Regroupement d'auteurs : Qui a écrit cet ensemble de romans ?

Mirco Kocher, Jacques Savoy

Institut d'informatique, Université de Neuchâtel
rue Emile Argand 11, 2000 Neuchâtel (Suisse)
{Mirco.Kocher, Jacques.Savoy}@unine.ch

RÉSUMÉ. *Cet article présente le problème du regroupement d'auteurs c'est-à-dire étant donné un ensemble* n *d'écrits, retournez le nombre* k *d'auteurs et regroupez dans* k *classes les textes par auteur. Liée au problème de l'attribution d'auteur, cette question possède toutefois la propriété d'être non-supervisée. Sur la base de deux collections de documents, une écrite en français, la seconde en anglais, diverses mesures de distance sont proposées et évaluées. Au niveau du choix des attributs, les* m *(avec* m = *50 à 2 000) mots les plus fréquents ou les* m *unigrammes et bigrammes de lettres sont étudiés. Les résultats indiquent que la représentation par mots présente habituellement une meilleure performance que celle basée sur les bigrammes de lettres. La distance calculée par le cosinus offre une qualité moindre que des fonctions basées sur la norme L_1 (à l'exemple de Canberra). Toutefois, le choix de la meilleure mesure de distance ne peut être fixée avec précision. Enfin, nous proposons d'appliquer une forme de ré-échantillonnage aléatoire (bootstrap) afin de tenir compte des variations lexicales. Nos résultats indiquent une variabilité importante des résultats face à des variantes lexicales. Enfin, une analyse détaillée révèle les difficultés et les raisons d'affectations erronées.*

ABSTRACT. *This paper describes the author clustering problem where, based on a set of* n *texts, the number* k *of distinct authors must be determined and the texts must be regrouped into* k *classes according to their author. Using two test collections, one written in French, the second in English, different distance measures are described and evaluated. To define the needed features, the* m *most frequent words (e.g.,* m *between 50 to 300) or the* m *letters and bigrams of letters have been used. Our experiments show that word-based representations offer usually the best performance. Using the cosine distance function does not produce a better F_1 value compared to functions based on the L_1 norm (e.g., Canberra). However, the best distance measure for all cases cannot be defined precisely. Applying a bootstrap approach, we show that the performance measures owns a relatively large variability. Finally, a deeper analysis indicates the difficulties and reasons explaining incorrect assignments.*

MOTS-CLÉS : *Classification automatique, apprentissage non-supervisé, attribution d'auteur.*

KEYWORDS: *Text clustering, unsupervised learning, authorship attribution.*

1. Introduction

Les études en vérification automatique de l'identité d'un auteur (Love, 2002), (Juola, 2006) ont connu un succès grandissant avec la multiplication des canaux électroniques. La présence de messages anonymes ou pseudo-anonymes soulève de nombreux défis en criminalité (Olsson, 2008) à l'exemple des chats calomnieux ou des courriels menaçants. Pourtant des questions plus traditionnelles demeurent sans réponse comme, par exemple, la véritable identité d'Elena Ferrante connue pour ses romans à succès. Est-ce, par exemple, Anita Raja (Gatti, 2016) ou Domenico Starnone (Cortelazzo *et al*., 2016) ? En littérature anglaise, on s'interroge toujours sur les relations de Shakespeare et de ses co-auteurs (Michell, 1996), (Craig & Kinney, 2009) et, en France, sur les liens entre Molière et Corneille (Labbé, 2009).

Face à cette pluralité d'interrogations, quatre grandes familles de problèmes peuvent être définies. Premièrement, connaissant un certain nombre d'écrits d'un ensemble d'auteurs, on doit déterminer quel est le véritable auteur d'un nouveau texte (Stamatatos, 2009). Deux variantes de ce problème ont été proposées. Dans un espace fermé, le véritable auteur est l'un des écrivains donnés. Par contre, dans un cadre ouvert, l'auteur peut être un des noms mentionnés ou un autre, encore inconnu.

Deuxièmement, l'attribution d'auteur peut correspondre à une vérification (Koppel *et al*., 2007) (Koppel & Winter, 2014), (Kocher & Savoy, 2017a). Dans ce cas, on connaît un certain nombre d'écrits d'un seul auteur. Face à un nouveau document (par exemple, un testament, une revendication), on doit déterminer si ce nouvel écrit provient de la plume de l'auteur donné. Cependant, la réponse ne saurait se limiter à une valeur booléenne et une estimation du degré de certitude (ou probabilité) que la réponse donnée soit correcte permet de juger de la confiance dans l'attribution proposée (Savoy, 2016).

Troisièmement, on parle de profilage lorsque l'on ne cherche pas à déterminer le nom de l'auteur mais à identifier certaines de ses caractéristiques démographiques comme son âge approximatif, son sexe, sa langue maternelle, son origine sociale, voire ses traits psychologiques (Argamon *et al*., 2009), (Pennebaker, 2011). Dans cette perspective, on fait l'hypothèse que ces caractéristiques ont une influence sur le style et que des traces peuvent être détectées. Sans grande surprise, on peut estimer qu'un blog parlant de *shopping* sera plus probablement rédigé par une femme tandis qu'un autre discutant de la composition de l'équipe France sera l'œuvre d'un homme. Mais d'autres éléments stylistiques peuvent révéler des renseignements sur l'auteur. Ainsi, Pennebaker (2011) indique qu'en moyenne une femme emploie un plus grand nombre de pronoms, de mots désignant la famille ou les relations sociales et que son texte contient plus d'émotions.

Quatrièmement, le problème du regroupement d'auteurs ne cherche pas directement à déterminer l'auteur de chaque écrit. Disposant d'un ensemble de n textes, on doit déterminer le nombre k d'auteurs. De plus, on doit regrouper en k

classes les *n* textes, chacune contenant tous les écrits du même auteur. Ce problème a été posé dernièrement lors de la campagne d'évaluation PAN-CLEF 2016 (Stamatatos *et al*., 2016) et constitue le cœur de cette communication.

