La classification des textes

Cyril Labbé, Dominique Labbé

La classification des textes

Comment trouver le meilleur classement possible au sein d'une collection de textes ?

Le 28 mars 2011, par **Cyril Labbé** et **Dominique Labbé**

Le texte principal de cet article ne demande aucune connaissance en mathématique et statistique. Quelques dépliants contiennent des formules qu'un lecteur pressé peut sauter sans inconvénient.

> *IL n'est point de secret que le temps ne révèle.*
>
> Jean Racine, *Britannicus* (1669), IV, 4.

1. INTRODUCTION

Comment identifier l'auteur d'un texte d'origine douteuse ou inconnue ? Les anglo-saxons se passionnent pour cette question connue sous le nom de « Authorship attribution » [1]. Depuis la première étude que le statisticien américain Mendehall a consacrée en 1887 à la longueur des mots chez Shakespeare, Bacon et Marlowe [2], les statistiques appliquées tiennent une place importante [3].

Nous proposons ici de considérer la recherche de l'auteur d'un texte comme un cas particulier d'une question plus générale : **Comment trouver le meilleur classement possible au sein d'une vaste collection de textes écrits dans une même langue ?**

Pour répondre à cette question, deux outils sont nécessaires :

- un calcul de « distance » entre les textes afin de mesurer précisément la plus ou moins grande proximité (similarité) de chacun des textes par rapport à tous les autres ;
- des procédures de classification qui, à l'aide des distances, repèrent les « meilleurs groupements possibles » au sein de cette population.

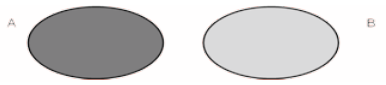
Cet article est consacré à la distance. La seconde question fera l'objet d'un article ultérieur.

2. LA DISTANCE INTERTEXTUELLE

La distance entre deux textes (« intertextuelle ») est mesurée comme on mesure la distance séparant deux objets dans l'espace [4]. L'unité de mesure n'est pas le mètre mais le « mot ».

2.1 Le calcul

Représentons deux textes A et B comme deux groupes de mots :



On superpose ces deux groupes et on compte le nombre de mots différents (zones grisées dans le schéma ci-dessous).

Par exemple, deux « textes » :

- A La secrétaire lui dit : « je suis le président ».
- B « Je suis le président », lui dit le président.

Ils semblent identiques à un mot près (secrétaire/président), à condition de considérer « la » et « le » comme un même mot, de négliger les majuscules initiales (*La* et *Je*), la ponctuation et l'ordre des mots. Cela revient à considérer que B n'est pas différent de : 'Le président lui dit : « je suis le président »' [5].

Cependant, la première phrase est ambiguë. On peut comprendre que la secrétaire *accompagne* le président ou qu'elle *assure* la présidence. En effet, l'usage en France admet que certaines fonctions soient désignées par le masculin, même lorsque le titulaire est une femme (*Madame le ministre*, *Madame le président*)…

Ces ambiguïtés sont nommées « homographies » : une même graphie mais plusieurs sens différents. En général, le contexte permet de lever ces ambiguïtés, en suivant des conventions strictes.

(cliquer pour déplier)

Les normes de dépouillement

Dans les deux textes donnés en exemple et en suivant l'ordre alphabétique, l'analyse doit résoudre quelques difficultés.

En premier lieu, deux mots sont écrits avec des majuscules initiales (La et Je) malgré le fait que ce ne sont pas des noms propres (ou parties de noms propres comme dans La Fontaine, La Bruyère…) il faut donc y reconnaître le pronom et l'article et considérer les deux je comme équivalents.

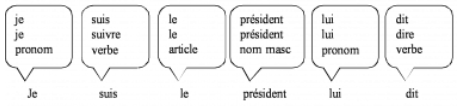
Deuxièmement, en se reportant aux entrées d'un dictionnaire de langue (meilleure représentation du lexique français) :

- *dit : verbe dire (troisième personne du singulier de l'indicatif présent ou du passé simple ou participe passé masculin singulier) ; adjectif masculin singulier (dit) ; nom masculin singulier (dit) ;*
- *la : article défini féminin singulier (le) ; pronom personnel objet ou attribut de la troisième personne féminin singulier (le) ;*

- *le : article défini masculin singulier ; pronom personnel objet ou attribut de la troisième personne masculin singulier ;*
- *lui : verbe luire au participe passé masculin singulier ; pronom personnel de la troisième personne du singulier (des deux genres quand il est employé comme complément) ;*
- *président : verbe présider à l'indicatif présent troisième personne du pluriel ; nom masculin singulier.*
- *secrétaire : nom féminin singulier ; nom masculin singulier ;*
- *suis : verbe être à l'indicatif présent première personne du singulier ; verbe suivre à l'indicatif présent première ou seconde personnes du singulier.*

Cette nomenclature fait consensus parmi les usagers de la langue. Par exemple, en français, toutes les flexions d'un verbe (modes, temps, personnes) sont regroupées sous l'infinitif de ce verbe ; de même les substantifs sont identifiés par leur genre (président/présidente). Ainsi, « un secrétaire » peut être un meuble…

L'informatique permet de résoudre ces difficultés [6]. L'opération consiste à ajouter une étiquette à chacun des mots du texte (figure ci-dessous).



L'étiquette vient s'ajouter au texte sans l'altérer. Elle comporte trois informations :

- *la **graphie standard** : majuscule initiale des mots communs ramenée en minuscule (comme pour « Je »), réduction des formes multiples à une graphie standard (puis et peux, événement et évènement…), correction des fautes d'orthographe…*
- *puis le **vocable**, c'est-à-dire l'entrée où se trouve la graphie dans le dictionnaire et la **catégorie grammaticale**, telle qu'elle figure en seconde position dans cette entrée de dictionnaire.*

*Un l'automate est chargé de standardiser les graphies et d'attacher une étiquette à chaque mot. Il utilise une nomenclature **systématique** (par exemple, si les substantifs se distinguent par le genre, alors tous les substantifs doivent se voir affecter le masculin ou le féminin), **exhaustive** (tous les mots y trouvent leur place), **univoque** (une seule entrée par mot), sans double compte ni catégorie ad hoc, ou fourre-tout, etc. Enfin l'opération est réversible : on peut retrouver le texte original, sans altération, à partir du fichier étiqueté.*

Tous les résultats présentés dans cet article sont obtenus avec des textes dépouillés en suivant strictement ces normes et conventions.

Le calcul de la distance entre A et B

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 : il y a 7 vocables différents dans A et autant dans B. $V_{(A,B)}$ 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. Dans l'exemple, les effectifs sont tous de 1 sauf pour l'article « le » (employé 2 fois dans A et autant dans B), pour « président » (employé une fois dans A et deux fois dans B) et « secrétaire » (absent de B) ;*
- *$\left|F_{iA} - F_{iB}\right|$ 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. Dans l'exemple ci-dessus, cette différence absolue est de 1 pour « président » et « secrétaire ».*
- *$D_{(A,B)}$: la distance entre A et B. Cette distance est le nombre de mots différents dans A par rapport à B (ou réciproquement). Pour la calculer cette distance, on utilise le tableau suivant :*

i	*Vocabulaire de A et B*	F_{iA}	F_{iB}	$\left\|F_{iA} - F_{iB}\right\|$
1	*dire (verbe)*	*1*	*1*	*0*
2	*être (verbe)*	*0*	*1*	*1*
3	*je (pronom)*	*1*	*1*	*0*
4	*le (article)*	*2*	*2*	*0*
5	*lui (pronom)*	*1*	*1*	*0*
6	*président (nom masculin)*	*1*	*2*	*1*
7	*secrétaire (nom féminin)*	*1*	*0*	*1*
8	*suivre (verbe)*	*1*	*0*	*1*
	Total	*8*	*8*	*4*

Les 8 vocables constituant le vocabulaire employé dans A et B sont rangés par ordre alphabétique (colonnes 1 et 2). Dans la troisième colonne : l'effectif du vocable d'indice i dans A ; dans la quatrième colonne, son effectif dans B, et, en dernière colonne, la différence absolue de ces 2 effectifs. La dernière ligne donne les résultats. La longueur de A (N_A), comme de B (N_B) est de 8 mots. La distance absolue entre A et B est égale à 4 mots.

Ces opérations sont résumées par formule (1).

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

Et, comme il y a 16 mots dans A et B, la distance relative est égale à 1/4 :

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

$D_{(A,B)}$ 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),*
- *l'inégalité triangulaire (le chemin direct entre A et B est toujours plus court qu'en passant par un point C non situé sur le segment AB).*

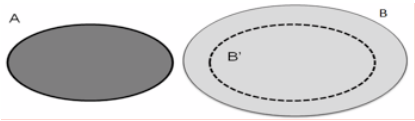
Ces propriétés ont d'importantes conséquences. Par exemple, on peut réaliser une représentation graphique de toutes les distances au sein d'une vaste population de textes, comme on dresse la carte d'une ville ou d'un quartier…

En admettant que « suis » est le verbe « suivre », dans A, et le verbe « être » dans B, **La distance absolue entre A et B est de 4 mots sur 16, soit 0.25**

2.2 La distance entre deux textes de longueurs inégales

Dans l'exemple ci-dessus, les deux textes ont la même longueur (le même nombre de mots). Comment procéder quand les deux textes ont des longueurs différentes, tout en conservant les propriétés de la distance ?

Supposons que B soit plus long que A ($N_A < N_B$). Il est proposé d'estimer leur distance en réduisant B à la longueur de A et en superposant cette réduction B' sur le texte A.



Autrement dit : si B avait la même longueur que A, combien ces deux textes compteraient-ils de mots différents ?

Le calcul pour deux textes de longueurs inégales

Nommons :

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

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

- $E_{iA(u)}$ *: l'effectif théorique, dans un texte 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 :*

$$(3)\ E_{iA(u)} = F_{iB} * U$$

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 :

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

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' :

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

Aux arrondis près, N_B' est égal à N_A. La formule (2) devient :

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

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

Premièrement, les effectifs dans A sont des entiers naturels alors que les effectifs théoriques dans B' sont des rationnels approchant des entiers naturels (inconnus). Autrement dit, le résultat de la soustraction - au numérateur de (4) et de (5) - comportera des décimales sans signification mais qui entreront pourtant dans 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.

Pour limiter cet effet, il est proposé d'éliminer du calcul :

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

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

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

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

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

Deuxièmement, le résultat de (5) est une estimation à cause des postulats qui fondent le calcul de l'effectif théorique d'un vocable dans B' (formule 3bis). Cette formule suppose que :

- *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 [7] ;*
- *l'apparition des vocables nouveaux se fait toujours au même rythme. En fait, 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.*

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. Une série d'expériences – dont certaines sont évoquées plus loin dans cet article - indiquent que :

- *les deux textes doivent avoir plus de 1 000 mots, et que, en-dessous de 3000 mots le résultat de (5) peut être instable [8],*
- *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% (avec des textes de longueurs supérieures à 5 000 mots et avec U < 2) et ±5% (lorsque U atteint 1 : 5).*

Les textes utilisés dans la suite de cet article comptent pour la plupart au moins 10 000 mots et leurs longueurs sont proches. Les résultats sont donc présentés avec quatre décimales. Du fait de l'incertitude introduite par l'estimation de la distance, la dernière décimale indique dans quel sens arrondir la troisième décimale qui est la dernière significative.

Cette distance intertextuelle se révèle un bon outil pour reconnaître l'auteur d'un texte douteux ou d'origine inconnue.

3. COMMENT IDENTIFIER L'AUTEUR ?

Depuis 1980, nous avons constitué de grandes collections de textes – les corpus - dépouillés selon les normes exposés ci-dessous.

Un exemple : le corpus du théâtre français au XVIIe

Tous les textes ont été dépouillés en suivant strictement les mêmes normes [9]

1. Les pièces de Pierre Corneille

N°	Titre	Création	Genre	Longueur en mots
01	Mélite	1630 ?	Comédie	16 690
02	Clitandre	1631	Tragi-comédie	14 402
03	La Veuve	1631	Comédie	17 661
04	La Galerie du Palais	1632	Comédie	16 140
05	La Suivante	1633	Comédie	15 160
06	Comédie des Tuileries	1634	Comédie	3 627
07	Médée	1635	Tragédie	14 269
08	La Place Royale	1634	Comédie	13 801
09	L'illusion comique	1636	Comédie	15 428
10	Le Cid	1636	Tragi-comédie	16 677
11	Cinna	1639	Tragédie	16 126
12	Horace	1640	Tragédie	16 482
13	Polyeucte	1641	Tragédie	16 472
14	Pompée	1642	Tragédie	16 492
15	Le Menteur 1	1642	Comédie	16 653
16	Suite du Menteur	1643	Comédie	17 675
17	Rodogune	1644	Tragédie	16 842
18	Théodore	1645	Tragédie	17 121
19	Héraclius	1647	Tragédie	17 433
20	Andromède	1650	Tragédie	15 514
21	Don Sanche	1650	Comédie héroïque	16 947
22	Nicomède	1651	Tragédie	16 923
23	Pertharite	1651	Tragédie	17 121
24	Œdipe	1659	Tragédie	18 618
25	Toison d'Or	1661	Tragédie	20 343
26	Sertorius	1662	Tragédie	17 675
27	Sophonisbe	1663	Tragédie	16 858
28	Othon	1664	Tragédie	16 971
29	Agésilas	1666	Tragédie	18 227
30	Atilla	1667	Tragédie	16 788
31	Tite et Bérénice	1670	Comédie héroïque	16 697
32	Pulchérie	1672	Tragédie	16 630
33	Suréna	1674	Tragédie	16 545
34	Psyché Corneille	1671	Comédie en vers	10 067
35	Psyché Molière	1671	Comédie en vers	4 816
36	Psyché Quinault	1671	Comédie en vers	1 399

Sources : Charles Marty-Laveaux. Œuvres complètes de P. Corneille. Paris : Hachette 1862. Collection Les Grands écrivains de la France. Muller Charles. Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille. Paris : Larousse, 1967 (réédition : Genève-Paris, Slatkine-Champion, 1979).

