



# STYLOMÉTRIE : ATTRIBUTION D'AUTEUR ET PROFILAGE

## Projet Scientifique Collectif *Rapport final*

2016-2017

---

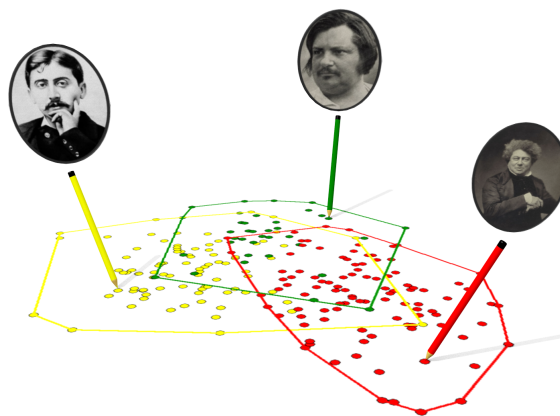
Guillaume Dalle, Jasmine Gamblin, Maxime Godin,  
Clément Mantoux, Wang Sun, Lucile Vigué



# TABLE DES MATIÈRES

<b>1</b>	<b>Fondements théoriques</b>	<b>4</b>
1.1	Stylométrie : mesure du style . . . . .	4
1.2	Apprentissage statistique . . . . .	5
1.2.1	Algorithmes de classification non supervisée . . . . .	6
1.2.2	Algorithmes de classification supervisée . . . . .	7
1.3	Méthodes de vérification . . . . .	8
1.3.1	Average Similarity . . . . .	9
1.3.2	Unmasking . . . . .	10
<b>2</b>	<b>Notre outil et son fonctionnement</b>	<b>12</b>
2.1	Présentation chronologique du projet . . . . .	12
2.2	Formulation des problèmes . . . . .	15
2.3	Fonctionnement interne . . . . .	16
2.3.1	Prétraitement des textes . . . . .	16
2.3.2	Analyse . . . . .	18
2.3.3	Classification . . . . .	18
2.4	Analyse et représentation des résultats . . . . .	19
2.4.1	Sortie textuelle . . . . .	19
2.4.2	Sortie graphique . . . . .	19
2.4.3	Importance des composantes . . . . .	21
<b>3</b>	<b>Applications littéraires</b>	<b>23</b>
3.1	Les époques et le genre . . . . .	23
3.1.1	Étude des périodes . . . . .	24
3.1.2	Étude du genre des romans . . . . .	26
3.1.3	Étude du sexe des auteurs . . . . .	27
3.1.4	Bilan des études de cas . . . . .	27
3.2	Un corpus naturaliste . . . . .	28
3.3	Vérité et mensonges . . . . .	31
3.3.1	Récolte du corpus . . . . .	31
3.3.2	Résultats . . . . .	32
3.3.3	Interprétation . . . . .	32
<b>4</b>	<b>Evaluation</b>	<b>35</b>
4.1	Performance des algorithmes . . . . .	35
4.1.1	Segmentation des oeuvres . . . . .	35
4.1.2	Classifieurs . . . . .	35
4.1.3	Analyseurs . . . . .	36
4.1.4	Autres critères . . . . .	37
4.2	Critique de nos modèles . . . . .	39

# INTRODUCTION



A la croisée des statistiques et de la linguistique, la stylométrie (voir [1] ou [2]) est la discipline qui cherche à caractériser le style d'un texte, permettant d'identifier son auteur, son époque ou encore son genre. De manière duale, on peut également chercher à établir les caractéristiques stylistiques distinctives de tel auteur, époque ou genre. Le développement récent des techniques d'apprentissage statistique remet au goût du jour cette discipline, avec des algorithmes intelligents capables d'analyser la masse d'information présente dans un texte.

L'attribution littéraire est un problème très ancien : depuis Homère jusqu'à Shakespeare, Molière ou Dumas, nombreux sont les auteurs dont certaines oeuvres sont contestées ou partagées. Il existe des cas où la paternité d'un texte n'a jamais pu être clairement établie par des méthodes historiques conventionnelles. Au cours de ce projet, nous nous sommes ainsi initialement intéressés à l'application des outils de la stylométrie à ce problème. Nous nous sommes donc rapidement fixé comme objectif la mise au point d'un outil informatique d'analyse stylistique de textes. En parallèle, pour valider notre outil, nous avons souhaité le tester sur des auteurs connus, puis sur des textes dont la paternité est contestée (voir [3] pour une introduction à ce type d'études).

L'outil, que nous avons présenté à l'occasion du rapport intermédiaire, a été finalisé. Un corpus de textes issus d'époques et de genres divers a été recueilli. L'interface graphique est désormais opérationnelle et nous avons couplé les fonctionnalités à notre corpus via une base de données. Parallèlement, alors que nous avons livré dans le rapport intermédiaire nos conclusions concernant quatre controverses littéraires variées, nous présentons ici des applications différentes. Il ne s'agit en effet plus seulement de déterminer l'auteur d'un texte, mais désormais d'étudier également les caractéristiques générales d'un style. Nous présentons donc trois études portant respectivement sur la détermination du genre et de l'époque de l'auteur, sur les caractéristiques stylistiques de l'école naturaliste et sur la détection de mensonges.

Nous présentons ainsi dans ce rapport les fondements théoriques de nos travaux, une description détaillée du fonctionnement de notre outil, puis trois études de cas et enfin une analyse critique des performances de nos algorithmes et des points faibles de nos modèles.

# 1

## FONDEMENTS THÉORIQUES

---

### 1.1 STYLOMÉTRIE : MESURE DU STYLE

---

Commençons par présenter le paradigme central de notre projet. Nous disposons à l'origine d'un certain nombre de textes que nous souhaitons pouvoir classer dans des catégories. Pour cela, la première étape de notre travail relève du domaine de la stylométrie : nous cherchons à associer un vecteur à partir de chacun des textes, les coordonnées du vecteur devant retenir le maximum d'informations de nature stylistique sur le texte considéré. Une fois cette étape effectuée, le problème sera réduit à une question de classification de vecteurs dans un espace stylistique donné. La qualité de cette étape préalable détermine donc le succès de nos classifications futures, car celles-ci dépendent des informations contenues dans les coordonnées des vecteurs représentant les textes.

Ce travail stylistique comporte deux étapes principales : le pré-traitement du texte puis l'application d'analyseurs. Le pré-traitement permet de passer d'un texte brut à un objet plus structuré. Les mots sont séparés et leurs racines indentifiées, les catégories grammaticales sont étiquetées. On obtient ainsi un objet nettement plus complet qu'un simple texte brut et qui se prête davantage à l'application d'analyseurs. Ce travail de prétraitement n'a pas été un sujet d'études en soi de notre projet et nous nous sommes reposés sur des outils préexistants : NLTK [4] et TreeTagger [5].

La stylométrie repose comme on l'a dit sur le calcul d'informations quantitatives portant sur le texte à étudier. C'est ainsi que nous pouvons représenter un texte comme un vecteur d'un espace stylistique de grande dimension. La richesse de la langue appelle nécessairement une grande variété dans les fonctions d'analyse des textes. Afin d'appréhender la globalité de cette richesse, nous avons donc développé des fonctions d'analyse dans quatre grands domaines.

#### **Grammaire et syntaxe :**

Dans cette catégorie d'analyseurs, on s'intéresse à la nature grammaticale des mots du texte. Par souci d'universalité, nous avons retenu les "parts of speech" du lemmatiseur TreeTagger pour catégoriser les mots.

Un premier analyseur évident est le vecteur des fréquences des catégories grammaticales des mots. L'esprit de cet analyseur est donc de capter les préférences de l'auteur du texte en terme de choix grammatical : utilisation des adjectifs, des pronoms, des systèmes de temps, etc.

Afin d'enrichir cette approche fréquentiste, nous avons également cherché à modéliser les enchaînements des différentes catégories grammaticales : il s'agit de se concentrer sur la syntaxe. L'approche retenue a été la modélisation par une chaîne de Markov [6] des transitions dans l'espace des catégories grammaticales. Un second analyseur est donc l'estimation par maximum de vraisemblance de la matrice de transition (plongée canoniquement dans  $\mathbb{R}^{N^2}$ ). Nous avons également cherché à étudier les effets à distance supérieure à un : plutôt que de s'intéresser

aux transitions entre  $n$ -uplets de catégories grammaticales, approche qui mène directement à une explosion du nombre de dimensions, nous avons introduit un paramètre de saut  $s$  qui va permettre de considérer les transitions à distance supérieure à un. La suite des catégories grammaticales est donc lue en ignorant  $s - 1$  mot(s) sur  $s$ .

Cette modélisation markovienne de la syntaxe permet également d'envisager des mesures de la complexité de la langue propre au texte du point de vue grammatical. Par exemple, l'écart en norme ( $L^\infty$  ou  $L^2$ ) de l'estimation de la matrice de transition par rapport à l'isobarycentre des matrices de transition estimées sur le corpus entier permet de mesurer un écart au style canonique du corpus. Nous pouvons ainsi identifier des styles qui utilisent des enchainements de catégories grammaticales peu courants.

### Ponctuation :

La ponctuation est également riche d'information sur le style d'un auteur. Nous avons donc choisi de prendre en compte les fréquences des différents signes de ponctuation de chaque langue pour former un analyseur.

La ponctuation est aussi un délimiteur de phrases. Nous avons donc cherché à modéliser la longueur des phrases. A ce titre, nous avons noté avec intérêt qu'une loi log-normale semble approcher la distribution des longueurs des phrases. Dans la majorité des cas, l'écart à cette distribution se fait avec une asymétrie vers la gauche (prédominance légère des phrases plus courtes). Ainsi, au-delà des statistiques de base sur la longueur des phrases (moyenne, écart-type, quantiles, minimum, maximum), nous avons aussi pris en compte estimé les paramètres d'une loi log-normale sur cette distribution. Afin de prendre en compte l'asymétrie empirique constatée, nous avons également estimé le coefficient d'asymétrie et le kurtosis de la distribution.

### Graphèmes :

Les graphèmes (plus petites entités du système d'écriture, lettres ou signes) sont également porteurs d'information. Dans ce registre, nous avons formé un analyseur qui mesure les fréquences relatives des différents graphèmes. Nous avons également généralisé cette approche aux fréquences des  $n$ -grammes ( $n$ -uplets de graphèmes).

### Vocabulaire :

Pour ce qui est du vocabulaire, nous avons retenu la méthode des stop-words, qui consiste à comparer les fréquences relatives de listes de mots prédéterminés et dépendants de la langue (listes établies par NLTK). Nous avons aussi utilisé différents indices synthétiques de la richesse du vocabulaire qui se basent sur le nombre de mots différents rapporté au nombre de mots total du texte.

## 1.2 APPRENTISSAGE STATISTIQUE

---

L'objet de notre étude est de représenter quantitativement le style d'un texte. Pour ce faire, on dispose d'un certain nombre de caractéristiques stylistiques susceptibles de différencier deux styles. Ces caractéristiques numériques, regroupées, forment un vecteur qui localise l'œuvre

dans un espace stylistique donné. Dire que deux textes sont écrits dans le même style revient ainsi à dire que leurs vecteurs stylistiques sont proches selon une certaine norme. Le choix des composantes étudiées, ainsi que l'importance accordée à chacune de ces composantes, sont naturellement susceptibles d'influer sur la proximité de deux œuvres.

Étant donné un espace stylistique, on s'intéresse maintenant aux méthodes statistiques et algorithmiques nous permettant de résoudre le problème de regroupement en catégories stylistiques : dans le formalisme précédent, un style s'apparente à une région de l'espace stylistique. Dès lors, il s'agit d'identifier ces régions pour un corpus d'œuvres donné, afin d'étudier l'appartenance, ou non, de certains textes à ces régions. On peut ainsi se demander si le point correspondant à l'œuvre *Les Trois Mousquetaires* se trouverait plutôt dans la région stylistique d'Alexandre Dumas ou d'Auguste Maquet. Nous avons ainsi utilisé différents outils de classification automatique (voir notamment [7] ou [8]) afin de les comparer et de s'adapter à des situations multiples.

### 1.2.1 • ALGORITHMES DE CLASSIFICATION NON SUPERVISÉE

L'objectif des algorithmes de classification non supervisée est, étant donné un ensemble de points, d'en trouver la partition la plus vraisemblable ([9], [10]). Pour ce faire, les points sont regroupés par ensembles partageant des caractéristiques communes. Initialement portés sur différents types d'algorithmes de clustering (clustering hiérarchique, algorithme EM, ...), nous avons centré nos travaux sur les algorithmes de classification par partitionnement. Ces algorithmes trouvent, par itérations successives, des optima locaux de certains critères de satisfaction. Stylistiquement parlant, il s'agit donc de trouver quels différents styles sont présents au sein d'un même corpus, et de les différencier.

#### **Algorithmes k-moyennes, k-médoïdes et méthode des nuées dynamiques :**

L'algorithme k-moyennes propose de représenter chaque cluster par son centroïde (isobarycentre). À partir d'un partitionnement initial arbitraire, on effectue des réallocations des données entre les clusters en choisissant de minimiser la distance point-centroïde. Il s'agit d'un algorithme simple, intuitif et qui présente de bonnes propriétés théoriques. À ce titre, il est l'un des algorithmes de classification automatique les plus répandus.

Des alternatives sont possibles en variant le représentant des clusters. On peut choisir, comme dans l'algorithme des k-médoïdes, un représentant central parmi les points du cluster ou alors, comme dans la méthode des nuées dynamiques, de remplacer le représentant unique par un noyau représentatif. Ces variantes de la classification par partitionnement permettent de détacher des clusters plus complexes (eg. non linéairement séparables).

#### **Partition floue et pseudo-partition :**

Une autre voie d'amélioration des algorithmes de partitionnement consiste à affaiblir les hypothèses sur le schéma attendu : il s'agit des partitions floues et des pseudo-partitions. Dans le premier cas, chaque point de l'ensemble des données est supposé appartenir à chaque cluster à un certain degré. On parle de *fuzzyfication*. On obtient alors une représentation matricielle stochastique de la classification en clusters. Les pseudo-partitions quant à elles n'exigent pas

des clusters disjoints, ce qui permet l'appartenance de certains points à plusieurs clusters simultanément.

L'intérêt de cette approche est de fournir un premier moyen de gestion des données aberrantes qui ne vont plus venir perturber la convergence des algorithmes.

### Approche basée sur la densité :

Sur un principe similaire aux méthodes de réallocation des données (eg. méthode des nuées dynamiques), l'approche basée sur la densité cherche à constituer des clusters au sein desquels les points peuvent être reliés en respectant une certaine condition de densité le long du chemin à des objets centraux – les noyaux, qui constituent des points de forte densité. L'algorithme DBSCAN suppose un niveau de densité homogène mais des améliorations comme OPTICS permettent de s'affranchir de ces contraintes. On peut au choix obtenir une partition stricte ou une pseudo-partition des données.

## 1.2.2 • ALGORITHMES DE CLASSIFICATION SUPERVISÉE

Contrairement aux algorithmes de clustering présentés ci-avant, les algorithmes de cette catégorie suppose que l'utilisateur dispose déjà d'une classification a priori d'un certain nombre de points de données. A partir de la connaissance de ces objets et des catégories auxquelles ils appartiennent, ces algorithmes vont identifier les « caractéristiques » de chaque catégorie et s'appuyer sur celles-ci pour classer de nouveaux objets : c'est l'apprentissage supervisé. Ce sont donc des algorithmes qui procèdent en deux phases : une première phase d'analyse des données déjà classées et une seconde phase d'attribution d'une catégorie à des objets inconnus. Tout comme pour les algorithmes de clustering, nous avons étudié et codé plusieurs algorithmes de cette catégorie ( $k$  plus proches voisins, algorithme *a priori*, ...) pour finalement nous focaliser sur trois d'entre eux :

### Classification naïve bayésienne :

Cet algorithme repose sur la formule de Bayes et permet d'en déduire un classifieur. L'algorithme suppose que les différentes coordonnées des vecteurs représentant chaque point de données sont indépendantes, hypothèse raisonnable si les caractéristiques ont été choisies pertinentes.