Dans la suite de cet article, nous présenterons un survol des connaissances en attribution d'auteur (section 2). La troisième section expose les grandes lignes des deux corpus utilisés dans nos expériences ainsi que la méthodologie d'évaluation. La quatrième décrit notre approche pour résoudre ce problème. Une cinquième présente une analyse de sensibilité et ouvre des perspectives afin de définir par apprentissage quelques seuils. Enfin, une dernière section analyse quelques exemples afin de mieux comprendre les erreurs d'affectation.

2. État des connaissances

Afin de résoudre les problèmes de l'attribution d'auteur, trois grandes familles d'approches ont été proposées (Juola, 2006). En premier lieu, on admet que le style demeure invariant pour une personne donnée (ou, pour le moins, dès que l'on atteint l'âge de 25 à 30 ans). Sur ce constat, on a proposé de recourir à des mesures stylométriques supposées invariantes comme la longueur moyenne des phrases, le nombre moyen de syllabes par mots, voire la taille du vocabulaire V (notée |V|) par rapport à la longueur du document (rapport TTR, *type-token ratio*). Comme variantes, on a suggéré le rapport entre le nombre de *hapax legomena* (mots apparaissant une seule fois) (notée V_1) et la taille du vocabulaire (soit $|V_1| / |V|$), ou le rapport entre le nombre de mots apparaissant deux fois (noté $|V_2|$) et la taille du vocabulaire (Rexha *et al*., 2016). Toutes ces mesures possèdent l'inconvénient d'être difficiles à interpréter et instable face à des textes de tailles différentes (Baayen, 2008).

Une deuxième famille d'approches se fonde sur le vocabulaire. Dans cette perspective, Mosteller & Wallace (1964) proposent de sélectionner de manière semi-automatique les vocables les plus pertinents. Ces travaux mettent en lumière l'importance des mots fréquents et, en particulier, des mots fonctionnels (déterminants, prépositions, conjonctions, pronoms et verbes auxiliaires). En poursuivant cette voie, Burrows (2002) propose de sélectionner les mots en se basant sur la fréquence d'occurrence. Ainsi la liste des attributs à retenir comprendra entre 50 à 150 vocables les plus fréquents, ensemble comprenant une forte proportion de mots fonctionnels. Ce seuil sera repoussé à 800 puis à 4 000 (Hoover, 2007) avec l'inclusion de mots lexicaux fréquents (noms, adjectifs, adverbes et verbes).

Les études menées par Zhao & Zobel (2007) proposent de définir *a priori* les vocables à retenir. Dans ce cas, on retient essentiellement les mots fonctionnels en ignorant les mots lexicaux qui reflètent plus les thèmes traités. Pour la langue anglaise, ces auteurs suggèrent une liste de 363 formes, un ensemble correspondant au contenu d'une liste de mots-outils d'un moteur de recherche. Dans une

perspective similaire, Hughes *et al*. (2012) proposent de retenir 307 mots (fonctionnels) afin de décrire les styles littéraires couvrant une période de 350 ans.

Dès lors, chaque texte peut être représenté par les attributs définis. Ensuite, une mesure de distance (ou de similarité) permet d'estimer la proximité de deux textes. Par exemple, Labbé (2007) propose une mesure de distance lexicale basée sur l'ensemble du vocabulaire et les fréquences d'occurrences. L'attribution s'établit selon la règle du plus proche voisin.

Comme troisième famille d'approches, nous pouvons signaler le recours à des techniques d'apprentissage automatique (*machine learning*) que l'on retrouve également dans la catégorisation automatique (Sebastiani, 2002), (Stamatatos, 2009). Dans ce cas, le système doit d'abord sélectionner les attributs (mots, bigrammes de mots ou de lettres, partie du discours, émoticons, abréviations, URL, présence de salutation, etc.) possédant le meilleur pouvoir discriminant, puis entraîner un classifieur. Ce type d'approche requiert un ensemble d'entrainement, données qui ne sont pas disponibles dans le cadre du regroupement d'auteurs.

Dès lors, pour résoudre ce problème, des approches proposent de déterminer en premier lieu le nombre *k* d'auteurs sur l'ensemble *n* d'écrits. Une fois cette valeur fixée, on applique un algorithme de classification *k-means* afin d'identifier les différents groupes de documents. Par itération successive, le nombre *k* d'auteurs peut être affiné. Comme second paradigme, la distance entre chaque écrit est calculée, puis on applique un algorithme de classification hiérarchique pour former les grappes de documents. Par exemple, Labbé (2007) recourt à un algorithme de classification hiérarchique (lien complet). Dans cette étude, nous suivrons cette seconde stratégie de résolution, choix qui nous a permis d'obtenir le deuxième rang lors de la dernière campagne d'évaluation PAN-CLEF 2016.

3. Corpus d'évaluation et méthode d'évaluation

Comme en recherche d'information, l'évaluation empirique tient une place importante en catégorisation de textes. Dans le cadre du regroupement d'auteurs, la création de collections tests ne s'avère pas une tâche trop lourde. Toutefois, les corpus créés lors de la campagne PAN-CLEF 2016 n'ont pas été rendus publics (certainement en vue d'une réutilisation ultérieure). Nos évaluations seront donc basées sur deux collections extraites d'œuvres littéraires couvrant la fin du XVIIIe au début du XXe siècle. Chaque texte retenu a été rédigé par un seul auteur. Si on estime que cette affirmation est erronée, quelques techniques récentes proposent de déterminer quel passage a été écrit par quel auteur (Rybicki *et al*., 2014), (Eder, 2015), (Rexha *et al*., 2016).

Pour la langue anglaise, nous disposons du corpus nommé Oxquarry1 composé de 52 fragments de romans, chacun d'environ 10 400 vocables (sans la ponctuation). La distribution par auteurs et titres des œuvres retenues est indiquée dans l'annexe, tableau A.1. Ce corpus a été écrit par neuf auteurs et chaque écrivain apparaît au

moins avec trois textes (c'est le cas pour Chesterton, Forster ou Tressel). Hardy s'avère l'auteur le plus fréquent avec 12 extraits, suivi par Conrad (8), Stevenson (7), Morris (6), Orczy (6) et Butler (4). Pour chaque œuvre retenue, au moins deux extraits apparaissent dans ce corpus. Cette contrainte a été introduite afin de faciliter les affectations (à l'époque, on n'était pas certain de pouvoir résoudre un tel problème avec une performance élévée).

