1835)	15 066	1 773
Roman : *Cinq-Mars* (1826)	135 609	6 758
Servitude et grandeur militaires (1835)	62 549	4 928
Total Vigny	265 192	9 459

Les textes utilisés sont extraits d'une bibliothèque électronique du français moderne comportant au total 30 millions de mots, dont 11,5 millions de mots pour les textes littéraires (XVIIe – XXe siècle, voir annexe). Il s'agit des éditions de référence (voir bibliographie à la fin). Chaque œuvre est précisément datée, détail dont on verra l'importance dans la suite de la discussion. Cette expérience porte donc au total sur 1 432 722 mots.

Avant de décrire les résultats de cette expérience, deux précisions sont nécessaires. Ce corpus pose quelques problèmes qui sont résolus à l'aide de découpages et de regroupements. Avant d'être traités par informatique, les textes ont fait l'objet d'un certain nombre de traitements préalables indispensables avant de les soumettre à la procédure d'attribution d'auteur.

Découpages et regroupements

Plus de la moitié du corpus est occupé par le seul Hugo et notamment par son roman fleuve (*Les Misérables*).

Nous avons vu que, dans l'expérience Brunet, ce déséquilibre joue un rôle important. En pratique, le calcul présenté ci-dessous exige que les longueurs des textes soient comprises dans une échelle de 1 : 7. Les 564 292 mots des *Misérables* ne sont donc pas directement comparables avec les 12 000 d'*André del Sarto*.

Par découpages et regroupements, on a ramené les dimensions des textes dans la fourchette 5 000-35 000 mots, en respectant les segmentations naturelles : par exemple, les livres et les tomes des *Contemplations*, les livres de *Notre-Dame de Paris*, les tomes et les livres des *Misérables*, etc. A titre d'exemple, le tableau 2 donne le détail de ces découpages pour *Contemplations*.

Tableau 2. Détail des *Contemplations* (V. Hugo)

Tome	Livre	Titre	Longueur (mots)
1 Autrefois (1830-1843)	1	Aurore	14 426
	2	L'âme en fleur	6 634
	3	Les luttes et les rêves	18 972
2 Aujourd'hui (1843-1855)	4	Pauca meae	6 213
	5	En marche	15 787
	6	Au bord de l'infini	29 858

Naturellement, ces découpages doivent non seulement respecter les césures indiquées par l'auteur mais aussi être sans incidences sur les résultats.

Cela donne 27 extraits pour les *Misérables*, 11 pour *Notre-Dame de Paris*, 7 pour *Cinq-Mars*, etc. En tout, il y a au moins deux textes pour chaque auteur dans chaque genre et 76 textes au total.

Comparer chacun de ces textes aux 75 autres aboutit à 2 850 comparaisons, portant non seulement sur les 1,4 millions de mots mais aussi les combinaisons les plus fréquentes, les phrases… Une telle masse dépasse les capacités humaines mais elle peut être traitée en quelques secondes par l'ordinateur. Mais, pour lui permettre d'effectuer ces opérations, un certain nombre de traitements préalables sont indispensables.

Ces opérations sont décrites dans : Labbé & Labbé 2013b, Labbé 2002, Labbé 1990. Elles comportent :

- Le balisage. En tête du texte, on place 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… Ainsi l'analyse ne porte que sur ce qu'entend le spectateur, selon un principe admis par tous mais rarement appliqué !

- Correction orthographique et standardisation des graphies. Par exemple : M., Mr., Monsieur, monsieur… Un automate peut reconnaître le même mot dans les trois dernières formes mais la première doit être identifiée à la main : monsieur, Marcel, Maurice, Marie, mètre(s)…? La question n'est pas anecdotique : dans les *Misérables*, il y a 1 507 "M." (soit 2,6 pour mille mots). En fait, il s'agit toujours de "monsieur". C'est le deuxième substantif. Si l'on n'avait pas attaché à chacun de ces "M." une étiquette indiquant sa véritable identité, on aurait fait une cascade d'erreurs : V. Hugo n'emploie pas « monsieur », la phrase de V. Hugo est plus courte qu'on le pensait, etc… Idem pour les majuscules initiales de vers dans la poésie ou dans le théâtre. Ces variantes graphiques 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 dans la poésie).

- Etiquetage : chaque mot du texte se voit doter d'une étiquette où figure sa graphie standard, son entrée de dictionnaire et sa catégorie grammaticale. A "M." on associe une étiquette "monsieur, nom masculin". Ou encore "est" peut recevoir deux étiquettes : "être, verbe indicatif présent" ou "est, nom masculin". Ces étiquettes ne se substituent pas au texte, elles s'y ajoutent et servent à établir le vocabulaire d'un 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.

III. L'attribution d'auteur par ordinateur

Rappelons qu'un lecteur est désarmé quand il lui faut identifier l'auteur d'un texte. De plus, les érudits ne parviennent pas à fonder leurs intuitions – parfois exactes mais toujours invérifiables - sur des critères précis.

Les lunettes pallient à la myopie, le télescope et le microscope permettent à l'œil de voir plus loin ou plus près. De même, l'ordinateur peut reconnaître l'auteur, là où notre cerveau est désarmé. En effet, le cerveau humain a de grandes capacités mais pas celle de garder simultanément en mémoire des centaines de milliers de mots, pour comparer un grand nombre de textes, ce que l'ordinateur peut faire aisément. Cette idée est ancienne (résumé dans Love 2002). Beaucoup de méthodes et d'indices ont été proposés (présentation d'ensemble dans : Stamatatos 2009 ; Koppel & Al. 2009).

Nous avons présenté une méthode originale pour la première fois il y a 13 ans (Labbé & Labbé 2001 et en français: Labbé & Labbé 2003). 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 2011a). Voir également : Savoy 2012.

Méthode

Soit deux écrivains (A et B). On demande à l'ordinateur de comparer chaque texte de A avec chaque texte de B et de compter les différences au sein de chacun des couples ainsi formés. Le nombre des différences forme la distance qui varie uniformément entre 0 (tous les mots sont communs) 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.

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 (ce qui permet un certain nombre d'opérations qui seront évoquées dans la suite de cet exposé).

En dessous d'une certaine distance, on peut conclure que les deux textes ont été écrits par le même auteur et que – s'ils ont publié sous des noms différents - l'un des deux a été la plume de l'ombre de l'autre.

Cette méthode a été mise au point selon les protocoles les plus rigoureux, comportant notamment de nombreuses expériences en aveugle (les textes sont choisis par des tiers, anonymés, l'auteur étant dévoilé après l'expérience). Plusieurs de ces expériences ont été publiées (Monière et Labbé 2006 ; Labbé 2007 ; *Images des Mathématiques* 2011). Cette méthode a déjà permis de résoudre un certain nombre de cas (outre le théâtre du XVIIe). Par exemple, elle détecte aisément la plume de R. Gary dans les 4 romans publiés sous le nom d'E. Ajar (Labbé 2004 ; Lafon & Peeters 2006). Récemment, elle a permis d'identifier plus d'une centaine de faux articles scientifiques - au milieu de onze millions de références bibliographiques - publiés par l'IEEE et Springer, deux des plus grands éditeurs scientifiques mondiaux (Van Noorden 2014, Labbé & Labbé 2012).

Quels sont les facteurs qui influencent la distance ?

Ces multiples expériences ont permis d'identifier et de mesurer l'importance des principaux facteurs qui déterminent la distance entre textes. Par importance décroissante :

- le genre : oral et écrit, prose, vers, comédie et tragédie…
- l'auteur,
- l'époque où a été rédigé le texte car chaque époque a un vocabulaire particulier,
- le thème (personnages, lieux, principaux motifs).