On dispose d'un vecteur  $x = (x_1, \dots, x_n)$  caractérisant un objet inconnu, et d'un groupe  $C_1, \dots, C_k$  de classes disjointes. Sur l'ensemble d'apprentissage on calcule empiriquement :

- $\mathbb{P}(x \in C_i)$  la probabilité d'appartenir à la classe  $C_i$  ;
- $\mathbb{P}(x_j = a_j \mid x \in C_i)$  la probabilité que la  $j$ -ème coordonnée d'un vecteur ait la valeur  $a_j$  sachant qu'il appartient à la classe  $C_i$ .

Pour classer un nouveau point de données  $(y_1, \dots, y_n)$ , on choisit la catégorie qui maximise la probabilité simultanée d'apparition des coordonnées du nouveau vecteur et d'appartenance à

la classe :  $\mathbb{P}(x \in C_i) \prod_{j=1}^n \mathbb{P}(x_j = y_j \mid x \in C_i)$ .



**Support Vector Machine (SVM) :**

Le SVM ou Séparateur à Vaste Marge cherche à déterminer un hyperplan affine de l'espace  $\mathbb{R}^d$  qui sépare un échantillon de données classées en deux catégories. Un tel hyperplan n'est en général pas unique, le parti pris de l'algorithme consiste à maximiser la distance entre l'hyperplan et le point le plus proche de lui. Une fois l'hyperplan défini, de nouveaux points de données peuvent être classés en déterminant s'ils appartiennent au demi-espace positif ou négatif de l'hyperplan orienté. Les algorithmes de SVM, dans le cas général, appliquent une transformation non linéaire (fonction noyau) permettant de ramener le problème initial à une séparation linéaire. Ils permettent également de séparer plus de deux catégories en utilisant le nombre requis d'hyperplans.

**Perceptron multicouche :**

Dans le modèle du perceptron multicouche [11], les neurones fonctionnent sur le modèle du perceptron. La fonction du perceptron simple est de déterminer de quel côté d'un certain hyperplan le vecteur d'entrée se situe. On retrouve ici le même paradigme que pour l'algorithme SVM. Les éléments du vecteur d'entrée  $x = (x_1, \dots, x_n)$  sont pondérés par les poids  $w = (w_1, \dots, w_n)$ , puis sommés. Pour obtenir la valeur de sortie, on soustrait le biais du résultat obtenu, puis on applique une fonction d'activation  $f$  à valeurs réelles. Pour un vecteur d'entrée  $x$  la valeur de sortie est donc  $f(x \cdot w - b)$ . La fonction  $f$  renvoie généralement 0 pour les valeurs négatives et 1 pour les valeurs positives, de telle sorte que, pour l'hyperplan affine orienté d'équation  $x \cdot w - b = 0$ , le neurone permet de savoir dans quel demi-espace se situe le vecteur d'entrée.

Dans le modèle du perceptron multicouche, les neurones sont alignés par couches, et le vecteur des sorties d'une couche constitue le vecteur d'entrée de la couche suivante. Les poids  $w$  du réseau sont initialement distribués au hasard. L'apprentissage se fait de façon supervisée : on fournit au réseau des données déjà classées, puis on modifie les poids et les biais des neurones en fonction de sa réponse grâce un algorithme de rétro-propagation des erreurs depuis la dernière couche vers la première (méthode de descente du gradient). A la fin de l'apprentissage, le réseau a convergé vers les valeurs optimales des poids et des biais et est donc capable de classer de nouveaux points de données.

## 1.3 MÉTHODES DE VÉRIFICATION

---

La problématique de la vérification de catégorie est un peu différente de celle de l'attribution, puisque souvent on ne dispose pas de candidats alternatifs auxquels comparer le texte controversé. Il s'agit alors de trouver une méthode pouvant délivrer un résultat binaire, éventuellement accompagné d'un indice de qualité, le tout sans données autres qu'un corpus de base rédigé par l'auteur à vérifier.

Nous nous sommes intéressés à deux méthodes, mais à en juger par les tests effectués, la première (Average Similarity) est plus facile d'emploi, son interprétation est plus simple et ses résultats plus fiables dans de nombreuses situations, c'est donc celle que nous privilégions.



### 1.3.1 • AVERAGE SIMILARITY

Cette première méthode, assez élémentaire, est tirée de l'article [12] (et un peu adaptée par la suite). Elle repose sur une mesure de similarité entre les textes vectorisés. Connaissant un corpus appartenant à une catégorie, on calcule la similarité moyenne entre les paires de textes de ce corpus, notée  $AGS$  (Average Group Similarity). Etant donné un texte inconnu, on calcule ensuite la moyenne de sa similarité avec chacun des textes du corpus, notée  $AS_{\text{Unk}}$  (Average Similarity of the Unknown Text).

On effectue ensuite le test suivant : si  $AS_{\text{Unk}} > AGS - \epsilon$ , on juge que le texte est suffisamment proche du corpus pour l'attribuer à la même catégorie, sinon on refuse et on postule l'existence d'une autre catégorie. Notre apport par rapport à l'article original est l'ajout de cette marge  $\epsilon$  qui permet une meilleure souplesse du modèle, via le calibrage.

L'idée de cette méthode est en effet qu'elle peut fonctionner seule, mais que pour la régler au mieux il est utile de lui fournir un ensemble de textes de calibrage. Cet ensemble sera constitué de textes de catégories différentes mais provenant éventuellement du même milieu littéraire, servant à "préparer" l'algorithme à ce qu'il pourra rencontrer lors de la phase de vérification en elle-même. Ici le calibrage consiste à régler  $\epsilon$  de façon optimale, mais on peut envisager des réglages plus complexes (par exemple déterminer d'autres paramètres du modèle) afin d'adapter le programme à une situation précise.

Restent alors à préciser :

- La fonction de similarité : après avoir testé une similarité en cosinus, sans grand succès, nous avons choisi une mesure correspondant à un noyau gaussien, car elle donnait de bons résultats dans nos applications

$$\text{sim}(x, y) = \exp(-\gamma N(x - y)^2) \quad \text{avec} \quad N(x - y) = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2}$$

- La valeur de  $\epsilon$  : celui-ci est fixé à un certain pourcentage  $a$  de l'écart-type sur les similarités entre paires du corpus, plus précisément le pourcentage qui donne la meilleure performance sur le corpus de calibrage. Cette performance est mesurée de telle sorte à minimiser à la fois les faux positifs et les faux négatifs, sans privilégier l'un des deux à outrance (minimisation de l'erreur à deux queues). On a donc choisi de la définir comme :

$$\text{qualite}(\text{fp}, \text{fn}) = (1 - \text{fp}^2)(1 - \text{fn}^2) \frac{\min(\theta, \alpha) \min(\pi/2 - \theta, \alpha)}{\alpha^2}$$

avec les valeurs

$$\theta = \arctan(\text{fn}/\text{fp}) \text{ et } \alpha = 0,4$$

A partir des résultats sur les textes issus du découpage, on peut remonter à l'attribution d'une œuvre entière par démocratie : si une majorité de textes est attribuée à une catégorie par l'algorithme, c'est ce résultat qu'on renvoie, sinon on suppose que l'œuvre est issue d'une autre catégorie. Cela permet en outre de juger du degré de confiance de la réponse (en voyant si l'élection est serrée). La figure 1 présente un exemple de résultat, obtenu sur un corpus composé de romans de Dumas et de Zola.

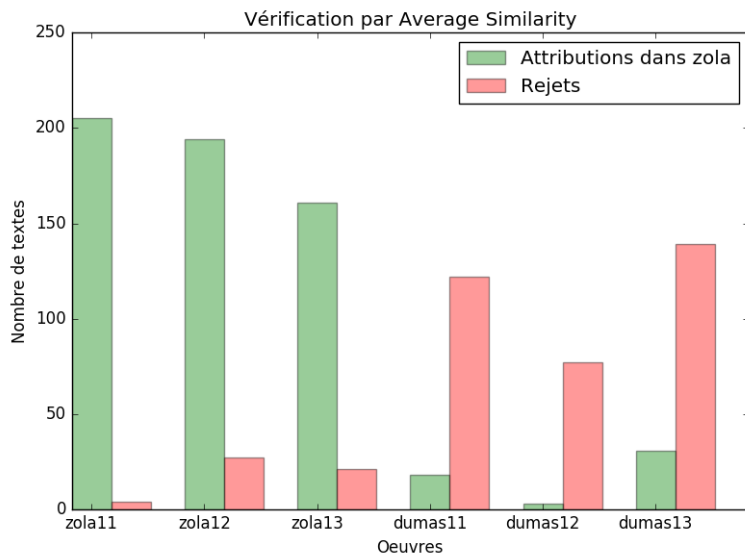


FIGURE 1 – Graphique des résultats obtenus par AS sur un corpus Zola-Dumas

### 1.3.2 • UNMASKING

Une deuxième méthode provient de l'article [13]. Elle vise à valider l'hypothèse selon laquelle les différences entre deux textes de la même catégorie sont superficielles et ne portent que sur quelques composantes stylistiques, tandis que les différences entre deux textes de catégories distinctes sont plus profondes et portent sur un ensemble plus large de caractéristiques.

L'idée consiste donc, étant donné un corpus d'une catégorie donnée et un corpus inconnu, à trier les composantes par ordre d'importance (au sens où celles qui séparent le mieux les paquets dans l'espace stylistiques seront en premier), puis à les retirer au fur et à mesure en évaluant la qualité de la séparation entre paquets.

Notre premier essai fut d'évaluer ladite séparation par des mesures de distance ou de variance élémentaires, mais cette approche ne fut pas fructueuse. Nous avons alors appliqué la méthode proposée dans l'article, à savoir itérer un classifieur en lui fournissant de moins en moins d'informations (au sens du nombre de composantes), et étudier la façon dont sa précision se dégrade. Et en effet, nous avons obtenu :

- Lorsque les catégories étaient identiques, des courbes de dégradation très pentues, avec un rapprochement rapide des deux clusters (le classifieur les distinguant de moins en moins) ;
- Lorsque les catégories étaient différentes, des courbes de dégradation très plates, suggérant une distance qui reste grande même si on enlève les facteurs de séparation a priori les plus importants.

En appliquant ensuite une méthode simple de distance en norme  $L^2$ , on peut classer une courbe de dégradation dans l'une des deux catégories, et ainsi déterminer pour nos textes à vérifier si leur courbe correspond à deux catégories différentes ou à une seule et même catégorie.

Là aussi, une phase de calibrage avec un troisième corpus est donc nécessaire, pour constituer un corpus de courbes permettant de reconnaître la situation qui se présentera dans la phase de vérification à proprement parler.

## 2

## NOTRE OUTIL ET SON FONCTIONNEMENT

### 2.1 PRÉSENTATION CHRONOLOGIQUE DU PROJET

La chronologie du projet (figure 2) s'est déroulée en quatre grandes étapes échelonnées de septembre 2016 à avril 2017. Dans un premier temps, nous avons cherché à définir nos objectifs principaux au fur et à mesure de notre exploration de l'état de l'art de la stylométrie. Dans un second temps, nous avons développé les premières versions de l'outil conçu au cours de notre projet pour répondre aux objectifs posés. Par la suite, nous avons amélioré l'outil en lui ajoutant des fonctionnalités supplémentaires, tout en poursuivant une évolution de grande ampleur sur le fond pour permettre le traitement de problèmes de stylométrie plus généraux. Enfin, au cours de la dernière partie du projet, nous avons mené de front la finalisation de l'outil, sa mise en application à des problèmes de stylométrie et l'évaluation de ses performances et de ses limites.

Dates	Grandes étapes		
septembre 2016 – octobre 2016	Recherches initiales et formulation des objectifs principaux		
novembre 2016 – décembre 2016	Développement du cœur de l'outil		
janvier 2017 – février 2017	Améliorations de l'outil	Evolution vers des problèmes plus généraux	
mars 2017 – avril 2017	Finalisation de l'outil	Applications littéraires	Evaluation de l'outil

FIGURE 2 – Tableau chronologique du projet

*septembre 2016 – octobre 2016*

#### Recherches initiales et formulation des objectifs principaux

La problématique initiale de notre projet scientifique collectif était la vérification de la paternité littéraire des *Trois Mousquetaires* d'Alexandre Dumas. Auguste Maquet a notamment participé à la rédaction de cette œuvre, mais jusqu'à quel point ? Plus généralement, le problème est donc de déterminer l'auteur d'un texte connaissant d'autres textes de cet auteur. Nous avons choisi d'attaquer ce problème par les méthodes de la stylométrie : l'analyse statistique assistée par ordinateur du style littéraire. Au cours de cette première étape, nous avons donc effectué des recherches initiales dans ce domaine, notamment en termes de lemmatisation

(analyse lexicale), d'analyse stylistique quantitative (statistiques caractéristiques d'un style) et de méthodes d'apprentissage statistique de classification. Parallèlement à ces recherches, nous avons entrepris un travail préliminaire d'implémentation des méthodes mises au jour par nos recherches (notamment des algorithmes de classification), en établissant un premier cadre global de résolution des problèmes en cinq étapes : *analyser* – *classifier* – *évaluer* – *interpréter* – *représenter*. Ce travail préliminaire, ainsi que l'analyse de l'état de l'art, ont été présentés dans la proposition détaillée ainsi que lors de la réunion de lancement du projet. En plus de l'objectif principal d'attribution d'auteurs, des objectifs secondaires, dans le domaine du profilage sociologique et de la détection de mensonge, ont été mentionnés. Leur prise en compte effective dans le projet aura lieu plus tardivement, à partir de janvier 2017.

*novembre 2016 – décembre 2016*

### **Développement du cœur de l'outil**

Suite à la réunion de lancement du projet, une phase de développement importante du projet a commencé. Nous avons développé une structure modulaire pour notre outil, sur laquelle nous avons pu par la suite brancher les différents ajouts et modifications ultérieurs. Plus précisément, nous avons mis en place des méthodes d'interprétation des composantes d'analyse (comparaison des valeurs et de leur dispersion entre les catégories, pondération des réseaux de neurones) qui nous ont permis d'affiner notre analyse du style des auteurs classifiés. Nous avons, toujours dans cette optique analytique, également développé un affichage graphique des résultats produits par notre outil : il s'agit concrètement de visualiser dans le plan la classification des textes et de permettre de dilater les composantes d'analyse afin d'observer visuellement leur effet sur la séparation des catégories. Enfin, nous avons mis en place de nouvelles techniques complémentaires de la classification : les méthodes de vérification qui permettent de vérifier qu'un texte appartient à la même catégorie qu'un ensemble de textes en calibrant l'outil avec des textes de catégories différentes.

Cette période s'est conclue par la remise du rapport intermédiaire. Afin de mettre en pratique le bagage théorique et les techniques implémentées que nous avons accumulés, nous avons réalisés quatre études de cas (en anglais, en français et en chinois) sur des problématiques d'attribution de paternité littéraire. Grâce à elles, nous avons pu vérifier que notre outil fournissait des réponses concrètes aux problèmes posés, et nous en avons profité pour juger de manière non quantitative des performances de notre outil.

De manière plus globale, ce bilan d'étape nous a permis de mettre en évidence des lacunes dans notre outil : certains problèmes étaient mal couverts par la structure de notre outil (profilage sociologique par exemple), des composantes d'analyse stylistiques supplémentaires étaient nécessaires, l'évaluation quantitative des performances n'était pas suffisante.

*janvier 2017 – février 2017*

### **Améliorations de l'outil**

Suite au rapport intermédiaire, nous avons donc cherché à combler les lacunes mises en évidence. Nous avons développé des fonctions d'analyse stylistique supplémentaires, basées par exemple sur des modélisations markoviennes de la syntaxe grammaticale ou encore des analyses de séries temporelles de ponctuation.

Dans un autre registre, nous avons cherché à compléter l'interface graphique d'affichage des résultats du traitement par une interface d'entrée conviviale qui permet aux utilisateurs d'utiliser les différentes fonctionnalités de notre outil : définition du problème, des fonctions d'analyse et de la technique de résolution (classification par réseaux de neurones, par SVM ou encore emploi d'une méthode de vérification, etc.)