Pour la langue française, le corpus nommé Brunet comprend 44 textes, chacun comprenant environ 7 650 lemmes. La distribution par auteur est reprise dans le tableau A.2. On y retrouve onze auteurs et pour chacun d'eux, deux œuvres ont été sélectionnées desquelles deux extraits apparaissent dans le corpus. La solution pour ce corpus se compose donc de onze groupes, chacun comprenant quatre textes. Toutefois, le roman *Le Secret de Wilhelm Storitz* n'a pas une paternité claire et pourrait être attribué à Michel Vernes, fils de Jules Vernes, qui l'a remanié.

Pour définir le regroupement parfait en onze groupes, nous avons besoin de six liens par auteur (quatre documents à relier entre eux = (3 x 4) / 2 = 6) et ceci pour les onze auteurs (nombre de liens intertextuels minimum = 66). Pour le corpus anglais, le nombre minimal de liens s'élève à 163 principalement en raison du nombre important de textes écrits par Hardy (12), Conrad (8) et Stevenson (7).

Comme mesure d'évaluation, nous reprenons les mesures proposées lors de la campagne d'évaluation PAN-CLEF 2016. En premier lieu, nous définissons la fonction binaire *cor*() entre deux documents d_i et d_j selon la formulation suivante :

$$cor(d_i, d_j) = \begin{cases} 1, \text{ si } A(d_i) = A(d_j) \wedge C(d_i) = C(d_j) \\ 0, \text{ autrement} \end{cases} \quad (1)$$

avec $A(d_i)$ indiquant le véritable auteur du document d_i, et $C(d_i)$ est la classe dans laquelle le document d_i est assigné. Sur cette fonction la précision (notée *prec*(d_i)) et le rappel (*rap*(d_i)) se définissent comme suit :

$$prec(d_i) = \sum_{d_j \in C(d_i)} cor(d_i, d_j) \Big/ |C(d_i)| \quad rap(d_i) = \sum_{d_j \in C(d_i)} cor(d_i, d_j) \Big/ |A_i| \quad (2)$$

avec $|C(d_i)|$ donnant le nombre de textes dans la classe $C(d_i)$ et $|A_i|$ le nombre de documents écrit par l'auteur A_i dans l'ensemble du corpus.

Pour illustrer l'application de ces deux mesures, prenons l'exemple d'une classe comprenant trois textes dont deux sont rédigés par le même auteur (soit d_1 et d_2) et le troisième (d_3) par un auteur différent. Ce regroupement contient donc une erreur. Pour calculer la précision assignée au document d_1, on débute par le calcul de la fonction *cor*(d_1, d_i) pour tous les membres de la classe. Pour *cor*(d_1, d_2) on obtient la valeur 1, de même que *cor*(d_1, d_1) tandis que *cor*(d_1, d_3) indique 0. La valeur de *prec*(d_1) = (1+1+0) / 3 = 2/3. Pour le rappel associé au document d_1, on obtient 2 / $|A_i|$, valeur dépendant du nombre de textes de l'auteur du document d_1.

Dans le cadre du regroupement d'auteurs, une précision de 1 (ou 100 %) s'avère aisée à obtenir ; il suffit de créer un groupe pour chaque document. Par contre la valeur du rappel restera très faible.

Sur la base de *n* documents composant un corpus, la précision et le rappel global correspondent la moyenne arithmétique des précisions et rappels selon l'équation suivante :

$$\text{prec} = {}^{1}/_{n} \sum_{i=1}^{n} prec(d_i) \qquad \text{rappel} = {}^{1}/_{n} \sum_{i=1}^{n} rap(d_i) \tag{3}$$

Finalement, la valeur F_1 définie comme (2 x prec x rappel) / (prec + rappel) permet d'obtenir une valeur de performance unique pour chaque modèle de classification. Cette valeur sert de clé de tri pour classer les différentes solutions proposées.

4. Choix des attributs et mesure de distance

Afin de regrouper les documents selon leur auteur, nous devons les représenter en fonction de leur style et non en fonction des thèmes qu'ils abordent. Comme mentionné précédemment, plusieurs études ont démontré que les vocables les plus fréquents ou les mots fonctionnels constituent des attributs pertinents pour détecter le style d'un auteur. Le vocabulaire lié aux thèmes s'avère pertinent en recherche d'information ou pour classer les textes selon des vedettes-matières. Dans le cadre du regroupement d'auteurs, le thème peut perturber les bonnes affectations, par exemple, lorsque deux auteurs abordent des sujets similaires.

Pour cerner les aspects stylistiques, une étude récente a démontré que tenir compte des 200 à 300 mots et signes de ponctuation les plus fréquents (Savoy, 2015) apporte de très bonnes performances en attribution d'auteur comparé à d'autres fonctions de sélection (rapport des cotes, gain d'information, chi-carré, etc.). Nous avons repris cette stratégie mais en tenant compte des lemmes (entrée dans le dictionnaire) et en ignorant les signes de ponctuation et les nombres. Certes, tenir compte de la ponctuation permet d'améliorer légèrement la performance globale (en moyenne de 2 % à 8 %). Toutefois, elle rend un peu plus complexe la comparaison entre une représentation par lemmes et celle par lettres et bigrammes de lettres. Donc, nous avons exclu la ponctuation de nos expériences. Pour la langue française, les mots les plus fréquents de notre corpus sont : le (38 293 occurrences), de (27 421), à (10 005), il (9 381), et (8 949).

Comme alternative, plusieurs études proposent de recourir aux fréquences des lettres et des bigrammes de lettres afin de distinguer les différents styles (Kjell, 1994), (Juola, 2006). Dans cette étude, la distinction entre majuscules et minuscules est ignorée et les signes de ponctuation sont éliminés. Par contre, on tiendra compte du fait qu'une lettre débute ou termine un mot. Le nombre maximal d'attributs s'élève à (27 x 27) + 27 = 756. Pour la langue française, on retrouve 576 (ou 76,2 %) combinaisons possibles dans notre corpus (580 (ou 76,7 %) pour l'anglais).

Les lettres françaises les plus fréquentes sont : e (289 269), r (168 416), i (154 846), n (111 312), et o (99 677). En indiquant par _ l'espace, les bigrammes de lettres les plus usuels sont : e_ (168 416), r_ (58 976), le (53 129), _l (47 452), _d (44 544).