2. Les pièces présentées par Molière

N°	Titre	Création	Genre	Longueur en mots
37	La jalousie	Avant 1659	Comédie prose	3 501
38	Médecin volant	Avant 1659	Comédie prose	3 876
39	L'étourdi*	1659	Comédie vers	18 671
40	Dépit amoureux*	1659	Comédie vers	16 242
41	Précieuses ridicules	1660	Comédie prose	6 648
42	Sganarelle*	1660	Comédie vers	6 042
43	Dom Garcie*	1661	Comédie héroïque vers	17 049
44	L'école des maris*	1661	Comédie vers	10 536
45	Les fâcheux*	1661	Comédie vers	7 922
46	L'école des femmes*	1662	Comédie vers	16 625
47	Critique de l'école	1663	Comédie prose	8 610
48	L'impromptu	1663	Comédie prose	7 168
49	Mariage forcé	1664	Comédie prose	6 058
50	Princesse d'Elide*	1664	Comédie vers & prose	11 333
51	Le Tartuffe*	1664	Comédie vers	18 271
52	Dom Juan*	1665	Comédie prose	17 452
53	L'amour médecin	1665	Comédie prose	6 147
54	Le Misanthrope*	1666	Comédie vers	17 180
55	Médecin malgré lui	1666	Comédie prose	9 317
56	Mélicerte*	1666	Comédie vers	5 540
**	Comédie pastorale	1667	Comédie vers libres	732
57	Le sicilien	1667	Comédie prose	5 375
58	Amphytrion*	1668	Comédie vers libres	15 117
59	Georges Dandin	1668	Comédie prose	11 009
60	L'avare*	1668	Comédie prose	21 033
61	M. de Pourceaugnac	1669	Comédie prose	11 803
62	Amants magnifiques*	1670	Comédie vers & prose	11 983
63	Bourgeois gentilhomme*	1670	Comédie prose	17 132
64	Fourberies de Scapin	1671	Comédie prose	14 245
65	Escarbagnas	1671	Comédie prose	5 564
66	Femmes savantes*	1672	Comédie vers	16 863
67	Malade imaginaire*	1673	Comédie prose	19 919

** Pièce écrite, en tout ou partie, par P. Corneille ** Pièce retirée des expériences à cause de sa petite taille.*

Sources : Eugène Despois. Œuvres complètes de Molière. Paris : Hachette, 1876. Collection Les Grands écrivains de la France. Kylander Britt-Mary. Le vocabulaire de Molière. Goteborg : Acta Universitatis, Gothoburgensis, 1995

3. Les pièces de Jean Racine.

N°	Titre	Création	Genre*	Longueur en mots
01	La Thébaïde	1664	Tragédie	13 813
02	Alexandre	1665	Tragédie	13 864

03	Andromaque	1667	Tragédie	15 076
04	Les Plaideurs	1668	Comédie	8 041
05	Britannicus	1669	Tragédie	15 387
06	Bérénice	1670	Tragédie	13 242
07	Bajazet	1672	Tragédie	15 297
08	Mithridate	1673	Tragédie	15 091
09	Iphigénie	1674	Tragédie	15 782
10	Phèdre	1677	Tragédie	14 394
11	Esther	1689	Tragédie	11 147
12	Athalie	1691	Tragédie	15 492

** Toutes les pièces de J. Racine sont en alexandrins.*

Source : Paul Mesnard. Œuvres de J. Racine. Paris : Hachette, 1885 (Les Grands écrivains de la France). Bernet Charles. Le vocabulaire des tragédies de Racine (Analyse statistique). Genève-Paris : Slatkine-Champion, 1983.

4. Les pièces de J. Mairet

N°	Titre	Création	Genre	Longueur en mots
01	Sylvie	1621	Tragi-comédie (alexandrins)	19 813
02	Sophonisbe	1634	Tragédie (alexandrins)	16 166

Sources : La Sylvie. Texte établi par Jules Masan. Paris : Société nouvelle d'édition, 1905. Sophonisbe. Texte établi par Charles Dédéyan. Paris : Nizet, 1969.

5. Les pièces de P. Quinault

N°	Titre	Création	Genre*	Longueur en mots
01	La mère coquette	1665	Comédie (vers libre)	16 130
02	Thésée	1675	Tragédie (vers libres)	8 284
03	Atis	1676	Opéra (vers libres)	7 959

Sources : La mère coquette. Edition d'Etienne Gros. Paris : Champion, 1926. Livrets d'opéra. Edition de Norman Buford. Paris : Champion, 1979.

La distance intertextuelle, entre deux textes disjoints - c'est-à-dire dont l'un n'est pas contenu en tout ou partie dans l'autre - est clairement influencée par le genre, l'auteur, l'époque et le thème de ces textes.

Le modèle ci-dessous systématise ces constats :

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

Le symbole $f\{\}$ signifie que la distance est fonction des termes, les **variables**, indiquées entre accolades.

Pour donner un poids à chacune de ces variables, on recherche des cas où toutes les autres sont fixes (raisonnement « toutes choses égales par ailleurs »). Cela permettra d'estimer successivement la distance en fonction du genre ($D_{Genre(A,B)}$), de l'auteur ($D_{Auteur(A,B)}$), de la chronologie ($D_{Chron(A,B)}$) et du thème ($D_{Thème(A,B)}$).

- Ces variables ne sont évidement pas indépendantes. On verra plus loin que auteur, temps et thèmes sont relativement liés (certains auteurs évoluent plus que d'autres au cours de leur vie créatrice). Cependant, les comportements globaux et les ordres de grandeurs entre ces différentes distances dépendent peu des autres variables.
- En pratique, on peut s'attendre à l'existence d'une borne « minimale » pour une distance entre deux textes différents, même quand les quatre dimensions qui interviennent sont neutralisées : même auteur, même époque, même thème et même genre.

3.1 Existence d'une borne minimum pour deux textes différents

Deux textes - écrits, à la même époque, par un seul auteur, sur un même thème et dans un même genre - auront une distance minimale non nulle, à cause de l'oubli, au moins partiel, et surtout à cause de la contrainte sociale qui condamne la répétition (le « copier-coller »).

Estimation du minimum

En se limitant à la littérature, la plus faible distance rencontrée sépare deux tragédies de Corneille : Tite et Bérénice (1670) et Pulchérie (1672) : 0.1535 (ou 1 535 mots différents pour 10 000). Leur écriture est contemporaine, le thème est le même (l'amour contrarié par la raison d'Etat), le genre (tragédie en alexandrins) est particulièrement contraignant. Pour l'instant, on peut considérer que, en littérature, 0.15 est la valeur minimale de la distance intertextuelle chez un même auteur.

Dans un genre moins contraignant, les deux comédies en alexandrins le Menteur et la Suite du menteur, du même Corneille, sont aussi intéressantes parce que la seconde est présentée comme la suite de la première et qu'elle a été écrite quasiment dans la foulée. Elles sont séparées par une distance de 0,1797. Molière fournit aussi quelques exemples de distances très faibles entre comédies en alexandrins proches dans le temps : Tartuffe (1664) – Misanthrope (1666) : 0.1707 ; Etourdi - Dépit amoureux (peu avant 1659) : 0.1740.

L'histoire littéraire offre d'autres exemples intéressants. Par exemple, deux romans de Marivaux : la Vie de Marianne – écrit entre 1727 et 1740 et inachevé à la mort de l'auteur – et le Paysan parvenu (1734-1735). Ces romans portent sur les mêmes thèmes. Ils sont séparés par une distance de 0.1733. Gary donne un autre cas avec deux romans parus sous le nom de Ajar : la Vie devant soi (1975) et l'Angoisse du roi Salomon (1979) : distance 0.1756. A la proximité des thèmes, s'ajoutait, pour Gary, la contrainte de reproduire le style et le vocabulaire de sa créature de papier…

Dans un autre genre, les lettres d'exil de V. Hugo fournissent un exemple intéressant puisque thèmes et destinataires sont à peu près identiques d'une année sur l'autre (voir section suivante). La plus petite distance entre lettres d'années différentes est 0.1845 (entre les années 1867-68).

Lorsque deux textes sont écrits par un auteur unique, à la même époque et dans un même genre, sur des thèmes proches ou semblables (et sans que cet auteur fasse du « copier-coller »), **on observe que la distance minimale varie entre 0.15 et 0.19**.

Les différences d'auteurs, de genres, de thèmes et d'époques ajoutent leurs effets à ce minimum. L'influence de chacune de ces variables est estimée en suivant la même méthode. Voici deux exemples à propos de la variable « temps ».

3.2 La distance en fonction de la chronologie

La plupart des auteurs changent de vocabulaire au cours de leur vie créatrice. De plus, la langue est un organisme vivant dont le lexique évolue constamment. Pour mesurer le poids de ce phénomène chronologique, il faut trouver des corpus où les trois autres variables (genre, auteur, thème) sont fixées. Dans ces corpus, on cherche une distance telle que :

$D_{Chron(A,B)} \simeq f\{(Epoque_A, Epoque_B)\}$

Par exemple :

La correspondance d'exil de V. Hugo

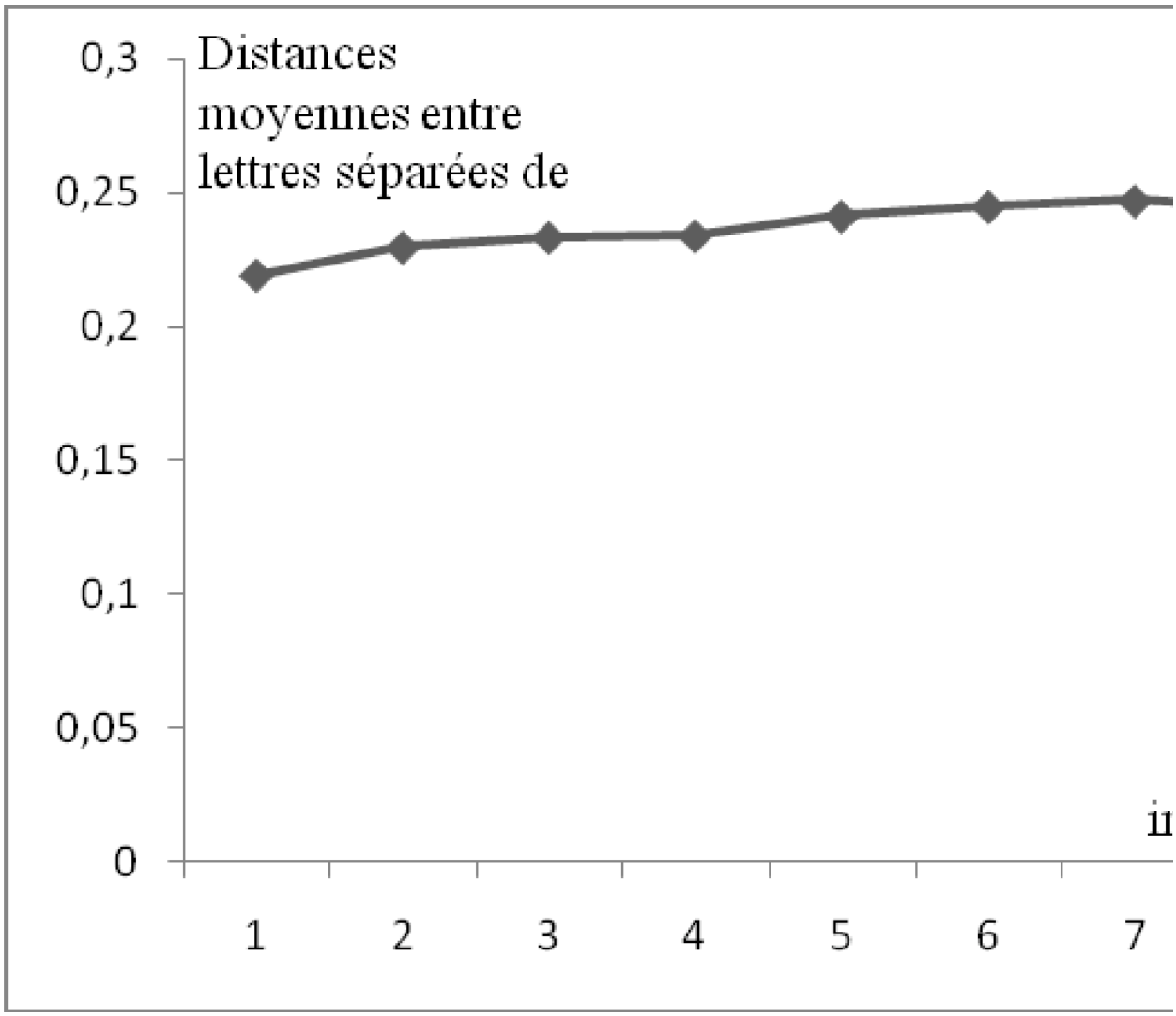
Hugo a été exilé dans les Iles anglo-normandes entre 1853 et 1870. La correspondance était son seul lien avec la France. Pendant tout son exil, les mêmes sujets sont abordés dans ces centaines de lettres envoyées à ses amis, sa famille, ses éditeurs, des journalistes et des écrivains : les aléas du courrier, la santé, la famille, l'argent, les épreuves de ses œuvres, les articles de presse… [10]

Les lettres sont classées chronologiquement, groupées par années, ce qui permet de mesurer la distance entre celles séparées d'un an, deux ans, trois ans… Ce corpus offre donc la possibilité de mesurer 16 moyennes pour le décalage d'un an, 15 pour le décalage de 2 ans (…) et seulement 8 pour le décalage de 10 ans. C'est pourquoi on ne va pas au-delà (ce serait donner aux fluctuations accidentelles un poids trop lourd). On obtient le tableau ci-dessous.

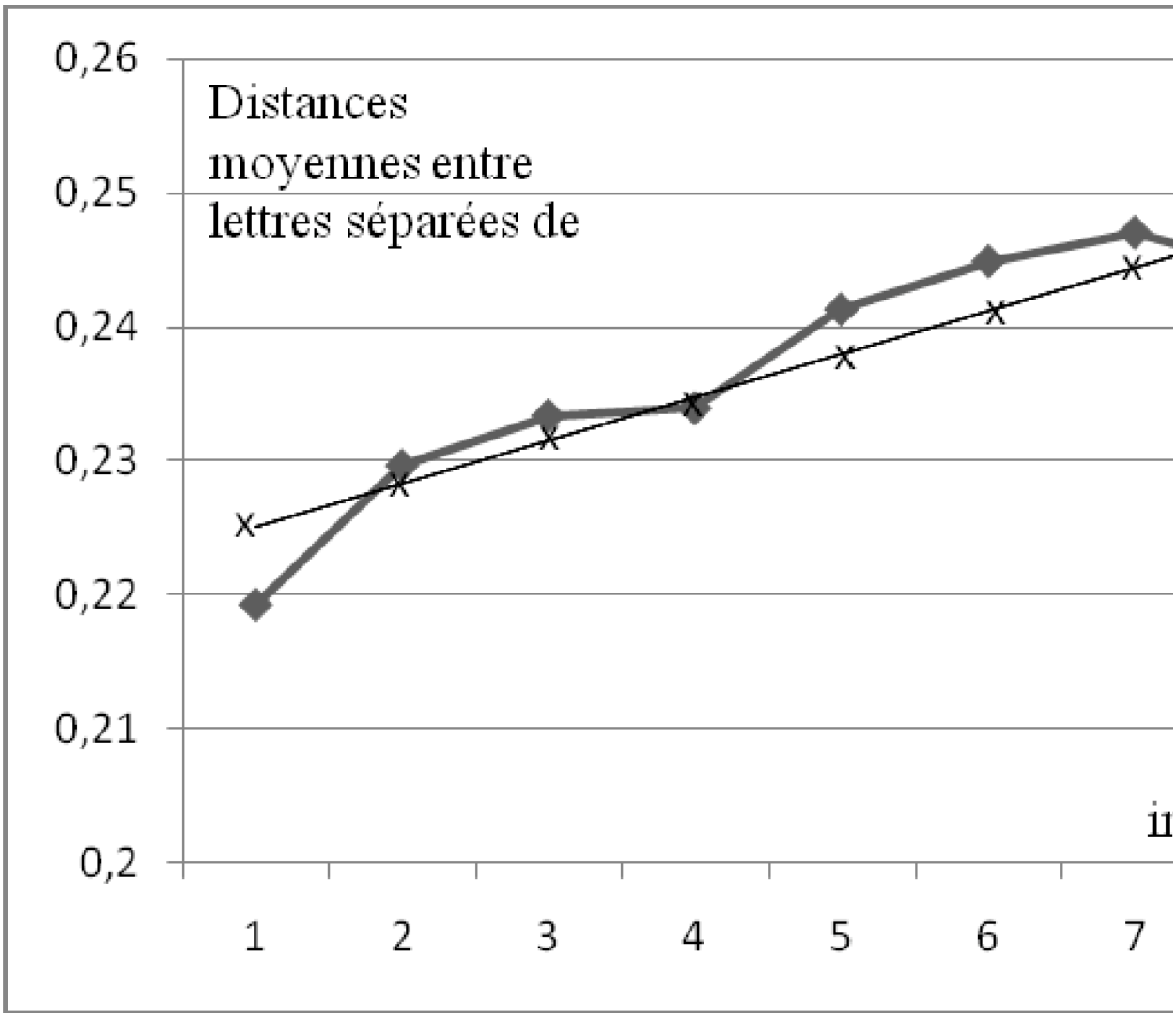
Intervalle temporel	*Distance moyenne*
1 an	*0,2193*
2 ans	*0,2297*
3 ans	*0,2334*
4 ans	*0,2340*
5 ans	*0,2413*
6 ans	*0,2449*
7 ans	*0,2470*
8 ans	*0,2442*
9 ans	*0,2497*
10 ans	*0,2522*

Ce tableau suggère une relation fonctionnelle qui peut être représentée dans un graphique.