L'importance relative de chaque facteur a été déterminée en utilisant le raisonnement « toutes choses égales par ailleurs ». Par exemple, en utilisant des textes appartenant au même genre (théâtre, poésie, roman, correspondance, etc.), écrits à la même époque, on peut isoler l'importance relative de l'auteur et du thème. Parfois, on a la chance que deux auteurs contemporains traitent du même thème, en aveugles, dans le même genre (par exemple les deux *Bérénice* de Corneille et Racine) : alors il ne reste plus que le facteur auteur.

La conclusion essentielle est la suivante : **dans un genre et à une époque donnée, la variable « auteur » l'emporte sur le thème.** Dès lors, pour déterminer l'auteur d'un texte d'origine douteuse ou inconnue, il suffit de le confronter à d'autres – dont l'origine n'est pas douteuse - écrits dans un même genre et à la même époque. NB : le théâtre doit être comparé au théâtre et, autant que possible, les tragédies entre elles, les comédies entre elles, etc.

Ces multiples expériences ont également permis de calibrer une échelle d'attribution d'auteur. Cette échelle s'applique aux textes dont les longueurs sont comprises entre 5 000 et 25 000 mots. Pour les textes plus longs, on utilise des extraits selon la méthode présentée dans Labbé 2007 :

- une valeur inférieure ou égale à 0.20 : auteur, genre et époque sont les mêmes, les thèmes sont très proches ;

- entre 0.20 et 0.25, l'auteur est probablement le même. Sinon, les deux textes ont été écrits à la même époque, dans un même genre, sur un sujet identique et avec des arguments semblables. Ce cas se rencontre souvent dans les articles de presse, à propos d'un même événement, parce que les journalistes travaillent à partir des mêmes sources et citent les mêmes noms de lieux et de personnes... Dans le cas d'œuvres littéraires appartenant à deux auteurs différents, le second s'est "inspiré" du premier (dans l'ordre chronologique). En tous cas, ce genre de "collision" peut difficilement se produire plusieurs fois entre deux auteurs distincts.

- au-dessus de 0.25, on entre dans une zone "grise" où deux hypothèses sont envisageables : un même auteur mais une époque et des thèmes différents ou deux auteurs contemporains traitant, dans un même genre, un thème proche… Plus la distance s'élève, plus la seconde hypothèse est probable ;

- au-dessus de 0.30, pour un même auteur, le genre est différent ou les dates de composition et les thèmes sont très éloignées ;

- au-dessus de 0.45 les auteurs sont différents ou bien, pour un même auteur, les textes sont de genres très éloignés, par exemple : oral et écrit.

Deux limites doivent être signalées. Premièrement, les résultats dépendent de la qualité des traitements préalables. Par exemple, ne pas harmoniser les graphies, ne pas réduire les majuscules initiales de phrase ou de vers interdit toute attribution d'auteur. Deuxièmement, en cas de textes de longueurs différentes, seuls les deux premiers chiffres de l'indice sont pertinents (dans les tableaux ci-dessous, le troisième chiffre est donné pour indiquer dans quel sens se fait l'arrondi).

IV. Quatre auteurs

Comparer chacun des 76 textes aux 75 autres du corpus implique au total 2 850 comparaisons différentes. Et le nombre des comparaisons texte par texte augmente exponentiellement avec l'augmentation du corpus. Cependant, dans le cas d'une attribution d'auteur, le modèle et l'échelle présentés ci-dessous permettent de ne considérer que les plus proches voisins de chaque texte douteux, lorsque leurs distances sont inférieures aux valeurs seuils de l'échelle ci-dessus.

Les plus proches voisins

Le tableau 3 ci-dessous indique le plus proche voisin de chaque texte (classement alphabétique). Les œuvres et les genres sont isolés par un cadre.

Tableau 3. Plus proche voisin de chaque texte et distance

Texte	Voisin	Distance
Hugo Contemplations1	Hugo Contemplations5	0,241
Hugo Contemplations2	Hugo Contemplations5	0,243
Hugo Contemplations3	Hugo Contemplations6	0,198
Hugo Contemplations4	*Hugo Contemplations5*	*0,253*
Hugo Contemplations5	Hugo Contemplations6	0,219
Hugo Contemplations6	Hugo Contemplations3	0,198
Hugo Misérables01	Hugo Misérables12	0,234
Hugo Misérables02	Hugo Misérables08	0,213
Hugo Misérables03	Hugo Misérables12	0,249
Hugo Misérables04	Hugo Misérables06	0,207
Hugo Misérables05	Hugo Misérables06	0,220
Hugo Misérables06	Hugo Misérables04	0,207
Hugo Misérables07	*Hugo Misérables23*	*0,269*
Hugo Misérables08	Hugo Misérables02	0,213
Hugo Misérables09	Hugo Misérables08	0,226
Hugo Misérables10	*Hugo Misérables12*	*0,275*
Hugo Misérables11	*Hugo Misérables19*	*0,255*
Hugo Misérables12	Hugo Misérables13	0,223
Hugo Misérables13	Hugo Misérables12	0,223
Hugo Misérables14	Hugo Misérables17	0,232
Hugo Misérables15	Hugo Misérables08	0,233
Hugo Misérables16	Hugo Misérables23	0,240
Hugo Misérables17	Hugo Misérables08	0,232
Hugo Misérables18	Hugo Misérables08	0,225
Hugo Misérables19	Hugo Misérables18	0,235
Hugo Misérables20	Hugo Misérables23	0,242
Hugo Misérables21	Hugo Misérables23	0,222
Hugo Misérables22	Hugo Misérables23	0,238
Hugo Misérables23	Hugo Misérables21	0,222
Hugo Misérables24	Hugo Misérables23	0,243
Hugo Misérables25	Hugo Misérables19	0,237
Hugo Misérables26	Hugo Misérables27	0,204
Hugo Misérables27	Hugo Misérables26	0,204
Hugo Notre-Dame01	*Hugo Notre-Dame10*	*0,274*
Hugo Notre-Dame02	Hugo Notre-Dame07	0,245
Hugo Notre-Dame03	*Hugo Notre-Dame05*	*0,348*
Hugo Notre-Dame04	*Hugo Notre-Dame06*	*0,279*
Hugo Notre-Dame05	***Hugo Misérables16***	***0,321***
Hugo Notre-Dame06	*Hugo Notre-Dame02*	*0,256*
Hugo Notre-Dame07	Hugo Notre-Dame08	0,222

Hugo Notre-Dame08	Hugo Notre-Dame11	0,212
Hugo Notre-Dame09	Hugo Notre-Dame08	0,243
Hugo Notre-Dame10	Hugo Notre-Dame07	0,245
Hugo Notre-Dame11	Hugo Notre-Dame08	0,212
Hugo RuyBlas	Hugo Hernani	0,236
Hugo Hernani	Hugo RuyBlas	0,236
Lamartine Méditations1	Lamartine Méditations2	0,222
Lamartine Méditations2	Lamartine Méditations1	0,222
Lamartine Graziella1	*Lamartine Graziella2*	*0,254*
Lamartine Graziella2	Lamartine Graziella3	0,250
Lamartine Graziella3	Lamartine Graziella2	0,250
Lamartine Saül	*Lamartine Toussaint*	*0,278*
Lamartine Toussaint	*Lamartine Saül*	*0,278*
Musset Poésies1	Musset Poésies2	0,245
Musset Poésies2	Musset Poésies1	0,245
Musset Confession1	Musset Confession2	0,195
Musset Confession2	Musset Confession1	0,195
Musset Confession3	Musset Confession4	0,176
Musset Confession4	Musset Confession3	0,176
Musset Confession5	Musset Confession4	0,178
Musset Lorenzaccio	Musset Sarto	0,225
Musset Sarto	Musset Lorenzaccio	0,225
Vigny Poèmes1	Vigny Poèmes2	0,245
Vigny Poèmes2	Vigny Poèmes1	0,245
Vigny Servitude1	Vigny Servitude3	0,190
Vigny Servitude2	Vigny Servitude3	0,205
Vigny Servitude3	Vigny Servitude1	0,190
Vigny Cinq-Mars1	Vigny Cinq-Mars2	0,200
Vigny Cinq-Mars2	Vigny Cinq-Mars1	0,200
Vigny Cinq-Mars3	Vigny Cinq-Mars7	0,196
Vigny Cinq-Mars4	Vigny Cinq-Mars5	0,194
Vigny Cinq-Mars5	Vigny Cinq-Mars4	0,194
Vigny Cinq-Mars6	Vigny Cinq-Mars7	0,197
Vigny Cinq-Mars7	Vigny Cinq-Mars4	0,196
Vigny Ancre	Vigny Chatterton	0,227
Vigny Chatterton	Vigny Ancre	0,227