Dès que chaque document est représenté par un nombre *m* de lemmes (ou de bigrammes de lettres), on peut calculer sa distance intertextuelle avec les autres entités du corpus. Le choix de cette fonction de distance peut s'opérer selon des critères théoriques (par exemple, symétrie, inégalité triangulaire) ou empiriques. Basée sur le profilage d'auteur, une étude récente (Kocher & Savoy, 2017b) indique qu'aucune mesure de distance s'avère toujours la meilleure. Par contre un groupe restreint permet d'obtenir de bonnes performances comme la distance de Canberra basée sur la norme L_1, ou Matusita (norme L_2) ou JDivergence (basée sur l'entropie). Nous avons repris ces mesures en y ajoutant la distance du cosinus et celle de Labbé (une variante de la distance de Manhattan). La définition de ces mesures de distance est reprise dans l'annexe.

La fonction étant choisie, le système calcule les distances entre tous les textes. Enfin, en partant de la distance la plus faible, on regroupe par paires les documents (lien) en faisant l'hypothèse qu'ils sont écrits par le même auteur.

Rang	Distance	Texte 1	Texte 2
1	0,192	14 Flaubert	36 Flaubert
2	0,205	2 Marivaux	24 Marivaux
3	0,205	4 Voltaire	26 Voltaire
4	0,211	6 Rousseau	28 Rousseau
5	0,216	2 Marivaux	23 Marivaux
6	0,223	23 Marivaux	24 Marivaux
7	0,227	1 Marivaux	23 Marivaux
...	…	…	…
12	0,253	15 Maupassant	35 Flaubert

Tableau 1. Exemple d'une liste ordonnée selon la distance intertextuelle

À titre illustratif, le tableau 1 indique les distances de Manhattan les plus faibles pour le corpus français en représentant les documents avec les 300 lemmes les plus fréquents. Dans ce bref extrait, on constate que lorsque les distances diminuent, le risque d'erreur s'amoindrit. De plus, l'établissement du premier lien s'effectue, habituellement, avec un second extrait de la même œuvre (e.g., *Bouvard et Pécuchet* pour Flaubert (n° 14 et n° 36), *Le paysan parvenu* avec Marivaux (n° 2 et n° 24), *Candide* pour Voltaire (n° 4, n° 26)). Par contre, au cinquième rang, le lien entre les textes n° 2 et n° 23 s'effectue sur deux œuvres différentes. En tenant compte uniquement des sept distances les plus courtes, nous pouvons former une classe pour Flaubert (extrait n° 14 et n° 36), une pour Voltaire (n° 4 et n° 26), une pour Rousseau (n° 6, n° 28) et un regroupement plus large pour Marivaux (n° 1, n° 2, n° 24, n° 23). En consultant l'annexe, on constatera que le système a découvert toutes les œuvres de Marivaux indiquant que son style se distingue clairement des autres écrivains. Le premier lien erroné apparaît en 12e position avec les extraits n° 15 (Maupassant, *Une vie*) et n° 35 (Flaubert, *Madame Bovary*).

Évidemment, lorsque l'on génère un nouveau regroupement, il convient de recalculer la distance entre le groupe nouvellement formé et les autres. Ainsi, sur le tableau 1, une fois le groupe « Flaubert » créé (n° 14 et n° 36), nous devons mesurer les distances entre cette grappe et les autres. Afin de favoriser des groupes homogènes, nous avons opté pour le *lien complet*, signifiant que la distance entre deux grappes correspond à la distance la plus grande entre toutes les paires d'individu appartement aux deux groupes.

Dans le tableau 2, nous avons repris les mesures de performance F_1 obtenues avec le corpus anglais en faisant varier les fonctions de distance, le nombre de lemmes et le nombre de liens entre documents (colonne « Param. ») (le nombre de liens est fixé à 40 dans le cas présent). Le haut du tableau correspond à une approche basée sur les lettres et leurs bigrammes. La partie inférieure repose sur une représentation à base des lemmes les plus fréquents. Comme ce corpus possède trois auteurs avec un nombre relativement important de textes, le nombre de liens à prendre en compte s'avère plus élevé. Dans le tableau 3, nous avons repris les mêmes mesures de performance obtenues cette fois avec le corpus français. Toutefois, le nombre minimal de liens pour définir ce corpus étant plus faible, nous avons réduit cette valeur à 30 dans notre analyse.

Param.	Manhattan	Canberra	Labbé	Matusita	Cosinus	JDivergence
50 / 40	0,567	0,590	0,567	0,607	0,656	0,607
100 / 40	0,665	0,659	0,665	0,614	0,699	0,652
150 / 40	0,675	0,705	0,675	0,812	0,705	**0,812**
200 / 40	0,694	0,745	0,694	0,695	0,705	0,694
500 / 40	**0,787**	**0,787**	**0,787**	**0,917**	**0,829**	0,791
50 / 40	0,589	0,721	0,695	0,688	0,604	0,688
200 / 40	0,773	0,854	0,863	0,840	0,679	0,840
300 / 40	0,773	**0,866**	0,770	0,868	0,679	0,868
500 / 40	0,770	0,863	**0,865**	0,901	**0,743**	0,822
2000 / 40	**0,803**	0,818	**0,865**	**0,913**	**0,743**	**0,901**

Tableau 2. Mesure de performance F_1 avec le corpus anglais selon plusieurs mesures de distance (lien complet, lettres en haut, lemmes en bas)

Les valeurs reprises dans le tableau 2 indique que la représentation par des lettres et bigrammes de lettres apportent habituellement des valeurs de performance inférieures à une représentation basée sur des lemmes. Cette dernière possède l'avantage supplémentaire d'être plus compréhensible. De plus, la fonction cosinus ne fournit pas de résultats probants par rapport aux autres mesures et celle de Manhattan offre une performance légèrement inférieure aux autres. Les quatre autres formulations donnent des performances similaires pour le corpus anglais (tableau 2) tandis que pour le corpus français (tableau 3), Canberra propose une valeur F_1 légèrement supérieure. Enfin, la dernière ligne de chaque partie des tableaux 2 et 3 est obtenue avec une représentation tenant compte à la fois du style (avec les 200 à 500 lemmes les plus fréquents) et des thèmes (avec les lemmes ayant des fréquences d'occurrence entre 501^e et 2 000^e rang).

