Stylométrie : analyse des caractères typographiques utilisés par un auteur

Olivier Dulcy

Plan

1 Préambule

L'objectif de cette étude est d'obtenir une signature pour un auteur à partir d'une analyse des caractères typographiques utilisés par ce dernier. Dans son article intitulé *Empirical evaluations of language-based author identification techniques* publié en 2001, Carole Chaski étudia 10 hypothèses pour tenter de différencier des auteurs : elle conclut que seule l'analyse de la syntaxe des phrases et de la ponctuation constituent la meilleure signature possible des auteurs, afin de pouvoir les distinguer par l'analyse de leur texte.

> Nous nous concentrerons dans ce travail sur l'analyse de la ponctuation.

2 Introduction

De nos jours, on peut identifier une personne avec des techniques qui ont fait leur preuve, comme par exemple l'empreinte digitale. Le problème d'identification demeure actuellement plus difficile pour ce qui est de l'écrit puisqu'on a pour seul support le texte. Il est en effet arrivé qu'une telle identification soit requise dans le domaine de la criminologie (affaire Roy Van Wyk en 2001) [1]. Des expertises ont également été réalisées par le passé pour attribuer une œuvre à un auteur, notamment pour savoir si certaines œuvres de Shakespeare étaient le fruit d'une collaboration, tel que *Pericles, Prince of Tyre*.

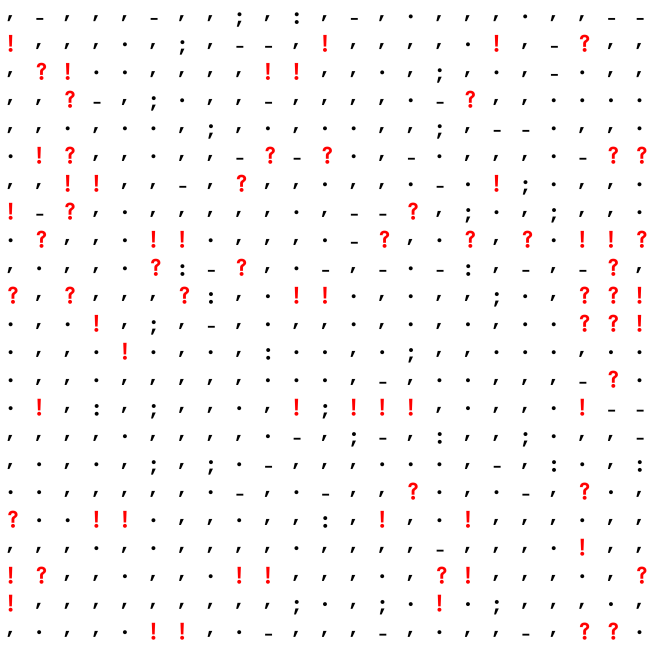
3 Modalités d'action

3.1 Approche qualitative : Étude de faisabilité

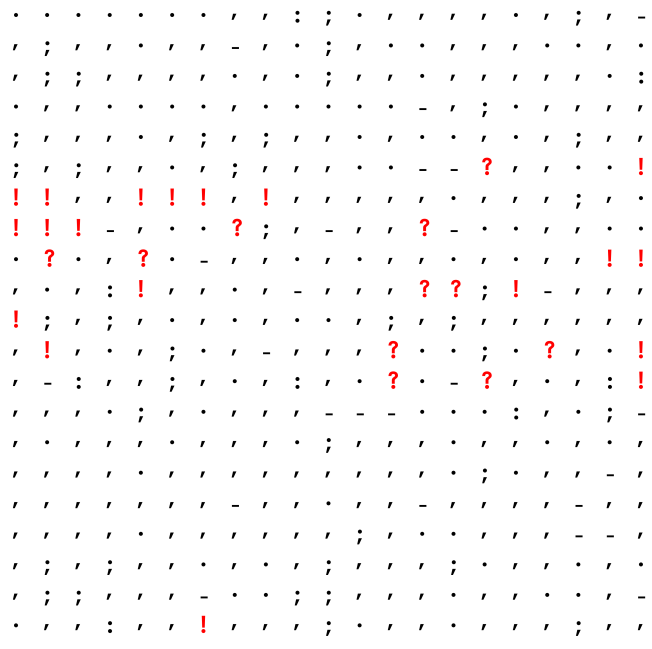
3.1.1 Carte de ponctuation

Une première approche consiste à observer s'il existe bel et bien une tendance, que l'on nommera signature par la suite. Pour cela, on peut relever la ponctuation chez deux auteurs et établir une « carte » qui reflète visuellement le relevé de la ponctuation utilisée par ces auteurs. Voici ce que donne un échantillon :