### Evolution vers des problèmes plus généraux

La conception de notre outil au cours des deux premières périodes s'était concentrée exclusivement sur l'objectif principal d'attribution d'auteurs, en négligeant les objectifs secondaires de profilage sociologique ou de détection de mensonge. Ce manque d'anticipation a par conséquent nécessité un travail d'adaptation de tout notre outil aux problèmes plus généraux vers lesquels nous souhaitions évoluer pour la fin du projet. Le saut conceptuel sous-jacent à ce changement a été de reformuler tous ces nouveaux problèmes sous la forme de problèmes d'attribution de catégories, et plus seulement d'auteurs. Par exemple, déterminer si un style d'écriture féminin existe revient à se demander si l'on peut classer des textes en deux catégories (auteurs masculins et auteurs féminins) en se basant sur un corpus de textes comportant les deux genres d'auteurs. Le degré d'existence de cette écriture sera quantifié par la précision de la classification obtenue. Plus particulièrement, nous avons restructuré notre outil autour de quatre grandes fonctions plus générales qui sont présentées figure 3.

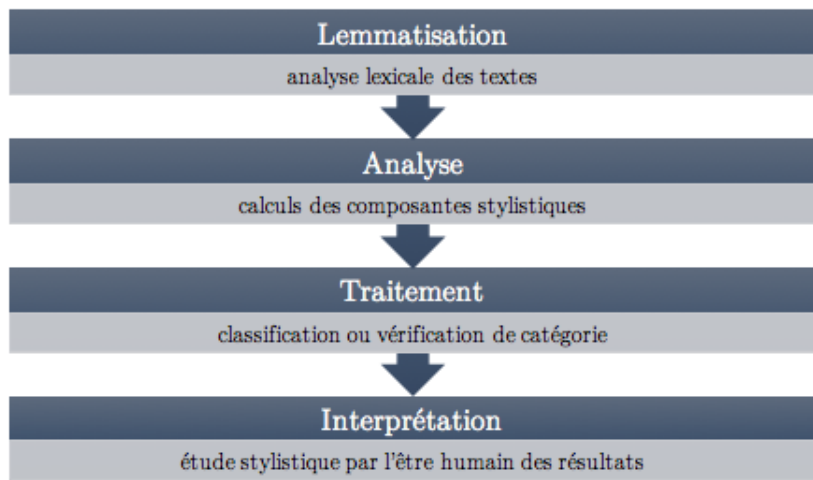


FIGURE 3 – Structure générale de l'outil

*mars 2017 – avril 2017*

### Finalisation de l'outil

La dernière étape du projet a bien entendu nécessité la finalisation de l'outil. Il s'agit d'une étape importante de tests de l'outil pour vérifier la coordination des différents modules qui le constituent et qui permettent une utilisation plus facile pour l'utilisateur. Dans cette optique, nous avons choisi de renforcer notre corpus intégré, en l'enrichissant et surtout en mettant en place une base de données qui permet de constituer plus facilement les corpus spécifiques à

chaque problème que l'on souhaite étudier. Cette base de données permet de rechercher facilement des textes correspondant à un certain nombre de critères comme la période, le pays, la langue ou encore le genre. Afin de rendre l'outil plus personnalisable, il est également possible d'insérer dans la base de données de nouveaux textes qui proviendront des bibliothèques numériques personnelles des utilisateurs. Nous présenterons lors de la soutenance une démonstration de l'outil qui montrera la version finale à laquelle nous aurons abouti.

### Applications littéraires

En parallèle de cette finalisation de l'outil, nous avons choisi de mettre en œuvre ce dernier sur une sélection de problèmes (études du style littéraire en fonction du genre et de l'époque, étude stylistique des naturalistes et mise en évidence des caractéristiques du mensonge écrit). La méthodologie et les résultats auxquels nous sommes parvenus sont présentés dans ce rapport final.

### Evaluation de l'outil

Pour revenir sur une des lacunes soulevées à l'issue du rapport intermédiaire, l'évaluation quantitative des performances de notre outil, nous avons mis en place au cours de cette période un processus de validation croisée. Ce dernier nous permet d'évaluer quantitativement la fiabilité de nos résultats. De manière plus large, nous avons cherché à mener une réflexion sur les qualités et les limites de notre outil dont les conclusions sont disponibles dans ce rapport.

## 2.2 FORMULATION DES PROBLÈMES

---

Nous avons développé une fenêtre d'entrée (figure 4) en utilisant la package tkinter de Python. Deux types de traitement sont disponibles : la classification et la vérification, correspondant aux deux premiers onglets. Le troisième onglet permet l'accès à un outil de recherche de textes couplé à la base de données de l'outil.

Les paramètres nécessaires sont fournis par l'utilisateur dans les différents champs de la fenêtre. Nous pouvons choisir la langue des textes, le classifieur, le nombre des mots dans chaque segment divisant les oeuvres, et les fonctions d'analyse stylistique. Certains analyseurs paramétrisables comme la modélisation markovienne de la syntaxe nécessitent de remplir des champs supplémentaires dans une fenêtre pop-up.

De plus, nous devons décider si nous segmentons les textes ou non, si les textes du corpus d'apprentissage doivent être équilibrés (pas de biais en faveur d'un auteur par exemple) ou encore si le vecteur associé à chaque texte est normalisé. Les textes de l'ensemble d'entraînement et de l'ensemble d'application sont également précisés dans cette fenêtre.

La base de données (onglet 3) permet d'effectuer des sélections rapides de corpus. Nous pouvons choisir les textes sur la tableau avec l'aide de Ctrl et Shift. En double cliquant sur un texte, on le sélectionne. Les boutons *To Training Set* ou *To Evaluation Set* ajoutent la sélection courante dans les ensembles appropriés. À la fin, un bouton Okay permet de lancer la classification. Une fonction *Reset* permet d'effacer les champs de l'onglet. Nous pouvons enfin fermer cette fenêtre en cliquant le bouton *Cancel*. L'onglet Vérification (figure 7) présente une



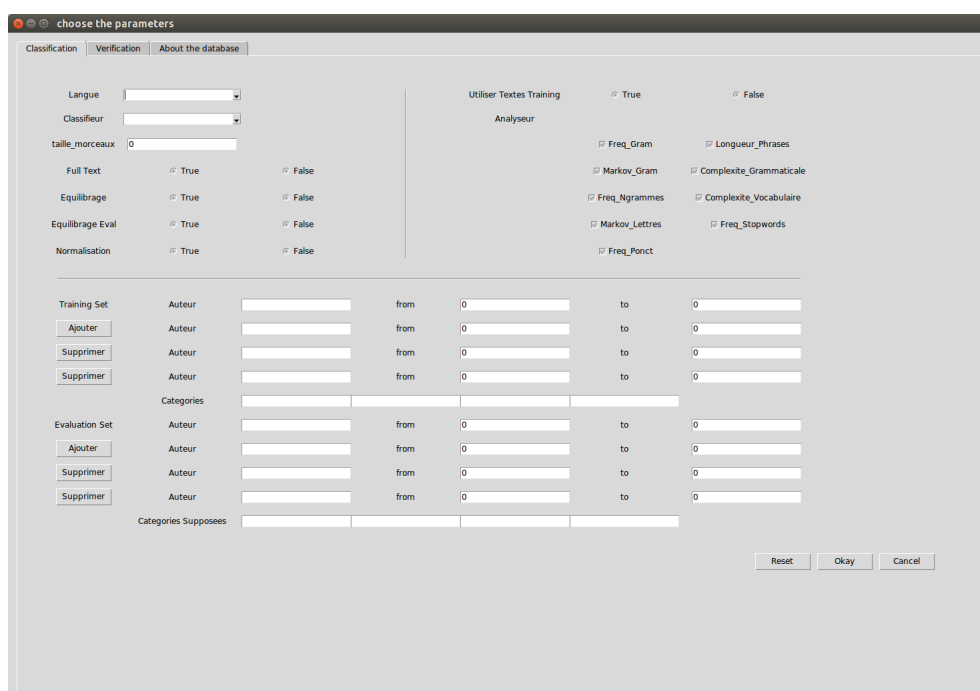


FIGURE 4 – Fenêtre d'accueil de l'outil

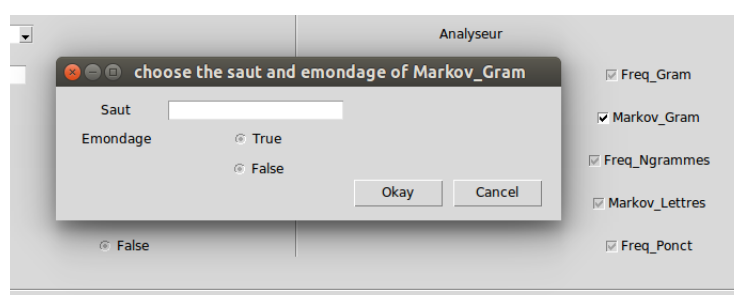


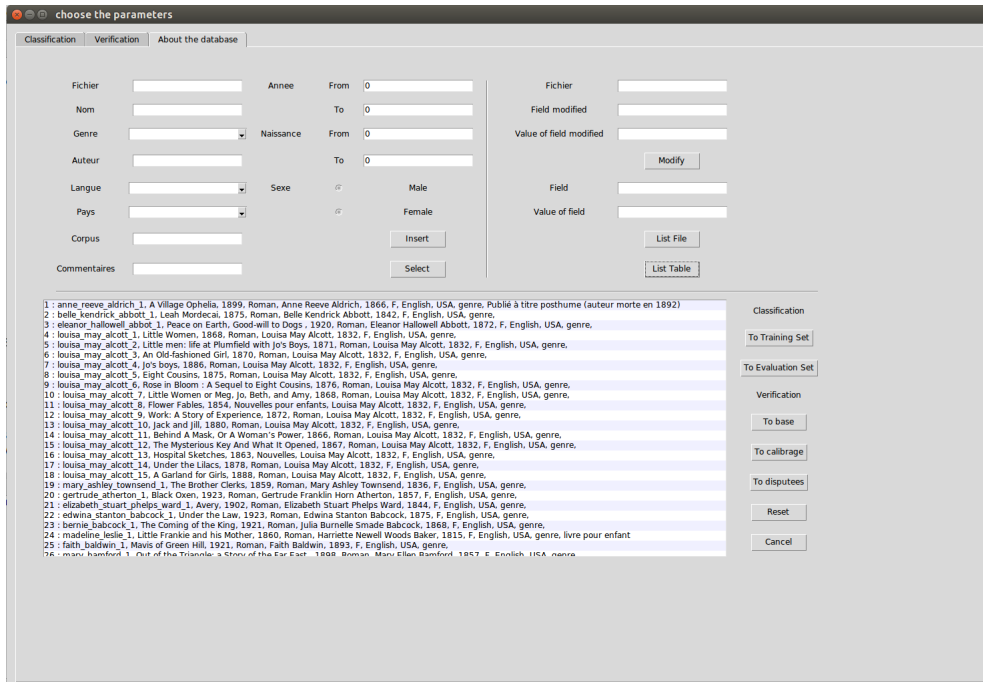
FIGURE 5 – Paramétrisation des analyseurs

configuration similaire. Nous pouvons également utiliser la base de données pour faciliter son utilisation.

## 2.3 FONCTIONNEMENT INTERNE

### 2.3.1 • PRÉTRAITEMENT DES TEXTES

Une fois les œuvres choisies, le programme crée une liste d'objets `Œuvre`, qui contient les textes lemmatisés. La lemmatisation est effectuée une seule fois, les objets `Œuvre` sont ensuite sérialisés à l'aide du module `pickle` dans des fichiers stockés par le programme. Les objets `Œuvre` ainsi créés sont ensuite divisés en segments de tailles égales : ces segments sont stockés



choose the parameters

Classification Verification About the database

Fichier  Année From  To

Nom  Field modified

Genre  Naissance From  To  Value of field modified

Auteur  Sexe  Male  Female

Langue  Pays  Field

Corpus  Insert  Value of field

Commentaires  Select  List File  List Table

Classification

To Training Set

To Evaluation Set

Verification

To base

To calibrage

To disputes

Reset

Cancel

1 : anne\_reeve\_aldrich\_1, A Village Ophelia, 1899, Roman, Anne Reeve Aldrich, 1866, F, English, USA, genre, Publié à titre posthume (auteur morte en 1892)

2 : belle\_kendrick\_abbott\_1, Leah Mordecai, 1875, Roman, Belle Kendrick Abbott, 1842, F, English, USA, genre,

3 : eleanor\_hallowell\_abbot\_1, Peace on Earth, Good-will to Dogs, 1920, Roman, Eleanor Hallowell Abbott, 1872, F, English, USA, genre,

4 : louisa\_may\_alcott\_1, Little Women, 1868, Roman, Louisa May Alcott, 1832, F, English, USA, genre,

5 : louisa\_may\_alcott\_2, Little men: life at Plumfield with Jo's Boys, 1871, Roman, Louisa May Alcott, 1832, F, English, USA, genre,

6 : louisa\_may\_alcott\_3, An Old-fashioned Girl, 1870, Roman, Louisa May Alcott, 1832, F, English, USA, genre,

7 : louisa\_may\_alcott\_4, Jo's boys, 1886, Roman, Louisa May Alcott, 1832, F, English, USA, genre,

8 : louisa\_may\_alcott\_5, Eight Cousins, 1875, Roman, Louisa May Alcott, 1832, F, English, USA, genre,

9 : louisa\_may\_alcott\_6, Rose in Bloom : A Sequel to Eight Cousins, 1876, Roman, Louisa May Alcott, 1832, F, English, USA, genre,

10 : louisa\_may\_alcott\_7, Little Women or Meg, Jo, Beth, and Amy, 1868, Roman, Louisa May Alcott, 1832, F, English, USA, genre,

11 : louisa\_may\_alcott\_8, Flower Fables, 1854, Nouvelles pour enfants, Louisa May Alcott, 1832, F, English, USA, genre,

12 : louisa\_may\_alcott\_9, Work: A Story of Experience, 1872, Roman, Louisa May Alcott, 1832, F, English, USA, genre,

13 : louisa\_may\_alcott\_10, Jack and Jill, 1880, Roman, Louisa May Alcott, 1832, F, English, USA, genre,

14 : louisa\_may\_alcott\_11, Behind A Mask, Or A Woman's Power, 1866, Roman, Louisa May Alcott, 1832, F, English, USA, genre,

15 : louisa\_may\_alcott\_12, The Mysterious Key And What It Opened, 1867, Roman, Louisa May Alcott, 1832, F, English, USA, genre,

16 : louisa\_may\_alcott\_13, Hospital sketches, 1863, Nouvelles, Louisa May Alcott, 1832, F, English, USA, genre,

17 : louisa\_may\_alcott\_14, Under the Lilacs, 1878, Roman, Louisa May Alcott, 1832, F, English, USA, genre,

18 : louisa\_may\_alcott\_15, A Garland for Girls, 1888, Roman, Louisa May Alcott, 1832, F, English, USA, genre,

19 : mary\_ashley\_townsend\_1, The Brother Clerks, 1859, Roman, Mary Ashley Townsend, 1836, F, English, USA, genre,

20 : gertrude\_atherton\_1, Black Oven, 1923, Roman, Gertrude Franklin Horn Atherton, 1857, F, English, USA, genre,

21 : elizabeth\_stuart\_phelps\_ward\_1, Avery, 1902, Roman, Elizabeth Stuart Phelps Ward, 1844, F, English, USA, genre,

22 : edwina\_stanton\_babcock\_1, Under the Law, 1923, Roman, Edwina Stanton Babcock, 1875, F, English, USA, genre,