Tous les plus proches voisins sont des textes de même genre et de même auteur et lorsqu'il y a plusieurs textes extraits d'un même ouvrage, ils sont mariés ensemble à une exception près en gras sur le tableau 3.

- En s'en tenant au seuil de 0.25, 53 textes (83%) sont mariés sans aucune erreur. Si l'on observe que **l'objectif n'est pas d'attribuer tout texte mais de le faire avec un haut degré de certitude**, le résultat est plus que satisfaisant ;

- Si l'on souhaite affecter tous les textes, il en reste 13 (en italiques sur le tableau) que l'on peut rapprocher "faute de mieux" de l'auteur du plus proche voisin mais il ne s'agit pas d'une attribution au plein sens du terme. Dans le cas présent, ces rapprochements sont tous exacts en ce qui concerne les auteurs mais le cinquième livre de *Notre-Dame de Paris* est "marié" au seizième livre des *Misérables* (l'explication de cette anomalie est donnée plus bas). En tous cas, ceci est sans conséquence quant à l'auteur : V. Hugo dans les deux cas.

Notre méthode attribue correctement tous les textes et identifie sans erreur les quatre auteurs.

Rappelons que trois des auteurs et la plupart des textes n'ont pas été choisis par nous mais par Brunet et Muller.

La discussion pourrait s'arrêter là. Toutefois, il n'est pas sans intérêt de poursuivre l'expérience.

Il est possible d'opérer des regroupements partiels non-supervisés et non-hiérarchiques selon le procédé présenté dans notre article *d'Images des mathématiques*. Par exemple, en utilisant les mariages croisés et le seuil de 0.25, les *Contemplations* 1, 2, 3, 5 et 6 forment un seul groupe (six premières lignes du tableau). Les trois quart des textes peuvent ainsi être groupés également sans aucune erreur.

On peut s'épargner ces opérations manuelles en utilisant des classifications automatiques.

Classification automatique

La classification hiérarchique ascendante est la méthode la plus répandue. Le résultat de cette classification est retracé dans le dendrogramme ci-dessous (tableau 4).

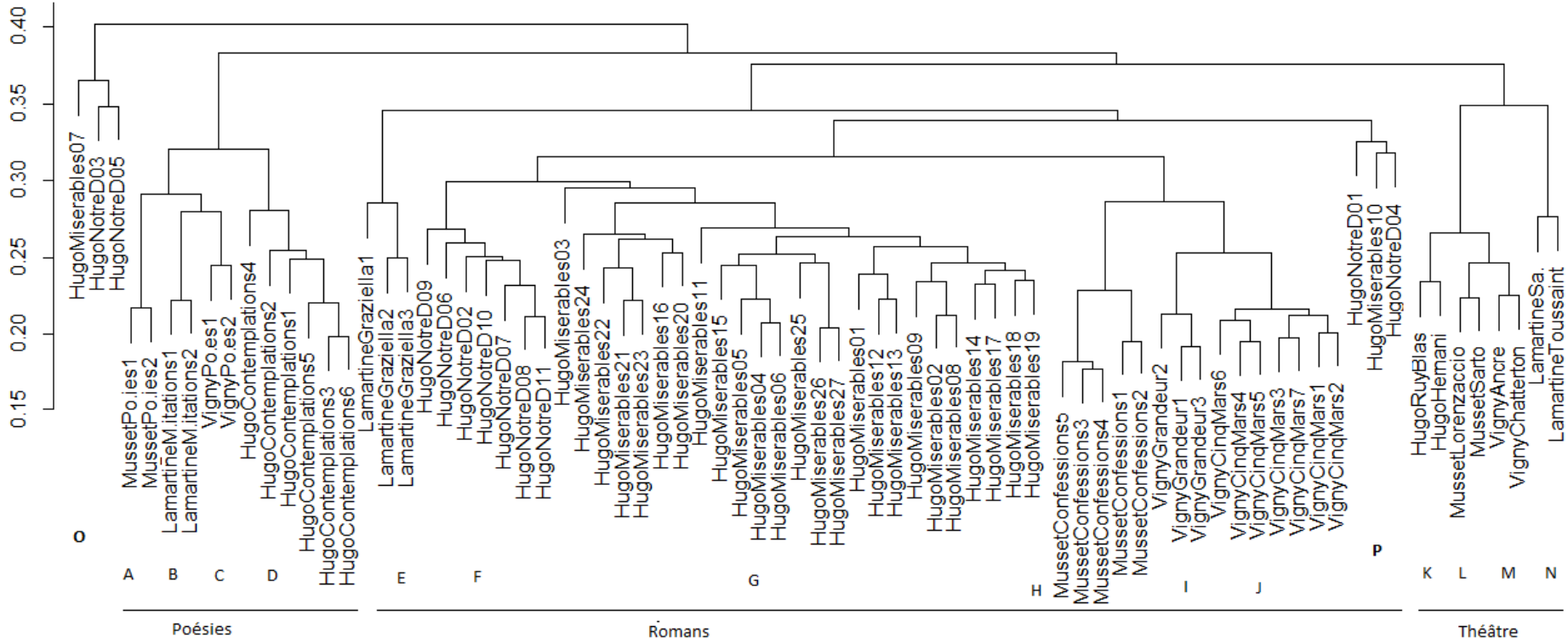
L'algorithme procède à la construction d'une classe en regroupant les deux textes séparés par la distance la plus faible - les troisième et quatrième parties de la *Confession* de Musset (0.176) -, puis il calcule - par la moyenne arithmétique - les distances de ce nouvel ensemble à tous les autres textes, etc. Et ceci jusqu'à la constitution d'un ensemble unique. Ces regroupements successifs sont représentés par un dendrogramme avec, en ordonnées, les distances correspondantes aux différents niveaux d'agrégation.

En coupant le graphe, horizontalement et au plus près de l'un des seuils mentionnés ci-dessus, on peut isoler les groupes de textes très proches, relativement proches, etc.

Il ne faut pas attacher trop d'importance au classement des textes de gauche à droite. Leur proximité est mesurée par la longueur du tracé qui les joint. Plus le trait horizontal est situé bas, plus le groupement est homogène. Ainsi la majorité des couples, formés au niveau le plus bas, se trouvent en dessous de 0.25 et concernent tous un même auteur, ce qui est conforme au tableau 3 ci-dessus.