Alexandre Dumas
Le collier de la reine



Jules Verne
Cinq semaines en ballon



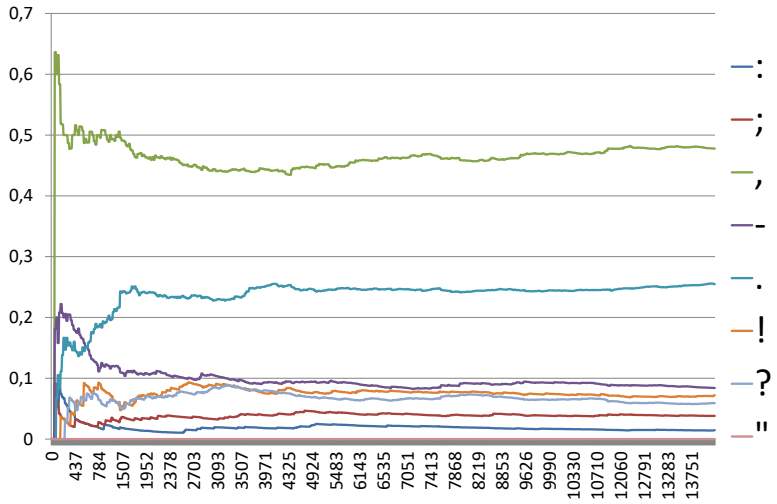
Blanc . , « » ; - : \ () []

Rouge ? !

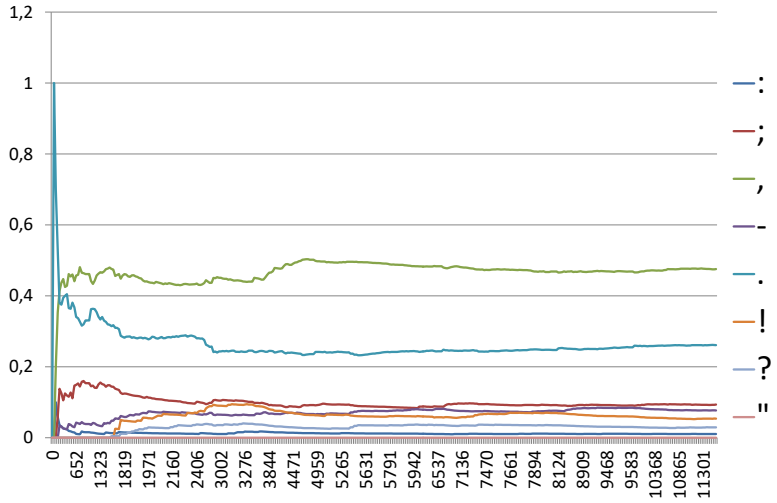
On peut constater chez Alexandre Dumas une présence plus prononcée de points d'interrogation et d'exclamation par rapport à Jules Verne. D'une manière générale, il a été possible d'observer des cartes très variées entre les différents auteurs parmi tous les livres.

3.1.2 Stabilité des proportions de la ponctuation en fonction du nombre de mots

Ceci nous amène alors à chercher à partir de combien de mots on peut observer cette tendance. Voici ce que donne chez nos deux auteurs le relevé des proportions de la ponctuation en fonction du nombre de mots (on peut, par analogie, comparer ce relevé à celui de la mesure d'une grandeur en fonction du temps) :



(a) Alexandre Dumas - *Le collier de la reine*



(b) Jules Verne - *Cinq semaines en ballon*

FIGURE 1 – Distribution de la ponctuation (en %) en fonction du nombre de mots

A partir de 1000 mots environ, on observe le début d'une stabilité dans la distribution de la ponctuation. Par analogie avec la physique, il y a un « un régime transitoire ». Empiriquement, ce seuil des 1000 mots est survenu plusieurs fois au cours de l'analyse des œuvres. Cette analyse montre l'existence d'une signature stable et qu'un échantillon de 1000 mots suffit pour proposer une signature d'œuvre d'un auteur.