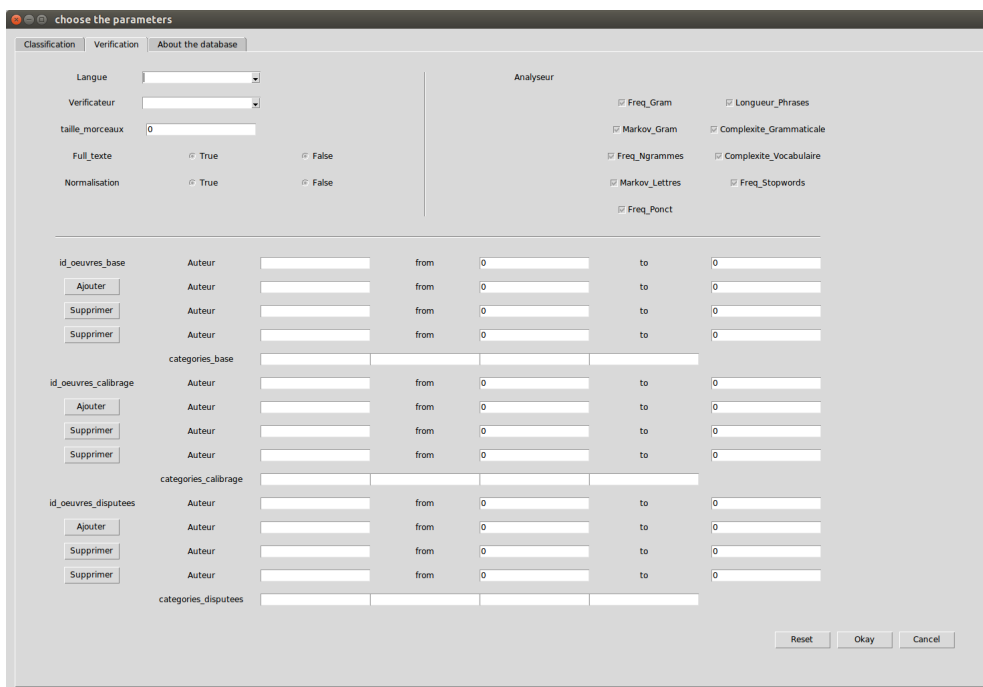
23 : bernie\_babcock\_1, The Coming of the King, 1921, Roman, Julia Burnelle Smade Babcock, 1868, F, English, USA, genre,

24 : madeline\_ladie\_1, Little Frankie and his Mother, 1860, Roman, Harriette Newell Woods Baker, 1815, F, English, USA, genre, livre pour enfant

25 : faith\_baldwin\_1, Mavis of Green Hill, 1921, Roman, Faith Baldwin, 1893, F, English, USA, genre,

26 : mary\_hamford\_1, Part of the Trisomie 3 Choir of the Eu Est, 1808, Roman, Mary Ellen Hamford, 1847, F, English, USA, genre,

FIGURE 6 – Exploration de la base de textes



choose the parameters

Classification Verification About the database

Langue

Verificateur

taille\_morceaux

Full\_text ☐ True ☐ False

Normalisation ☐ True ☐ False

Analyses

☐ Freq\_Gram ☐ Longueur\_Phrases

☐ Markov\_Gram ☐ Complexite\_Grammaticale

☐ Freq\_Ngrammes ☐ Complexite\_Vocabulaire

☐ Markov\_Letres ☐ Freq\_Stopwords

☐ Freq\_Ponct

id\_oeuvres\_base

Auteur  from  to

Ajouter

Supprimer

Supprimer

categories\_base

Auteur  from  to

Ajouter

Supprimer

Supprimer

id\_oeuvres\_calibrage

Auteur  from  to

Ajouter

Supprimer

Supprimer

categories\_calibrage

Auteur  from  to

Ajouter

Supprimer

Supprimer

id\_oeuvres\_disputees

Auteur  from  to

Ajouter

Supprimer

Supprimer

categories\_disputees

Reset  Okay  Cancel

FIGURE 7 – Onglet Vérification

dans des objets Texte. Ce découpage permet d'obtenir une région stylistique associée à une œuvre, plutôt qu'un seul point. La taille des morceaux est la même pour toutes les œuvres : il

devient ainsi possible d'étudier la précision de notre classification en fonction de la taille des morceaux. En effet, les longues œuvres donneraient des résultats plus précis que les œuvres courtes. Avec les morceaux de tailles égales, deux œuvres de taille différentes génèreront un nombre différent d'échantillons. Afin d'éviter que ce découpage n'engendre un biais dans le classifieur, le nombre d'échantillons par œuvre est fixé. Les échantillons des œuvres les plus longues sont choisis au pseudo-hasard : la graine utilisée pour la génération aléatoire est sauvegardée. Il est ainsi possible de renouveler une expérience sur un ensemble de Textes donné avec des paramètres différents.

### 2.3.2 • ANALYSE

L'étape suivante calcule le vecteur stylistique de chaque Texte. Pour ce faire, les Textes sont traités, et le programme calcule les composantes demandées par l'utilisateur, parmi celles présentées en première partie. Les Textes sont ainsi traités par un objet Analyseur. Celui-ci est composé d'un arbre d'objets Analyseur (figure 8) (par exemple, une branche « Grammaire » où figureront les fréquences grammaticales et les transitions markoviennes associées, et une branche « Vocabulaire » où se trouveront plusieurs indices d'étude de la richesse du vocabulaire présent dans un texte). Chaque feuille de l'arbre est une fonction d'analyse (comme les fréquences des mots les plus courants) et représente une série de composantes du vecteur stylistique. L'analyseur traite donc récursivement l'ensemble des textes, et enregistre dans chaque objet Texte son vecteur stylistique.

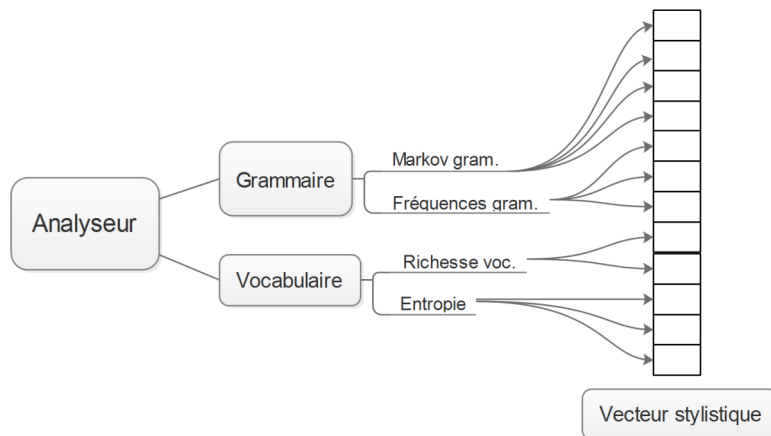


FIGURE 8 – Exemple de structure d'un analyseur

### 2.3.3 • CLASSIFICATION

Une fois formés les vecteurs des textes, un algorithme de classification analyse et traite le problème initialement posé. Cet algorithme est choisi par l'utilisateur parmi une liste incluant des algorithmes d'apprentissage et de clustering. Parmi les algorithmes que nous avons codés au

cours de l'élaboration de l'outil, nous avons gardé Bayes naïf, K-Means, K-Médoïdes, un SVM et un réseau de neurones, ce qui permet à l'utilisateur de tester un même corpus sous des angles différents. Le réseau de neurones utilisé est un perceptron multicouche à deux couches, de tailles respectives 10 et le nombre de catégories (paramètres choisis pour leurs bonnes performances sur de nombreux corpus). Les textes vectorisés sont séparés entre l'ensemble d'entraînement et l'ensemble d'application. Le classifieur est alors appliqué à l'ensemble de textes, et produit, sous forme standardisée, le résultat de sa classification. Notamment, on obtient en sortie deux matrices stockant les coefficients d'appartenance respectifs de chaque Texte à chaque catégorie. Ces coefficients peuvent différer des coefficients théoriques : par exemple, un Texte attribué initialement à la catégorie Dumas peut présenter un coefficient d'appartenance très faible à cette catégorie, ce qui mettrait en évidence une erreur possible d'attribution de l'œuvre. Dans la plupart de nos algorithmes de classification (sauf les algorithmes flous), ces coefficients sont binaires.

## 2.4 ANALYSE ET REPRÉSENTATION DES RÉSULTATS

### 2.4.1 • SORTIE TEXTUELLE

Une fois la classification effectuée, les sorties de l'algorithme sont affichées en console (figure 9) en vue d'être interprétées par l'utilisateur. Ce canal de sortie donne accès à des informations quantitatives précises sur les résultats de l'algorithme. On obtient notamment les résultats numériques de la classification, ainsi que les caractéristiques stylistiques ayant le plus contribué à la classification. Ces composantes peuvent être obtenues par plusieurs méthodes de calcul, que nous présentons plus loin.

### 2.4.2 • SORTIE GRAPHIQUE

Les sorties de l'algorithme sont également transmises à une interface d'affichage 2D (figure 10). Cette interface permet à l'utilisateur d'appréhender qualitativement les résultats de la classification. Elle résume le résultat de la classification en affichant la proportion de textes classés dans chaque catégorie, et affiche les textes classifiés regroupés par catégorie en 2D, entourés par leurs enveloppes convexes. Pour réduire la dimension du problème, nous appliquons une analyse en composantes principales (ACP) aux vecteurs stylistiques des Textes. On obtient ainsi une représentation des points susceptible de rendre compte de leur répartition dans un espace de dimension bien plus élevée (676 par exemple pour les bigrammes d'une langue en alphabet latin).

Néanmoins, l'ACP n'est pas suffisante pour obtenir des résultats satisfaisants : elle ne permet en effet pas de séparer les clusters, mais de repérer les axes les plus intéressants pour séparer les points dans leur ensemble. Deux clusters peuvent apparaître superposés en 2D après ACP, alors qu'ils sont nettement séparés en 3D, mais selon le mauvais axe. Pour pallier ce problème, l'interface comporte des curseurs permettant de régler l'importance des composantes les plus discriminantes, en les dilatant dans la matrice de passage rendue par l'ACP : concrètement, les

```

Résultats de la classification :
dumas1 est dans la catégorie dumas (70.0 %).
dumas2 est dans la catégorie dumas (76.6666666667 %).
proust3 est dans la catégorie proust (100.0 %).
proust4 est dans la catégorie proust (100.0 %).

=====

Composantes les plus importantes dans la classification :

1) Fréquence de '
Importance : 3.3347
  Ecart intra clusters pour cette composante : 0.0326
  Ecart inter clusters pour cette composante : 0.1087
    Moyenne parmi les textes de la categorie dumas : 0.2261
    Moyenne parmi les textes de la categorie proust : 0.3348

2) Fréquence de la catégorie grammaticale PUN|
Importance : 2.6347
  Ecart intra clusters pour cette composante : 0.0092
  Ecart inter clusters pour cette composante : 0.0243
    Moyenne parmi les textes de la categorie dumas : 0.1072
    Moyenne parmi les textes de la categorie proust : 0.0829

3) Fréquence de la catégorie grammaticale INT
Importance : 2.4648
  Ecart intra clusters pour cette composante : 0.0015
  Ecart inter clusters pour cette composante : 0.0037
    Moyenne parmi les textes de la categorie dumas : 0.0046
    Moyenne parmi les textes de la categorie proust : 0.0009

4) Fréquence de ?
Importance : 2.3049
  Ecart intra clusters pour cette composante : 0.0093
  Ecart inter clusters pour cette composante : 0.0214
    Moyenne parmi les textes de la categorie dumas : 0.0310
    Moyenne parmi les textes de la categorie proust : 0.0096

```

FIGURE 9 – Exemple de sortie texte en console

colonnes correspondantes sont multipliées par un facteur ajustable via le curseur.

Cette visualisation des résultats offre à l'utilisateur des perspectives d'interprétation humaine intéressantes : les curseurs permettent de distinguer finement comment les caractéristiques stylistiques dissocient les auteurs. Elle permet également de trier les caractéristiques, mettant en évidence celles qui séparent mieux un auteur particulier du groupe. On peut ainsi

dégager des lignes générales définissant le style d'un auteur par rapport à un autre.

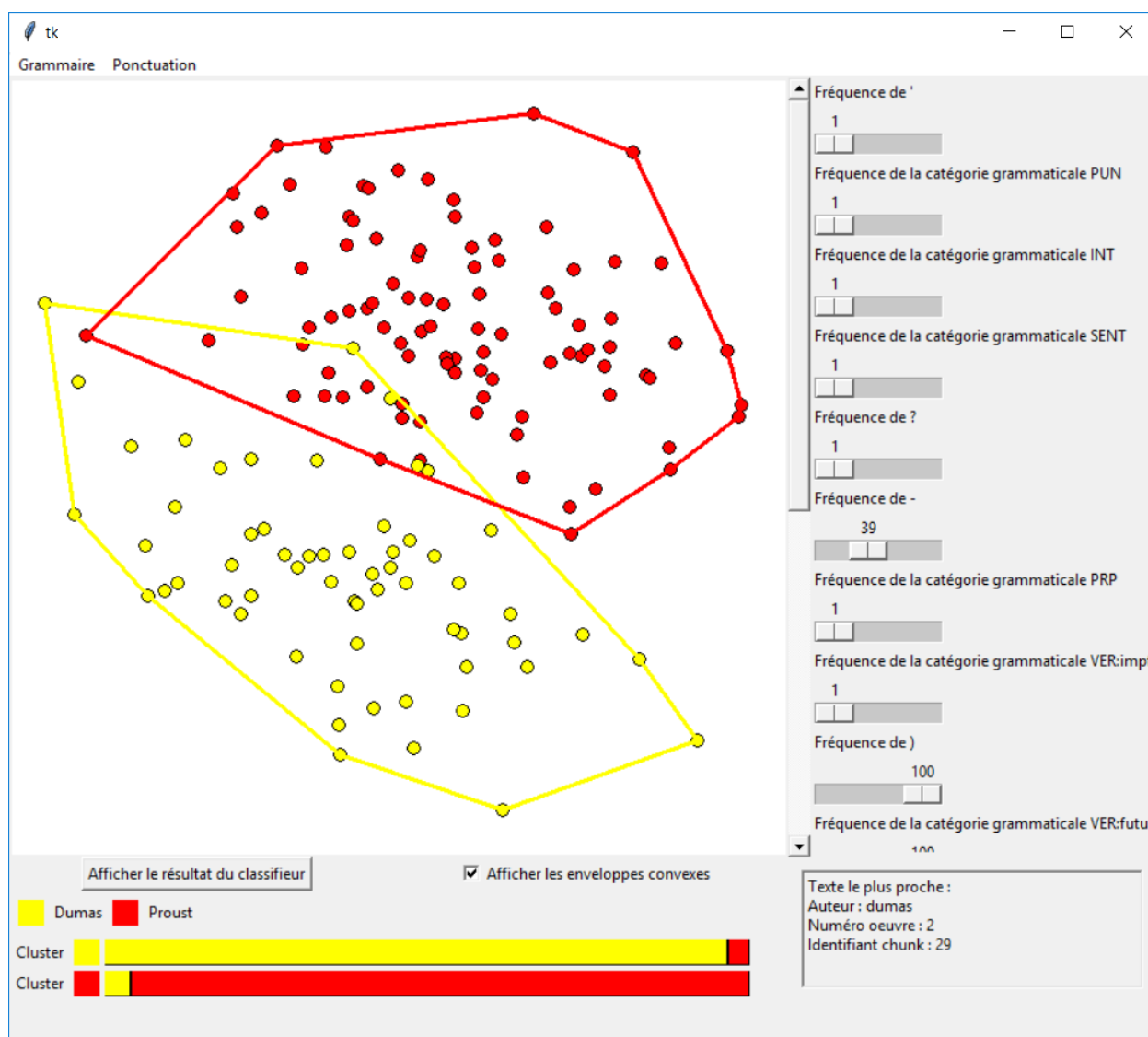


FIGURE 10 – Interface graphique d’affichage des résultats

### 2.4.3 • IMPORTANCE DES COMPOSANTES

L’importance des composantes dans la classification n’est pas une notion clairement définie. Nous avons implémenté plusieurs fonctions d’évaluation de cette importance. En premier lieu, nous avons utilisé le tri en norme euclidienne des colonnes de la matrice de l’ACP. Les colonnes les plus importantes représentent les caractéristiques stylistiques qui répartissent le mieux les Textes dans l’espace. Néanmoins, cette méthode comporte l’inconvénient mentionné précédemment : elle ne permet pas de sélectionner les composantes dissociant les catégories les unes des autres, mais seulement celles qui les dispersent le plus dans l’espace stylistique.