La technique produit parfois des "effets de chaîne". Certaines proximités entre textes ne sont pas discernables car les sommets qui les relient sont effacés par des agrégations effectuées à un niveau inférieur. Les niveaux les plus élevés de l'arbre doivent donc être considérés avec prudence. L'appartenance de chacun des textes à une classe donnée doit éventuellement être contrôlée sur la matrice des distances.

Tableau 4. Dendrogramme de la classification hiérarchique sur le corpus des quatre auteurs



De gauche à droite :
Poésies **A** : *Poésies* de Musset ; **B** : *Méditations* de Lamartine ; **C** : *Poésies* de Vigny ; **D** : *Contemplations* de V. Hugo
Romans **E** : *Graziella* de Lamartine (totalité) ; **F** : *Notre-Dame de Paris* (7 livres) ; **G** : *Les Misérables* (25 livres) ; **H** : *Confession* de Musset (totalité) ; **I** : Servitude et Grandeur militaires de Vigny en totalité ; **J** : Cinq-Mars de Vigny en totalité ; **P** : deux livres de *Notre-Dame de Paris* et un des *Misérables*.
Théâtre : **K** : les deux pièces de Hugo ; **L** : les deux pièces de Musset ; **M** : les deux pièces de Vigny ; **N** : les deux pièces de Lamartine.
A part : **O** : Septième livre des *Misérables*, Troisième et cinquième livre de *Notre-Dame de Paris*.

Graphique réalisé avec R (hcluster).

Le dendrogramme confirme toutes les conclusions tirées de l'examen des voisinages. Pour Lamartine, Musset et Vigny, tous les textes sont correctement classés par genres et par œuvres. Pour Hugo en revanche, les groupes P et surtout O signalent des difficultés. Un livre des *Misérables* (6 du tome 2) est groupé – dans les romans – avec les livres 1 et 4 de *Notre-Dame de Paris*. Dans les *Misérables*, ce livre est consacré à la description du couvent du Petit-Picpus dans lequel J. Valjean s'est réfugié, ce qui donne à Hugo l'occasion de longues digressions sur la religion et le clergé dans la société française. C'est aussi le thème du livre 1 de *Notre-Dame de Paris* (une fête religieuse) et du livre 4 (présentation de C. Frollo, archidiacre de Notre-Dame). Le groupe O (à gauche) met en échec la classification. Dans les *Misérables*, le premier livre du deuxième tome est entièrement consacré à Waterloo. Les livres 3 et 5 de *Notre-Dame* sont aussi remplis de digressions historiques…

Autrement dit, loin d'être des "erreurs", le calcul de la distance et la classification ont mis en valeur des caractéristiques singulières dans l'œuvre de Hugo : la présence de longues digressions – historiques, philosophiques ou politiques - sans véritables liens avec le thème de l'ouvrage. La même remarque peut être faite à propos du théâtre de Lamartine qui est décalé à droite de la figure et fort éloigné des pièces des trois autres.

Sur le graphique, le dernier niveau d'agrégation concerne les genres. L'expérience permet d'apprécier assez précisément le poids de ce facteur.

V. Trois genres…

Pour mesurer l'importance de ce facteur, on utilise le raisonnement "toutes choses égales par ailleurs" en groupant les œuvres par genre et par auteur.

Le tableau 5 ci-dessous donne une synthèse de ces calculs. Les œuvres sont regroupées par genre (donc trois lignes et trois colonnes par auteur). Cette agrégation est rendue possible par les propriétés de la distance intertextuelle : les chiffres sont les moyennes arithmétiques des distances entre chacun des textes composant un groupe.

Ce tableau est une matrice carrée (autant de lignes que de colonnes) dont la diagonale est nulle (première propriété de la distance intertextuelle : la distance d'un objet à lui-même est nulle). Elle est également symétrique (deuxième propriété de la distance).

Les quatre carrés centraux (3x3) sont les distances entre genres, internes à l'œuvre de chaque auteur.

Les valeurs en gras sont les distances entre les textes d'auteurs différents appartenant à un même genre (poésie, roman, théâtre). Toutes ces valeurs s'inscrivent dans l'échelle présentée ci-dessus, sauf deux valeurs remarquables qui sont soulignées dans le tableau.

On tire de ce tableau, deux conclusions principales :

- pour deux auteurs différents, les œuvres de un même genre sont généralement plus proches que celles de chacun dans deux genres différents (carrés centraux). Par exemple, en première ligne, les œuvres poétiques de Hugo sont séparées de celles de Lamartine, de Musset et de Vigny, respectivement par des distances de 0.323, 0.324 et 0.317, c'est-à-dire inférieures à celles qui, au sein du corpus Hugo séparent sa poésie de ses romans (0.384) ou de son théâtre (0.391). Quasiment toutes les lignes et les colonnes vérifient cette propriété. Musset est une exception notable : chez lui, théâtre et roman sont très proches, comme si Musset était parvenu à "théâtraliser" sa *Confession* ou à "romancer" ses deux tragédies ;

- pour chaque couple d'auteurs, les distances les plus faibles sont toujours celles séparant des œuvres appartenant à un même genre. Une distance se situe même au seuil de 0,25 : le théâtre de Musset et celui de Vigny (cette proximité remarquable est analysée plus bas).

Tableau 5. Distances entre les œuvres groupées par genre pour chaque auteur (moyenne arithmétique des distances individuelles)

	Hugo			Lamartine			Musset			Vigny		
	Poésie	Roman	Théâtre	Poésie	Roman	Théâtre	Poésie	Roman	Théâtre	Poésie	Roman	Théâtre
Hugo Poésie	0,000	0,384	0,391	**0,323**	0,360	0,366	**0,324**	0,376	0,414	**0,317**	0,373	0,446
Hugo Roman	0,384	0,000	0,396	0,412	**0,354**	0,409	0,348	**0,343**	0,389	0,382	**0,315**	0,414
Hugo Théâtre	0,391	0,396	0,000	0,429	0,419	**0,328**	0,332	0,310	**0,261**	0,405	0,336	**0,274**
Lamartine Poésie	**0,323**	0,412	0,429	0,000	0,351	0,331	**0,331**	0,409	0,441	**0,280**	0,392	0,483
Lamartine Roman	0,360	**0,354**	0,419	0,351	0,000	0,383	0,343	**0,354**	0,417	0,334	**0,325**	0,442
Lamartine Théâtre	0,366	0,409	**0,328**	0,331	0,383	0,000	0,318	0,350	**0,350**	0,340	0,346	**0,373**
Musset Poésie	**0,324**	0,348	0,332	**0,331**	0,343	0,318	0,000	0,301	0,339	**0,293**	0,312	0,384
Musset Roman	0,376	**0,343**	0,310	0,409	**0,354**	0,350	0,301	0,000	0,274	0,378	**0,286**	0,301
Musset Théâtre	0,414	0,389	**0,261**	0,441	0,417	**0,350**	0,339	0,274	0,000	0,425	0,330	**0,248**
Vigny Poésie	**0,317**	0,382	0,405	**0,280**	0,334	0,340	**0,293**	0,378	0,425	0,000	0,345	0,456
Vigny Roman	0,373	**0,315**	0,336	0,392	**0,325**	0,346	0,312	**0,286**	0,330	0,345	0,000	0,340
Vigny Théâtre	0,446	0,414	**0,274**	0,483	0,442	**0,373**	0,384	0,301	**0,248**	0,456	0,340	0,000

Sans être une loi absolue, cette expérience confirme donc que le genre l'emporte sur l'auteur. **Il faut donc neutraliser le genre pour voir apparaître l'auteur**. Ce qui signifie qu'une attribution d'auteur par ordinateur ne peut porter que sur des textes d'un même genre.