3.2 Approche quantitative : le test d'indépendance du χ^2 de Pearson

Après avoir mené une analyse qualitative, nous allons mener une approche quantitative. Les différents programmes développés en Python vont nous fournir plusieurs relevés de la ponctuation chez nos auteurs. Nous allons alors traiter ces données grâce à un outil statistique : le test du χ^2. On pose au préalable une

hypothèse, appelée hypothèse nulle et on souhaite la confronter à nos valeurs. Le test du χ^2 se décline en 3 versions, dont l'une est le test d'indépendance.

Dans le cas du test d'indépendance du χ^2, on souhaite déterminer si les lignes et les colonnes d'un tableau croisé (c'est à dire les deux variables étudiées) ne sont pas indépendantes. Ainsi, l'objet du test d'indépendance du χ^2 est de savoir à partir de quand on peut dire que les variations observées sont dues au hasard, et à partir de quand on peut estimer qu'elles sont dues à un lien entre les deux variables. On dispose de la formule suivante :

Test du χ^2 de Pearson

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

- $O_{i,j}$: observation de type i, j
- r : nombre de lignes, c : nombre de colonne (étiquettes et total exclus)
- p_i : probabilité d'être de type i
- p_j : probabilité d'avoir le caractère j
- $E_{i,j}$ = Effectif total$\times p_i p_j$ la fréquence théorique de type i, j sous l'hypothèse d'indépendance
- Degré de liberté $= (r-1) \times (c-1)$

Il a été démontré en mathématiques, grâce notamment au théorème Centrale Limite, que le test du χ^2 de Pearson tend vers la loi du χ^2 lorsque le nombre d'observations tend vers $+\infty$:

Loi du χ^2 à n degrés de liberté notée $\chi^2(n)$

Soient $X_1, \ldots, X_n$ variables aléatoires indépendantes, de même loi normale, de moyenne respective μ_i et d'écart-type σ_i. On considère leurs variables centrées réduites $Y_i = \frac{X_i - \mu_i}{\sigma_i}$.

$$\chi^2 = \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$$

Connaissant le résultat du χ^2 et le degré de liberté, il est alors possible de calculer la « p-value ». Il s'agit de la probabilité, pour un modèle statistique, d'obtenir un résultat supérieur ou égal sous l'hypothèse H_0 à celui fourni par les observations. Cette valeur est donnée par les tables du χ^2 et par des logiciels statistiques.

Ensuite, on va comparer cette valeur à un seuil préalablement défini fixé noté α (par convention, $\alpha = 0,05$) qui représente la fiabilité du test. Il existe deux types d'erreur : on peut rejeter l'hypothèse nulle alors que cette dernière est vraie (erreur de type I, c'est α) ou accepter l'hypothèse nulle alors qu'elle est fausse (erreur de type II). Ainsi, on compare la p-value à ce seuil :

- Si elle est inférieur à ce seuil, on peut rejeter l'hypothèse d'indépendance au risque de se tromper d'une probabilité de α.
- Dans le cas contraire, on ne peut pas rejeter l'hypothèse d'indépendance.

Remarque : La p-value n'est pas la probabilité que l'hypothèse nulle soit vraie, il s'agit d'un indicateur pour rejeter ou non l'hypothèse d'indépendance.

4 Résultats

4.1 Restitution des résultats

L'étude a été menée sur 11 auteurs français, en considérant pour chaque auteur 3 ou 4 œuvres. Les graphes sont regroupés à la fin du rapport.
Voici les données brutes :

Auteur	:	?	,	-	.	!	?	**Ponctuation total**	Nombre de Mots
Alexandre Dumas	70	236	2372	377	999	255	232	4541	33103
Gustave Flaubert	44	176	1618	172	768	66	44	2888	20189
Jules Verne	62	326	3622	774	1860	390	212	7246	48641
Maupassant	120	109	1948	153	916	127	100	3473	23779
Montesquieu	40	77	348	36	210	1	6	718	5941
Marcel Proust	245	329	6482	844	1984	145	118	10147	98542
George Sand	83	417	3101	373	1448	123	72	5617	49481
Stendhal	64	247	1315	306	885	71	45	2933	21806
Paul Verlaine	55	38	1488	175	323	119	51	2249	16785
Victor Hugo	43	108	905	104	509	187	48	1904	10957
Emile Zola	81	128	3432	365	1222	258	132	5618	33635

TABLE 1 – Signature des auteurs

4.2 Analyse

4.2.1 Peut-on différencier des auteurs via la ponctuation ?

La comparaison a été effectuée à chaque fois entre 2 auteurs. Prenons le cas d'Alexandre Dumas et comparons le aux autres auteurs. L'hypothèse zéro notée H_0 est la suivante : Est ce que la ponctuation est indépendante de l'auteur ? Les résultats sont regroupés dans le tableau suivant, pour un degré de liberté égal à 6.