Sur l'axe horizontal (ou abscisse) on inscrit le temps (ici l'intervalle temporel séparant les groupes de lettres) et sur l'axe vertical (l'ordonnée), l'échelle des distances. Chaque point représente une moyenne des distances entre deux années plus ou moins éloignées. Cette série est discrète (il n'y a aucune observation entre 1 an, 2 ans, etc.). Cependant, on suppose qu'il s'agit d'un phénomène « continu » : il pourrait aussi bien être observé à l'échelle du trimestre, du mois, de la semaine, du jour… Cette caractéristique autorise à joindre les points par un trait gras qui symbolise, en quelque sorte, l'écoulement du temps.



Ce trait semble à peu près plat, avec de petits accidents et une légère montée de gauche à droite. Afin de rendre ces deux phénomènes plus lisibles, on déplace l'axe horizontal pour qu'il ne coupe plus l'axe vertical à l'ordonnée 0 mais près de la distance la plus faible (ici 0.20). C'est un « zoom » qui grossit le phénomène à étudier. Naturellement, il faut se souvenir que, à l'échelle exacte, la courbe est moins accidentée et moins pentue.



A part une légère encoche pour les lettres séparées par huit années, la tendance à l'augmentation des distances moyennes est relativement régulière et elle présente un profil linéaire (figuré par le trait maigre sur le graphique). Autrement dit, les distances augmenteraient régulièrement en fonction du temps et les petites ruptures de pente dans la courbe seraient l'effet de légères perturbations que l'on propose de négliger (ici elles proviennent de ce que le poids des différents thèmes change un peu selon les années…).

Pour vérifier cette intuition, on procède en trois temps.

1. Calcul de la droite d'ajustement

Cette droite est figurée par le trait maigre sur le graphique. Elle passe « au plus près » des observations. Elle représente le profil du phénomène en négligeant les perturbations accidentelles. C'est pourquoi on parle « d'ajustement » - ou de « régression » - des distances (en fonction du temps).

Cette droite est calculée de la manière suivante (sous réserve de disposer d'un nombre suffisant d'observations). On note :

t : le rang des intervalles (t variant ici de 1 à T = 10 années)

$\bar{t}$ *: le « milieu » de la série temporelle (5,5 années)*

D_t *: la distance moyenne des lettres de l'année t (avec celles de l'année de base)*

$\bar{D}$ *: la moyenne de ces moyennes.* $\bar{D}$ *= 0.2396*

On fait correspondre à chaque valeur observée (D_t *), un point théorique (*D'_t*) de même abscisse t et dont l'ordonnée donne la valeur que l'on aurait observée s'il n'y avait pas eu les petites fluctuations aléatoires. Cette valeur théorique est donnée par l'équation suivante :*

$$D'_t = at + b$$

b est le point où la droite d'ajustement coupe l'ordonnée (ou année 0). Ici b = 0.2218. On notera que cette distance n'est pas égale à Dmin car la diversité thématique de la correspondance pour une année donnée s'y trouve incorporée. Simplement, cette diversité thématique se retrouve à peu près à l'identique chaque année.

a est la « pente » de la droite d'ajustement par rapport à l'horizontale (l'axe des abscisses), calculée pour minimiser la somme des carrés des écarts entre les valeurs observées (D_t*) et les valeurs théoriques correspondantes (*D'_t*) en utilisant la formule suivante :*

$$a = \frac{\sum_{t=1}^{T}(t-\bar{t})(D_t-\bar{D})}{\sum_{t=1}^{T}(t-\bar{t})}$$

Ce coefficient varie entre 0 (droite horizontale) et ± ∞ (droite verticale). Lorsqu'il est égal à 1, la droite est parallèle à la première diagonale ; et parallèle à la seconde diagonale quand a = -1.

Cette pente permet de mesurer directement l'influence de la variable temps : lorsque l'intervalle entre lettres augmente d'un an, la distance moyenne entre ces lettres augmente de 32 mots (pour 10 000). C'est peu mais sur 10 ans, cela donne une croissance de 323 mots soit 15% de la distance d'origine (année 0).

2. La qualité de l'ajustement

Cette qualité est mesurée par le « coefficient de détermination » (de la distance par le temps) :

$$R^2 = \frac{\text{somme des carrés de la régression}}{\text{total des carrés}}$$

$$R^2 = \frac{\sum_{t=1}^{T}(D_t - \bar{D})(D'_t - \bar{D})}{\sum_{t=1}^{T}(D_t - \bar{D})^2 \sum_{t=1}^{T}(D'_t - \bar{D})^2} = 0{,}907$$

Si le temps était seul en cause, ce coefficient serait égal à 1. A l'inverse, si le temps n'avait aucune relation avec la distance, ce coefficient serait égal à 0. Etant donné le nombre des valeurs de la série (10), ce coefficient peut être considéré comme excellent : la liaison est attestée.

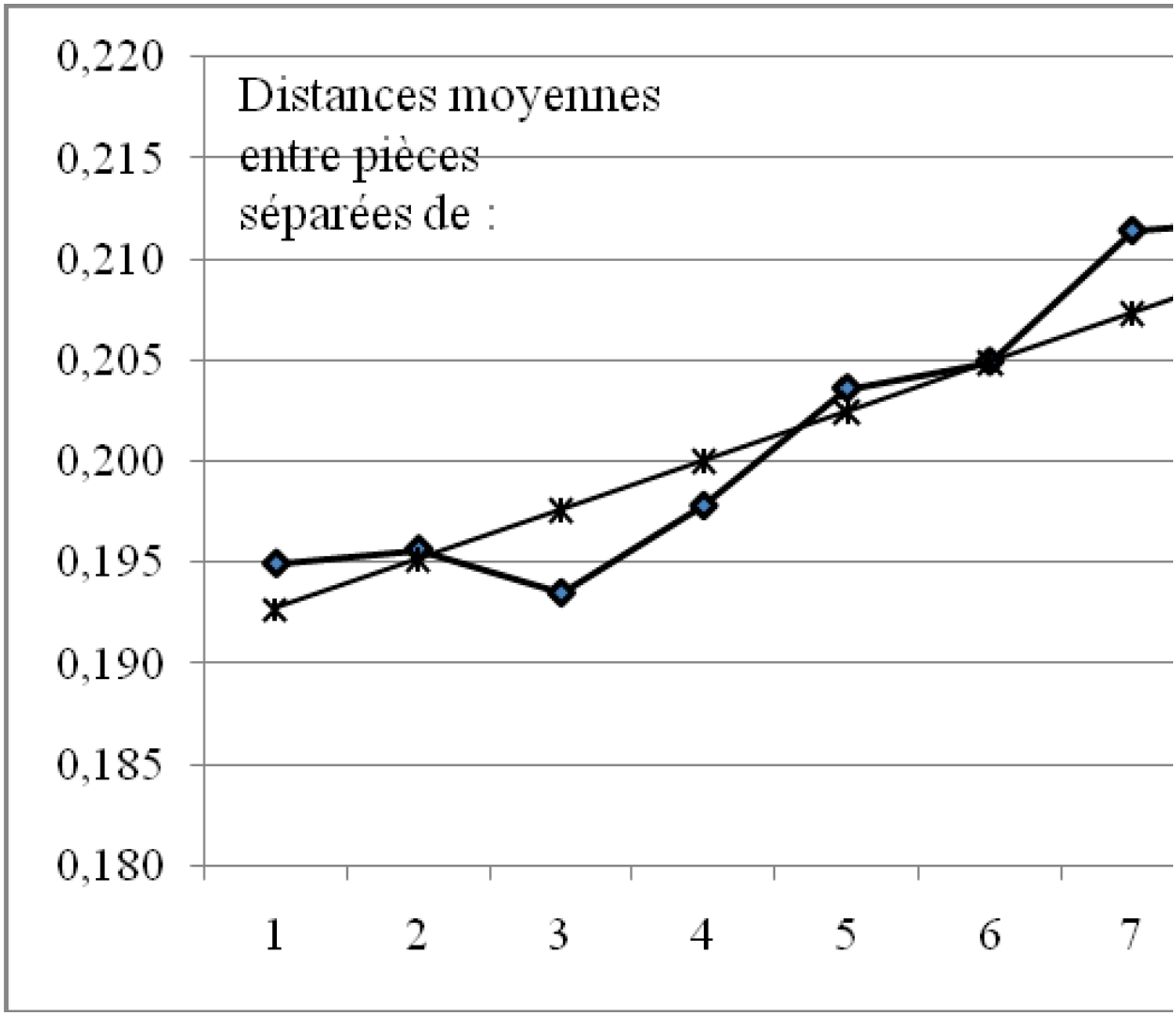
3. Interprétation de la relation

Le fait qu'une variable évolue régulièrement au cours du temps ne signifie pas que le temps est l'explication de cette variation. Le temps peut être simplement la condition permettant à une autre variable de s'exprimer. Par exemple, chez certains auteurs, la découverte de nouveaux thèmes ou des changements de « registre »… Autrement dit, il faut préalablement vérifier que les autres variables sont « neutralisées » pour pouvoir conclure que le temps est bien une variable autonome explicative d'une certaine proportion de distance entre des textes. Au cours de ses 18 ans d'exil, la correspondance de Victor Hugo remplit bien ces conditions. Cela permet d'affirmer que le temps est à l'origine de cette augmentation des distances entre ses lettres en fonction du nombre d'années qui les séparent.

Chez Hugo, la distance entre deux textes non contemporains - de même genre et de même thème -, augmente quasi constamment de 1.4% par an, soit 15% en dix ans.

Les tragédies de Corneille

Pour les dates de création des pièces, voir ci-dessus le corpus du théâtre classique. De Cinna (1639) à Suréna (1674), P. Corneille a donné 18 tragédies en alexandrins comportant 5 actes (on écarte Don Sanche, Psyché et Tite et Bérénice qui sont des « tragi-comédies »). Il faut remarquer qu'ici la date – imparfaitement connue pour certaines pièces – est la création et non pas l'écriture. La figure ci-dessous décrit l'évolution de la distance entre les pièces dont la création est séparée par des intervalles de temps croissants. Comme dans la figure précédente, on a déplacé l'origine de l'ordonnée pour obtenir un « zoom » ; le trait maigre est la droite d'ajustement.



Les formules ci-dessus donnent les valeurs suivantes :

- *moyenne des distances : 0.2037*
- *pente de la droite d'ajustement : a = 0.00240*
- *origine de la droite d'ajustement : b = 0.1904*
- *coefficient de détermination : 0.886*

Etant le nombre des valeurs (18), le coefficient de détermination est bon. La liaison est avérée. La tendance est une augmentation de la distance de 12% en 10 ans, soit un rythme moyen annuel de + 1.1%.

On peut retenir qu'une distance de 0.19 entre deux pièces contemporaines de Corneille équivaut à une distance supérieure à 0.21 entre deux pièces de cet auteur mais séparées par 10 ans et de 0.24 pour un intervalle de 20 ans.

Dans les pièces de Corneille, la tendance à l'augmentation de la distance est de 12% en 10 ans. Autrement dit, une distance de 0.19 entre deux pièces contemporaines de Corneille équivaut à une distance supérieure à 0.21 entre deux pièces de cet auteur mais séparées par 10 ans et de 0.24 pour un intervalle de 20 ans.

Corneille fait preuve d'une stabilité plus grande que la majorité des écrivains, notamment Hugo. On le vérifie en répétant la même expérience sur les romans, la poésie et le théâtre de Hugo : à chaque fois, on observe une augmentation importante des distances avec l'éloignement temporel des œuvres.

Plus deux textes d'un même auteur sont éloignés dans le temps, plus leur distance est forte. Pour la plupart des auteurs, et selon les genres, cet accroissement s'inscrit dans un intervalle **de 0.5 à 1.5 % par année**.

Ces expériences mettent également en valeur le poids prépondérant de la variable « genre ».

3.3 La distance en fonction du genre

Par « genre », on désigne le vocabulaire particulier et l'ensemble des règles - souvent implicites - qui caractérisent un mode particulier de communication. Ecrit et oral, prose, poésie, théâtre, fiction, correspondance… sont quelques-uns des principaux genres [11].

Le raisonnement est le même que ci-dessus. Pour estimer l'influence du genre, on utilise des corpus où la distance intertextuelle se réduit à une distance en genre telle que :

$$D_{Genre(A,B)} \simeq f\{(Genre_A, Genre_B)\}$$

On utilise des auteurs qui ont écrit à la fois du théâtre, de la poésie – en prose et/ou en vers -, des romans, des lettres… Ces œuvres permettent de préciser la notion de genre.

Estimation du poids du genre

Pour illustrer ce raisonnement, on utilisera à nouveau Hugo qui a exploité à peu près tous les genres. Le calcul est possible du fait que les textes utilisés sont séparés par des laps de temps pas trop différents et que le renouvellement thématique semble comparable dans les différents compartiments (sous réserve de ce qui a été dit de sa correspondance d'exil).

	Poésie	*Romans*	*Théâtre*	*Correspondance*
Poésie	*0.2540*	*0.3875*	*0.3907*	*0.4355*
Roman	*0.3875*	*0.3001*	*0.4010*	*0.4060*
Théâtre	*0.3907*	*0.4010*	*0.3099*	*0.3521*
Correspondance	*0.4355*	*0.4060*	*0.3521*	*0.2435*

La distance intertextuelle étant symétrique, la partie supérieure droite du tableau contient les mêmes informations que la partie inférieure gauche. La diagonale du tableau indique les distances moyennes « intra-genre » (entre les textes appartenant à un même genre) et les autres cases les distances moyennes « inter-genre » (entre textes de genres différents). Les premières sont comprises entre 0.24 et 0.30 ; les secondes sont plus élevées et toutes supérieures à 0.35.