VI. Auteurs, époque et thèmes

Pour connaître le poids respectif de l'auteur, du thème et de l'époque, le genre est neutralisé en découpant le corpus en trois sous-ensembles homogènes.

Les poésies des quatre auteurs

Le tableau 6 ci-dessous résume les résultats obtenus sur les poésies des quatre auteurs. Les distances internes à une œuvre sont en diagonale ("intra-auteur"), celles entre œuvres de différents auteurs sont sur les lignes et colonnes correspondantes ("inter-auteur").

Tableau 6. Distances entre les poésies des quatre auteurs

	Hugo		Lamartine		Musset		Vigny	
	Contempl.1 (1830-1843)	Contempl.2 (1843-1855)	Méditations1 (1815-1820)	Méditations2 (1820-1848)	Poésies1 (1829-35)	Poésies2 (1829-35)	Poèmes1 (1826)	Poèmes2 (1826)
Hugo								
Contemplations1	0,000	**0,253**	0,329	0,317	0,343	0,317	0,317	0,314
Contemplations2	**0,253**	0,000	0,325	0,322	0,330	0,307	0,325	0,313
Lamartine								
Méditations1	0,329	0,325	0,000	**0,222**	0,366	0,303	0,274	0,307
Méditations2	0,317	0,322	**0,222**	0,000	0,356	0,301	0,255	0,284
Musset								
Poésies1	0,343	0,330	0,366	0,356	0,000	**0,245**	0,338	0,288
Poésies2	0,317	0,307	0,303	0,301	**0,245**	0,000	0,290	0,255
Vigny								
Poèmes1	0,317	0,325	0,274	0,255	0,338	0,290	0,000	**0,245**
Poèmes2	0,314	0,313	0,307	0,284	0,288	0,255	**0,245**	0,000
Moyenne	**0,313**	0,311	0,304	0,294	**0,324**	0,288	0,292	0,287

Toutes les distances les plus faibles – et notamment celles inférieures à 0,25 - se trouvent sur la diagonale du tableau, c'est-à-dire entre divers poèmes d'un même auteur (intra-auteur). Toutes les distances sur ces diagonales sont systématiquement inférieures à celles entre auteurs différents sur les mêmes lignes ou colonnes (distances inter-auteurs). D'où la première conclusion : **Pour les poésies de ces quatre auteurs, le facteur auteur l'emporte sur la diversité des thèmes et des époques de composition des textes.**

Dans ce tableau, sont soulignées les deux valeurs inter-auteurs les plus courtes (0,255 qu'il faut arrondir à 0,26) qui indiquent une double influence probable. Si l'on s'en tient aux dates de composition et de publication, la première influence est celle des *Méditations* de Lamartine sur les premiers poèmes de Vigny. La seconde est celle des poèmes de Vigny sur ceux Musset.

La moyenne en dernière ligne résume les informations contenues dans les colonnes (et lignes) correspondantes et souligne le décalage de Hugo ainsi que des premières poèmes de Musset.

La même opération est opérée sur les pièces théâtre puis sur les romans.

Le théâtre

Le tableau 7 ci-dessous est construit de la même manière que le précédent et aboutit à la même conclusion : les distances sur la diagonale (cadre double) sont systématiquement inférieures à toutes les autres sur les mêmes lignes et colonnes et ceci même pour les deux pièces non contemporaines (cas de Lamartine). **Pour les pièces de théâtre, les quatre auteurs sont identifiables et, dans la distance entre textes, l'auteur l'emporte sur la diversité des thèmes et des époques de composition.**

Tableau 7. Distances intertextuelles séparant les 8 pièces de théâtre du corpus

	Hugo		Lamartine		Musset		Vigny	
	RuyBlas (1838)	Hernani (1830)	Saül (1818)	Toussaint (1850)	Lorenzaccio (1834)	Sarto (1833)	Ancre (1831)	Chatterton (1835)
RuyBlas	0,000	**0,236**	0,354	0,315	0,247	0,267	0,260	0,278
Hernani	**0,236**	0,000	0,330	0,314	0,251	0,277	0,273	0,287
Saül	0,354	0,330	0,000	**0,278**	0,357	0,371	0,383	0,389
Toussaint	0,315	0,314	**0,278**	0,000	0,316	0,355	0,358	0,363
Lorenzaccio	0,247	0,251	0,357	0,316	0,000	**0,225**	0,248	0,245
Sarto	0,267	0,277	0,371	0,355	**0,225**	0,000	0,258	0,244
Ancre	0,260	0,273	0,383	0,358	0,248	0,258	0,000	**0,227**
Chatterton	0,278	0,287	0,389	0,363	0,245	0,244	**0,227**	0,000
Moyenne	0,280	0,281	0,352	0,329	0,270	0,285	0,287	0,290

Trois distances remarquables sont soulignées (entre les pièces de Vigny et de Musset). Les deux premiers chiffres étant les seuls significatifs, les seules distances problématiques (0,24 et 0.25) séparent *Chatterton* (Vigny) d'*André Del Sarto* et de *Lorenzaccio* (Musset). Les pièces sont contemporaines et leur faible distance souligne la proximité des thèmes très romantiques, spécialement celui de l' "artiste maudit" : le poète Chatterton comme le peintre Del Sarto traversent une période d'impuissance créatrice, ils vivent un amour impossible et finissent par se suicider. D'après les dates de publication, l'influence serait celle de Musset sur Vigny…

Cet "accident", intéressant pour l'histoire littéraire, ne remet pas en cause la conclusion générale : les distances intra-auteurs sont toujours plus faibles que les distances inter-auteurs. Le tableau 8 ci-dessous permet de calculer cet écart.

Tableau 8. Distances intra-auteurs et inter-auteurs dans les 8 pièces de théâtre

	Intra-auteur	Inter-auteur
Hugo	0,236	0,288
Lamartine	0,278	0,350
Musset	0,225	0,286
Vigny	0,227	0,299
Moyenne	0,242	0,306

La dernière ligne indique que, pour le théâtre, les distances inter-auteurs sont d'environ 27% plus élevées en moyenne que celles internes aux œuvres d'un seul auteur.

On en déduit que, pour les pièces de théâtre, **la variable auteur l'emporte sur le thème et sur le temps** (du moins la durée d'une vie créatrice), et sur celui du sous-genre (vers et prose). En effet, les pièces de Hugo et de Lamartine sont en vers ; celles des deux autres en prose.

Les romans

Le tableau 9 ci-dessous récapitule les distances entre les romans des quatre auteurs. Dans le corpus, Lamartine et Musset n'ont qu'un roman mais ceux-ci sont découpés en plusieurs extraits. Pour Hugo à l'éloignement des thèmes s'ajoute un décalage de plus de 30 ans entre *Notre-Dame de Paris* et les *Misérables*.

Tableau 9. Distances intertextuelles entre les 6 romans.

	Hugo		Lamartine	Musset	Vigny	
	Notre-Dame (1831)	Misérables (1862)	Graziella (1852)	Confession (1836)	Servitude (1835)	Cinq Mars (1826)
Hugo Notre-Dame	0,000	**0,324**	0,353	0,363	0,351	0,324
Hugo Misérables	**0,324**	0,000	0,355	0,335	0,329	0,312
Lamartine Graziella	0,353	0,355	0,000	0,354	0,347	0,327
Musset Confession	0,363	0,335	0,354	0,000	0,267	0,288
Vigny Servitude	0,351	0,329	0,347	0,267	0,000	**0,270**
Vigny Cinq-Mars	0,324	0,312	0,327	0,288	**0,270**	0,000

La prédominance du facteur auteur (sur le thème et l'époque) serait vérifiée si les valeurs "intra" - cadres supérieur gauche et inférieur droit - étaient inférieures à toutes les autres valeurs. Deux valeurs sont divergentes par rapport à cette attente :

- les deux romans de Hugo sont très éloignés et dans un cas, plus proches du premier roman de Vigny qui passe pour l'invention du roman historique et qui a eu une influence considérable sur les autres écrivains de l'époque (spécialement A. Dumas). Mais ici les valeurs élevées (supérieures à 0.3) indiquent simplement une parenté. Ce cas montre donc que, **dans le genre romanesque, lorsque thèmes et temps ajoutent leurs effets ceux-ci peuvent contrebalancer le poids du facteur auteur**.