Param.	Manhattan	Canberra	Labbé	Matusita	Cosinus	JDivergence
50 / 30	0,626	0,592	0,626	0,727	0,700	0,727
100 / 30	**0,709**	0,724	**0,709**	0,765	0,705	0,765
150 / 30	0,705	0,693	0,705	**0,782**	**0,711**	**0,782**
200 / 30	0,705	**0,745**	0,705	0,727	**0,711**	0,727
500 / 30	0,705	0,736	0,705	0,659	**0,711**	0,659
50 / 30	0,610	0,702	0,605	0,639	0,607	0,575
200 / 30	0,636	0,761	0,664	0,705	0,620	0,705
300 / 30	0,627	0,758	**0,725**	0,721	**0,682**	**0,770**
500 / 30	0,634	**0,829**	0,666	**0,770**	0,620	0,724
2000 / 30	**0,659**	0,766	0,679	**0,770**	0,626	0,724

Tableau 3. Mesure de performance F_1 avec le corpus français selon plusieurs mesures de distance (lien complet, lettres en haut, lemmes en bas)

Fixer le nombre de liens à retenir lors de la construction des différentes classes constitue le paramètre affectant le plus la performance (voir tableau 4). Il est également celui qui s'avère le plus difficile à estimer car il dépend de la distribution des textes sur le nombre *k* d'auteurs. Sachant que le corpus anglais se compose d'un nombre plus élevé de textes et plus faible d'auteurs (52 extraits, 9 auteurs), le nombre minimal de liens requis sera plus important que pour le corpus français (44 textes, 11 auteurs).

	Corpus anglais			Corpus français		
Param.	Canberra	Manhattan	Matusita	Canberra	Manhattan	Matusita
300 / 10	0,402	0,387	0,385	0,533	0,554	0,533
300 / 20	0,501	0,503	0,503	0,687	0,646	0,646
300 / 25	0,553	0,576	0,573	**0,760**	**0,651**	**0,755**
300 / 30	0,669	0,656	0,656	0,758	0,627	0,721
300 / 35	0,763	0,737	0,735	0,638	0,552	0,621
300 / 40	**0,866**	**0,773**	**0,868**	0,395	0,370	0,372
300 / 50	0,432	0,313	0,377	0,167	0,167	0,167
300 / 60	0,242	0,242	0,242	0,167	0,167	0,167

Tableau 4. Valeur de la mesure F_1 avec le corpus anglais (à gauche) et français (à droite) et avec une variation du nombre de liens (300 lemmes)

En faisant varier tous les paramètres et pour le corpus anglais, la meilleure performance s'élève à 0,901. Elle s'obtient en s'appuyant sur une représentation basée sur les 400 lemmes les plus fréquents, avec la fonction de Matusita, et en tenant compte de 40 liens. Pour la collection française, la valeur F_1 la plus forte s'élève à 0,881 (400 lemmes, la fonction de Canberra, et 30 liens). En comparant ces valeurs extrêmes avec les performances indiquées dans le tableau 4, on constate que la différence s'avère relativement faible (anglais : 0,901 vs. 0,868 (-3.7 %) ; français : 0,881 vs. 0,760 (-13.7 %)). Ces valeurs indiquent que les lemmes fonctionnels reflètent bien le style distinct de chaque auteur et que l'ajout de lemmes liés aux thèmes des documents apportent une amélioration faible de la performance.

Le nombre de liens (40 pour le corpus anglais et 30 pour le français) est plus faible que le minimum requis pour obtenir un parfait regroupement (soit 163 pour la collection anglaise et 66 pour le français). Ces valeurs optimales indiquent bien que prendre en compte un nombre plus élevé de paires de documents conduit à inclure un nombre croissant de liens intertextuels erronés.

Implicitement, nous avons fait l'hypothèse que le lien complet permettait de définir des groupes homogènes et favoriserait la discrimination entre auteurs. Or la distance entre deux groupes peut également se fonder sur la distance la plus courte entre tous les éléments individuels appartement aux deux grappes (lien simple) voire sur la moyenne des distances entre les paires d'éléments appartenant à chacun des deux groupes.

Les performances reportées dans le tableau 5 indiquent que le choix du lien complet n'est pas forcément le meilleur dans tous les cas. Comme alternative intéressante, nous pouvons suggérer le lien moyen apportant une meilleure performance pour le corpus anglais.

Corpus anglais			Corpus français		
Param.	Lien	F_1	Param.	Lien	F_1
300 / 40	complet	0,866	300 / 30	complet	0,758
300 / 40	moyen	**0,923**	300 / 30	moyen	**0,803**
300 / 40	simple	0,571	300 / 30	simple	0,769
400 / 40	complet	0,863	400 / 30	complet	**0,881**
400 / 40	moyen	**0,904**	400 / 30	moyen	0,822
400 / 40	simple	0,580	400 / 30	simple	0,717

Tableau 5. Variation de la mesure F_1 en appliquant la fonction de distance Canberra et en variant le calcul de distance entre groupes

5. Analyse de sensibilité

Afin de pouvoir estimer la variabilité sous-jacente aux mesures de performance, nous avons appliqué une ré-échantillonnage aléatoire avec remplacement (*bootstrap*). Pour chaque texte, l'ordinateur générera 200 nouveaux extraits, tous possédant la même longueur. Dans chaque copie, la probabilité de choisir un lemme dépend uniquement de sa fréquence relative dans le texte original.

Pour la langue anglaise, la représentation est basée sur les 300 lemmes les plus fréquents, avec la fonction Canberra et en tenant compte des 40 paires de documents les plus similaires. Avec ces paramètres, la valeur F_1 se monte à 0,866 (voir tableau 4). Sur l'ensemble des 200 copies, la valeur F_1 moyenne s'élève à 0,888 (min : 0,766, max : 0,934, stdev : 0,032). Pour le corpus français, les mêmes valeurs des paramètres ont été choisies, excepté pour le nombre de liens limité à 30. Dans le tableau 4, la performance F_1 obtenue est de 0,758. Avec les 200 échantillons, la distribution des valeurs F_1 varie de 0,654 (min) à 0,892 (max) pour une moyenne de 0,781 (stdev : 0,042).