On en tire que les genres les plus homogènes et les plus décalés sont la poésie et la correspondance (augmentation moyenne de 60 à 70% de la distance avec un texte d'un autre genre), les romans et le théâtre sont les plus hétérogènes mais les plus centraux (augmentation moyenne de 20 à 33% en passant à un autre genre).

Toutefois, la notion de genre se prête moins à la mesure que la chronologie. En effet, comme le suggère la diagonale du tableau, les principaux genres ne sont pas homogènes. Il en est ainsi du théâtre, même en alexandrins. Il faut distinguer deux sous-genres principaux : la comédie et la tragédie, comme on peut le constater avec le cas des deux dernières comédies de Corneille : le Menteur et la Suite du menteur. D'après une lettre de Corneille, on sait en effet qu'il a composé le Menteur et Pompée (tragédie) dans le même hiver. On peut donc comparer ces deux comédies avec les tragédies contemporaines (Horace, Polyeucte, Pompée et Rodogune).

	Horace	*Polyeucte*	*Pompée*	*Menteur1*	*Menteur2*	*Rodogune*
Horace (1640	*0*	*0,2121*	*0,2127*	*0,2829*	*0,2828*	*0,2089*
Polyeucte (1641)	*0,2121*	*0*	*0,2134*	*0,2385*	*0,2395*	*0,2016*
Pompée (1642)	*0,2127*	*0,2134*	*0*	*0,2806*	*0,2801*	*0,2151*
Menteur (1642)	*0,2829*	*0,2385*	*0,2806*	*0*	*0,1797*	*0,2650*
Menteur2 (1643)	*0,2828*	*0,2395*	*0,2801*	*0,1797*	*0*	*0,2583*
Rodogune (1645)	*0,2089*	*0,2016*	*0,2151*	*0,2650*	*0,2583*	*0*

La distance moyenne entre tragédies contemporaines est : 0.2106 (les différences de thèmes ajoutent ici leur influence à la distance minimale) ; entre les deux comédies : 0.1797 (distance minimale) et entre tragédies et comédies de 0.2660. La différence est donc nette même si elle est beaucoup moins élevée que celle observée en comparant le théâtre et la poésie, la prose et les vers, le roman et la correspondance, etc. On est donc parfois obligé d'introduire la notion de « sous-genre » au sein des principaux genres.

Une différence de genre, toutes choses égales par ailleurs, **augmente la distance entre deux textes d'au moins 20%** (pour des genres proches ou entre les principaux sous-genres) **et jusqu'à 70%** pour des genres décalés comme la poésie en vers ou la correspondance par rapport aux autres. **Le genre est donc le facteur prépondérant expliquant la distance entre les textes**. Ensuite vient l'auteur.

3.4 La distance en fonction des auteurs

L'influence de la variable auteur est étudiée de la même manière, en utilisant des textes contemporains écrits, sur un même thème, par deux auteurs différents, situation où la distance intertextuelle se résume à une distance entre auteurs :

$$D_{Auteur(A,B)} \simeq f\{(Auteur_A, Auteur_B)\}$$

Estimation du poids d'une différence d'auteur

Il est proposé d'examiner ici deux cas qui se rapprochent beaucoup d'une expérience de laboratoire.

Premièrement, la correspondance que Flaubert (GF) et Maupassant (GM) ont échangée de 1874 à la mort de Flaubert (1880). Il s'agit d'un véritable dialogue, les mêmes thèmes étant abordés par les deux hommes. Les lettres sont classées par ordre chronologique et, dans le tableau ci-dessous, elles sont rassemblées en deux ensembles de longueur équivalente.

	GF à GM (1874-78)	*GF à GM (1879-80)*	*GM à GF (1875-78)*	*GM à GF (1879-80)*
GF à GM (1874-78)	*0*	***0.2534***	*0.2886*	*0.2716*
GF à GM (1879-80)	***0.2534***	*0*	*0.2972*	*0.2659*
GM à GF (1875-78)	*0.2886*	*0.2972*	*0*	***0.2322***
GM à GF (1879-80)	*0.2716*	*0.2659*	***0.2322***	*0*

En gras, les distances observées chez un même auteur (distances « intra ») et en maigre celles entre les deux (« inter »). Les distances inter sont en moyenne de 16% plus élevées que les distances intra. Mais, cet écart est un minimum car il faut tenir compte du mimétisme considérable de Maupassant envers Flaubert.

La seconde « expérience de laboratoire » est fournie par Corneille et Racine. En 1670, ces deux auteurs ont chacun écrit une tragédie sur l'amour impossible entre un empereur romain (Titus) et

une reine orientale (Bérénice). Ils l'ont fait au même moment, en aveugle, en alexandrins et en respectant les fameuses règles qui enserraient la grande tragédie au moins depuis la célèbre « querelle du Cid ». Et tous deux avaient en tête la liaison entre Louis XIV et sa belle-sœur Henriette d'Angleterre. Le lieu de l'action, les personnages, les thèmes étaient donc identiques (d'où un vocabulaire commun important…). La distance entre les deux Bérénice est de 0.2561. Le tableau ci-dessous donne les principales distances caractéristiques entre Corneille et Racine à l'époque de Tite et Bérénice.

	Tite et Bérénice (Corneille, 1670)	*Bérénice (Racine, 1670)*
Corneille :		
Agésilas (1666)	*0.1585*	*0.2780*
Attila (1667)	*0.1801*	*0.2892*
Tite et Bérénice (1670)	*-*	*0.2561*
Pulchérie (1672)	*0.1535*	*0.2712*
Suréna (1674)	*0.1559*	*0.2643*
Racine :		
Andromaque (1667)	*0.2587*	*0.2266*
Britannicus (1669)	*0.2507*	*0.2095*
Bérénice (1670)	*0.2561*	*-*
Bazajet (1672)	*0.2620*	*0.2195*
Mithridate (1673)	*0.2488*	*0.2061*

Entre toutes les tragédies de Corneille et de Racine entre 1664 (première pièce de Racine et 1674 (dernière tragédie de Corneille), on constate que les distances « inter » sont de 20 à 70% plus élevées que les distances « intra ». Mais, sauf pour Bérénice, la variable thème ajoute ses effets à la variable auteur.

La distance entre deux textes contemporains écrits par deux écrivains différents, dans un même genre et sur un thème semblable, est de 15% à 40% plus élevée qu'entre deux textes contemporains d'un même auteur et ceci, même dans le cas où l'un des deux éprouve un fort mimétisme envers l'autre.

Après le genre, l'auteur est la variable prépondérante expliquant la distance entre les textes.

3.5 La distance en fonction du thème

Comme précédemment, on recherche les situations où l'on se rapproche le plus de la formulation suivante :

$$D_{Thème(A,B)} \simeq f\{(Thème_A, Thème_B)\}$$

Estimation du poids des différences de thèmes

On peut utiliser à nouveau la comparaison Corneille et Racine présentée ci-dessus à propos des deux Bérénices. La comparaison de toutes les tragédies des deux auteurs entre 1664 (première pièce de Racine) et 1674 (retraite de Corneille) permet d'estimer qu'une différence de thèmes – entre textes contemporains écrits dans un même genre - augmente la distance de 3 à 8% alors que le poids qu'une différence d'auteur l'augmente de 15 à 40%.

Cette expérience a un autre intérêt : elle montre l'influence mutuelle entraînée par la rivalité entre Corneille et Racine et la grande proximité de leurs tragédies des années 1664-1674. L'influence mutuelle des deux rivaux n'est pas seule en cause. Les contraintes de la tragédie classique sont fortes : alexandrins, durée, découpage en scènes et actes, règles de bienséance, d'unité de temps et de lieu, mêmes thèmes antiques… Autrement dit, les distances observées entre ces tragédies contemporaines peuvent être considérées comme les minima possibles entre deux auteurs différents.

Au total, ces expériences conduisent à établir les relations suivantes :

- **La distance entre deux textes est le produit de quatre variables : le genre, l'auteur, le thème et l'époque** ;
- pour l'échelle de la vie humaine, **la variable genre pèse plus lourd que l'auteur qui l'emporte sur le thème et sur la chronologie**.

Dès lors, **pour attribuer des textes d'origine inconnue ou douteuse, il faut les comparer à d'autres, écrits dans le même genre et à la même époque par des auteurs incontestés**.

4. UNE MISE À L'ÉPREUVE

Comment vérifier l'aptitude de la distance intertextuelle à identifier les auteurs ? La procédure s'apparente aux expériences « en aveugle » utilisées en biologie ou en médecine. En voici une inédite.

4.1 Sélection des textes et des auteurs

Les 12 auteurs et les 56 textes sélectionnés

Les extraits, tirés de différents romans du XIXe siècle, sont de même longueur (10 000 mots). Ils sont disjoints (ne portent pas sur les mêmes portions de texte).

N°	*Auteur*	*Titre et date de publication*
1	*Balzac*	*Père Goriot (1835)*
2	*Dumas(père)*	*Monte Cristo (1845)*
3	*Dumas(fils)*	*Dame aux camélias (1848)*
4	*Flaubert*	*Bouvard et Pécuchet (1880)*
5	*Gautier*	*Avatar (1856)*
6	*Hugo*	*Misérables (1862)*
7	*Lamartine*	*Graziella (1852)*
8	*Maupassant*	*Bel ami (1885)*
9	*Stendhal*	*Chartreuse de Parme (1839)*
10	*Verne*	*Tour du monde (1873)*
11	*Vigny*	*Servitude et grandeur militaires (1835)*
12	*Zola*	*Bête humaine (1890)*
13	*Balzac*	*Père Goriot (1835)*
14	*Dumas(père)*	*Monte Cristo (1845)*
15	*Dumas(fils)*	*Dame aux camélias (1848)*
16	*Flaubert*	*Bouvard et Pécuchet (1880)*
17	*Gautier*	*Avatar (1856)*
18	*Hugo*	*Misérables (1862)*
19	*Lamartine*	*Graziella (1852)*
20	*Maupassant*	*Bel ami (1885)*

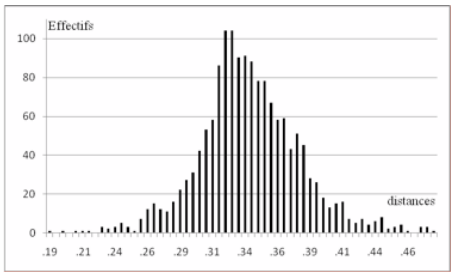
21	Stendhal	Chartreuse de Parme (1839)
22	Verne	Tour du monde (1873)
23	Vigny	Servitude et grandeur militaires (1835)
24	Zola	Bête humaine (1890)
25	Dumas(père)	Trois mousquetaires (1845)
26	Dumas(fils)	Dame aux camélias (1848)
27	Flaubert	Madame Bovary (1857)
28	Hugo	Misérables (1862)
29	Maupassant	Fort comme la mort (1889)
30	Stendhal	Chartreuse de Parme (1839)
31	Vigny	Servitude et grandeur militaires (1835)
32	Zola	Germinal (1885)
33	Dumas(père)	Trois mousquetaires (1845)
34	Flaubert	Madame Bovary (1857)
35	Hugo	Notre-Dame de Paris (1831)
36	Maupassant	Fort comme la mort (1889)
37	Stendhal	Rouge et Noir (1830)
38	Zola	Germinal (1885)
39	Dumas(père)	Trois mousquetaires (1845)
40	Flaubert	Madame Bovary (1857)
41	Hugo	Notre-Dame de Paris (1831)
42	Maupassant	Mont Oriol (1887)
43	Stendhal	Rouge et Noir (1830)
44	Zola	l'Argent (1891)
45	Flaubert	Education sentimentale (1869)
46	Hugo	Notre-Dame de Paris (1831)
47	Maupassant	Mont Oriol (1887)
48	Stendhal	Rouge et Noir (1830)
49	Zola	l'Argent (1891)
50	Flaubert	Education sentimentale (1869)
51	Maupassant	Pierre et Jean (1888)
52	Zola	l'Argent (1891)
53	Flaubert	Education sentimentale (1869)
54	Maupassant	Pierre et Jean (1888)
55	Maupassant	Une vie (1883)
56	Maupassant	Une vie (1883)

Ils ont été choisis pour avoir :

- un nombre important d'auteurs contemporains différents travaillant dans un même genre (ici la fiction romanesque),
- pour certains auteurs, des textes décalés dans le temps, dans le style ou par les thèmes (Hugo, Flaubert, Zola)…
- pour différents auteurs, des textes supposés proches (Dumas père et fils, la *Bovary* de Flaubert et *Une vie*, le premier roman de Maupassant, etc) ou des textes contemporains portant sur les mêmes thèmes.
- un nombre impair d'extraits pour une majorité d'auteurs (ou de livres) afin de déjouer les algorithmes de classification qui regroupent les individus par paires.

Les 56 textes sont anonymés et on calcule les distances entre les 1 540 couples différents. Ces 1 540 distances sont rangées par ordre croissant. Puis elles sont regroupées par classes d'intervalle égal (dans la figure ci-dessous l'intervalle est de .010).

Histogramme des distances classées par ordre croissant dans des classes d'intervalle .01



Dès qu'il y a au moins 8 auteurs et plus de 40 textes, on obtient ce profil (« courbe en cloche » ou « courbe de Laplace-Gauss ») qui indique que, pour l'essentiel, la population étudiée est gouvernée par le hasard, mais aussi qu'elle comporte à ses deux extrémités des individus singuliers.

Les caractéristiques singulières de la distribution

Avec 56 objets différents, on peut former 1540 couples différents et mesurer pour chacun de ces couples, la distance séparant les textes. Notons Dj chacune de ces distances, avec j variant de 1 (la plus petite) à J (la plus grande). J étant ici égal à 1540. Quatre valeurs centrales sont à considérer :

- *la moyenne (notée $\overline{D}$) est égale à : .3359 La longueur des textes étant de 10 000 mots, on peut lire ce résultat de la manière suivante : « en moyenne, les textes sont séparés par une distance de 3 359 mots ». Ou encore : « il y a en moyenne 3 359 mots différents dans chaque texte comparé à tous les autres ».*
- *le mode : distance la plus fréquente, par convention fixée au centre de la classe la plus peuplée (.32-.33) soit .3250*
- *la médiane : valeur qui partage la population en deux parts égale, soit ici la 770e valeur. Me = .03328. La moitié des distances sont supérieures à cette valeur et la moitié lui sont inférieures.*
- *la médiale (notée Ml) : partage la distance totale en deux parties égales. La différence entre la médiane et la médiale donne une idée de la concentration du caractère étudié sur les fortes valeurs (et de l'asymétrie de la distribution) : Ml = .3371*

Dans une distribution gaussienne, le mode, la moyenne et la médiane sont confondues. Dans le cas présent, on a une légère asymétrie à gauche (mode < médiane < moyenne) qui peut être négligée. La grande proximité des quatre valeurs indique que la distribution est à peu près symétrique autour de la moyenne et que l'on est donc bien en présence d'une population "gaussienne".