- le second roman de Vigny (*Servitude et grandeur militaires*) est aussi proche de la *Confession* de Musset que du premier roman de Vigny (*Cinq-Mars*) et cette proximité est remarquablement faible (0.27). D'une part, *Servitude et grandeur militaires* est à la limite entre les mémoires et le roman. Le livre ne comporte pas d'intrigue : c'est un recueil de récits indépendants les uns des autres. D'autre part, la *Confession* de Musset ne dissimule pas son caractère autobiographique. Enfin, les dates de publication des deux ouvrages ne laissent pas de doute quant au sens de cette influence.

Les passages concernés par cette influence peuvent être précisément localisés grâce à la technique de la "fenêtre glissante" (Labbé 2007). Pour ne pas alourdir, on se contente de donner la comparaison par extraits dans le tableau 10 ci-dessous.

Tableau 10. Distances intertextuelles entre les extraits de *Servitude et grandeur militaires* (Vigny) et *Confession d'un enfant du siècle* (Musset).

Vigny	Musset Confession1	Confession2	Confession3	Confession4
Servitude1	**0,239**	0,249	0,266	0,273
Servitude2	0,252	0,271	0,292	0,296
Servitude3	**0,236**	**0,245**	0,260	0,270
Moyenne	**0,242**	0,255	0,273	0,280

Les valeurs en gras indiquent que les proximités significatives sont localisées au début et que Musset a puisé son inspiration principalement dans le dernier tiers et au début du roman de Vigny, spécialement dans la sorte de "confession" en préface et dont est extraite la citation qui ouvre cette conférence. Mais ce démarquage s'atténue dès le deuxième quart et disparaît à la moitié de la *Confession* de Musset. Les hypothèses concurrentes (plagiat, collaboration à certains passages) peuvent donc être éliminées sans qu'il soit besoin de pousser l'analyse.

Les historiens de la littérature savent combien a été forte l'influence des romans puis du théâtre de Vigny sur ses cadets romantiques, notamment Musset ou Dumas. Nous ne faisons donc que confirmer cette influence, la mesurer et la localiser précisément.

VII. Autres indices

Pour une attribution d'auteur, il existe de nombreux indices auxiliaires qui permettent de valider et de préciser la paternité d'un texte ou d'une œuvre – ou à l'inverse de mettre en doute cette paternité lorsqu'on se trouve dans la zone où deux hypothèses sont envisageables (plume de l'ombre ou influence passagère). Outre les classifications évoquées au début de cette communication il s'agit de : sens des mots (Labbé & Labbé 2005, Labbé 2010), combinaison des mots les plus fréquents, longueur et structure de la phrase (Labbé & Labbé 2010).

Ce dernier point est particulièrement intéressant. En effet, la longueur et la construction des phrases est l'une des dimensions essentielles du style (Molinié 1986, 53-78). Les principales caractéristiques *théoriques* de la phrase françaises sont bien connues, notamment depuis les travaux de Le Goffic (1999) (Pour une actualisation : Charolles 2007). Mais, hormis quelques travaux pionniers - comme ceux de Richaudeau (1988), de Milly (1975, 1986) sur Proust ou de Garette (1995) sur Racine -, les études *empiriques* sur de vastes corpus manquent. Cela s'explique notamment par les problèmes auxquels se heurte l'étude des phrases par informatique.

Premièrement, la phrase est l'espace de texte compris entre deux ponctuations fortes. La longueur de la phrase est mesurée par le nombre de mots compris dans cet espace. Une ponctuation forte est l'un des signes suivants : '.' '…' '?' '!', quand ils sont suivis d'un mot dont l'initiale est en majuscule. Si un nom propre suit un point, l'opérateur doit se substituer à l'automate et trancher entre deux possibilités : début d'une nouvelle phrase ou simple abréviation (par exemple "M. Madeleine"). Le respect de ces conventions est indispensable comme l'a montré la controverse autour de la longueur des phrases chez Proust (Milly 1975 & 1986 : 165-167). Par exemple, dans les *Misérables* de V. Hugo, compter tous les points derrière « M. » comme des fins de phrases augmente faussement le nombre de ces phrases de 5% ! Les cas de ce genre sont innombrables et expliquent pourquoi la quasi-totalité des dépouillements informatiques sont inutilisables pour une étude précise de la phrase…

Deuxièmement, pour étudier la construction de la phrase, il faut d'abord identifier chacun de ses composants (notamment les groupes verbaux et nominaux) et donc pour cela étiqueter sans erreur tous les mots qui la composent…

Troisièmement, pour s'en tenir à la seule longueur – meilleur indice dans le cas d'une attribution d'auteur – il ne faut pas se contenter d'une mesure unique, par exemple la longueur moyenne, mais prendre en compte une série de valeurs significatives comme celles présentées dans le tableau 11 ci-dessous.

Le **mode** est la longueur la plus fréquente. Par exemple, dans le théâtre de Hugo, les phrases les plus fréquentes ne comportent qu'un seul mot (essentiellement, des adverbes et des interjections). A l'opposé, dans *Cinq-Mars* de Vigny, ce mode est de 17 mots, ce qui est considérable. Le mode de la poésie est compris entre 8 et 10 mots (nombre moyen de mots composant un vers de 12 pieds ce qui vérifie le poids du genre.

La **médiane** sépare en deux parts égales la population des phrases rangée par longueurs croissantes. Dans le théâtre de Hugo, la moitié des phrases comportent 4 mots et moins. A l'opposé, dans *Cinq-Mars*, cette moitié est atteinte avec 22 mots.

Contrairement à une opinion répandue, la **moyenne** n'est pas "au milieu", du moins lorsque la répartition du caractère (ici les mots) est effectuée de manière très inégalitaire entre les individus (ici les phrases). Dans ce cas, la moyenne est "tirée" vers le haut par quelques individus très riches (comme pour les revenus). C'est ce qui se passe dans la poésie et les romans de Vigny où figurent quelques phrases très longues. L'écart entre la médiane et la moyenne est un indice simple permettant de mesurer cette inégalité de répartition. Par exemple, cet indice distingue nettement Musset de Vigny.

Enfin, dernière valeur centrale : la **médiale** sépare le texte en deux parts égales (en fonction de la longueur des phrases). Dans le théâtre de Hugo, la moitié du texte est occupé par des phrases de moins de 12 mots – ou encore : la moitié du temps, le spectateur entend des phrases extrêmement courtes, dans une sorte d'accumulation haletante qui est assez caractéristique du spectacle romantique, dominé par les passions portées à leur paroxysme. A l'opposé, cette valeur est de 40 mots dans *Cinq-Mars* et excède même ce seuil dans les premiers poèmes de Vigny qui sont manifestement d'une nature singulière.

Enfin l'**écart-type** donne une mesure synthétique de la dispersion des longueurs autour de la moyenne. Cette dispersion est considérable dans les poésies de Hugo et de Vigny, minimale dans le théâtre de ces deux mêmes auteurs.

Tableau 11. Longueurs des phrases (en mots) dans les œuvres des trois auteurs classées par genre. Valeurs centrales et dispersion.