Actuellement, trois méthodes peuvent être utilisées et donnent, en fonction des corpus, des résultats plus ou moins satisfaisants : tout d'abord, le rapport de la variance inter-cluster sur la variance intra-cluster pour chaque composante donnée. Cette grandeur mesure donc la propension qu'ont les clusters formés par le classifieur à se distinguer entre eux sur l'axe étudié. Nous avons aussi essayé de trier les composantes par gain d'information, c'est-à-dire le rapport de l'entropie de l'ensemble des vecteurs par l'information intrinsèque du cluster. Cette méthode donne des résultats très variables en fonction des analyseurs utilisés et du corpus étudié. Nous utilisons également, lorsque le classifieur utilisé est le réseau de neurones, les sommes de poids du réseau : ceux-ci déterminent en effet l'importance que joue chaque composante dans le processus d'apprentissage et de classification du réseau. Ces deux méthodes permettent souvent d'obtenir de bons résultats. Dans le cas contraire, l'interface de sortie permet de sélectionner manuellement les composantes stylistiques dont l'utilisateur souhaite ajuster l'importance dans la matrice de l'ACP, et donc dans l'affichage 2D.

Par exemple, sur la figure 11, les réglages des curseurs ne figurent pas sur les images, mais ont été choisis manuellement de manière à tenter de maximiser la dissociation. On constate que, quelle que soit la méthode de tri, la modification de l'importance des composantes permet nettement d'améliorer la séparation des clusters. Sur cet exemple, le réseau de neurones et le gain d'information sont les deux méthodes de tri les plus performantes. Sur plusieurs tests réalisés sur différents corpus, les performances du tri par le rapport d'écart inter/intra-cluster et par gain d'information varient inégalement, bien que le premier donne plus souvent de meilleurs performances. Il ressort que, en moyenne, le réseau de neurones offre la performance la plus stable, et, en cas de bon fonctionnement, les meilleures bonnes performances. Néanmoins, il arrive parfois que les deux autres méthodes de tri permettent de meilleurs résultats, et ont l'avantage de ne pas demander le temps de calcul nécessaire à la convergence du réseau de neurones, parfois longue dans des espaces de grande dimension. Il est donc intéressant de pouvoir utiliser les trois méthodes si besoin : c'est d'autant plus vrai que les tris donnés par les trois indicateurs se recoupent assez souvent.



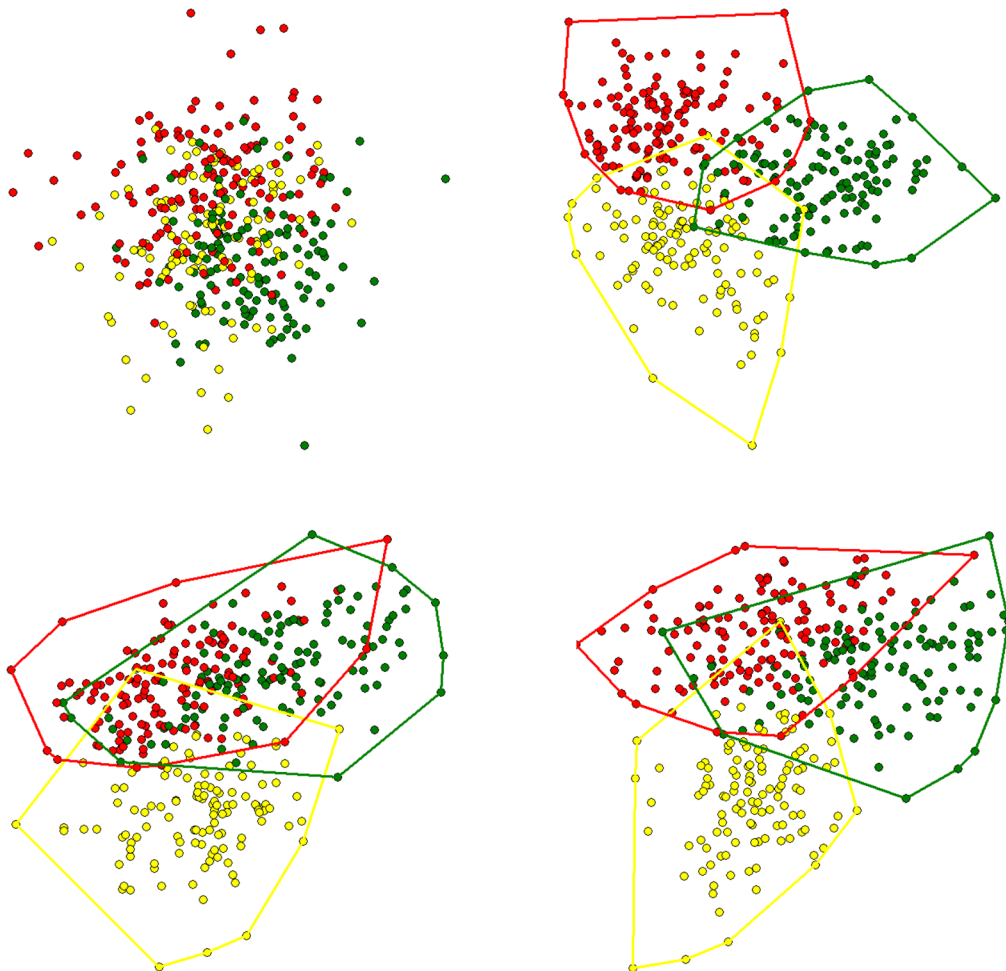


FIGURE 11 – Test de dissociation visuelle des clusters pour un même corpus d’œuvres de Proust, Balzac et Dumas. Dans cet ordre : sans modification, tri par réseau de neurones, tri par importance inter/intra-clusters, tri par gain d’information

### 3

## APPLICATIONS LITTÉRAIRES

En complément des quatre premières études de cas présentées dans le rapport intermédiaire, nous avons mené trois nouvelles études différentes, avec notamment des efforts conséquents sur l’établissement des corpus, la clarification de la méthodologie et l’analyse des résultats. Cette partie présente donc trois applications littéraires supplémentaires de notre outil.

## 3.1 LES ÉPOQUES ET LE GENRE

Nous avons jusqu'ici utilisé notre outil pour reconnaître un auteur donné, nous voulons dorénavant l'utiliser pour reconnaître des « catégories » d'auteurs ou de textes, catégories formées sur un critère tel que l'âge de l'auteur, son sexe, le genre du texte, ... Nous ne nous attendons pas a priori à obtenir des classifications aussi précises que celles obtenues lorsqu'il s'agissait de reconnaître un auteur donné. L'objectif est davantage d'étudier la capacité de notre outil à distinguer ces catégories plus générales, et, le cas échéant, d'étudier sur quels critères une telle classification s'opère.

Afin de mener ces études de cas nous avons élaboré un corpus de 65 romans américains écrits entre 1851 et 1923 par 52 auteurs différents. Nous les avons regroupés dans une base de données avec des informations sur les textes en question (titre, date de publication, langue, etc.) et sur leur auteur (date de naissance, pays, sexe, etc.) pour pouvoir ensuite facilement sélectionner les textes relevant d'une catégorie donnée.

Nous avons mené plusieurs études de cas en sélectionnant des textes de notre corpus selon des critères tels que : le sexe de l'auteur, la période d'écriture du texte, la date de naissance de l'auteur et le genre du roman (roman pour enfants ou pour adultes). Nous avons à chaque fois mis en place la même procédure exposée en détail dans l'étude de cas sur les périodes. Dans un premier temps, nous avons utilisé la validation croisée avec un type de fonctions d'analyse à chaque fois pour détecter ceux capables d'effectuer une classification meilleure que le hasard, i.e. plus performante que 50% de textes correctement classifiés (pour une classification entre deux catégories). Puis nous avons assorti les types de fonctions d'analyse ainsi sélectionnés pour trouver les meilleures combinaisons en terme de précision de classification. Enfin nous nous sommes penchés sur les critères que ces fonctions emploient prioritairement pour classifier les textes afin de déterminer quelles sont les caractéristiques qui permettent de différencier nos catégories de textes et nous avons tenté, dans la mesure du possible, d'interpréter ces caractéristiques.

### 3.1.1 • ÉTUDE DES PÉRIODES

Nous avons utilisé notre base de données pour créer un corpus de romans américains comprenant deux catégories : 11 romans écrits entre 1851 et 1875 d'une part et 12 romans écrits entre 1910 et 1923 d'autre part, les auteurs de ces romans étant tous différents. Dans toute cette étude nous utilisons l'algorithme SVM.

Dans un premier temps, nous utilisons la validation croisée pour déterminer les fonctions d'analyse les plus intéressantes (nous choisissons comme paramètres une taille de fragment de 6000 mots et 30 essais). Pour cela, nous lançons plusieurs validations croisées avec à chaque fois une seule fonction d'analyse et nous éliminons les fonctions d'analyse qui ont donné une précision d'environ 50% dans la classification (précision comparable à une attribution au hasard des textes dans chaque catégorie). Il nous reste :

- la complexité du vocabulaire (précision de 60% )
- la longueur des phrases (précision de 70% )

- la fréquence des 1-grammes (précision de 75% )
- les transitions de lettres (précision de 80%)
- la fréquence de la ponctuation (précision de 85%)

En additionnant les fonctions d'analyse, on trouve que les combinaisons les plus efficaces sont les deux dernières fonctions d'analyse ou les trois dernières, ce qui paraît logique car ce sont celles qui classifient le mieux seules. On obtient dans ces deux cas une précision d'environ 85%.

### **Classification avec les fonctions d'analyse des transitions de lettres et de fréquence de la ponctuation**

Les fonctions d'analyse de la longueur des phrases mettent en évidence une tendance claire pour des phrases plus longues avant 1875 qu'après 1910. Dans l'ordre décroissant de priorité, notre outil utilise la valeur : du premier quartile (16 avant 1875, 12 après 1910), de la médiane (26 avant 1875, 19 après 1910), du troisième quartile (40 avant 1875, 30 après 1910), de la moyenne (31 avant 1875, 23 après 1910) et de la moyenne du logarithme (3,2 avant 1875, 2,9 après 1910) de la longueur de la phrase.

Les fonctions d'analyse concernant la fréquence des 2-grammes et les transitions de lettres sont directement liées au vocabulaire utilisé. Il est donc plus difficile d'en tirer des conclusions sans étudier les textes du corpus plus en détail. On peut émettre des suppositions, par exemple la transition s -> h est plus fréquente après 1910 qu'avant 1875, ce qui pourrait s'expliquer par une augmentation du nombre de personnages féminins dans les romans écrits après 1910 et donc du pronom she. En étudiant en détail les textes du corpus, on remarque en effet que le pronom she est en moyenne nettement plus fréquent après 1910 qu'avant 1875, cependant la très grande variation du nombre d'occurrences de ce pronom d'un texte à l'autre et la taille relativement faible de notre corpus d'étude ne nous permet pas de tirer de conclusions qui seraient trop hâtives mais cela peut constituer une piste à approfondir.

Les fonctions d'analyse de la ponctuation utilisent, par ordre décroissant de priorité, la fréquence : de la virgule (0,5 avant 1875, 0,3 après 1910), du point-virgule (0,06 avant 1875, 0,02 après 1910), du point (0,25 avant 1875, 0,33 après 1910), de l'apostrophe (0,06 avant 1875, 0,14 après 1910). L'utilisation plus fréquente de la virgule et du point-virgule avant 1875 est en accord avec l'usage de phrases plus longues donc plus étoffées, la fréquence élevée du point après 1910 est en cohérence avec des phrases plus courtes, donc un retour plus fréquent du point. L'usage plus important d'apostrophes dans les textes plus récents peu s'expliquer par un langage plus oral, l'apostrophe en anglais servant aux contractions (les textes les plus anciens utilisent peut-être davantage des formes telles que I am ou cannot plutôt que I'm et can't).

Nous nous sommes ici demandé s'il était possible de classer les textes en fonction de leur période d'écriture, mais nous pouvons nous poser la question suivante : est-ce la période à laquelle l'auteur écrit qui influe sur son style ou est-ce la période à laquelle appartient l'auteur en question ? En d'autres termes, est-il plus facile de classer un texte en fonction de la période à laquelle il a été écrit ou en fonction de la période à laquelle son auteur est né ? Dans le premier cas cela revient à considérer que c'est le style contemporain à l'écriture du texte qui a la plus grande influence sur le style du texte, dans le second cas cela mène à penser que c'est le contexte historique dans lequel a grandi l'auteur et dans lequel il a appris à écrire qui prime. Dans l'étude précédente nous avons considéré des textes répartis en deux catégories selon la

date de leur publication indépendamment de la date de naissance de leur auteur. Nous pouvons maintenant constituer un nouveau corpus avec des textes choisis selon la date de naissance de leur auteur. Nous avons ainsi une catégorie de 14 textes écrits par des auteurs nés avant 1840 et l'autre de 13 textes écrits par des auteurs nés après 1870. Nous effectuons de nouveau la validation croisée avec les mêmes paramètres et une seule fonction d'analyse à chaque fois. Il apparaît que ce sont les mêmes fonctions d'analyse qui permettent de classer en deux catégories mais que cette fois-ci, à l'exception de l'analyse de la complexité du vocabulaire qui offre toujours une précision d'environ 60%, toutes les autres fonctions d'analyse offrent une précision comprise entre 70% et 75% environ. Même en combinant les fonctions d'analyse, on ne parvient pas à atteindre une précision de 80%. Nous n'allons pas refaire une analyse détaillée car à l'évidence ce sont à peu près les mêmes critères qu'utilise notre outil pour classer les textes mais nous pouvons noter que la précision de la classification a clairement chuté. Avec les outils d'analyse qui sont les nôtres, il est plus facile de déterminer la période d'écriture d'un texte que la période à laquelle est né l'auteur de ce texte. On pourrait évidemment pousser notre étude plus loin pour étudier l'influence de l'âge de l'auteur sur la classification : un auteur de 20 ans et un autre de 60 ans écrivant à la même époque n'ont sans doute pas le même style. Cependant, avec notre corpus de textes, on risquerait de ne pas avoir suffisamment de textes dans chaque catégorie pour mener une étude quelque peu sérieuse : il faudrait largement étoffer le corpus.

### 3.1.2 • ÉTUDE DU GENRE DES ROMANS

Notre corpus contient 18 romans pour enfants et 30 romans pour adultes d'auteurs tous différents, nous avons voulu savoir s'il était possible de les discriminer sur des critères statistiques.

Les types de fonctions d'analyse qui apparaissent capables de classer les deux catégories de romans à l'issue de la validation croisée sont :

- la fréquence de la ponctuation (précision de 60%)
- la complexité du vocabulaire (précision de 60%)
- la longueur des phrases (précision de 60%)
- la fréquence des 2-grammes (précision de 65%)
- les transitions de lettres (précision de 65%)

On note que la précision de ces fonctions d'analyse n'est pas comparable à celle de l'étude de cas sur les périodes. En combinant les deux dernières avec une ou plusieurs des trois premières, on peut obtenir une précision de classification d'environ 70%. Dans toutes les combinaisons de ce type, on s'aperçoit que les critères prioritairement utilisés pour classer sont issus des deux derniers types de fonctions de classification : l'analyse de la fréquence des 2-grammes et celle des transitions de lettres. Cela peut paraître logique car ce sont les fonctions qui classent le mieux seules. Cependant cela rend notre interprétation littéraire plus complexe car ces deux types de fonctions reposent sur la nature du vocabulaire utilisé et ne peuvent pas donner lieu à une interprétation simple sans étude plus approfondie des textes en eux-mêmes. On peut néanmoins supposer que les champs lexicaux employés pour un public adulte et un public enfant sont différents. Ainsi on note que la transition  $b \rightarrow o$  est plus fréquente dans les romans pour enfants (0,15) que dans les romans pour adultes (0,12) de même que la transition  $i \rightarrow r$  (0,046

dans les romans pour enfants et 0,040 dans les romans pour adultes) : cela peut s'expliquer si les termes boy et girl sont plus fréquents dans les romans pour enfants que dans ceux pour adultes, ce qui paraîtrait plausible. On note par ailleurs qu'en ce qui concerne l'analyse de la longueur des phrases, un élément utilisé prioritairement pour classifier est l'écart-type logarithmique de la longueur des phrases qui est plus faible dans les romans pour enfants (0,64) que dans ceux pour adultes (0,70). Cela traduit des phrases de longueur plus diverse dans les romans pour adultes que dans ceux pour enfants.