Cela permet d'associer à la moyenne, une mesure standard de la dispersion des observations : l'écart type (noté σ). C'est la racine carrée de la variance (moyenne quadratique des écarts des observations à la moyenne arithmétique) :

$$\sigma = \sqrt{\frac{\sum_{j=1}^{J}(D_j - \bar{D})^2}{J}} = .0380$$

Une population gaussienne présente plusieurs propriétés intéressantes, résumées dans le schéma ci-dessous :

Les valeurs indiquées sur le schéma fournissent des repères commodes :

- *environ les deux tiers des observations sont groupées autour de la moyenne dans un intervalle égal à plus ou moins un écart type ;*
- *moins de 5% de ces observations se situent en dehors de l'intervalle de deux écarts-types autour de la moyenne (2.5 en dessous et 2.5 au dessus), dit « intervalle de fluctuation standard » ;*
- *moins de 1% de ces observations se situent en dehors de l'intervalle de trois écarts-types (0.5 de chaque côté de l'intervalle).*

Dans l'expérience présente, les valeurs observées correspondent aux valeurs prédites par le schéma théorique ci-dessus : 5% des distances sortent de la « norme », c'est-à-dire qu'elles sont situées au-delà des bornes de l'intervalle de fluctuation standard autour de la moyenne (±1.96σ).

L'hypothèse à tester est la suivante : à une époque donnée et dans un genre donné, la variable auteur pèse plus lourd que la variable thème. Dans ce cas,

- *les distances « anormalement » courtes (inférieures ou égales à .260) doivent s'observer entre textes d'un même auteur, lorsqu'ils portent sur des thèmes peu éloignés,*
- *les distances les plus longues (égales ou supérieures à .411) doivent s'observer entre textes d'auteurs différents portant sur des thèmes éloignés.*

L'attribution d'auteur utilise ces distances singulières.

4.2 Attribution d'auteur en aveugle

L'hypothèse à tester est la suivante : à une époque donnée et dans un genre donné, les différences d'auteurs pèsent plus lourd que les différences de thèmes dans la distance entre deux textes. Cette hypothèse est vérifiée :

Les distances les plus courtes

Le tableau ci-dessus récapitule les 42 distances remarquablement faibles (inférieures à la moyenne diminuée de deux écarts-types).

Rang	*Texte 1*	*Texte2*	*Distance*	*Auteur 1*	*Auteur 2*
01	*15*	*26*	*0,17773*	*Dumas fils*	*Dumas fils*
02	*03*	*26*	*0,19095*	*Dumas fils*	*Dumas fils*
03	*03*	*15*	*0,20389*	*Dumas fils*	*Dumas fils*
04	*11*	*31*	*0,20811*	*Vigny*	*Vigny*
05	*12*	*24*	*0,21089*	*Zola*	*Zola*
06	*32*	*38*	*0,22008*	*Zola*	*Zola*
07	*23*	*31*	*0,22010*	*Vigny*	*Vigny*
08	*11*	*23*	*0,22011*	*Vigny*	*Vigny*
09	*51*	*53*	*0,22824*	*Maupassant*	*Maupassant*
10	*02*	*14*	*0,22916*	*Dumas*	*Dumas*
11	*55*	*56*	*0,23012*	*Maupassant*	*Maupassant*
12	*45*	*50*	*0,23274*	*Flaubert*	*Flaubert*
13	*34*	*40*	*0.23465*	*Flaubert*	*Flaubert*
14	*25*	*39*	*0,23579*	*Dumas*	*Dumas*
15	*42*	*47*	*0,23639*	*Maupassant*	*Maupassant*
16	*27*	*34*	*0,23718*	*Flaubert*	*Flaubert*
17	*33*	*39*	*0,23756*	*Dumas*	*Dumas*
18	*50*	*53*	*0,23957*	*Flaubert*	*Flaubert*
19	*37*	*43*	*0,24243*	*Stendhal*	*Stendhal*
20	*54*	*55*	*0,24407*	*Maupassant*	*Maupassant*
21	*18*	*28*	*0,24412*	*Hugo*	*Hugo*
22	*08*	*20*	*0,24608*	*Maupassant*	*Maupassant*
23	*54*	*56*	*0,25002*	*Maupassant*	*Maupassant*
24	*45*	*53*	*0,25019*	*Flaubert*	*Flaubert*
25	*20*	*51*	*0,25056*	*Maupassant*	*Maupassant*
26	*14*	*33*	*0,25242*	*Dumas*	*Dumas*
27	*08*	*51*	*0,25356*	*Maupassant*	*Maupassant*
28	*06*	*28*	*0,25412*	*Hugo*	*Hugo*
29	*04*	*16*	*0,25500*	*Flaubert*	*Flaubert*
30	*21*	*30*	*0,25527*	*Stendhal*	*Stendhal*
31	*29*	*54*	*0,25535*	*Maupassant*	*Maupassant*
32	*20*	*54*	*0,25569*	*Maupassant*	*Maupassant*
33	*05*	*17*	*0,25661*	*Gautier*	*Gautier*
34	*29*	*36*	*0,25746*	*Maupassant*	*Maupassant*
35	*20*	*29*	*0,25850*	*Maupassant*	*Maupassant*
36	*36*	*54*	*0,25873*	*Maupassant*	*Maupassant*
37	*44*	*52*	*0,25891*	*Zola*	*Zola*
38	*27*	*40*	*0.25892*	*Flaubert*	*Flaubert*
39	*01*	*13*	*0,25897*	*Balzac*	*Balzac*
40	*41*	*46*	*0,25928*	*Hugo*	*Hugo*
41	*27*	*45*	*0,25950*	*Flaubert*	*Flaubert*
42	*20*	*56*	*0.26010*	*Maupassant*	*Maupassant*

NB : les noms des auteurs et des œuvres ont été dévoilés après l'expérience.

La distance la plus faible (0.1777) sépare les textes n° 15 et 26 ; la seconde (0.1910) les textes 03 et 26. Les textes 03,15 et 26 constituent donc un même groupe (G1) et, si notre hypothèse est exacte, ils doivent avoir été écrits par un même auteur, à la même époque et sur un même thème. On répète l'opération pour les 39 autres distances inférieures ou égales à la borne inférieure de l'intervalle de fluctuation standard. Ces distances sont examinées une par une en constituant des groupes comme celui donné en exemple ci-dessus (tableau ci-dessous).

Etape	*Textes joints*	*Groupe*	*Auteur*
01	*15 26*	*G1 (15 - 26)*	*Dumas fils (Dame aux camélias)*
02	*03 26*	*G1 (03 - 15 - 26)*	*Dumas fils (Dame aux camélias)*
03	*03 15*	*id*	*Dumas fils (Dame aux camélias)*
04	*11 31*	*G2 (11 - 31)*	*Vigny (Servitude et grandeur militaires)*
05	*12 24*	*G3 (12 - 24)*	*Zola (Bête humaine)*
06	*32 38*	*G4 (32 - 38)*	*Zola (Germinal)*

07	23 31	G2 (11 - 23 - 31)	Vigny (Servitude et grandeur militaires)
08	11 23	id	Vigny (Servitude et grandeur militaires)
09	51 54	G5 (51 - 54)	Maupassant (Pierre et Jean)
10	02 14	G6 (02 - 14)	Dumas père (Monte Cristo)
11	55 56	G7 (55 - 56)	Maupassant (Une vie)
12	45 50	G8 (45 - 50)	Flaubert (Education sentimentale)
13	34 40	G9 (34 - 40)	Flaubert (Bovary)
14	25 39	G10 (25 - 39)	Dumas père (Trois mousquetaires)
15	42 47	G11 (42 - 47)	Maupassant (Mont Oriol)
16	27 34	G9 (27 – 34 - 40)	Flaubert (Bovary)
17	33 39	G10 (25 - 33 - 39)	Dumas père (Trois mousquetaires)
18	50 53	G8 (45 - 50 - 53)	Flaubert (Education sentimentale)
19	37 43	G13 (37 - 43)	Stendhal (Rouge et Noir)
20	54 55	G5 (51 - 54 - 55 - 56)	Maupassant (Pierre et Jean, Une vie)
21	18 28	G14 (18 - 28)	Hugo (Misérables)
22	08 20	G15 (08 - 20)	Maupassant (Bel ami)
23	54 56	Id	Maupassant (Pierre et Jean, Une Vie)
24	45 53	Id	Flaubert (Education sentimentale)
25	20 51	G5 (08 - 20 - 51 - 54 - 55 - 56)	Maupassant (Bel ami, Pierre et Jean, Une Vie)
26	14 33	G6 (02 - 14 -25 - 33 - 39)	Dumas (Monte Cristo, Trois mousquetaires)
27	08 51	Id	Maupassant (Bel ami, Pierre et Jean, Une Vie)
28	06 28	G14 (06 -18 - 28)	Hugo (Misérables)
29	04 16	G16 (04 - 16)	Flaubert (Bouvard et Pécuchet)
30	21 30	G17 (21 - 30)	Stendhal (Chartreuse de Parme)
31	29 54	G5 (08 - 20 - 29 - 51 - 54 - 55 - 56)	Maupassant (Bel ami, Fort comme la mort, Pierre et Jean, Une Vie)
32	20 54	Id	Maupassant (Bel ami, Fort comme la mort, Pierre et Jean, Une Vie)
33	05 17	G18 (05 - 17)	Gautier (Avatar)
34	29 36	G5 (08 - 20 - 29 - 36 - 51 - 54 - 55 - 56)	Maupassant (Bel ami, Fort comme la mort, Pierre et Jean, Une Vie)
35	20 29	Id	Maupassant (Bel ami, Fort comme la mort, Pierre et Jean, Une Vie)
36	36 54	Id	Maupassant (Bel ami, Fort comme la mort, Pierre et Jean, Une Vie)
37	44 52	G19 (44 - 52)	Zola (l'Argent)
38	27 40	G 9 (27 - 34 - 40)	Flaubert (Bovary)
39	01 13	G20 (01 - 13)	Balzac (Père Goriot)
40	41 46	G21 (41 - 46)	Hugo (Notre-Dame de Paris)
41	27 45	G8 (27 – 34 – 40 - 45 - 50 - 53)	Flaubert (Bovary, Education sentimentale)
42	20 56	Id	Maupassant (Bel ami, Fort comme la mort, Pierre et Jean, Une Vie)

NB : les noms des auteurs et des œuvres ont été dévoilés après l'expérience

---

Les distances « anormalement » courtes séparent des textes d'un même auteur, portant sur des thèmes peu éloignés (généralement extraits d'un même ouvrage),

Les distances les plus longues

Les distances remarquablement élevées (supérieures à la moyenne augmentée de deux écarts-types)

Rang	Texte 1	Texte2	Distance	Auteur 1	Auteur 2
1494	07	48	0,41725	Lamartine	Stendhal
1495	07	28	0,41751	Lamartine	Hugo
1496	07	10	0,41776	Lamartine	Verne
1497	02	04	0,41838	Dumas(père)	Maupassant
1498	05	26	0,41962	Gautier	Dumas(fils)
1499	03	50	0,42015	Dumas(fils)	Flaubert
1500	03	10	0,42044	Dumas(fils)	Verne
1501	10	15	0,42097	Verne	Dumas(fils)
1502	07	18	0,42151	Lamartine	Hugo
1503	16	33	0,42195	Flaubert	Dumas(père)
1504	02	07	0,42303	Dumas(père)	Lamartine
1505	15	42	0,42419	Dumas(fils)	Maupassant
1506	15	27	0,42529	Dumas(fils)	Flaubert
1507	26	42	0,42575	Dumas(fils)	Maupassant
1508	26	27	0,42714	Dumas(fils)	Flaubert
1509	03	27	0,42920	Dumas(fils)	Flaubert
1510	07	33	0,43062	Lamartine	Dumas(père)
1511	03	22	0,43183	Dumas(fils)	Verne
1512	10	26	0,43219	Verne	Dumas(fils)
1513	15	49	0,43303	Dumas(fils)	Zola
1514	03	49	0,43364	Dumas(fils)	Zola
1515	03	05	0,43481	Dumas(fils)	Gautier
1516	26	50	0,43511	Dumas(fils)	Flaubert
1517	03	53	0,43527	Dumas(fils)	Flaubert
1518	26	49	0,43544	Dumas(fils)	Zola
1519	02	16	0,43565	Dumas(père)	Flaubert
1520	15	50	0,43584	Dumas(fils)	Flaubert
1521	15	22	0,43642	Dumas(fils)	Verne
1522	03	45	0,43778	Dumas(fils)	Flaubert
1523	03	32	0,43797	Dumas(fils)	Zola
1524	22	26	0,44132	Verne	Dumas(fils)
1525	15	53	0,44238	Dumas(fils)	Flaubert
1526	05	15	0,44579	Gautier	Dumas(fils)
1527	26	53	0,44583	Dumas(fils)	Flaubert
1528	15	45	0,44714	Dumas(fils)	Flaubert
1529	03	04	0,45118	Dumas(fils)	Flaubert
1530	04	15	0,45186	Flaubert	Dumas(fils)
1531	26	32	0,45260	Dumas(fils)	Zola
1532	15	32	0,45317	Dumas(fils)	Zola

1533	*26*	*45*	*0,45502*	*Dumas(fils)*	*Flaubert*
1534	*04*	*26*	*0,46506*	*Flaubert*	*Dumas(fils)*
1535	*03*	*16*	*0,46711*	*Dumas(fils)*	*Flaubert*
1536	*07*	*26*	*0,46854*	*Lamartine*	*Dumas(fils)*
1537	*15*	*16*	*0,47232*	*Dumas(fils)*	*Flaubert*
1538	*03*	*07*	*0,47321*	*Dumas(fils)*	*Lamartine*
1539	*07*	*15*	*0,47369*	*Lamartine*	*Dumas(fils)*
1540	*16*	*26*	*0,47717*	*Flaubert*	*Dumas(fils)*

NB : les noms des auteurs ont été dévoilés après l'expérience

La plus forte distance sépare les n°16 (Flaubert, Bouvard et Pécuchet) et 26 (Dumas fils, La dame aux camélias). Puisque le n° 16 a été classé dans le groupe n°9 (avec le n° 04), on en tire que les auteurs des groupes 1 et 9 ne sont pas les mêmes. L'opération est répétée jusqu'à la borne supérieure de l'intervalle de fluctuation normale (tableau ci-dessous).