	Mode	Médiane	Moyenne	Médiale	Ecart-type
Hugo					
Contemplations (1830-1843)	8	15,1	23,6	35,3	25,5
Contemplations (1843-1855)	10	13,8	22,6	35,9	23,6
Lamartine					
Méditations (1815-1820)	10	19,3	25,2	34,0	21,5
Méditations (1820-1848)	9	18,3	24,5	34,8	20,7
Musset					
Poésies (1829-1835	10	14,3	17,9	24,1	14,7
Poésies (1829-1835)	8	15,0	19,6	27,9	16,1
Vigny					
Poèmes (1826)	9	18,6	28,2	42,3	26,3
Poèmes (1826)	9	18,2	26,7	36,3	23,0
Hugo					
Hernani (1830)	1	4,7	7,8	11,1	9,1
Ruy Blas (1838)	1	4,6	8,0	11,6	9,6
Lamartine					
Saul (1818)	9	9,4	14.9	20,7	14,2
Toussaint Louverture (1850)	4	9,4	14,6	20,5	15,9
Musset					
André del Sarto (1833)	6	7,2	10,3	13,7	9,3
Lorenzaccio (1834)	7	10,5	14,3	18,7	12,0
Vigny					
Ancre (1831)	4	7,3	10,0	13,1	8,5
Chatterton (1835)	6	8,5	11,7	15,0	9,7
Hugo					
Notre-Dame de Paris (1831)	6	12,0	18,5	26,0	18,1
Misérables (1862)	6	11,4	17,0	23,7	17,5
Lamartine					
Graziella (1852)	9	19,3	25,1	31,8	22,6
Musset					
Confession (1836)	16	20,0	24,3	29,7	18,0
Vigny					
Cinq-Mars (1826)	17	21,1	28,4	39,7	24,1
Servitude (1835)	12	20,5	25,8	33,5	20,1

Le tableau souligne d'abord la séparation des genres : la phrase théâtrale est nettement plus courte et moins variable que celles des deux autres genres. Le théâtre repose essentiellement sur le dialogue et, à l'oral, la phrase est brève. A l'opposé, la poésie, surtout quand elle explore l'intériorité, utilise des phrases amples et complexes.

Au sein de ces trois cadres chaque auteur se singularise nettement sauf pour le théâtre de Vigny et de Musset. Dans tous les autres cas, même quand l'une des valeurs centrales est commune, les autres sont distinctes de même que leur dispersion.

Chez Hugo, les mêmes valeurs se retrouvent au cours du temps. Il en est de même pour Lamartine. Ceci est d'autant plus remarquable qu'il s'écoule 32 ans entre *Notre-Dame de Paris* et les

Misérables – avec une très légère tendance à la baisse (4 valeurs sur 5) entre ces deux romans qui s'explique par une proportion un peu plus importante des dialogues dans les *Misérables*. Pour la poésie, l'écriture des *Contemplations* (Hugo) s'étend sur 25 ans ou celle des *Méditations* (Lamartine) sur 33 ans.

A l'inverse, sur des périodes pourtant beaucoup plus courtes, on observe :

- une nette tendance à l'alourdissement de la phrase chez Musset, tant dans ses poésies que dans son théâtre. Cela suggère un repli sur soi et un goût pour l'introspection (dont Proust serait le cas le plus classique) ;

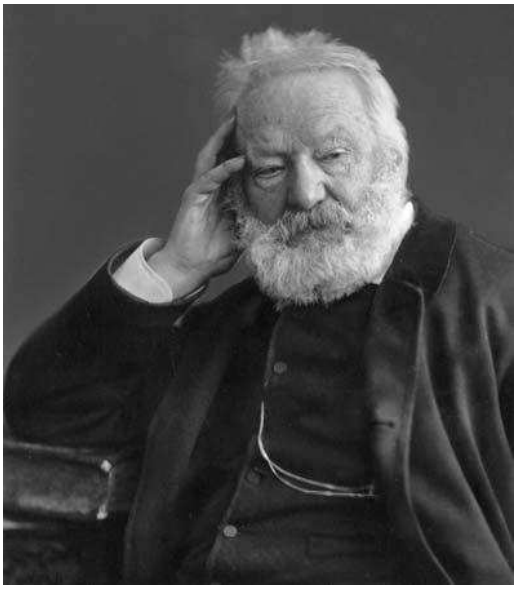
- une tendance au resserrement de la phrase chez Vigny, spécialement dans ses romans, comme si ses succès théâtraux l'amenaient à simplifier son expression et à privilégier l'action et le dialogue sur les descriptions, les portraits et la psychologie des personnages.

Pour ce qui concerne l'attribution d'auteur, on déduit de ces deux derniers cas que **les caractéristiques de la phrase – et plus largement les indices stylistiques - ne peuvent être utilisés que lorsqu'on s'est auparavant assuré que l'auteur supposé fait preuve d'une certaine stabilité stylistique**, ce qui n'est pas toujours le cas.

Conclusions

Premièrement, Hugo, Lamartine, Musset (et Vigny) sont des auteurs bien différents et parfaitement identifiable, non pas à la seule lecture – même érudite - mais à l'aide des outils que les mathématiques appliquées apportent aux sciences humaines.

Nous pouvons donc restituer un visage aux silhouettes placées au début de cet exposé :



Victor Hugo (1802-1885)

Alfred de Musset (1810-1857)

Alphonse de Lamartine (1790-1869)

Alfred de Vigny (1797-1863)

Qu'il se soit trouvé des chercheurs pour prétendre le contraire et d'autres pour les croire montre simplement que des idées fausses peuvent s'emparer de certains milieux intellectuels et comment elles les dupent, selon les termes de Vigny. En effet, l'expérience évoquée au début de ce papier enfonce une porte ouverte. Une langue ne comporte pas simplement une phonétique, une syntaxe et un lexique. Elle fournit aussi à ses usagers des "genres", c'est-à-dire des règles pour échanger les signes linguistiques en fonction des situations de communication et du « médium » choisi pour cette communication. En français, il existe une série de genres : la fiction romanesque, la poésie, le théâtre, la correspondance... Les principales caractéristiques de certains genres sont maintenant connues : la poésie (Labbé & Labbé 2009), la correspondance (Labbé & Labbé 2013a), le discours politique (Labbé & Labbé 2012). Ces codes sont assez lâches et durent tant qu'ils sont adaptés à la société et aux nécessités de la communication mais un auteur ne peut s'en échapper.

Dans un genre donné, chaque auteur est identifiable par un certain vocabulaire mais aussi par un style et des thèmes particuliers. On en conclut que :

- pour reconnaître l'auteur d'un texte dont l'auteur est douteux ou inconnu, il faut le confronter à d'autres textes – dont les auteurs sont certains – écrits dans le même genre ;

- la distance intertextuelle est l'outil essentiel pour une attribution d'auteur par ordinateur. Dans les limites fixées, l'échelle de la distance est un outil efficace. Elle est complétée par des classifications – non supervisées – l'étude de la combinaison et du sens des mots ainsi que du style.

Une dernière question vient logiquement : si l'on a identifié l'auteur, qu'apprend-on sur lui ? Par exemple, la méthode a permis de repérer les passages singuliers des romans de Hugo et de rapprocher ces passages entre deux oeuvres. On a vu également que la distance intertextuelle révèle des proximités (ou des éloignements) et en donne une mesure précise.

C'est pourquoi Vigny a été introduit. Sans lui, l'expérience n'avait guère d'intérêt tant Hugo, Lamartine et Musset sont différents les uns des autres. En revanche, il existe des proximités intéressantes entre Musset et Vigny.