Dans le tableau 6, d'autres exemples sont indiqués avec les fonctions de distance Canberra (Can.) ou Manhattan (Man). La moyenne obtenue avec un échantillon de 200 copies est reprise ainsi que les percentiles 5 % et 95 %. Ces dernières valeurs indiquent que les mesures de performance possèdent une variabilité importante. Dès lors, on peut considérer que de petites différences entre des valeurs de performance ne seront pas statistiquement significatives.

Corpus anglais				Corpus français			
Param.	Distance	Moy.	Percentile	Param.	Distance	Moy.	Percentile
300 / 35	Can	0,751	0,714-0,793	300 / 20	Can	0,678	0,646-0,706
300 / 40	Can	0,888	0,832-0,923	300 / 25	Can	0,761	0,712-0,803
300 / 50	Can	0,375	0,338-0,432	300 / 30	Can	0,781	0,712-0,836
300 / 40	Man	0,741	0,661-0,820	300 / 30	Man	0,631	0,573-0,705

Tableau 6. Ré-échantillonnage aléatoire sur la mesure F_1

La détermination du nombre de paires de documents à prendre en compte joue un rôle essentiel dans la performance. En fixant la même représentation pour les deux corpus, nous estimons que nous pouvons apprendre à distinguer si une valeur de distance reflète plutôt un appariement entre textes rédigés par le même auteur ou par deux écrivains distincts. En effet, une faible distance possède une probabilité plus forte d'être calculée entre deux textes écrits par le même auteur.

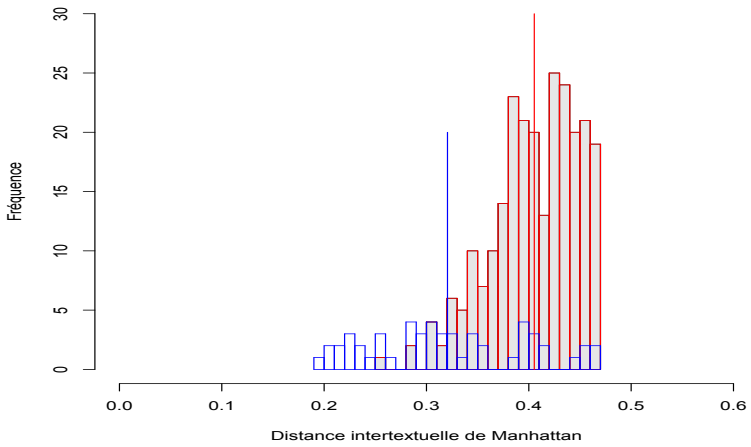

Figure 1. Distribution des distances avec, en clair (bleu), les distances entre paires du même auteur, et en gris (rouge) entre paires de deux auteurs différents

En vue d'illustrer cette hypothèse, la fonction Manhattan a été utilisée pour calculer les distances entre tous les textes du corpus français. Dans la figure 1, la distribution des distances entre paires de documents rédigés par le même écrivain est visualisée en clair (bleu). De plus, la moyenne de ces valeurs est signalée par une ligne verticale (bleue) placée sur la gauche. La distribution entre textes écrits par

deux auteurs distincts est représentée en gris (rouge) et leur moyenne est indiquée par une barre verticale rouge (à droite du graphique). En sélectionnant une autre fonction de distance, une figure similaire peut être obtenue.

6. Analyse des erreurs

Une analyse des erreurs permet de connaître les raisons expliquant une faible distance entre deux textes rédigés par deux auteurs distincts (ou la présence d'une distance importante entre deux textes écrits par le même écrivain). Le tableau 7 indique les dix premières erreurs de notre stratégie (300 lemmes les plus fréquents) avec une distance calculée par la fonction Manhattan (corpus français).

Rang	Distance	Texte 1	Texte 2
1	0,192	14 Flaubert	36 Flaubert
…	…	…	…
12	0,253	15 Maupassant	35 Flaubert
20	0,286	18 Zola	38 Maupassant
21	0,287	13 Flaubert	37 Maupassant
26	0,303	14 Flaubert	15 Maupassant
30	0,307	9 Balzac	19 Vernes
32	0,309	13 Flaubert	17 Zola
33	0,314	13 Flaubert	40 Zola
35	0,316	16 Maupassant	40 Zola
38	0,321	10 Balzac	19 Vernes
39	0,322	13 Flaubert	15 Maupassant

Tableau 7. Exemple des erreurs sur le corpus français

Dans ce tableau, la plus faible distance de Manhattan (0,192) se situe entre deux extraits de Flaubert (*Bouvard et Pécuchet*). La première erreur se situe au 12^{e} rang avec un lien erroné entre Maupassant (n^{o} 15, *Une vie)* et Flaubert (n^{o} 35, *Madame Bovary*). Les prochaines affectations incorrectes se situent entre des œuvres de Flaubert, Maupassant, Zola et, pour un cas, entre Balzac et Vernes.

Flaubert n^{o} 13		Maupassant n^{o} 37		Zola n^{o} 17		Zola n^{o} 40		Marivaux n^{o} 2	
Prob.	lem.	Prob.	lem.	Prob.	lem.	Prob.	lem.	Prob.	lem.
15,7%	le	14,8%	le	15,1%	le	14,7%	le	7,1%	je
9,9%	de	8,3%	de	11,6%	de	10,0%	de	7,1%	le
5,4%	il	4,4%	il	5,0%	un	5,2%	il	6,9%	de
4,3%	à	3,8%	un	4,7%	il	4,1%	un	4,8%	que
3,5%	son	3,8%	et	3,7%	à	3,2%	à	4,4%	être

Tableau 8. Les cinq lemmes les plus fréquents de différents extraits d'œuvre

Pour le corpus anglais, la première erreur se situe au 69^{e} rang et parmi les dix premières erreurs, Hardy se retrouve dix fois avec ces œuvres de Orczy (5 fois), Forster (3), Conrad (1) et Tressel (1).