Etape	*Textes disjoints*	*Groupes disjoints*
01	*16 26*	*G9 ≠ G1*
02	*07 15*	*07 ≠ G1*
03	*03 07*	*id*
04	*15 16*	*G9 ≠ G1*
05	*07 26*	*G9 ≠ G1*
06	*03 16*	*G9 ≠ G1*
07	*04 26*	*G9 ≠ G1*
08	*26 45*	*G6 ≠ G1*
09	*15 32*	*G3 ≠ G1*
10	*26 32*	*G3 ≠ G1*
11	*04 15*	*G9 ≠ G1*
12	*03 04*	*G9 ≠ G1*
13	*15 45*	*G6 ≠ G1*
14	*26 53*	*G4 ≠ G1*
15	*05 15*	*G11 ≠ G1*
16	*15 53*	*G6 ≠ G1*
17	*22 26*	*22 ≠ G1*
18	*03 32*	*G3 ≠ G1*
19	*03 45*	*G6 ≠ G1*
20	*15 22*	*22 ≠ G1*
21	*15 50*	*G6 ≠ G1*
22	*02 16*	*G5 ≠ G9*
23	*26 49*	*49 ≠ G1*
24	*03 53*	*G6 ≠ G1*
25	*26 50*	*G6 ≠ G1*
26	*03 05*	*G11 ≠ G1*
27	*03 49*	*49 ≠ G1*
28	*15 49*	*49 ≠ G1*
29	*10 26*	*10 ≠ G1*
30	*03 22*	*22 ≠ G1*
31	*07 33*	*7 ≠ G5*
32	*03 27*	*G6 ≠ G1*
33	*26 27*	*G6 ≠ G1*
34	*26 42*	*G7 ≠ G1*
35	*15 27*	*G6 ≠ G1*
36	*15 42*	*G7 ≠ G1*
37	*02 07*	*07 ≠ G5*
38	*16 33*	*G9 ≠ G8*
39	*07 18*	*07 ≠ G8*
40	*10 15*	*10 ≠ G1*
41	*03 10*	*10 ≠ G1*
42	*03 50*	*G6 ≠ G1*
43	*05 26*	*G11 ≠ G1*
44	*02 04*	*G9 ≠ G5*
45	*07 10*	*07 ≠ 10*
46	*07 28*	*07 ≠ G8*
47	*07 48*	*07 ≠ 48*

Conclusions :

- *l'auteur du groupe n°1 n'est pas celui des groupes 3, 4, 6, 5, 7 et 11 ni celui des textes 07, 22 et 49 (qui restent à classer),*
- *l'auteur du groupe n°5 n'est pas celui du groupe 9,*
- *l'auteur du groupe n°8 n'est pas celui du groupe 9,*
- *le texte 07 (qui reste à classer) ne peut avoir été écrit par les auteurs des groupes 1, 5 et 8,*
- *les textes 7 et 10 (qui restent à classer) ne sont pas du même auteur,*
- *les textes 7 et 48 (qui restent à classer) ne sont pas du même auteur.*

Les distances « anormalement » longues séparent effectivement des textes d'auteurs différents portant sur des thèmes éloignés.

Entre les deux

Certains textes avaient été choisis parce qu'il semblait difficile d'y déceler la même plume, par exemple : Bouvard et Pécuchet et les deux autres romans de Flaubert ou, pour Hugo : Notre-Dame de Paris et Les Misérables. A l'inverse, d'autres textes étaient sélectionnés parce qu'ils semblaient remarquablement proches bien que d'auteurs différents (Bovary de Flaubert et Une vie de Maupassant ou les Trois Mousquetaires de Dumas et Servitude et grandeur militaires de Vigny).

Ces intuitions n'étaient pas toutes injustifiées comme on le voit dans les tableaux ci-dessous ne sont mentionnées que les distances inférieures à 0.3000. En effet, pour des distances comprises entre 0.260 et 0.300, on rencontre surtout des textes d'un même auteur, généralement d'ouvrages différents. Mais il y a également un certain nombre d'intrus. Les 2 tableaux ci-dessous présentent les deux cas remarquables : l'extrait n°40 (Bovary puis le n°31 (Grandeur et servitude militaires).

	Du n° 40 (Flaubert Bovary) au…	
34	*Flaubert Bovary (extrait 2)*	*0,2346*
27	*Flaubert Bovary (extrait 1)*	*0,2589*
45	*Flaubert Education sentimentale (extrait 1)*	*0,2636*

56	Maupassant Une vie (extrait 2)	0,2667
08	Maupassant Bel ami (extrait 1)	0,2669
50	Flaubert Education sentimentale (extrait 2)	0,2689
55	Maupassant Une vie (extrait 1)	0,2737
54	Maupassant Pierre et Jean (extrait 2)	0,2746
36	Maupassant Fort comme la mort (extrait 2)	0,2783
20	Maupassant Bel ami (extrait 2)	0,2789
47	Maupassant Mont Oriol (extrait 2)	0,2820
53	Flaubert Education sentimentale (extrait 3)	0,2832
51	Maupassant Pierre et Jean (extrait 1)	0,2863
24	Zola La bête humaine (extrait 2)	0,2913
12	Zola La bête humaine (extrait 1)	0,2927
29	Maupassant Fort comme la mort (extrait 1)	0,2948
38	Zola Germinal (extrait 2)	0,2952
42	Maupassant Mont Oriol (extrait 1)	0,2974
44	Zola l'Argent (extrait 1)	0,2980
52	Zola l'Argent (extrait 3)	0,2981
(…)		
04	Flaubert Bouvard et Pécuchet (extrait 1)	0,3148
16	Flaubert Bouvard et Pécuchet (extrait 2)	0,3337

L'influence de Flaubert (celui de Bovary et même plus précisément de certains passages de ce livre) sur Maupassant était déjà connue. Elle se vérifie ici. De plus, beaucoup de distances entre ces passages de Bovary et les romans de Maupassant sont inférieures à celles entre Bovary et les autres romans de Flaubert… Mais l'impact de Bovary sur l'histoire littéraire ne se limite pas à Maupassant, Zola, également, semble ne pas y être insensible, notamment dans certaines scènes de la Bête humaine… En revanche, à la fin de sa vie, Flaubert avait tourné la page et explorait d'autres thèmes dans un style différent (Bouvard et Pécuchet paraît 23 ans après Bovary).

De cet exemple extrême, on conclura que lorsque les variables « thème » et « temps » additionnent leurs effets, elles peuvent excéder le poids de la variable auteur.

Le second cas remarquable est celui du roman de Vigny Servitude et grandeur militaires (1835), spécialement la fin (extrait 3). Mais ici, il faut tenir compte des dates de publication : le roman de Vigny est antérieur à Graziella ou Monte Cristo et à la Chartreuse de Parme, mais postérieur à Notre-Dame de Paris (1831).

	Du n°31 (Vigny Servitude et grandeur... extrait 3) au...	distance
11	Vigny Servitude et grandeur... (extrait 1)	0,2081
23	Vigny Servitude et grandeur... (extrait 2)	0,2201
19	Lamartine Graziella (extrait 2)	0,2778
36	Maupassant Fort comme la mort (extrait 1)	0,2817
14	Dumas père Monte Cristo (extrait 2)	0,2829
02	Dumas père Monte Cristo (extrait 1)	0,2855
25	Dumas père Trois Mousquetaires (extrait 1)	0,2875
39	Dumas père Trois Mousquetaires (extrait 3)	0,2904
21	Stendhal Chartreuse de Parme (extrait 2)	0,2985
46	Hugo Notre-Dame de Paris (extrait 3)	0,2988

V. Hugo et A. de Vigny passent pour les inventeurs du « roman historique », Vigny ayant l'antériorité avec Cinq-Mars (1826), absent de cette expérience. Celle-ci a pour principal intérêt de révéler un « sous-genre » au sein de la fiction romanesque.

Lorsque les variables « thème » et « temps » additionnent leurs effets, elles peuvent excéder le poids de la variable auteur.

Les distances entre des textes d'un même auteur portant sur des thèmes différents sont inférieures à celles observées entre textes d'auteurs différents portant sur un même thème (sauf quand le temps ajoute ses effets aux différences de thèmes). Ces caractéristiques aboutissent à une première classification

48 textes sont attribués au bon auteur

Les jonctions opérées grâce aux distances remarquablement faibles et les disjonctions (grâce aux distances remarquablement fortes) sont résumées dans le tableau ci-dessous. Les auteurs et les titres ont été révélés après l'expérience. Les groupes sont classés par ordre d'agrégation : les premiers textes agrégés sont ceux de la Dame au Camélias, puis ceux de Vigny, etc.

N°	Textes	Auteurs
01	03 15 26	Dumas fils Dame aux camélias
02	11 23 31	Vigny Servitude et grandeur militaires
03	12 24 32 38	Zola la Bête humaine, Germinal
04	08 20 29 36 51 54 55 56	Maupassant Bel Ami, Fort comme la mort, Pierre et Jean, Une vie
05	02 14 25 33 39	Dumas père Monte Cristo, Trois Mousquetaires
06	27 34 40 45 50 53	Flaubert Bovary, Education sentimentale
07	42 47	Maupassant Mont Oriol
08	37 43	Stendhal Rouge et Noir
09	06 18 28	Hugo Misérables
10	04 16	Flaubert Bouvart et Pécuchet
11	21 30	Stendhal Chartreuse de Parme
12	05 17	Gautier Avatar
13	44 52	Zola l'Argent
14	01 13	Balzac Père Goriot
15	41 46	Hugo Notre Dame de Paris

- *Huit textes n'ont pas pu être associés à un autre, ni rattachés à un groupe déjà existant : 07 et 19 (Lamartine), 09 et 48 (Stendhal), 10 et 22 (Verne), 35 (Hugo), 49 (Zola).*
- *De plus, après dévoilement des auteurs et des titres, il apparaît que quatre auteurs figurent dans plus d'un groupe (Flaubert, Hugo, Maupassant, Stendhal, Zola), mais qu'il s'agit toujours de livres différents (thèmes éloignés) écrits à des époques éloignées.*

A ce point de l'expérience, il faut rappeler que la procédure ne vise pas à retrouver à tout coup l'auteur d'un texte mais à tester une hypothèse et à conclure avec un degré de certitude raisonnable. L'hypothèse a bien résisté, puisque toutes les distances remarquablement faibles séparent des textes écrits par un même auteur et, pour la plupart, il s'agit d'extraits d'un même ouvrage, donc portant également sur un même thème. A l'opposé, toutes les distances les plus fortes séparent des auteurs différents.

Que faire avec les 8 textes restés non classés ? On passe à une autre méthode la « recherche du plus proche voisin » : on associe deux textes à un même groupe si l'un est le plus proche voisin de l'autre, même si leur distance mutuelle dépasse la borne inférieure de l'intervalle de fluctuation standard.

Chacun des 8 restants est attribué au groupe de son plus proche voisin

Sachant que tout auteur a au moins deux textes, on recherche quelle est la distance la plus courte concernant les 8 textes demeurant non classés (tableaux ci-dessous). Pour éclairer le raisonnement, on donne aussi les distances immédiatement plus élevées.

N°	N°	Distance	Auteurs
07	19	0.3120	Lamartine - Lamartine

	27	0.3411	Lamartine – Flaubert (Bovary)
	04	0.3444	Lamartine – Flaubert (Bouvard)
	34	0.3500	Lamartine – Flaubert (Bovary)
	05	0.3543	Lamartine – Gautier

N°	N°	Distance	Auteurs
09	37	0.2730	Stendhal (Chartreuse – Rouge et Noir)
	21	0.2741	Stendhal (Chartreuse – Chartreuse)
	30	0.2814	Stendhal (Chartreuse – Chartreuse)
	43	0.2903	Stendhal (Chartreuse – Rouge et Noir)
	34	0.3101	Stendhal (Chartreuse) – Flaubert (Bovary)

N°	N°	Distance	Auteurs
10	22	0.2790	Verne - Verne
	25	0.3369	Verne – Dumas (Trois Mousquetaires)
	09	0.3371	Verne - Stendhal (Chartreuse)
	43	0.3376	Verne - Stendhal (Rouge et Noir)
	34	0.3449	Verne – Flaubert (Bovary)

N°	N°	Distance	Auteurs
35	41	0.2787	Hugo - Hugo (Notre-Dame)
	46	0.2819	Hugo - Hugo (Notre-Dame)
	35	0.3059	Hugo - Hugo (Notre-Dame)
	40	0.3085	Hugo (Notre-Dame) – Flaubert (Bovary)
	08	0.3129	Hugo (Notre-Dame) – Maupassant (Bel Ami)

N°	N°	Distance	Auteurs
48	43	0.2640	Stendhal (Rouge et Noir)
	30	0.2760	Stendhal (Rouge et Noir) – Stendhal (Chartreuse)
	37	0.2800	Stendhal - Stendhal (Rouge et Noir)
	21	0.2955	Stendhal (Rouge et Noir) – Stendhal (Chartreuse)
	18	0.2964	Stendhal (Rouge et Noir) – Hugo (Misérables)

N°	N°	Distance	Auteurs
49	44	0,2635	Zola - Zola (l'Argent)
	38	0,2999	Zola (l'Argent) – Zola (Germinal)
	52	0,3059	Zola - Zola (l'Argent)
	34	0,3110	Zola (l'Argent) – Flaubert (Bovary)
	24	0,3160	Zola (l'Argent) – Zola (la Bête humaine)

Il n'y a que 5 tableaux car, parmi les 8 textes restants à classer, 6 sont groupés par couples (7 et 19), (9 et 48) et (10 et 22). Chaque première ligne confirme le principe suivant : ***pour des auteurs contemporains, travaillant dans un même genre, la variable auteur joue un rôle prépondérant - par rapport aux thèmes - dans la distance entre deux textes****. Ce qui peut aussi se dire de la manière suivante : les textes produits par deux auteurs différents - travaillant sur un même thème, à une même époque - sont plus éloignés que ceux que chacun de ces deux auteurs produit sur des thèmes différents (ex : Lamartine et Flaubert ou Verne et Dumas…).*

Le classement complet est obtenu à l'aide de la première ligne de chacune des 6 cases ci-dessus. Tous les textes groupés ensemble sont d'un même auteur.

On aboutit au tableau suivant*.

N°	*Textes*	*Auteurs***
01	03 15 26	Dumas fils (*Dame aux camélias*)
02	11 23 31	Vigny (*Servitude et grandeur militaires*)
03	12 24 32 38	Zola (*la Bête humaine*, *Germinal*)
04	08 20 29 36 51 54 55 56	Maupassant (*Bel Ami*, *Fort comme la mort*, *Pierre et Jean*, *Une vie*)
05	02 14 25 33 39	Dumas père (*Monte Cristo*, *Trois Mousquetaires*)
06	27 34 40 45 50 53	Flaubert (*Bovary*, *Education sentimentale*)
07	42 47	Maupassant (*Mont Oriol*)
08	**09** 37 43 **48**	Stendhal (*Chartreuse de Parme et Rouge et Noir*)
09	06 18 28	Hugo (*Misérables*)
10	04 16	Flaubert (*Bouvart et Pécuchet*)
11	21 30	Stendhal (*Chartreuse de Parme*)
12	05 17	Gautier (*Avatar*)
13	44 **49** 52	Zola (*l'Argent*)
14	01 13	Balzac (*Père Goriot*)
15	**35** 41 46	Hugo (*Notre Dame de Paris*)
16	**07 19**	Lamartine (*Graziella*)
17	**10 22**	Verne (*Tour du monde*)

* En gras les n° des textes groupés par la méthode du « plus proche voisin ». ** Titres et auteurs ont été révélés à la fin de l'expérience.

Tous les textes sont associés à au moins un autre et attribués au bon auteur, mais il y a une erreur de livre : dans le huitième groupe, l'extrait n° 9 (*la Chartreuse de Parme*) est groupé avec le n° 37 (et donc avec les n° 43 et 48, c'est-à-dire *le Rouge et Noir*) alors qu'il aurait dû l'être avec les n° 21 et 30. Mais il s'agit d'un résultat fourni « faute de mieux », à la dernière étape de la classification.

Certains auteurs figurent dans plusieurs groupes - Flaubert, Hugo, Maupassant, Stendhal, Zola – mais il s'agit de livres différents écrits à des époques souvent éloignées. On vérifie ainsi que, lorsque thème et temps additionnent leurs effets, leur influence peut être plus forte que celle de l'auteur.