Les cas d'influence d'un auteur sur un autre sont nombreux dans l'histoire littéraire. La première tragédie de Racine (*la Thébaïde*) est "cornélienne" ; le début d'*Une vie* (premier roman de Maupassant) est singulièrement proche de *Madame Bovary* (Flaubert). Certains passages des *Trois Mousquetaires* (Dumas) signalent l'influence de *Cinq Mars* de Vigny, etc. Ces cas n'enlèvent rien à la singularité de chacun des auteurs mais ils sont intéressants pour l'histoire littéraire, comme le sont les marques d'influences entre Vigny et Musset.

En revanche, quand il y a des proximités systématiques et s'étendant sur toute la durée de la création d'une œuvre, on peut rejeter les influences passagères – ou un démarquage localisé - et conclure qu'un des deux auteurs a été la plume de l'ombre de l'autre. C'est le cas, notamment, pour P. Corneille qui a écrit les principales œuvres présentées sous le nom de Molière.

Naturellement, pour un panorama d'ensemble de cette époque – dite « romantique » - il aurait fallu introduire dans l'expérience les précurseurs (Chateaubriand, Constant, Mme de Staël…) et les contemporains comme Balzac, Dumas ou Gautier. Les outils sont là. Nous espérons avoir suggéré combien ils pourraient être utiles pour l'étude de notre histoire littéraire.

Remarque terminale

Les logiciels – développés depuis 40 ans et connus sous le nom générique de "lexicométrie" – sont une œuvre collective. Cyril Labbé et moi-même avons été les animateurs d'un réseau qui a compris notamment : Edward Arnold, Guy Bensimon, Jean-Guy Bergeron, Mathieu Brugidou, Pierre Hubert, Nelly & Jean Leselbaum, Thomas Merriam, Denis Monière, Jacques Picard, André Pibarot, Jacques Savoy… Nous avons également bénéficié du soutien de nos laboratoires respectifs : PACTE-CNRS et Laboratoire d'Informatique de Grenoble (LIG-IMAG).

Références

Editions des œuvres étudiées dans cet article.

Hugo Victor :
Les Contemplations. Paris : M. Levy frères, 1856.
Les Misérables (5 tomes). Paris : Émile Testard, 1890.
Notre-Dame de Paris. Paris : Ollendorff, 1904.
Hernani. Paris : Hetzel, 1889.
Ruy Blas. Paris : L. Conquest, 1889.
Lamartine Alphonse de :
Méditations poétiques - Paris : Firmin Didot – 1849.
Graziella. Paris : Librairie Nouvelle, 1852.
Saül. Paris : Chez l'auteur, 1860.
Toussaint Louverture. Paris : Michel Lévy Frères, 1857
Musset Alfred de. *Oeuvres complètes*. Paris : Charpentier et Fasquelle, 1888.
Vigny Alfred de. *Oeuvres complètes*. Paris : Larousse, 1913.

Les travaux publiés par notre réseau de recherche sont consultables en ligne (notamment sur le site hal.archives-ouvertes.fr). Une liste plus complète peut être consultée sur la page personnelle de D. Labbé.

Arnold Edward (2008). Le sens des mots chez Tony Blair (people et Europe). In Heiden Serge et Pincemin Bénédicte (Eds). *9e Journées internationales d'analyse statistique des données textuelles* (Lyon, 12-14 mars 2008). Lyon : Presses universitaires de Lyon, 2008, volume 1, p 109-119.

Barthes Roland (1968). La mort de l'auteur. *Œuvres complètes*. II. Paris : Seuil, 1994, p 491-495.

Brunet Etienne (1988). Une mesure de la distance intertextuel1e : la connexion lexicale. *Informatique et Statistique dans les Sciences humaines*. XXIV, 1 à 4, p 81-84.

Brunet Etienne & Muller Charles (1988). La statistique résout-elle les problèmes d'attribution ? *Strumenti critici,* III, n°3, p. 367-387.

Charolles Michel, Fournier Nathalie, Fuchs Catherine, Lefeuvre Florence (2007). *Parcours de la phrase: Mélanges offerts à Pierre Le Goffic*. Paris : Ed. Ophrys.

Chepiga Valentina (2009). Méthodologies croisées pour l'attribution des textes la place de la génétique. Les cas Gary/Ajar. *Modeles Linguistiques*. Vol. 30, N°. 1, p. 101-132.

Foucault Michel (1969). Qu'est-ce qu'un auteur ? *Bulletin de la Société française de philosophie* – 63-3, juillet-septembre 1969, p. 73-104.

Garette Robert (1995). *La phrase de Racine. Etude stylistique et stylométrique*. Toulouse : Presses universitaires du Mirail.

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 2001, 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*. 2005, 39, p. 335-351.

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.

Dominique Cyril & Labbé (2011b). Baudelaire, Rimbaud et Verlaine. In Banks David (Ed). *Aspects linguistiques du texte poétique*. Paris : l'Harmattan, p. 17-45.

Labbé, Cyril & Labbé, Dominique (2011c) : "Existe-t-il un langage propre à la politique ?" Communication aux XIIe Journées de l'ERLA, Brest, 18-19 novembre 2011.

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

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 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 (2004). *Romain Gary et Emile Ajar*. Grenoble : Cerat-IEP, mai 2004.

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

Labbé Dominique (2010). Le calcul du sens des mots. La lexicologie assistée par ordinateur. Communication au séminaire *Mathématiques et société*. Neuchâtel, 3 novembre 2010.

Labbé Dominique (2014). *Les plumes de l'ombre. Molière a-t-il écrit ses pièces* ? Université Inter-âges, Grenoble.

Lafon Michel & Peeters Benoît (2006). *Nous est un autre*. Paris, Flammarion.

Le Goffic Pierre (1999). *Grammaire de la phrase française*. Paris : Hachette.

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

Milly Jean (1975). *La Phrase de Proust*. Paris : Larousse (Réédition Paris : Champion, 1983).

Milly Jean (1986). *La longueur des phrases dans "Combray"*. Paris-Genève : Champion-Slatkine.

Molinié Georges (1986). *Eléments de stylistique française*. Paris : PUF.

Monière D. & Labbé D. (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.

Pibarot André, Picard Jacques & Labbé Dominique (1998). Les syntagmes répétés dans l'analyse des commentaires libres. In Mellet Sylvie (ed). *4e Journées d'analyse des données textuelles*. Nice, 1998, p. 507-516.

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

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.

Sur le romantisme en littérature.

Gautier Théophile. *Histoire du romantisme*. Paris : G. Charpentier et Cie, 1874.

Millet Claude. *Le romantisme : du bouleversement des lettres dans la France postrévolutionnaire*. Paris : Librairie générale française, 2007.

Richard Jean-Pierre (1999). *Études sur le romantisme*. Paris : Seuil.

Vaillant Alain (2005). *La crise de la littérature : romantisme et modernité*. Grenoble : ELLUG.

Vaillant Alain (dir) (2012). *Le romantisme : dictionnaire*. Paris : CNRS éditions.

Annexe

Bibliothèque électronique du français moderne (1er mars 2014)

	Longueur (mots)	Vocabulaire
Discours politique	**11 646 202**	**42 885**
Littérature (XVIIe –XXe siècles)	**11 534 542**	**61 211**
Romans et nouvelles	7 322 531	49 365
Théâtre	2 891 292	15 551
Poésie	975 187	18 810
Correspondance	345 542	11 070
Romans policiers	**548 682**	**17 274**
Presse	**2 939 632**	**58 690**
Sciences	**774 514**	**18 523**
Français oral	**2 978 122**	**18 429**
Total	**29 674 341**	**99 921**