	Flaubert	Jules Verne	Maupassant	Montesquieu	Marcel Proust	George Sand	Stendhal	Paul Verlaine	Victor Hugo	Émile Zola
χ^2	142	72	125	120	516	249	207	172	89	164
p	10^{-28}	10^{-13}	10^{-24}	10^{-23}	10^{-108}	10^{-51}	10^{-42}	10^{-34}	10^{-17}	10^{-33}

Les valeurs obtenues de p sont toutes inférieures à $0,05$, on peut donc rejeter l'hypothèse d'indépendance.

> Le test de Pearson permet de différencier systématiquement les auteurs, en calculant la signature sur 3 ou 4 œuvres complètes, au seuil de 5%

4.2.2 Peut-on associer objectivement un texte à son auteur ?

❑ ENTRE 2 PAGES D'UNE MÊME ŒUVRE :

Tout d'abord, vérifions si le test du χ^2 nous permet bien d'associer 1 page « inconnue » à son œuvre correspondante (donc à son auteur). Cet exemple repose sur la comparaison de deux pages « similaires », c'est à dire même style de texte : dialogue, description... et environ de même longueur de texte. Prenons par exemple le texte de Jules Verne *Voyage au centre de la Terre*.

Hypothèse zéro : Existe-t-il un lien entre 2 pages d'une même œuvre de Jules Verne ? (*Voyage au centre de la Terre*) : $\chi^2 = 3,88$ avec 6 degrés de liberté $\Rightarrow p = 0,69 \geqslant 0,05$

Auteur	:	;	,	-	.	!	?	Total
J.V Page 1	1	2	78	9	53	23	6	172
J.V Page 2	2	6	112	23	99	38	15	295
	3	8	190	32	152	61	21	467

> L'hypothèse d'indépendance ne peut donc a priori par être rejetée au seuil de 5%, il y a donc un lien entre les 2 pages

❑ Entre une œuvre et la signature d'un auteur :
Maintenant, qu'en est-il si on considère une œuvre et la signature de l'auteur (c'est à dire les nombres qui caractérisent l'auteur d'après le tableau 1 à la page 5)

Hypothèse zéro : Existe-t-il un lien entre une œuvre et la signature d'un même auteur ? (ici, Paul Verlaine)
$\chi^2 = 32,96$ avec 6 degrés de liberté $\Rightarrow p = 1 \times 10^{-5} \leqslant 0,05$

Auteur	:	;	,	-	.	!	?	Total
Mes Hôpitaux	5	4	90	13	48	2	1	163
Signature	55	38	1488	175	323	119	51	2249
	60	42	1578	188	371	121	52	2412

On rejette l'hypothèse d'un lien entre une œuvre et la signature d'un même auteur, au seuil de 5% (Échec d'association)

Cet échec d'association montre les limites du test du χ^2. En effet, cela montre l'existence de liens avec d'autres paramètres non pris en compte tels que le type du passage (est-ce une description, un dialogue... ?). L'association entre deux pages d'une même œuvre a pu être un succès puisqu'elles étaient d'un même style, ce qui n'est pas le cas ici.