Revenons à la littérature française avec l'extrait n° 13 (*Madame Bovary*) de Flaubert qui est associé par erreur avec les textes n° 37 (Maupassant, *Une vie*), n° 17 (Zola, *Thérèse Raquin*) et n° 40 (Zola, *La bête humaine*). Afin de comprendre ces rapprochements erronés, le tableau 8 présente les cinq lemmes les plus fréquents extraits de ces œuvres. Ces informations indiquent bien que le style (limité à ces cinq caractéristiques) de ces textes s'avère similaire, en particulier en comparaison avec l'extrait n° 2 (*Le paysan parvenu*) de Marivaux.

7. Conclusion

Le regroupement d'auteurs constitue un nouveau problème dans le contexte de l'attribution d'auteur et une nouvelle piste a été créée en 2016 dans le cadre des campagnes PAN-CLEF. Ce problème se définit comme suit : ayant *n* textes écrits par *k* auteurs, le système doit regrouper, dans *k* classes distinctes, les textes rédigés par la même personne. Pour résoudre ce défi, nous proposons de représenter les textes selon leur style en recourant aux lemmes *m* les plus fréquents (avec *m* = 200 à 500). En recourant à la fréquence des lettres et de leurs bigrammes, les performances obtenues s'avérèrent habituellement moins élevées que la langue soit anglaise ou française (voir tableau 2 et 3).

Inclure plus de lemmes permet, parfois, d'augmenter la performance mais implique de tenir compte des mots reliés aux thèmes abordées dans les documents. Le risque s'avère plus élevé de faire un sur-apprentissage (voir tableau 4). Tenir compte des lemmes lexicaux implique que les thèmes sont liés aux auteurs et que cette relation n'est peut-être valide que pour une collection de test donné.

Afin de calculer la distance intertextuelle, nos évaluations empiriques indiquent que la fonction du cosinus offre des performances moindres que des distances basées sur la norme L_1 (comme celle de Canberra). Il convient toutefois de signaler qu'aucune fonction de distance apporte, dans tous les cas, la meilleure efficacité (voir tableau 2 et 3).

Pour les fonctions de distance les plus efficientes, notre étude démontre que les valeurs les plus faibles indiquent des appariements entre textes rédigés par le même écrivain (voir figure 1). Toutefois la distinction entre deux groupes de valeurs de distance (valeurs faibles ou fortes) n'est pas très marquée. Au niveau de la mesure de distance entre groupes de textes, le lien complet peut être vue comme une solution conservatrice. Toutefois, le lien moyen tend à apporter de meilleures performances moyennes.

L'analyse détaillée des affectations proposées par la machine indique que certains auteurs possèdent un style très particulier qui permet plus aisément de les distinguer des autres, à l'exemple de Marivaux en littérature française. Pour d'autres écrivains (e.g., Maupassant, Flaubert, Zola ou Hardy et Orczy), les styles demeurent assez proches, du moins lorsqu'on les analyse par les lemmes fonctionnels les plus fréquents.

Remerciements

Nos remerciements à D. Labbé pour nous avoir permis d'obtenir les deux corpus de documents utilisés lors de nos évaluations et, en particulier, la collection française lemmatisée. Les auteurs remercient les relecteurs pour leurs commentaires constructifs. Cette recherche a été subventionnée, en partie, par le Fonds National Suisse pour la Recherche Scientifique, subvention n° 200021_149665/1.

8. Bibliographie

Argamon, S., Koppel, M., Pennebaker, J.W., & Schler, J. 2009. Automatically Profiling the Author of an Anonymous Text. *Communications of the ACM*, 52(2), 119-123.

Baayen, H.R. 2008. *Analysis Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press, Cambridge.

Burrows J.F. 2002. Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3), 267-287.

Craig, H., & Kinney, A.F. 2009. *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge University Press, Cambridge.

Cortelazzo, M.A., Nadalutti, P., Ondelli, S., & Tuzzi, A. 2016. Authorship Attribution and Text Clustering for Contemporary Italian Novels. *Proceedings Qualico 2017*, 7-8.

Eder, M. 2015. Rolling Stylometry. *Digital Scholarship in the Humanities*, 31(3), 457-469.

Gatti, C. 2016. La véritable identité d'Elena Ferrante révélée. BublioObs, 2 octobre 2016.

Hoover D.L. 2007. Corpus Stylistics, Stylometry, and the Styles of *Henry James*. *Style*, 41(2), 160-189.

Hughes, J.M., Foti, N.J., Krakauer, D.C., & Rockmore, D.N. 2012. Quantitative Patterns of Stylistic Influence in the Evolution of Literature. *Proceedings of the PNAS,* 109(20), pp. 7682-7686.

Juola, P. 2006. Authorship Attribution. *Foundations and Trends in Information Retrieval*.

Kjell, B. 1994. Authorship Determination Using Letter Pair Frequency Features with Neural Network Classifier. *Literary and Linguistics Computing*, 9(2), 119-124.

Kocher, M., & Savoy, J. 2017a. A Simple and Efficient Algorithm for Authorship Verification. *Journal of the American Society for Information Science and Technology*, 68(1), 259-269.

Kocher, M., & Savoy, J. 2017b. Distance Measures in Author Profiling. *Information Processing & Management,* minor revisions.

Koppel, M., Schler, J., & Bonchek-Dokow, E. 2007. Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of Machine Learning research*, 8(6), 1261-1276.

Koppel, M., & Winter, Y. 2014. Determining If Two Documents are by the Same Author. *Journal of American Society for Information Science & Technology,* 65(1), 178-187.

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

Labbé, D. 2009. *Si deux et deux font quatre, alors Molière n'a pas écrit Don Juan.* Max Milo, Paris.

Love, H. 2002. *Attributing Authorship: An Introduction.* Cambridge University Press.

Michell, J. 1996. *Who Wrote Shakespeare?* Thames and Hudson: New York (NY).

Mosteller, F., & Wallace, D.L. 1964. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Addison-Wesley, Reading.