### 3.1.3 • ÉTUDE DU SEXE DES AUTEURS

Notre corpus contient 25 romans écrits par des femmes et 27 romans écrits par des hommes, tous les auteurs étant différents. Le but est de déterminer si notre outil peut distinguer un texte écrit par une femme d'un texte écrit par un homme.

Les types de fonctions d'analyse qui apparaissent capables de classifier les deux catégories de romans à l'issue de la validation croisée sont :

- la fréquence de la ponctuation (précision de 60%)
- la fréquence des 2-grammes (précision de 75%)
- les transitions de lettres (précision de 75%)

On peut combiner les deux derniers types de fonctions ou l'ensemble des trois et on obtient une précision de classification de 75%. Dans tous les cas l'apport de l'analyse de la complexité du vocabulaire demeure très marginale, les types de fonctions utilisés sont prioritairement les deux derniers. Comme dans l'étude de cas précédente, cela nous confronte au problème qui est que l'on ne peut pas tirer de conclusions simples des critères de classification employés par ces deux types de fonctions. On peut, comme précédemment supposer que les deux types de romans ne s'adresseraient pas à la même audience, ce qui engendrerait l'emploi de champs lexicaux différents. On peut également supposer que les écrivains féminins n'ont pas eu la même éducation, donc pas d'accès aux mêmes sources littéraires, que les écrivains masculins ce qui se ressentirait dans leur style. Enfin on peut aussi conjecturer que la situation sociale des femmes étant différente de celles des hommes à cette époque-là, il existerait une situation d'énonciation féminine qui se caractériserait par une focalisation sur des éléments négligés par les auteurs masculins, comme par exemple le quotidien de femmes au foyer. Cela peut être corroboré par l'utilisation de la transition  $s \rightarrow h$  par notre outil afin de discriminer la catégorie des auteurs féminins (fréquence de 0,14) de celle des auteurs masculins (fréquence de 0,11), cela pourrait s'expliquer par une plus grande utilisation du pronom personnel she, marqueur d'une plus grande présence de personnages féminins. Pour vérifier cela nous calculons le rapport du nombre d'occurrences du pronom he sur celui du pronom she dans les textes des deux catégories : la moyenne est de 4,2 dans la catégorie des auteurs féminins (elle est de 1,4 si on en retire un roman où le rapport vaut 71) et de 6,4 dans la catégorie des auteurs masculins (elle est de 5,2 si on en retire le roman où ce rapport est le plus élevé et vaut alors 38). On remarque donc que dans notre corpus le pronom she est plus fréquent dans les romans écrits par des femmes que dans ceux écrits par des hommes ce qui accrédite notre hypothèse que les personnages féminins y sont plus fréquents que dans les romans écrits par des hommes. On note par ailleurs que dans 11 des 25 romans écrits par des femmes ce rapport est plus petit que 1, ce qui signifie que le

pronom *she* y est plus présent que *he*, alors que dans les romans écrits par des hommes cette situation n'a lieu que 5 fois sur 27.

### 3.1.4 • BILAN DES ÉTUDES DE CAS

En conclusion, nous remarquons que notre outil a permis de distinguer les différentes catégories que nous avons constituées avec une précision plus ou moins bonne selon les études de cas. L'obstacle majeur à l'interprétation des critères de différenciation des catégories est l'utilisation omniprésente des analyseurs basés sur la fréquence des N-grammes et des transitions de lettres car ceux-ci reposent sur la nature du vocabulaire utilisé. Nous avons cependant pu, au cas par cas, faire certaines hypothèses interprétatives. La critique principale que l'on peut apporter à un tel travail est l'usage d'un corpus de taille restreinte et inhomogène, que ce soit à cause de la période d'écriture qui s'étale sur environ 70 ans ou sur les différents genres de romans qu'il comporte. Cela s'explique par la difficulté de rassembler un corpus important et homogène d'oeuvres appartenant au domaine public sans y consacrer un temps démesuré.

## 3.2 UN CORPUS NATURALISTE

Nous avons analysé un corpus composé d'auteurs naturalistes afin de déterminer quelles caractéristiques stylistiques propres à chacun nos algorithmes étaient capables de nous fournir. Notre outil étant globalement capable de différencier tous ces auteurs, nous nous sommes en effet intéressés à la question de savoir si les critères qu'il employait pouvaient nous donner à nous humains des indices pour distinguer ces auteurs "à l'œil nu", c'est-à-dire si les fréquences des composantes renvoyées variaient assez entre les différents auteurs pour qu'on espère voir une différence en lisant les textes.

### Corpus

Le corpus que nous avons utilisé pour étudier l'école naturaliste a été composé en grande partie sur les conseils de N. Wanlin, et est constitué des textes suivants provenant de huit auteurs :

- Paul Alexis : *Trente romans* - *La Fin de Lucie Pellegrin* - *Après la bataille* - *Le Collage*
- Henry Céard : *La saignée*
- Frères Goncourt : *Germinie Lacerteux* - *La Fille Élisa* - *Madame Gervaisais*
- Léon Hennique : *Deux nouvelles* - *La dévouée* - *Un caractère* - *Poeuf* - *L'affaire du grand 7*
- Joris-Karl Huysmans : *Marthe* - *Les soeurs Vatard* - *En ménage* - *A vau l'eau* - *A rebours* - *En rade* - *Là-bas* - *En route* - *L'oblat*
- Maupassant : *Boule de suif* - *La vaison Tellier* - *Une vie* - *Contes de la bécasse* - *Pierre et Jean* - *Le Horla* - *Fort comme la mort* - *Contes du jour et de la nuit*
- Méténier : *Barbe Bleue* - *Le Gorille* - *Zézette* - *La vocation*
- Zola : *L'argent* - *L'assommoir* - *Au bonheur des dames* - *La bête humaine* - *La conquête de Plassans* - *La curée* - *La débâcle* - *Le docteur Pascal* - *La faute de l'abbé Mouret* -



*Germinal – Lourdes – Nana – L'œuvre – Une page d'amour – Paris – Pot-Bouille – Le rêve – Rome – Son excellence – Eugène Rougon – La terre – Thérèse Raquin – Le ventre de Paris*

## Protocole

L'idée initiale était de déterminer avec nos algorithmes de classification les composantes significatives dans la distinction entre chaque auteur et tous les autres, donc en leur demandant de classer les textes en deux catégories : "auteur étudié" et "autres naturalistes". Cependant cette méthode s'est révélée peu satisfaisante du fait du manque de textes disponibles pour la plupart des auteurs. Il devenait compliqué d'avoir un nombre de textes équilibré dans les deux catégories si l'on voulait garder dans la catégorie "autres naturalistes" assez d'auteurs différents pour justifier son appellation. De plus, même pour Zola pour qui nous disposions de nombreux textes, la classification obtenue n'était pas très précise et ne donnait que peu de composantes significatives (entre 1 et 3 selon l'analyseur et la taille des morceaux).

Nous avons donc opté pour une comparaison de toutes les paires d'auteurs, avant de nous intéresser aux composantes qui apparaissaient plusieurs fois pour un même auteur. Nous avons utilisé le classifieur SVM pour des raisons de vitesse d'exécution, avec comme analyseur par défaut les 1-transitions grammaticales, les fréquences des catégories grammaticales, les fréquences des signes de ponctuation et les fréquences des stopwords (analyseur donnant la meilleure précision pour la plupart des paires d'auteurs). La taille des morceaux a été choisie entre 800 et 5000 mots en fonction du nombre d'œuvres disponibles pour les auteurs étudiés, de manière à avoir plus d'une centaine de morceaux par catégories après équilibrage du corpus d'apprentissage. Pour chaque paire d'auteurs, nous avons mis dans le corpus d'apprentissage toutes les œuvres disponibles pour ces auteurs sauf deux (une pour chaque auteur) que nous avons mis dans l'ensemble d'application. Puis nous avons permuté les œuvres pour vérifier qu'elles étaient toutes bien attribuées et faisaient apparaître des composantes significatives similaires.

## Résultats

Tout d'abord, presque toutes les œuvres ont été correctement attribuées, les exceptions étant des œuvres des frères Goncourt que le classifieur n'arrivait pas à distinguer de celles de Maupassant.

Deuxièmement, on constate que pour la plupart des paires d'auteurs la liste des composantes significatives affichées en sortie ne variait pas beaucoup selon la répartition des œuvres entre le corpus d'apprentissage et l'ensemble d'application.

Par ailleurs les résultats concernant les auteurs Hennique et Alexis sont faussés en raison de textes de mauvaise qualité (présence de la mention "Digitized by VjOOQIC" à de nombreuses reprises, mots comportant des fautes ou coupés en deux, symboles à des endroits improbables), ce qui a notamment entraîné l'apparition régulière parmi les composantes significatives de la fréquence des abréviations. Même en retirant de l'analyseur les fréquences des catégories grammaticales, les classifications et les composantes significatives demeurent probablement affectées par ce problème.

Pour chaque auteur, nous avons sélectionné les composantes apparaissant comme étant significatives pour le distinguer d'au moins deux autres auteurs, puis nous avons conservé celles



dont la fréquence variait suffisamment entre les auteurs (nous avons imposé une différence supérieure à 1%, car nous recherchons a priori des caractéristiques stylistiques pouvant être détectées par un humain ou à la limite de la perception humaine).

Les résultats par auteur sont les suivants :

Alexis utilise :

- peu d’apostrophes (8,6% en moyenne pour les textes d’Alexis contre 14% en moyenne pour les autres)
- peu de points-virgules (1,3% contre 3%)

Goncourt utilise :

- plus de noms communs (20% contre 17%)
- plus de prépositions (12% contre 10%)

Hennique utilise :

- peu de conjonctions (2,7% contre 4%), notamment peu de « et » (4% contre 5,6%)

Huysmans utilise :

- beaucoup plus de virgules (62% contre 40%)
- plus de « et » (7,3% contre 5,1%)
- plus de transitions nom commun-nom commun (2% contre 0,5%)
- très peu d’apostrophes (0,1% contre 15%)
- peu de transitions pronom relatif-pronom personnel (25% contre 50%)

Maupassant utilise :

- beaucoup plus d’apostrophes (23% contre 12%)
- plus de deux points (3% contre 1,4%)

Méténier utilise :

- plus de pronoms personnels (9% contre 7,7%)
- peu de transitions pronom personnel-imparfait (17% contre 35%)
- peu de virgules (30% contre 47%)
- peu d’adjectifs (4,2% contre 5,4%)

Zola utilise :

- beaucoup plus d’apostrophes (22% contre 12%)
- plus de transitions interjection-ponctuation (70% contre 30%)
- plus de transitions adverbe-ponctuation (22% contre 15%)

On constate qu’à l’exception de l’abondance de virgules chez Huysmans, qui pouvait être repérée sans faire de statistiques, la plupart de ces caractéristiques stylistiques sont difficilement perceptibles pour un lecteur humain, même une fois averti. Cependant certains de ces écarts statistiques peuvent se traduire par des caractéristiques mieux repérables par un lecteur : par exemple il est probable qu’une fréquence élevée d’apostrophe traduise une énonciation à la première personne, ce qui induit beaucoup de « j’ », ou du moins (car dans ce corpus peu de textes sont écrits à la première personne) de nombreux dialogues dans lesquels les personnages

parlent d'eux-mêmes. La fréquence un peu élevée de deux points chez Maupassant pourrait aussi traduire une plus grande place accordée aux dialogues, mais ceux-ci ne sont pas toujours introduits par la même ponctuation, ce qui fausse également toute tentative de « mesurer » la place des dialogues et regardant la fréquence des tirets ou des guillemets. En conclusion, notre outil est assez précis pour distinguer des auteurs proches comme ceux de ce corpus, mais il n'est pas d'une grande utilité pour indiquer les caractéristiques sur lesquelles il base sa classification à un lecteur désireux d'apercevoir lui-même ces différences.

### 3.3 VÉRITÉ ET MENSONGES

---

Inspirés par des articles anglophones évoquant cette possibilité, nous avons aussi souhaité appliquer nos outils à une perspective un peu moins littéraire : celle de détecter les mensonges. Plus spécifiquement, il s'agissait de voir si nos algorithmes et nos fonctions statistiques étaient à même de faire la différence entre le langage d'une personne qui dit la vérité et le langage d'une personne qui ment, sans s'attacher au contenu mais uniquement en examinant la forme stylistique.

#### 3.3.1 • RÉCOLTE DU CORPUS

La plus grande difficulté dans ce projet était de recueillir un corpus cohérent. L'idéal aurait été de disposer d'un corpus déjà annoté, rassemblé par des professionnels (sociologues, juristes), comme cela a déjà été fait en d'autres langues. Cependant nous n'en avons pas trouvé en langue française qui soit facilement accessible, et nous avons donc décidé de le créer nous-mêmes.

Notre première idée était de rassembler un corpus oral. En effet, la langue orale est beaucoup plus caractéristique et moins consciemment contrôlable que la langue écrite. Même sans se préoccuper des marqueurs physiologiques du mensonge, et même si le langage oral est ensuite retranscrit pour être analysé par ordinateur, il ne fait pas de doute qu'une interview fournit beaucoup plus d'éléments pour déterminer la véracité des affirmations qu'un simple texte. Malheureusement, la contrainte temporelle nous a arrêtés : il nous semblait impossible de conduire un nombre suffisant d'interviews (chacune devant durer au moins 3 minutes pour produire un contenu d'une longueur raisonnable), puis de numériser le texte de toutes ces interviews à la main (les assistants de type dictée vocale n'étant pas suffisamment fiables pour se reposer dessus à 100%).

Nous nous sommes donc tournés vers un corpus écrit. Il fut récolté grâce à un sondage en ligne, que nous avons partagé via Facebook avec nos amis et connaissances. Ce sondage posait d'abord une question simple : "Racontez votre journée d'hier". Sur la deuxième page, il demandait quelques informations personnelles à l'internaute (statut professionnel, âge, sexe, etc.). Enfin sur la dernière page il demandait de nouveau à l'internaute de raconter une journée, fictive cette fois-ci. Le but était de comparer des expressions sincères et mensongères sur le même sujet, dans des conditions aussi semblables que possibles.

Au total, près d'une centaine de contributions ont été recueillies. Certaines étaient incomplètes (les gens répondaient à la première question puis renonçaient devant la dernière), d'autres

étaient trop courtes pour être prises en compte, d'autres encore étaient complètement fantaisistes. Le choix effectué fut donc de se restreindre à un corpus aussi homogène et utile que possible, en éliminant (de façon assez arbitraire) les contributions jugées inadaptées. Parmi les réponses retenues, on trouve 63 élèves polytechniciens, 13 étudiants d'autres institutions, et quelques catégories marginales que nous laissons de côté en raison des effectifs insuffisants. Seuls 4 répondants n'avaient pas le français comme langue maternelle, dont 3 le parlaient parfaitement, nous avons donc seulement exclu le dernier.

### 3.3.2 • RÉSULTATS

Pour garder une population aussi homogène que possible, nous nous sommes restreints dans un premier temps au cas des élèves polytechniciens, ce qui donne un total de 62 "vérités" et 48 "mensonges" à exploiter, soit 10800 mots et 7500 mots respectivement dans chaque catégorie.

La méthode utilisée fut d'abord une exploration quelque peu empirique des réglages permettant d'obtenir la meilleure efficacité d'attribution. Cette exploration s'est effectuée à l'aide du classifieur SVM, car pour la validation croisée 'leave-one-out' nécessite un algorithme assez rapide. Ensuite, la phase d'interprétation s'est faite grâce à l'estimation de l'importance des composantes.