5 Conclusion

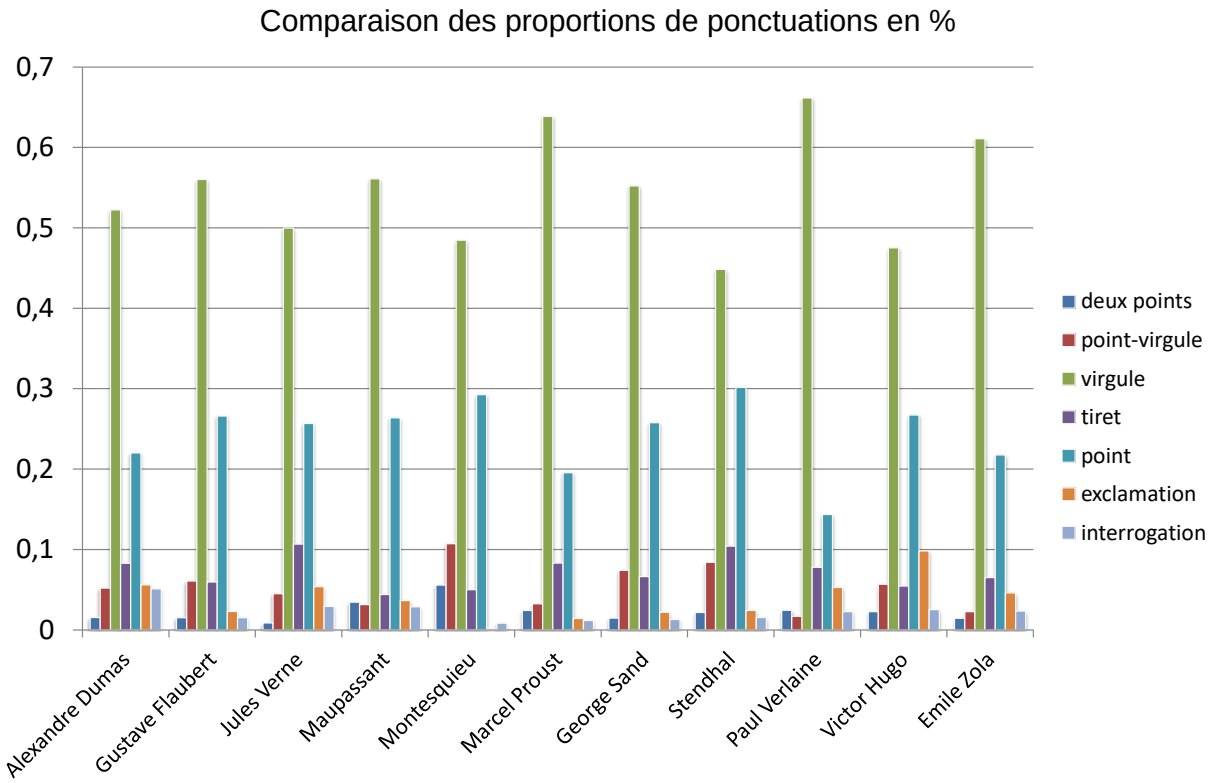
Différentiation des auteurs via la ponctuation : ✔

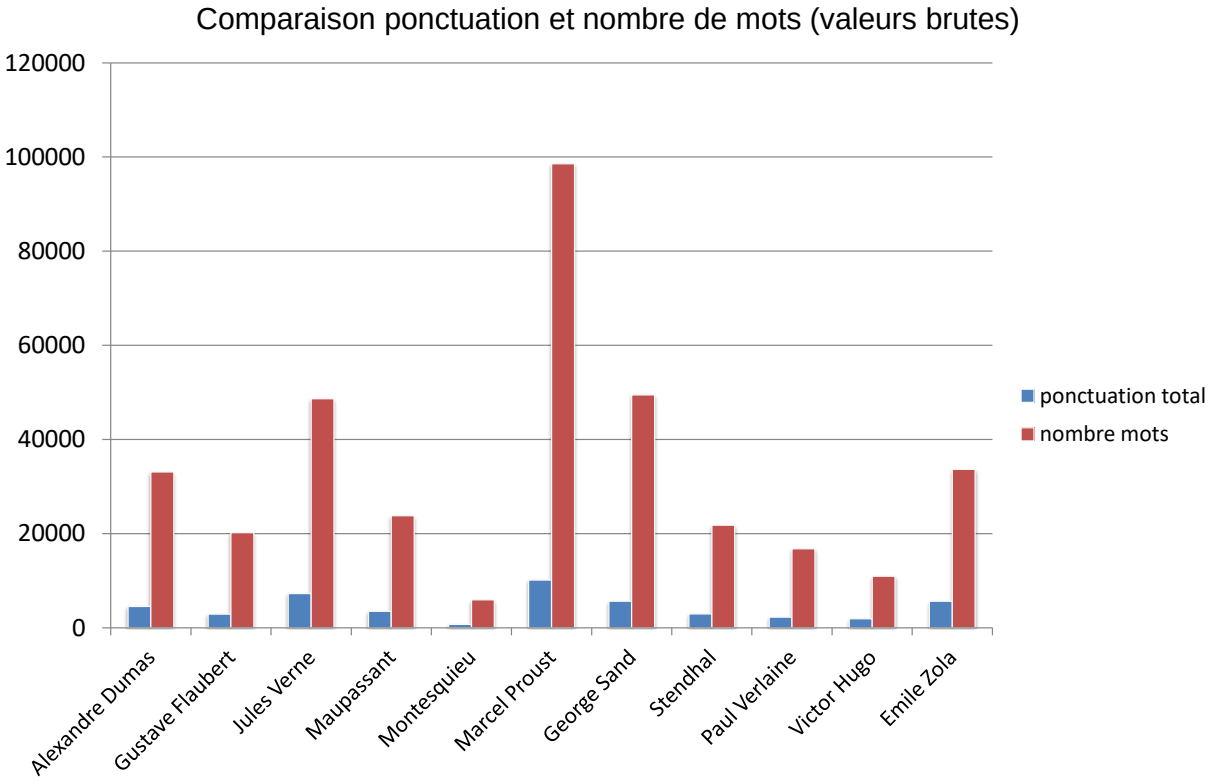
Corrélation entre 2 pages d'un même auteur : ✔