Dans cette expérience, rien n'a pu « fausser » les résultats : aucun jugement a priori, pas d'intervention manuelle, ni d'« apprentissage ». Il n'y a pas non plus de « risque d'erreur » à proprement parler, puisqu'on ne travaille pas sur échantillons mais sur des populations entières. En revanche, le risque d'un « faux positif » existe, même s'il est extrêmement faible, du moins lorsqu'on utilise un nombre suffisant de textes.

4.3 D'autres expériences

Une expérience unique permet de conclure que l'hypothèse n'est pas invalidée. Pour la considérer comme utilisable, il faut que personne n'ait pu la mettre en échec. Les tests se déroulent comme dans l'expérience précédente : les textes sont choisis par une autre personne que l'opérateur et ils sont « anonymés » avant de lui être soumis. Puis l'ensemble est mis dans le domaine public afin que l'expérience puisse être reproduite. Citons notamment :

- Une expérience en aveugle sur des textes anglais [12].
- Plusieurs travaux sur le théâtre élisabéthain [13].
- Sur les premiers ministres italiens [14].
- L'identification d'une « plume de l'ombre » soupçonnée d'avoir travaillé pour deux Premiers ministres québécois successifs. Parmi tous les discours prononcés par ces deux chefs, la méthode ci-dessus isole ceux qui sont « trop proches » pour être de deux auteurs différents et ceux « trop éloignés » pour être d'un même auteur. [15].
- sur R. Gary, auteur, à la fin des années 1970, d'une supercherie littéraire (E. Ajar). L'intérêt réside ici dans la volonté de Gary de masquer sa propre écriture pour créer un « auteur de papier » (E. Ajar) [16].

4.3 Echelle de la distance intertextuelle

De nombreuses expériences, comme celles qui viennent d'être présentées, ont abouti à l'étalonnage d'une échelle des distances.

Présentation de l'échelle

Pour deux textes dont la longueur tourne autour de 10.000 mots, on observe que :

- *distances inférieures à 0.20 (un mot sur 5) : les deux textes sont écrits dans un même genre, par un seul auteur, à la même époque et sur des thèmes proches,*
- *entre 0.20 et .25 : l'auteur et le genre sont identiques mais les dates de composition et (ou) les thèmes sont plus éloignés. Si les textes sont d'auteurs différents, alors, non seulement ils sont écrits dans le même genre et sur le même thème, mais il y a une « influence » mutuelle ou le second auteur s'est « inspiré » du premier…*
- *entre 0.25 et 0.35, soit l'auteur est le même et alors plusieurs facteurs ont changé (temps et thème), soit ce sont deux auteurs différents, travaillant à la même époque dans un même genre sur des thèmes plus ou moins proches,*
- *au-dessus de .35, si les deux textes sont écrits dans un même genre, alors les auteurs sont différents. Si l'auteur est le même, alors les genres sont différents.*
- *au dessus de .45, auteur et genre sont différents (sauf s'il s'agit d'une transcription de l'oral comparée à de l'écrit).*

Cette échelle permet d'identifier rapidement les textes trop proches pour avoir été écrits par deux auteurs différents (ou trop éloignés pour être de la même main), sans avoir besoin de refaire à chaque fois des analyses longues et fastidieuses.

Trois remarques à propos de l'échelle de la distance intertextuelle

- *L'indice de la distance relative varie uniformément entre un minimum de 0 et un maximum de 1 (fonction dite « monotone »). Les valeurs indiquées dans l'échelle ne sont donc pas des seuils mais des bornes sur un continuum.*
- *Cette échelle est applicable à des textes dont les longueurs sont comprises entre 5 000 et 25 000 mots. En effet, pour les raisons exposées dans la section 2.2, le calcul est inapplicable sur des textes de longueur inférieure à 1 000 mots (instabilité trop forte de l'indice). Entre 1 000 et 5 000 mots, les distances tendent à être plus élevées (et d'autant plus que les textes sont plus courts) et à diminuer sensiblement avec l'allongement de textes. Ensuite, la diminution est beaucoup plus lente.*
- *Cette échelle a été établie à partir d'un grand nombre d'expériences comme celle présentée ci-dessus. Les bornes ne visent pas à identifier à coup sûr les textes mais à permettre des conclusions rapides et assurées, sans avoir besoin de refaire à chaque fois des analyses longues et fastidieuses. Naturellement, plus on s'approche des bornes, plus les conclusions doivent être prudentes.*

Cette échelle s'applique à des textes parfaitement corrigés et étiquetés. C'était le cas de certaines pièces de théâtre du XVIIe siècle (présentées au début de cet article).

5. LE THÉÂTRE DU XVIIe SIÈCLE

Appliquée aux pièces présentées au début de la troisième section, cette méthode classe correctement la plupart des textes, mais elle apporte deux surprises qui concernent toutes deux des pièces de Corneille et de Molière.

5.1 Corneille et Molière

Les œuvres de Corneille et Molière présentent deux anomalies uniques dans l'histoire littéraire :

- Deux comédies en alexandrins de P. Corneille - *le Menteur* (1642) et *la suite du Menteur* (1643) – sont classées avec 14 comédies présentées par Molière entre 1659 et 1673 : douze comédies en alexandrins et deux comédies en prose (*Dom Juan* et *l'Avare*). Toutes les distances séparant ces pièces sont remarquablement faibles.

Les Menteurs et les pièces de Molière

Le tableau ci-dessous présente les distances remarquables séparant les principales pièces présentées par Molière et le Menteur (Corneille 1642), la Suite du Menteur (Corneille 1643) et les Plaideurs (Racine 1668).

N°	Pièces	Genre	Le Menteur	Suite du Menteur	Les Plaideurs
15	Le Menteur (1642)	Vers	-	0,1797	0,2961
16	La suite du Menteur (1643)	Vers	0,1797	-	0,2933
39	L'étourdi (1658)	Vers	0,2048	0,2060	0,2691
40	Dépit amoureux (1658)	Vers	0,2154	0,2114	0,2702
42	Sganarelle (1660)	Vers	0,2585	0,2527	0,2928
44	L'école des maris (1661)	Vers	0,2231	0,2168	0,2791
45	Les fâcheux (1661)	Vers	0,2477	0,2476	0,3062
46	L'école des femmes (1662)	Vers	0,2256	0,2167	0,2608
50	Princesse d'Elide (1664)	Vers Prose	0,2518	0,2426	0,3135
51	Le Tartuffe (1664)	Vers	0,2416	0,2315	0,2753
54	Le Misanthrope (1666)	Vers	0,2524	0,2331	0,2824
56	Mélicerte (1666)	Vers	0,2569	0,2499	0,3217
58	Amphytrion (1668)	Vers libres	0,2525	0,2563	0,2966
60	L'Avare (1668)	Prose	0,2566	0,2439	0,2696
66	Femmes savantes (1672)	Vers	0,2598	0,2486	0,2829
	Moyenne œuvre Molière		0,2761	0,2680	0,3002
	Moyenne pièces en vers Molière		0,2386	0,2315	0,2850
	Moyenne œuvre Corneille		0,2513	0,2480	0,3532
	Moyenne œuvre Racine		0,3140	0,3129	0,3763

Ces valeurs sont d'autant plus remarquables que la Suite du Menteur est séparée de l'Etourdi par 15 ans et qu'il s'écoule encore 15 ans jusqu'à la dernière pièce en vers de Molière (les Femmes savantes). Même si Corneille fait preuve d'une plus grande stabilité que la moyenne des écrivains, l'influence de la chronologie dans son œuvre est incontestable (section 3.2 ci-dessus), de telle sorte que les distances affichées dans les deux premières colonnes du tableau 13 sont ***les plus faibles que l'on peut rencontrer dans l'œuvre d'un seul auteur, dans un même genre, quand cette œuvre s'étend sur trente ans****.*

La dernière colonne du tableau donne les résultats avec les Plaideurs, unique comédie de Racine, en alexandrins, qui est contemporaine des pièces de Molière et fortement influencée par elles. Les distances sont conformes à ce qui est attendu entre deux auteurs différents travaillant dans un même genre et sur des thèmes voisins.

Les dernières lignes du tableau peuvent expliquer l'étonnement qui a accueilli notre travail : les Menteurs sont peu connus et assez décalés par rapport au reste de l'œuvre de Corneille que l'on associe surtout au Cid, Cinna, Horace, Pompée…. De ce fait, la proximité des Menteurs avec les comédies en vers de Molière passe inaperçue.

- Deux « tragi-comédies » présentées par Molière, également en alexandrins, (*Dom Garcie* et *Psyché*) sont classées avec les tragédies et les « comédies héroïques » contemporaines de P. Corneille.

Psyché, Dom Garcie et les tragédies de Corneille

Le tableau ci-dessous présente les distances remarquables séparant Dom Garcie (Molière) et Psyché (Corneille et Molière) des dernières pièces de Corneille.

Corneille	*Dom Garcie(Molière,1661*	*Psyché (Molière 1671)*
Rodogune (1644)	*0,2448*	*0,2310*
Théodore (1645)	*0,2341*	*0,2445*
Héraclius (1647)	*0,2480*	*0,2729*
Andromède (1650)	*0,2412*	*0,2180*
Don Sanche (1650)	*0,2242*	*0,2514*
Nicomède (1651)	*0,2445*	*0,2644*
Pertharite (1651)	*0,2346*	*0,2632*
Œdipe (1659)	*0,2234*	*0,2264*
Toison d'or (1661)	*0,2212*	*0,2198*
Sertorius (1662)	*0,2229*	*0,2378*
Sophonisbe (1663)	*0,2276*	*0,2357*
Othon (1664)	*0,2346*	*0,2399*
Agésilas (1666)	*0,2342*	*0,2327*
Attila (1667)	*0,2349*	*0,2270*
Tite et Bérénice (1670)	*0,2275*	*0,2347*
Psyché (1671)	*0,2300*	*–*
Pulchérie (1672)	*0,2300*	*0,2260*
Suréna (1674)	*0,2165*	*0,2236*
Moyenne Corneille	*0,2431*	*0,2435*
Moyenne Molière	*0,2862*	*0,2974*

De nouveau, il faut tenir compte de ce que le tableau couvre 30 ans de création et que Dom Garcie et Psyché sont elles-mêmes séparées par un intervalle de 10 ans (d'où leur distance de 0.2300). Là encore ***les distances – entre deux auteurs supposés différents - sont les plus faibles que l'on peut rencontrer dans l'œuvre d'un seul auteur****, dans un même genre, quand cette œuvre s'étend sur trente ans.*

Les deux dernières lignes du tableau expliquent à nouveau l'étonnement qui a accueilli notre travail : Dom Garcie et Psyché sont peu étudiés et assez étranges par rapport au reste de l'œuvre présentée sous le nom de Molière alors qu'elles sont très « cornéliennes ».

Toutes ces distances – entre deux auteurs supposés différents - sont les plus faibles que l'on peut rencontrer dans l'œuvre d'un seul auteur, dans un même genre, quand cette œuvre s'étend sur trente ans.

5.2 Trois remarques

- Premièrement, la méthode a été testée sur des dizaines de milliers de textes, sans jamais rencontrer de tels croisements entre deux œuvres d'auteurs différents, sauf… quand les deux auteurs ne font qu'un, comme dans le cas de Gary et Ajar.
- Deuxièmement, le XVIIe siècle fournit plusieurs contre-épreuves. Les deux *Bérénice* de Corneille et Racine ont déjà été évoquées. La distance entre ces deux pièces contemporaines, sur un même thème, est plus élevée que toutes celles constatées entre les pièces en vers de Molière et les deux *Menteurs* de Corneille ou entre *Dom Garcie*, *Psyché* et les tragédies de Corneille, alors que toutes ces pièces sont séparées par un laps de temps important et que les thèmes sont très divers. Le XVIIe fournit d'autres exemples. Les *Sophonisbe* de Mairet et de Corneille se distinguent bien alors qu'ils ont travaillé dans le même genre, sur le même thème, les mêmes événements, les mêmes personnages, en suivant la même trame narrative. Les comédies de Quinault, ou celle de Racine (*les Plaideurs*) se distinguent bien de celles de Corneille et de Molière qui sont impossibles à départager.
- Troisièmement, plusieurs objections ont été présentées.

Les objections

Lorsque notre travail sur Corneille et Molière a été connu, plusieurs objections ont été présentées.

- *En premier lieu, ces objections portent sur la méthode. Elles ont notamment été exposées dans une revue internationale à laquelle nous avons répondu [17].*
- *En second lieu, on a affirmé que toute statistique comporte une « marge d'erreur » (sous-entendant que le cas Corneille-Molière entrait dans cette marge). Cette objection vient de la confusion entre statistique descriptive et statistique probabiliste. L'attribution d'auteur ne porte pas sur des échantillons mais sur des populations recensées exhaustivement et sur des effectifs connus à l'unité près. Il n'y a donc pas de marge d'erreur. En revanche, il existe bien une incertitude liée à l'estimation de la distance entre textes de longueurs différentes (section 2.3). Les textes de Molière et de Corneille sont de longueurs voisines, et toutes supérieures à 5 000 mots, de telle sorte que, dans ce cas précis, l'incertitude est très faible.*
- *Enfin, on a objecté que Molière a beaucoup joué Corneille, qu'il en connaissait donc des milliers de vers et que cela a pu « déteindre » sur lui. Nous avons-nous même signalé cet argument dans notre article de 2001. En effet, l'agenda de la troupe [18] indique que Corneille est le second auteur le plus joué, derrière… Molière. De plus, des cas de mimétisme existent dans l'histoire littéraire. Nous avons signalé l'influence que Madame Bovary a exercé sur la génération suivante, notamment Maupassant ou Zola, celle des premiers romans de Vigny et de Hugo sur le sous-genre du roman historique, ou encore le mimétisme de Racine envers Corneille. Mais les distances entre Corneille et Molière sont nettement plus faibles que celles observées dans ces autres cas et, surtout, elles sont systématiques : toutes les comédies en alexandrins de Molière sont les sœurs cadettes des deux Menteurs de Corneille ; les deux tragi-comédies de Molière (Dom Garcie et Psyché) sont sœurs de toutes les tragédies contemporaines de Corneille. Un tel cas est unique dans l'histoire littéraire.*

5.3 Les autres indices

Plusieurs autres caractéristiques des œuvres de Corneille et de Molière ne se rencontrent pas chez deux auteurs différents

Autres caractéristiques communes des deux œuvres

Les deux œuvres présentent beaucoup d'autres caractéristiques communes. Trois sont particulièrement intéressantes.