La première observation est que si l'on prend les contributions seules, le processus n'est pas optimal, et ce pour deux raisons :

- une moyenne de moins de 200 mots par réponse est trop courte pour générer des statistiques significatives ;
- les réponses sont de taille très hétérogène et de contenu extrêmement varié, ce qui tend à exacerber la variance des caractéristiques.

On ne dépasse pas, quel que soit l'analyseur utilisé, une précision d'environ 60%. Pour améliorer ce résultat, les réponses ont donc été regroupées dans un gros fichier, et mélangées phrase par phrase afin d'homogénéiser le style ; le but étant de mettre en valeur les composantes essentielles du mensonge et de la vérité. Ce fichier à son tour a été redécoupé en morceaux dont nous choisissons la taille. Mais forcément, plus les morceaux sont longs et statistiquement significatifs, moins il y a de points dans l'espace puisque le produit de la taille des morceaux par le nombre de morceaux est constant. La taille trop faible des données représente donc vraiment le facteur limitant de cette étude.

Concernant les analyseurs utilisés, on voit (figure 12) qu'en choisissant bien celui qu'on utilise (sans même les combiner) on peut obtenir une précision de 75% voire 80%. Chose étonnante, la majorité des analyseurs préfère disposer de peu de textes mais d'une longueur conséquente, cependant certains analyseurs sont plus efficaces avec beaucoup de petits textes. Deuxième remarque intéressante, combiner plusieurs analyseurs n'améliore pas souvent la qualité de la classification, car les algorithmes d'apprentissage sont aussi handicapés par le nombre de dimensions de l'espace. Nous y reviendrons ultérieurement.

On voit donc que notre outil est sensiblement plus performant que le hasard pour discerner les mensonges de la vérité. Passons maintenant aux enseignements pratiques sur la façon de le faire.

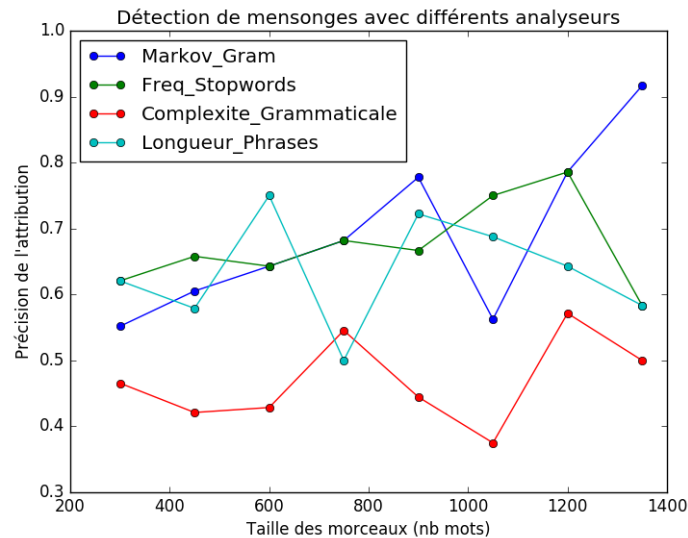


FIGURE 12 – Influence de la taille des morceaux sur l'efficacité de la détection

### 3.3.3 • INTERPRÉTATION

Pour juger de l'importance des composantes dans la classification, nous avons séparé les deux ensembles de réponses (vérité et mensonges) en un ensemble d'apprentissage et un ensemble d'application afin d'appliquer notre outil. La taille des morceaux a servi de variable d'ajustement selon les analyseurs, oscillant entre 300 et 800 afin de garantir un bon compromis entre une longueur suffisante et un nombre suffisant de textes.

Voici donc le résumé des caractéristiques les plus discriminantes du mensonge par écrit chez les polytechniciens :

1. Catégories grammaticales : les menteurs utilisent légèrement moins de noms communs mais plus de pronoms personnels, et moins de prépositions. On peut supposer que le manque de préposition traduit une moins grande facilité à situer les actions imaginaires dans le temps et l'espace, tandis que les pronoms personnels traduisent une volonté de rendre le récit plus vivant ou dynamique, bref de le romancer.
2. Transitions grammaticales : les probabilités de passage markoviennes valident les remarques précédentes, notamment au sujet de l'usage des pronoms personnels. Méfiance toutefois, car beaucoup de ces composantes sont identiquement nulles dans un groupe ou l'autre, ce qui fausse quelque peu les résultats, car cela peut être dû à un manque de données sur un échantillon aussi petit. On remarque aussi, chez les sincères, une forte présence des transitions entre préposition et adjectifs numéraux, ce qui indique une quantification plus importante dans leurs récits, chose sans doute difficile pour les menteurs.
3. Lettres : les menteurs utilisent plus souvent les lettres "j" et "m", ce qui peut suggérer que leurs récits font un usage plus prononcé de la première personne. Sur ce point nous restons toutefois méfiants, car étrangement les lettres les plus discriminantes variaient

fortement avec la taille des morceaux.

4. Ponctuation : les menteurs utilisent sensiblement plus de points et moins de virgules. Leur discours semblerait donc un peu plus saccadé et un peu moins construit, y compris à l'écrit.
5. Mots-outils : on y confirme l'intuition obtenue grâce aux lettres, les menteurs se servent plus des pronoms personnels de la première personne, tandis que les sincères utilisent beaucoup plus souvent "on" et "se". On confirme également que l'usage des prépositions telles que "sur" et "à" est nettement plus répandu chez les sincères.
6. Originalité grammaticale : l'écriture véridique est beaucoup plus originale que l'écriture mensongère, au sens de la déviation par rapport à la matrice de transition moyenne sur les catégories grammaticales. Peut-être les menteurs, occupés à trouver le contenu de leur invention, sont-ils moins concentrés sur l'enchaînement de leurs structures grammaticales.
7. Vocabulaire : les différences dans l'étendue du vocabulaire, quelle que soit la mesure adoptée, ne sont pas vraiment significatives.
8. Longueur des phrases : les menteurs font des phrases légèrement moins longues, ce qui reprend l'observation sur l'originalité grammaticale et la ponctuation. Cette différence est toutefois peu marquée.

## 4

# EVALUATION

---

### 4.1 PERFORMANCE DES ALGORITHMES

---

Afin de savoir quelles performances nous pouvons attendre de notre outil, il est intéressant de l'évaluer sur un corpus de base, notamment pour étudier l'influence des différents paramètres sur la qualité de l'attribution. Nous utiliseront pour cela deux corpus de test de taille différente, tirés tous deux du Projet Gutenberg.

1. Le premier est francophone, composé de romans réalistes du XIX<sup>ème</sup> :
  - Balzac : les trois premiers tomes de la *Comédie humaine*
  - Stendhal : *Le Rouge et le Noir*, *La Chartreuse de Parme*
  - Flaubert : *Madame Bovary*, *L'Education Sentimentale*Après la phase d'équilibrage, on obtient environ 350 000 mots par auteur.
2. Le second est anglophone, composé des trois oeuvres les plus célèbres des soeurs Brontë :
  - Anne Brontë : *Agnes Grey*
  - Charlotte Brontë : *Jane Eyre*
  - Emily Brontë : *Wuthering Heights*Après la phase d'équilibrage, on obtient environ 70 000 mots par auteur.

Sauf mention contraire, les tests seront, ici encore, effectués avec le classifieur SVM par défaut, mais nous comparerons également les performances des différents classifieurs. L'analyseur par défaut sera une combinaison des fréquences grammaticales et des fréquences de la ponctuation, permettant une dimension de l'espace assez réduite pour que les calculs se fassent rapidement. Là encore, d'autres analyseurs seront testés.

#### 4.1.1 • SEGMENTATION DES OEUVRES

Pour étudier l'effet de la taille des morceaux, on utilise une validation croisée, avec 30% des textes dans eval set et le reste dans training set. Contrairement au cas du très petit corpus de mensonges, où la taille était vraiment limitante, on voit (figure 13) qu'avec le corpus moyen 2 comme avec le grand corpus 1 on améliore globalement la précision en augmentant la taille des morceaux. La variance importante des résultats s'explique par le découpage en morceaux de tailles différentes, qui modifie sensiblement les positions des points dans l'espace stylistique.

#### 4.1.2 • CLASSIFIEURS

On peut aussi comparer les différents classifieurs que nous avons programmés : à la fois les méthodes d'apprentissage (SVM, réseaux de neurones, algorithme a priori, naive bayes) et les méthodes de clustering (k-means, k-medoids). En gardant la même méthode de validation

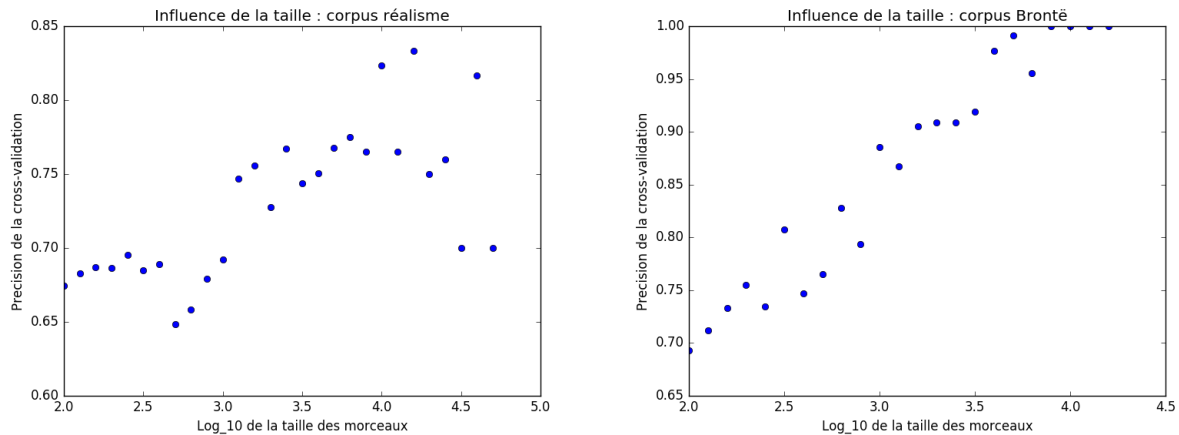


FIGURE 13 – Influence de la taille des morceaux sur la classification

croisée 70% / 30%, on remarque (figure 14) une assez grande disparité. SVM présente toujours d'excellentes performances, puisque c'est un algorithme commercial. Le réseau de neurones converge beaucoup mieux avec peu de points (donc de gros morceaux), dépassant même SVM. Bayes présente une performance moyenne, et les deux algorithmes de clustering ont environ un tiers de bonnes réponses voire moins.

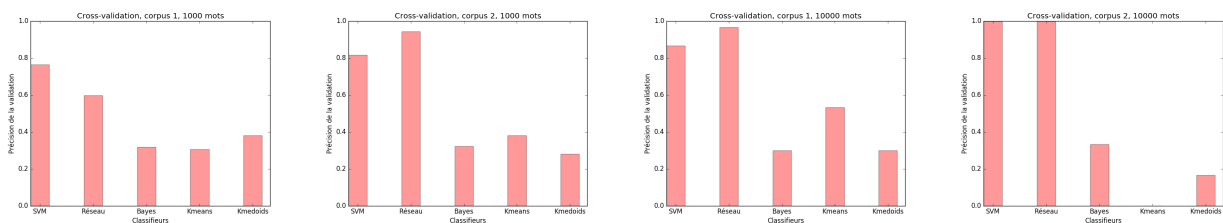


FIGURE 14 – Influence du classifieur, validation 70 / 30

En revanche, si on évalue la performance avec une validation de type "leave-one-out", on observe (figure 15) un comportement très différent, toujours avec 10 000 mots par morceau. SVM reste bon, le réseau est exclus car trop lent sur ce type de validation, et Bayes est inefficace. Mais du côté des algorithmes de clustering, on observe un score parfait.

On en conclut donc que les algorithmes d'apprentissage sont meilleurs pour généraliser à partir de peu de données, tandis que les algorithmes de clustering sont plus performants lorsqu'il s'agit d'utiliser un grand ensemble d'apprentissage pour catégoriser un seul texte.

#### 4.1.3 • ANALYSEURS

Afin de mesurer la performance des différents analyseurs, on se place sur une taille de morceaux de 10 000 mots, jugée à peu près optimale précédemment, en gardant en tête qu'elle était optimale pour notre analyseur par défaut. On va ainsi évaluer tous les analyseurs sur les deux corpus, en sachant que leurs performances dépendent bien évidemment du style de



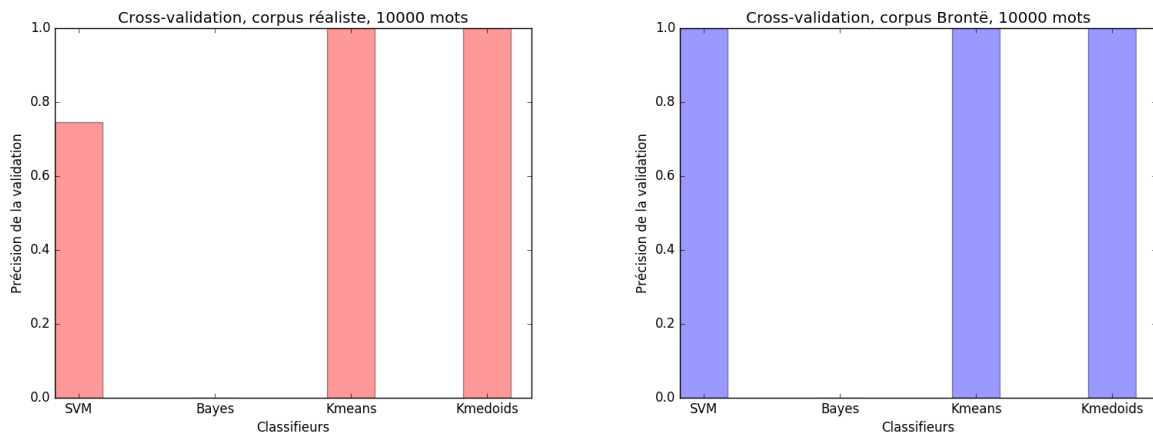


FIGURE 15 – Influence du classifieur, validation "leave-one-out"

chaque auteur et des caractéristiques saillantes permettant de trier le corpus. Le temps de calcul important requis pour ces tests nous a contraints à ne tester les analyseurs que seuls et 2 par 2, sans en agréger plus.

L'analyse portait sur la précision de l'attribution, toujours évaluée par validation croisée 70 % / 30 %. On voit (figure 16) que celle-ci n'est pas du tout une fonction monotone de la dimension : plus de composantes ne veut pas dire une meilleure analyse. Chaque point sur le graphe correspondant à un analyseur (ou à une combinaison), on constate en effet des performances extrêmement variées et pas forcément meilleures avec plusieurs analyseurs, ce que confirme un zoom sur les plus petites dimensions.

1. Pour le premier corpus, les analyseurs les plus performants (au sens de la précision) sont la fréquence des lettres (0,92) et la combinaison des fréquences de transition entre catégories grammaticales à l'ordre 1 et à l'ordre 2 (0,91)
2. Pour le second corpus, l'analyseur le plus performant est la fréquence de la ponctuation, qu'il soit seul (0,98) ou combiné avec d'autres tels que la fréquence des lettres, les catégories grammaticales, les mots-outils (stopwords) ou la complexité du vocabulaire (précision de 1,0)

On en déduit donc que si l'on veut mettre en valeur des différences entre les auteurs, il ne faut pas procéder de la même façon dans les deux cas. Pour différencier Balzac de Flaubert, une analyse grammaticale sera efficace, tandis que pour séparer les soeurs Brontë, il suffit souvent de regarder la ponctuation.

Le choix de l'analyseur devra donc s'effectuer au cas par cas, puisqu'il dépend, beaucoup plus que le reste des paramètres, du style des textes considérés. Voilà un des nombreux aspects où l'expertise littéraire n'est pas automatisable.