Association entre une œuvre et la signature d'un auteur : ✘

Il faut en moyenne plus de 1000 mots pour établir une signature, ce qui n'est pas toujours possible d'obtenir. Les logiciels développés ont un temps d'exécution de $13,5$ s pour traiter 42 fichiers (1000 à 40000 mots), ce qui demeure inférieur à celui nécessaire pour analyser et interpréter toutes les valeurs obtenues via ce processus. Il serait pertinent d'observer si l'on pourrait aboutir aux mêmes résultats indépendamment de la langue utilisée.

Graphes


Comparaison des proportions de ponctuations en %
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Alexandre Dumas
Gustave Flaubert
Jules Verne
Maupassant
Montesquieu
Marcel Proust
George Sand
Stendhal
Paul Verlaine
Victor Hugo
Emile Zola
deux points
point-virgule
virgule
tiret
point
exclamation
interrogation


Comparaison ponctuation et nombre de mots (valeurs brutes)
120000
100000
80000
60000
40000
20000
0
Alexandre Dumas
Gustave Flaubert
Jules Verne
Maupassant
Montesquieu
Marcel Proust
George Sand
Stendhal
Paul Verlaine
Victor Hugo
Emile Zola
ponctuation total
nombre mots

Liste des œuvres utilisées

Alexandre Dumas

- Le Collier de la reine
- Le Comte de Monte-Cristo, tome I
- Le Docteur mystérieux
- Le Maître d'armes

Gustave Flaubert

- L'éducation sentimentale
- Madame Bovary. Mœurs de province
- Œuvres de jeunesse I
- Trois contes

Jules Verne

- Vingt mille lieues sous les mers
- Autour de la Lune
- Cinq semaines en ballon
- De la Terre a la Lune

Maupassant

- Les Dimanches d'un bourgeois de Paris
- Le Horla
- Le Père Milon
- Le Rosier de Mme Husson

Montesquieu

- Essai sur le goût
- Lettres Familières Du Président de Montesquieu, Baron de La Brede, à Divers Amis
- Lysimaque

Marcel Proust

- À la recherche du temps perdu I
- À la recherche du temps perdu II
- À la recherche du temps perdu III
- À la recherche du temps perdu IV

George Sand

- La Dernière Aldini
- La Mare au diable
- La Marquise
- La Petite Fadette

Stendhal

- Le Chevalier de Saint-Ismier
- Le Coffre et le Revenant
- Le Juif
- Le Philtre

Paul Verlaine

— Mes hôpitaux
— Mes prisons
— Odes en son honneur
— Parallèlement

Victor Hugo

— La Fin de Satan
— Les Contemplations
— Le Roi s'amuse

Émile Zola

— Germinal
— « J'accuse… ! »
— La Bête humaine
— L'Assommoir

Références

[1] C. E. Chaski, Empirical evaluations of language-based author identification techniques, Forensic Linguistics, 8 (2001), pp. 1–65.