Olsson, J. 2008. *Forensic Linguistics.* Continuum, London.
Pennebaker, J.W. 2011. *The Secret Life of Pronouns. What our Words Say about us*. Bloomsbury Press, New York.
Rexha, A., Klampfl, S., Kröll, M., & Kern, R. 2016. Towards a More Fine Grained Analysis of Scientific Authorship: Predicting the Number of Authors using Stylometric Features. *Proceedings BIR@ECIR 2016*, 26–31.
Rybicki, J., Hoover, D., & Kestemont, M. 2014. Collaborative Authorship: Conrad, Ford, and Rolling Delta. *Literary and Linguistic Computing*, 29(3), 422-431.
Savoy, J. 2015. Comparative Evaluation of Term Selection Functions for Authorship Attribution. *Digital Scholarship in the Humanities*, 30(2), 246-261.
Savoy, J. 2016. Estimating the Probability of an Authorship Attribution. *Journal of American Society for Information Science & Technology,* 67(6), 1462-1472.
Sebastiani, F. 2002. Machine Learning in Automatic Text Categorization. *ACM Computing Survey*, 34(1), 1-27.
Stamatatos, E. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3), 433-214.
Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., & Potthast, M. 2016. Clustering by Authorship Within and Across Documents. *Working Papers, CLEF-2016*.
Zhao, Y., & Zobel, J. 2007. Searching with Style: Authorship Attribution in Classic Literature. *Proceedings ACSC2007*, Ballarat, 59-68.

9. Annexe

Dans les formules de distance données ci-dessous, les lettres majuscules indiquent les vecteurs représentants les documents. Les minuscules correspondent aux éléments de chaque vecteur (avec leur position). Ces derniers ont été normalisés selon la longueur du texte (noté n_A ou n_B).

$$dist_{Manhattan}(A,B) = \sum_{i=1}^{m}|a_i - b_i| \quad \text{(A.1)}$$

$$dist_{Canberra}(A,B) = \sum_{i=1}^{m}\left(\left|a_i - b_i\right| \Big/ a_i + b_i\right) \quad \text{(A.2)}$$

$$dist_{Matusita}(A,B) = \sqrt{\sum_{i=1}^{m}\left(\sqrt{a_i} - \sqrt{b_i}\right)^2} \quad \text{(A.3)}$$

$$Sim_{Cosine}(A,B) = \sum_{i=1}^{m} a_i b_i \Big/ \sqrt{\sum_{i=1}^{m} a_i^2}\sqrt{\sum_{i=1}^{m} b_i^2} \quad \text{(A.4)}$$

$$dist_{Cosinus}(A,B) = cos^{-1}(Sim_{Cosine}(A,B)) \Big/ \pi \quad \text{(A.5)}$$

$$dist_{JDivergence}(A,B) = \sum_{i=1}^{m}(a_i - b_i)\, log\left(a_i \Big/ b_i\right) \quad \text{(A.6)}$$

Pour le calcul de la distance de Labbé, on suppose que $n_A < n_B$ (sinon on intervertit les deux vecteurs). De plus, les composantes a_i (et b_i) reflètent la fréquence d'occurrence des mots, lemmes ou bigrammes de lettres. La normalisation sera introduite via le facteur 2 n_A.

$$dist_{Labbé}(A,B) = \sum_{i=1}^{m}\left|a_i - \hat{b}_i\right| \Big/ 2 \cdot n_A \quad \text{avec } \hat{b}_i = b_i \cdot \frac{n_A}{n_B} \quad \text{(A.7)}$$

n°	Auteur	Titre bref	n°	Auteur	Titre bref
A1	Hardy	Jude	A2	Butler	Erewhon
B1	Butler	Erewhon	B2	Morris	Dream of JB
C1	Morris	News	C2	Tressel	Ragged TP
D1	Stevenson	Catrionae	D2	Hardy	Jude
E1	Butler	Erewhon	E2	Stevenson	Ballantrae
F1	Stevenson	Ballantrae	F2	Hardy	Wessex Tales
G1	Conrad	Lord Jim	G2	Orczy	Elusive P
H1	Hardy	Madding	H2	Conrad	Lord Jim
I1	Orczy	Scarlet P	I2	Morris	News
J1	Morris	Dream of JB	J2	Hardy	Well beloved
K1	Stevenson	Catrionae	K2	Conrad	Almayer
L1	Hardy	Jude	L2	Hardy	Well beloved
M1	Orczy	Scarlet P	M2	Morris	News
N1	Stevenson	Ballantrae	N2	Conrad	Almayer
O1	Conrad	Lord Jim	O2	Forster	Room with view
P1	Chesterton	Man who was	P2	Forster	Room with view
Q1	Butler	Erewhon	Q2	Conrad	Almayer
R1	Chesterton	Man who was	R2	Stevenson	Catrionae
S1	Morris	News	S2	Hardy	Madding
T1	Conrad	Almayer	T2	Hardy	Well beloved
U1	Orczy	Elusive P	U2	Chesterton	Man who was
V1	Conrad	Lord Jim	V2	Forster	Room with view
W1	Orczy	Elusive P	W2	Stevenson	Catrionae
X1	Hardy	Wessex Tales	X2	Hardy	Well beloved
Y1	Tressel	Ragged TP	Y2	Orczy	Scarlet P
Z1	Tressel	Ragged TP	Z2	Hardy	Madding

Tableau A.1. Identificateur, nom de l'auteur et titre abrégé des œuvres composant le corpus Oxquarry1

n°	Auteur	Titre
1, 23	Marivaux	La vie de Marianne
2, 24	Marivaux	Le paysan parvenu
3, 25	Voltaire	Zadig
4, 26	Voltaire	Candide
5, 27	Rousseau	La nouvelle Héloise
6, 28	Rousseau	Emile
7, 29	Chateaubriant	Atala
8, 30	Chateaubriant	La vie de Rancé
9, 31	Balzac	Les Chouans
10, 32	Balzac	Le cousin Pons
11, 33	Sand	Indiana
12, 34	Sand	La mare au diable
13, 35	Flaubert	Madame Bovary
14, 36	Flaubert	Bouvard et Pécuchet
15, 37	Maupassant	Une vie
16, 38	Maupassant	Pierre et Jean
17, 39	Zola	Thérèse Raquin
18, 40	Zola	La bête humaine
19, 41	Verne	De la terre à la lune
20, 42	Verne	Le secret de Wilhelm Storitz
21, 43	Proust	Du côté de chez Swann
22, 44	Proust	Le temps retrouvé

Tableau A.2. Identificateur, nom de l'auteur et titre des œuvres composant le corpus Brunet