#### 4.1.4 • AUTRES CRITÈRES

Notons enfin que si la précision de l'attribution est le critère privilégié lorsqu'on connaît déjà les auteurs des oeuvres en question, elle n'est pas le seul. Dans d'autres situations, on peut

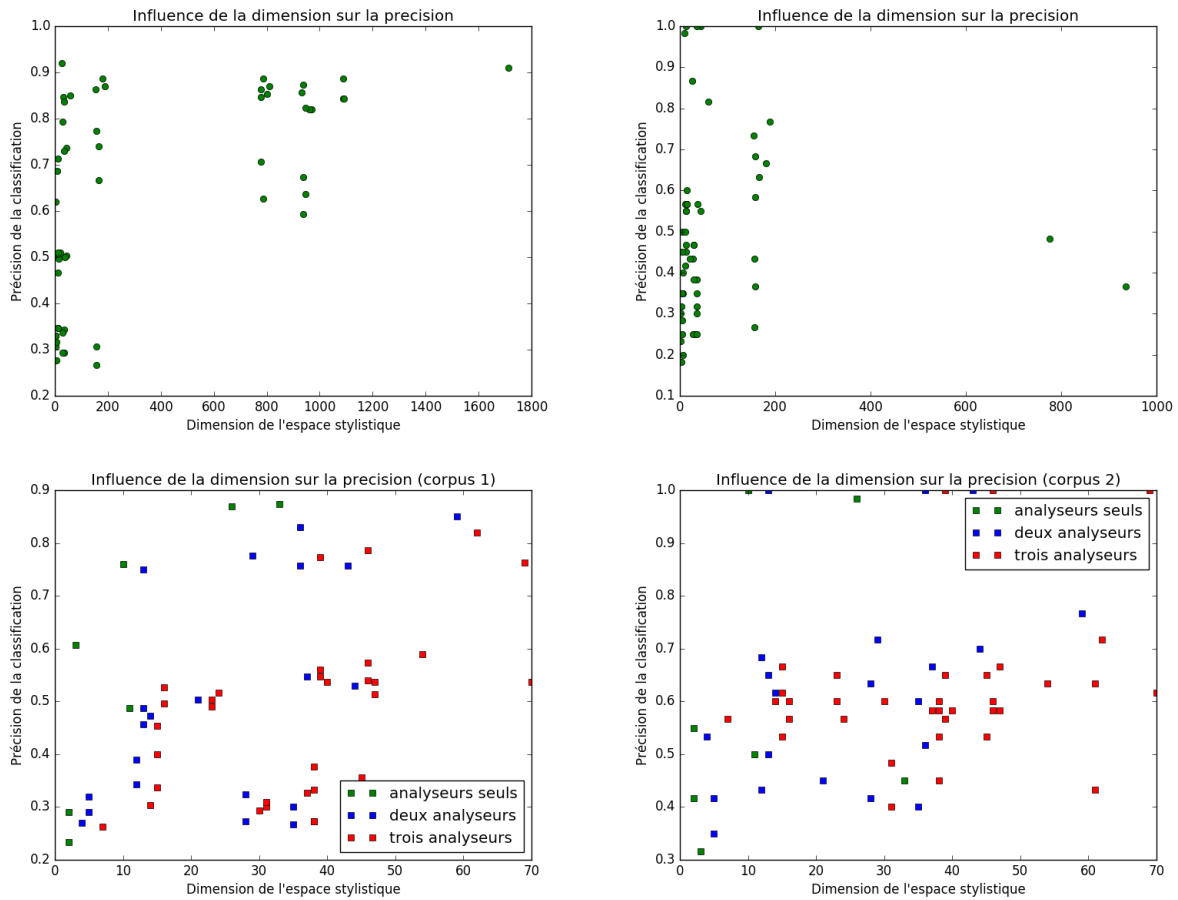


FIGURE 16 – Précision des différents analyseurs pour les corpus 1 et 2

avoir besoin de juger un peu plus finement ce qui fait une bonne ou une mauvaise classification. C'est pourquoi notre projet inclut d'autres outils d'évaluation, détaillés brièvement ci-dessous et implémentés dans l'outil [9] [10]..

**Schéma interne :** Le schéma interne repose uniquement sur les données disponibles, il s'agit par exemple d'évaluer la quantité d'information préservé par la classification par rapport à la dissimilarité de départ. La statistique de Huberts interne s'applique par exemple aux partitionnements tandis que le coefficient de corrélation cophénétique (CPCC) évalue les classifications hiérarchiques

**Schéma externe :** Le schéma externe repose sur une classification préétablie des données. Il s'agit donc de mesurer la qualité du partitionnement obtenu par l'algorithme en le comparant au partitionnement réel. En particulier, on peut s'intéresser aux nombre de liaisons correctes ou incorrectes, interclasses ou intraclasses. Les indices de Rand, de Jaccard ou encore de Fowlkes et Mallows mettent en pratique ce schéma d'évaluation.

**Schéma relatif :** Le schéma relatif s'intéresse à des critères d'inertie interclasse et intraclasse. Il valorise les classifications avec une petite inertie intraclasse et une grande inertie interclasses. L'indice de Dunn ou l'indice de Davies et Bouldin constituent des exemples souvent utilisés de ce critère d'évaluation des algorithmes de classification.

## 4.2 CRITIQUE DE NOS MODÈLES

---

Nous nous intéressons ici aux points faibles de notre démarche, qui sont principalement des problèmes que nous avons rencontrés et que nous n'avons pas résolus de manière totalement satisfaisante.

**Sources** Une de nos principales difficultés a été de rassembler les corpus sur lesquels nous allions travailler. Recherchant des œuvres libres de droits, nous étions déjà restreints aux auteurs antérieurs au XX<sup>ème</sup> siècle. Même pour ces auteurs, nous n'avons pas toujours accès à autant d'œuvres que nous l'aurions souhaité, et nous avons eu du mal à regrouper des corpus cohérents, c'est-à-dire assez conséquents et représentatifs d'un auteur, d'une période ou d'un courant littéraire. Le fait que les textes récoltés proviennent de sites différents et ne soient pas toujours dépourvus d'erreurs a ajouté une certaine dose d'incertitude dans nos applications. Par exemple les textes trouvés sur le site Archives.org comportaient à la fin de chaque page la mention "Digitized by VjOOQIC" ce qui faussait les fréquences d'abréviations et de mots étrangers.

**Analyseurs** Une autre limitation dans les études d'analyse de style et de différenciation d'auteurs que nous avons menées est inhérente à la stylométrie elle-même : en transformant les textes en vecteurs de nombres, on étudie seulement des statistiques et on ne peut rendre compte de phénomènes tels que l'humour, l'ironie, les registres de langue utilisés et la plupart des figures de style. D'autres caractéristiques auraient cependant pu être implémentées, comme le découpage en paragraphes ou même l'alternance entre récit et dialogue. A contrario, on ne peut pas être certain que telle ou telle caractéristique permettant de différencier deux auteurs soit bien une traduction de leurs différences stylistiques, et non une différence liée uniquement au contenu des œuvres que nous avons comparées. Notamment on ne peut pas dire qu'une fréquence anormalement élevée de la lettre w soit caractéristique du style de Marcel Proust ; elle indique plutôt la présence du personnage de Swann.

**Classifieurs** On peut reprocher à notre outil un côté un peu "boîte noire", qui fait qu'une fois les différents algorithmes de classification regroupés sous l'appellation "classifieurs", on peut se servir de l'un ou de l'autre en oubliant qu'ils fonctionnent selon des principes différents. Cet aspect est renforcé par le processus d'équilibrage des textes qui se fait de manière aléatoire.

**Sorties et Résultats** L'interprétation des résultats est l'étape la plus délicate, notamment celle des différentes composantes "significatives" données en sortie du programme. Ces dernières sont classées par défaut selon le rapport de leur variance inter-clusters sur leur variance intra-clusters en raison de la rapidité de cette méthode, mais nous avons vu précédemment que cette

méthode est un peu moins fiable que celle utilisant les poids du réseau de neurones. Il s'agit alors de déterminer quelles caractéristiques des textes sont à l'origine de ces statistiques jugées plutôt différentes entre les deux catégories. Par exemple, si il y a une fréquence élevée de virgules dans les textes de Huysmans, est-ce parce qu'il fait des phrases hachées ou des phrases très longues ? Si les textes de Maupassant comportent beaucoup de tirets, cela traduit-il un grand nombre de dialogues ou un format de texte dans lequel les chapitres sont séparés par une rangée de tirets ? Lors de cette étape nous devons faire appel à notre sens critique et à notre sens littéraire, et nous ne pouvons garantir d'avoir écarté tous les biais des statistiques, ni d'avoir exploité à fond tous les résultats dont nous disposons.

L'évaluation de la classification est aussi importante : si on veut distinguer Zola et Balzac et que l'on utilise comme ensemble d'apprentissage *L'assomoir* et *Le père Goriot*, peut-on dire que le résultat "*L'assomoir* est dans la catégorie Zola (70%), *Le père Goriot* est dans la catégorie Balzac (55%)" est plus précis que celui-ci : "*L'assomoir* est dans la catégorie Zola (80%), *Le père Goriot* est dans la catégorie Zola (55%)" ? Manquant de pistes pour savoir lequel de la petite dizaine d'indices que nous avons implémentés était le plus pertinent pour traduire la fiabilité du résultat, nous nous sommes principalement basés sur la précision de l'attribution (dans le cas, comme ici, où l'on connaît déjà les catégories des textes de l'ensemble d'entraînement). Cela peut être discutable, dans la mesure où les précisions des deux classifications ci-dessus sont égales alors qu'on aimerait traduire le fait que la première a attribué correctement les œuvres entières, ce qui n'est pas le cas de la seconde.

## CONCLUSION

---

Notre questionnement, initialement axé sur Dumas, nous a progressivement mené vers la problématique, plus générale, de la résolution de problèmes de paternité littéraire. L'implémentation d'algorithmes de classification et de vérification nous a conduit à élaborer un outil de traitement systématique de conflits d'attribution de paternité, et de vérification d'identité. Le développement de cet outil a mobilisé nos ressources tout au long du projet, et sa finalisation nous a permis d'étudier de façon quantitative plusieurs controverses littéraires et de nous prononcer à leur sujet, arguments numériques à l'appui.

En complément de ce développement, nous avons entrepris un travail indispensable d'évaluation des performances. D'abord qualitative (rapport intermédiaire), cette évaluation est devenue plus quantitative, avec une validation croisée systématique des fonctions d'analyse, des algorithmes de classification et de vérification et de l'influence de leurs paramètres respectifs. Cette évaluation valide notre travail et fournit des garanties scientifiques pour d'éventuels utilisateurs extérieurs.

A cet égard, le succès global de notre outil et de notre approche systématique dans les nombreuses études de cas (sept au total) que nous avons menées au cours du projet corrobore ce gage de qualité. Il est tout à fait remarquable de constater l'efficacité de l'approche par apprentissage statistique dans ces problématiques. En effet, les algorithmes employés sont simples en comparaison à la richesse du style présent dans un texte mais ceux-ci parviennent tout de même à en saisir l'essence.

Un des résultats les plus surprenant de notre étude concerne la détection de mensonges. Malgré les difficultés et les limites de notre corpus de mensonges et de vérités, une utilisation fine de nos outils a permis de donner une réponse affirmative à cette question. Nous avons ainsi mis en évidence huit facteurs qui caractérisent le style des menteurs, dressant ainsi leur portrait-robot et extrayant la structure de leur ADN stylistique.

En tout état de cause, il convient de nuancer nos réussites, en ce que nos modèles sont critiquables et que les performances de nos algorithmes laissent à penser que des améliorations et des optimisations supplémentaires permettraient d'exploiter plus avant notre modèle et son cadre théorique.

Toutefois, ces possibilités d'amélioration n'empêchent pas l'outil d'être aujourd'hui fonctionnel. Cet outil, fruit de notre travail, peut dès à présent être mis à disposition de la communauté stylométrique, lui proposant un éclairage plus mathématique et formel pour compléter leurs travaux.

## RÉFÉRENCES

---

- [1] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3) :538–556, 2009.
- [2] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1) :9–26, January 2009.
- [3] Patrick Juola. Authorship Attribution. *Found. Trends Inf. Retr.*, 1(3) :233–334, December 2006.
- [4] Steven Bird, Edward Loper, and Edward Klein. *Natural Language Processing with Python*. O’reilly media inc edition, 2009.
- [5] Helmut Schmid. TreeTagger.
- [6] Conrad Sanderson and Simon Guenter. Short text authorship attribution via sequence kernels, Markov chains and author unmasking : An investigation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 482–491. Association for Computational Linguistics, 2006.
- [7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2 edition, 2009.
- [8] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1) :1–37, January 2008.
- [9] Laurent Candillier. *Contextualisation, Visualisation et Evaluation en Apprentissage Non Supervisé*. phdthesis, Université Charles de Gaulle - Lille III, September 2006.
- [10] Guillaume Cleuziou. *Une méthode de classification non-supervisée pour l’apprentissage de règles et la recherche d’information*. PhD thesis, Université d’Orléans, 2004.
- [11] Eric Hervet. *Reseaux de neurones*, 2014.
- [12] María Pelaez Brioso and Rafael Muñoz Guillena. Authorship Verification, Average Similarity Analysis. *Proceedings of Recent Advances in Natural Language Processing*, pages 84–90, 2015.
- [13] Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. Measuring differentiability : Unmasking pseudonymous authors. In *Proceedings of the twenty-first international conference on Machine learning*, page 62. ACM, 2007.
- [14] Nikos Tsimboukakis and George Tambouratzis. A comparative study on authorship attribution classification tasks using both neural network and statistical methods. *Neural Computing and Applications*, 19(4) :573–582, June 2010.
- [15] Antonio Neme, J.R.G. Pulido, Abril Muñoz, Sergio Hernández, and Teresa Dey. Stylistics analysis and authorship attribution algorithms based on self-organizing maps. *Neurocomputing*, 147 :147–159, January 2015.

- [16] Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING'07)*, pages 263–272, 2007.
- [17] Rada Mihalcea and Carlo Strapparava. The lie detector : Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312. Association for Computational Linguistics, 2009.
- [18] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In *2012 IEEE Symposium on Security and Privacy*, pages 461–475. IEEE, 2012.
- [19] Michael Robert Brennan and Rachel Greenstadt. Practical Attacks Against Authorship Recognition Techniques. In *IAAI*, 2009.
- [20] Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. Overview of the 3rd Author Profiling Task at PAN 2015. In *CLEF*, 2015.
- [21] Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2) :119–123, 2009.
- [22] Julia Hirschberg, Stefan Benus, Jason M. Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, and others. Distinguishing deceptive from non-deceptive speech. In *INTERSPEECH*, pages 1833–1836, 2005.
- [23] Charles F. Bond, Jr. and Bella M. DePaulo. Accuracy of Deception Judgments. *Personality and Social Psychology Review*, 10(3) :214–234, July 2006.
- [24] A. Piolat, R.J. Booth, C.K. Chung, M. Davids, and J.W. Pennebaker. La version française du dictionnaire pour le LIWC : modalités de construction et exemples d'utilisation. *Psychologie Française*, 56(3) :145–159, September 2011.
- [25] Tommaso Fornaciari and Massimo Poesio. DeCour : a corpus of DEceptive statements in Italian COURts. In *LREC*, pages 1585–1590, 2012.
- [26] Marc Parizeau. *Réseaux de neurones*, volume 3. Université Laval, 2006.
- [27] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. A framework for authorship identification of online messages : Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3) :378–393, February 2006.
- [28] Thorsten Joachims. *Learning to classify text using support vector machines : Methods, theory and algorithms*. Kluwer Academic Publishers, 2002.
- [29] F Mosteller and D Wallace. *Inference and Disputed Authorship : The Federalist*. Addison-Wesley, 1964.
- [30] André Salem and Ludovic Lebart. *Méthodes de la statistique textuelle*. Paris 3, January 1993.
- [31] T. C. Mendenhall. The Characteristic Curves of Composition. *Science*, 9(214) :237–249, 1887.



- [32] James W. Pennebaker Martha E. Francis and Roger J. Booth. Linguistic Inquiry and Word Count. Technical report, Technical Report, Dallas, TX : Southern Methodist University, 1993.
- [33] F. Mosteller and D. Wallace. *Inference and Disputed Authorship : The Federalist*. Addison-Wesley, 1964.