1. Les combinaisons des verbes les plus fréquents

Ces combinaisons sont un indice utile pour confirmer une attribution d'auteur. Le tableau ci-dessous donne les combinaisons préférées de Corneille, de Molière et de Racine (fréquence pour 100.000 mots)

P. Corneille		*Molière*		*Racine*	
Syntagmes	*F*	*Syntagmes*	*F*	*Syntagmes*	*F*
faire voir	*33,8*	***faire voir***	*31,5*	*aller voir*	*12,0*
pouvoir être	*18,8*	***pouvoir être***	*25,5*	*pouvoir voir*	*9,6*
pouvoir faire	*18,4*	***pouvoir faire***	*25,5*	*faire entendre*	*9,0*
faire naître	*13,9*	*vouloir dire*	*24,9*	*pouvoir faire*	*8,4*
pouvoir voir	*13,4*	*vouloir faire*	*19,5*	*aller chercher*	*7,8*
devoir être	*12,7*	*pouvoir dire*	*14,5*	*faire parler*	*7,8*
pouvoir souffrir	*10,8*	*pouvoir avoir*	*13,7*	*pouvoir être*	*7,8*
vouloir faire	*9,9*	*aller faire*	*13,2*	*venir chercher*	*7,2*
faire connaître	*9,6*	*avoir faire*	*13,2*	*faire éclater*	*6,6*
devoir faire	*8,7*	*pouvoir voir*	*12,3*	*falloir partir*	*6,6*

Racine ne partage avec Corneille et Molière que trois combinaisons (soulignées dans le tableau) : « pouvoir voir », « pouvoir faire » et « pouvoir être », mais avec un classement et des densités très différentes. En revanche, Corneille et Molière en ont cinq en commun (en gras) dont les trois premières dans le même ordre et avec des densités voisines (en italiques). Etant donné le nombre des combinaisons possibles, la probabilité pour qu'une telle « coïncidence » survienne au hasard est négligeable.

Il existe un seul cas comparable dans les 4 derniers siècles de littérature française : Gary et Ajar. Depuis huit ans, personne n'a pu trouver un autre exemple concernant deux auteurs réellement différents.

2. Le sens spécifique des mots usuels

Le sens spécifique que chaque auteur donne aux principaux mots qu'il emploie peut être retrouvé grâce à l'étude des réseaux sémantiques. Cette étude révèle que, chez Corneille et Molière, les principaux vocables ont le même sens, ou plutôt, que ceux de Molière forment un sous-ensemble dans ceux de Corneille. Le plus évocateur est le mot « amour » – substantif le plus employé par Corneille comme par Molière [19]. Là encore, ces significations sont propres à Corneille et ne se retrouvent pas chez ses contemporains.

3. Les longueurs de phrase

Ces longueurs offrent également un outil auxiliaire pour l'attribution d'auteur [20].

De nombreux indices historiques vont dans le même sens.

Résumé du dossier historique

La clef est fournie par les témoignages de certains contemporains (de Corneille et de Molière) et par le fonctionnement du théâtre à leur époque. Voici résumés les points principaux de cette étude historique [21].

1. Les témoignages des contemporains

Trois éditeurs, contemporains de Molière, ont indiqué, dans les premières éditions de trois pièces, respectivement en 1662, 1671 et 1683, que Molière n'a pas écrit Le dépit amoureux, Psyché et Dom Juan. Dans deux cas, Corneille est mentionné comme l'auteur.

En 1663, les deux principaux critiques de théâtre - Donneau de Visé et Robinet - ont insinué que Molière n'écrivait pas les pièces qu'il présentait. En 1670, Robinet a attribué le Bourgeois gentilhomme à Corneille.

A trois reprises, Boileau a insinué que Molière n'était pas l'auteur des pièces qu'il présentait. Il a aussi accusé les frères Corneille d'être « affamés d'argent » et de se conduire en « mercenaires ». La Fontaine a fait connaissance avec Molière en 1662 et ne l'a plus jamais mentionné dans ses écrits, ni dans sa correspondance. Madame de Sévigné qui recevait beaucoup ne l'a jamais invité, etc.

Aucun des contemporains de Molière ne l'a traité comme un écrivain et lui-même ne s'est pas comporté comme tel.

En particulier,

- *Il ne reste aucun manuscrit de lui, à l'exception d'une vingtaine de signatures sur des actes officiels [22] : pas de lettre, aucune note de sa main, aucun témoignage qu'il ait entretenu une correspondance avec un de ses contemporains.*
- *Entre 1659 et 1673 - période pendant laquelle il est supposé avoir écrit ses chefs d'œuvre - son emploi du temps est connu grâce au « registre de La Grange » [23]. Cet emploi du temps ne lui permettait pas d'écrire une moyenne de deux pièces par an, comme il est supposé l'avoir fait.*

2. Le système du comédien poète

Durant la seconde moitié du XVIIe, 6 pièces de théâtre sur 10 ont été présentées par des acteurs qui ne les avaient pas écrites. C'est le cas de 9 comédies sur 10. On appelait ces acteurs des « comédiens poètes » . Thomas Corneille, le frère cadet de Pierre, était associé à Montfleury puis à Hauteroche, La Fontaine travaillait avec Champmeslé, Boursault (ami des frères Corneille) écrivait pour Poisson…

Le système s'explique de la manière suivante. Les parisiens aimaient beaucoup les comédies légères et satiriques comme celles de Molière. Plus de la moitié des recettes des troupes provenaient de ces pièces. Mais elles étaient condamnées par l'Eglise, la Cour, l'Académie française, de telle sorte que les écrivains qui les fournissaient aux troupes voulaient rester dans l'ombre.

Enfin, il n'y avait pas de protection du droit d'auteur et pas de statut légal pour les troupes. Pour protéger leur exclusivité, les comédiens faisaient acheter le texte par l'un d'entre eux (et limitaient strictement les copies). Cet acteur supervisait la création et l'exploitation, comme le font les metteurs en scène aujourd'hui, il subissait les critiques et, éventuellement, la censure comme cela est arrivé plusieurs fois à Molière…

En définitive, il reste deux solutions. Molière aurait éprouvé un fort mimétisme envers Corneille pendant toute sa vie créatrice. Ou bien Corneille et Molière auraient collaboré selon la procédure usuelle à cette époque et consistant à faire endosser certaines comédies par un « comédien poète ». La convergence de plusieurs indices statistiques - distances, combinaisons des verbes usuels, sens des principaux mots, longueurs de phrases -, avec de nombreux indices historiques, rend possible une conclusion en faveur de la seconde solution.

Plusieurs personnes en avaient eu l'intuition. Citons le poète P. Louÿs (au début du XXe siècle) [24], le romancier H. Poulaille [25] et H. Wouters [26].

6. CONCLUSIONS

Tous nos travaux sont consultables sur le site « archives en ligne » du CNRS (HAL-SHS). Programmes et données sont dans le domaine public. La plupart des documents historiques sont en ligne ou publiés. Tout est vérifiable, tout est reproductible.

Il paraît donc possible de reconnaître l'auteur d'un texte douteux ou d'origine inconnue en le comparant à des textes écrits dans un même genre et à la même époque par des écrivains incontestables.

La distance intertextuelle constitue l'outil principal, combinée avec diverses techniques de classification (qui seront évoquées dans un autre article). Les autres indices – comme les combinaisons des verbes fréquents, le sens des mots les plus usuels ou la longueur des phrases – offrent d'utiles compléments. Naturellement, il faut aussi consulter les documents d'époque.

Cette méthode permettra de résoudre certaines énigmes que l'histoire littéraire nous a léguées, et notamment celle que posent les centaines de comédies, à la mode au

XVIIe, présentées par des comédiens confrères de Molière.

La statistique a beaucoup d'autres applications dans l'analyse littéraire. On peut connaître le vocabulaire d'un auteur, d'une œuvre, d'une époque, mais aussi le sens spécifique des mots, les styles, les thèmes préférés. On peut décrire les évolutions dans une œuvre, les ruptures, les continuités. On reconstituera les principaux courants littéraires ; on éclairera les filiations et les influences mutuelles…

Enfin, il est maintenant possible de connaître la manière dont une langue est parlée et écrite par ses usagers.

P.S. :

Remerciements

Cette recherche est collective. Pierre Hubert et Denis Monière ont collaboré à la mise au point initiale de la distance intertextuelle. De nombreux chercheurs ont participé aux applications et aux développements, notamment : Jean-Guy Bergeron, Mathieu Brugidou, Paul Jolissaint, Gerard Ledger, Jean et Nelly Leselbaum, Xuan Luong, Tom Merriam, Mathieu Ruhlman, Jacques Savoy. Le présent article a bénéficié de l'aide de Nicolas Bedaride, Vincent Beffara, Jacques Istas, Christian Mercat dont les remarques ont permis de nombreuses améliorations.

Notes

[▲1] Pour un tableau d'ensemble de la question : Love Harold. *Attributing Authorship : an Introduction*. Cambridge. Cambridge University Press, 2002.

[▲2] Mendenhall Thomas C.The characteristic curves of composition. *Science*, 9, 1887, 237-249.

[▲3] Le véritable départ a été donné par l'étude de Mosteller & Wallace, en 1964, sur les "federalist papers », série d'articles anonymes du XVIIIe siècle qui passent pour avoir eu une grande influence sur les constituants américains (Mosteller Frederick & Wallace David L. *Inference and Disputed Authorship : The Federalist*. Reading : Addison-Wessley Publishing Company, 1964. Republié sous le titre : *Applied Bayesian and Classical Inference : The Case of the "Federalist Papers"*. New York : Springer-Verlag, 1984). On trouvera en ligne un tableau assez complet des différentes méthodes : Koppel Moshe, Schler Jonathan & Argamon Shlomo. Computational Methods in Authorship Attribution, *Journal of the American Society for Information Science and Technology*, 2009, 60-1, p. 9-26. Trois articles de synthèse comportent une bibliographie importante : Grieve Jack. Quantitative Authorship Attribution. An Evaluation of Techniques. *Literary and Linguistic Computing*. 22-3, 2007, 251-270 ; Rudman Joseph. The State of Authorship Attribution Studies : Some Problems and Solutions. *Computers and the Humanities*. 31, 1997, 351-365 ; Holmes David I. The Analysis of Literary Style – A Review. *The Journal of the Royal Statistical Society*. A, 148 4, 1985, 328–341. Reproduit dans Labbé Dominique, Serant Daniel, Thoiron Philippe. *Etudes sur la richesse et la structure lexicales*. Paris-Genève : Champion-Slatkine, 1988, 67-76.

[▲4] La méthode a été présentée dans trois articles consultables en ligne : **Labbé Cyril & Labbé Dominique. « Inter-Textual Distance and Authorship Attribution Corneille and Molière ». *Journal of Quantitative Linguistics*. 8-3, 2001, 213-231** ; **Labbé Cyril & Labbé Dominique (2003). « La distance intertextuelle ». Corpus. 2, 2003, 95-118** ; **Labbé Dominique. Experiments on Authorship Attribution by Intertextual Distance in English. *Journal of Quantitative Linguistics*, April 2007, 14-1. 33-80**.

[▲5] Pour une démonstration de cette équivalence, voir la scène du *Bourgeois gentilhomme* où le maître de rhétorique explique à monsieur Jourdain les manières de dire : « Belle Marquise, vos beaux yeux »…

[▲6] Muller Charles (1963). Le mot, unité de texte et unité de lexique en statistique lexicologique. Reproduit dans : *Langue française et linguistique quantitative*. Genève-Paris : Slatkine-Champion, 1979, 125-143 ; Muller Charles (1977). *Principes et méthodes de statistique lexicale*. Paris : Hachette. **Labbé Dominique. *Normes de saisie et de dépouillement des textes politiques*. Grenoble : *Cahier du CERAT*,1990**

[▲7] l'influence de la spécialisation du vocabulaire sur l'accroissement de celui-ci a été présentée dans : Hubert Pierre & Labbé Dominique (1988). Un modèle de partition du vocabulaire. In Labbé Dominique, Thoiron Philippe et Serant Daniel. *Etudes sur la richesse et la structure lexicale*. Paris-Genève : Slatkine-Champion, p 93-114.

[▲8] **Labbé Cyril & Labbé Dominique. « La distance intertextuelle ». *Corpus*. 2003 2, p. 95-118**.

[▲9] **Labbé Dominique. 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.**

[▲10] Ce corpus compte au total 1 126 lettres, 292 000 mots et 10 000 vocables différents (**Labbé Cyril & Labbé Dominique (2009). « Existe-t-il un genre épistolaire. Hugo, Flaubert et Maupassant ». *Xe journées de l'ERLA*. Brest novembre 2009)**.

[▲11] Les caractéristiques de certains de ces genres ont été présentées. Outre celle sur le genre épistolaire citée dans la note précédente, voir : **Labbé Cyril & Labbé Dominique. « Baudelaire, Rimbaud et Verlaine ». VIIIe journées de l'ERLA, *Aspects linguistiques du texte poétique*. Brest 16-17 novembre 2007**.

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

[▲13] Merriam Thomas. Intertextual Distances between Shakespeare Plays, with Special Reference to Henry V (verse). *Journal of Quantitative Linguistics*. 9-3. December 2002. p 260-273 ; **An Application of Authorship Attribution by Intertextual Distance in English. *Corpus*. 2. 2003. p 167-182** ; Intertextual Distance, Three Authors. *Literary and Linguistic Computing*. 18-4. November 2003. p. 379-388

[▲14] **Tuzzi Arjuna, Popescu Ioan-Iovitz & Altmann Gabriel. Quantitative Analysis of Italian Texts. *Studies in Quantitative Linguistics*. 6, 2010.**

[▲15] La plume de l'ombre a confirmé publiquement les deux listes. **Monière Denis & Labbé Dominique. 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**

[▲16] Lafon Michel & Peeters Benoît (2006). *Nous est un autre*. Paris, Flammarion ; **Labbé Dominique (2004a). *Romain Gary et Emile Ajar*. Grenoble : Cerat-IEP, mai 2004**.

[▲17] **Labbé Cyril & Labbé Dominique (2007b). *Corneille a écrit 16 pièces représentées sous le nom de Molière. Réponses à : Viprey Jean-Marie et Ledoux Claude-Nicolas, 'About Labbé's "Inter-textual Distance"*. Grenoble : PACTE-IEP**

[▲18] Young Bert E. & Grace P. *Le registre de La Grange*. Genève, Slatkine, 1997.

[▲19] **Labbé Cyril & Labbé Dominique (2005). « How to Measure the Meanings of Words ? Amour in Corneille's Work ». Language Resources Evaluation. 39, p. 335-351.**

[▲20] **Labbé Cyril & Labbé Dominique. « 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**

Dominique. Ce que disent les phrases de Corneille et Molière. Communication devant les Xe Journées Internationales d'Analyse des Données Textuelles. Rome : 11 juin 2010.

[▲21] Pour le détail : Labbé Dominique. *Qui a écrit Tartuffe ?* Montréal : Monière et Wollank, 2009. Réédition : *Si deux et deux sont quatre Molière n'a pas écrit Don Juan*. Paris : Max Milo.

[▲22] Dulait Suzanne. *Inventaire raisonné des autographes de Molière*. Genève : Droz, 1967.

[▲23] Young Bert E. & Grace P. *Op. cit.*

[▲24] Les articles de P. Louÿs sont reproduits en annexe de Boissier Denis. *L'affaire Molière*. Paris : Jean-Cyrille Godefroy, 2004.

[▲25] Poulaille Henry. *Corneille sous le masque de Molière*. Paris : Grasset, 1957.

[▲26] Wouters Hippolyte & Ville de Goyet Christine (de). *Molière ou l'auteur imaginaire ?* Bruxelles : Complexe, 1990.

Crédits images

Pour citer cet article : **Cyril Labbé** et **Dominique Labbé**, **La classification des textes**. *Images des Mathématiques*, CNRS, 2011. En ligne, URL : **http://images.math.cnrs.fr/La-classification-des-textes.html**