Catégorisation de documents : Applications en attribution d'auteur et analyse stylistique

Jacques SAVOY

Institut d'informatique, Université de Neuchâtel, Suisse
Jacques.Savoy@unine.ch

Résumé. La catégorisation de documents (attribution d'un texte à une ou plusieurs catégories prédéfinies) représente un problème possédant de multiples facettes. Ainsi, l'indexation automatique correspond à l'une d'entre elles qui se fonde sur la sémantique des documents. Cependant d'autres applications analysent les mots-outils, ces formes qui ne portent que peu ou pas de sens. Or ces dernières permettent, en grande partie, de décrire le style d'un auteur voire de déterminer quelques aspects de son profil. Sur la base de ces éléments, nous allons présenter comment identifier le véritable auteur d'un document, ou savoir si celui-ci a été écrit par un homme ou une femme. Afin d'illustrer nos propos, nous aborderons le cas *d'Elena Ferrante*, un pseudonyme mondialement connu depuis la parution de son roman *L'amie prodigieuse* (Gallimard, 2016). Comme autre exemple, nous analyserons les discours des présidents américains de G. Washington (1789) à D. Trump (2017) afin d'en découvrir quelques traces évolutives tant stylistiques que thématiques. Dans ce dernier cas, une synthèse sera extraite d'un corpus de discours sous la forme d'un graphique décrivant les rapprochements entre présidences.
Mots-clés : Classification automatique, apprentissage non-supervisé, attribution d'auteur.

1 Introduction

Le *big data*, avec ses défis et ses promesses, touche également les humanités. Cependant cette question n'est pas nouvelle car depuis longtemps les archivistes doivent faire face à des volumes considérables de documents. La mise en valeur de leurs fonds numériques constitue un souci quotidien. Et le document de demain sera uniquement numérique. Pour les bibliothécaires une nouvelle conception de la gestion de leurs fonds doit être entreprise afin de rendre leur consultation plus aisée. Pour nous, les chercheurs en science de l'information, la conception, l'implémentation et la vérification d'outils permettant de les aider doit devenir une priorité. Mes propos visent à illustrer quelques méthodes et applications dans cette perspective.

Pour être plus précis, des techniques de catégorisation de textes (Sebastiani, 2002) ont prouvé une certaine efficacité. Dans ce cadre, les catégories sont prédéfinies et, sur la base d'exemples, la machine s'avère capable d'apprendre à effectuer un appariement entre un texte d'une part et, d'autre part, une ou plusieurs catégories. Ainsi, l'indexation automatique fournit à des nouveaux documents les

vedettes-matières adéquates (à l'exemple des dépêches d'agence ou d'articles scientifiques dans le système Medline/PubMed).

Dans d'autres applications, le style plutôt que le contenu sémantique se situe au cœur de la catégorisation. Comme exemple typique, on rencontre l'identification du véritable auteur (Love, 2002), (Juola, 2006) d'un texte ou d'un extrait. En effet, 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. En littérature anglaise, on s'interroge sur des écrits que l'on peut ou non attribuer à Edgar Poe (Schöberlein, 2017) ou sur les relations de Shakespeare et de ses co-auteurs (Michell, 1996), (Craig & Kinney, 2009). Le monde francophone connait le débat entre Molière et Corneille (Marusenko & Rodionova, 2010).

Analyser le style et la rhétorique peut également nous amener à explorer son évolution au fil des ans, voire des décennies. Sur la Toile, plusieurs sites offrent un accès aux principaux discours des présidents américains de G. Washington (1789–1797) à D. Trump (2017–). Sur la base de ces documents numériques, comment la machine peut-elle les traiter afin de faire ressortir une synthèse de leurs similitudes et différences ?

Dans la suite de cet article, nous présenterons un survol des connaissances en attribution d'auteur (section 2) formant la pierre angulaire de nos différentes applications. La troisième section expose le problème du profilage de l'auteur d'un écrit. La quatrième cherche à déterminer qui se cache derrière la signature *Elena Ferrante.* Enfin, une dernière section décrit quelques applications fondées sur l'analyse de l'évolution stylistique et thématique des présidences américaines.

2 État des connaissances

Les approches modernes en attribution d'auteur visent à révéler, de la manière la plus fiable possible, qui est l'auteur d'un document ou d'un fragment de texte (Juola, 2006), (Stamatatos, 2009). Deux variantes de ce problème ont été proposées. Dans un espace fermé, le véritable auteur est l'un des écrivains proposés tandis que dans un cadre ouvert, l'auteur peut être un des noms mentionnés ou un autre, encore inconnu. La réponse ne saurait se limiter à un simple nom, et une justification doit être fournie. De plus, une estimation du degré de certitude (ou probabilité) que la réponse donnée soit correcte permet de juger de la crédibilité des calculs sous-jacents (Savoy, 2016).

Afin de résoudre les problèmes de l'attribution d'auteur, trois grandes familles d'approches ont été proposées (Juola, 2006), (Stamatatos, 2009). 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 (indiquée par n) (rapport TTR, *type-token ratio,* soit $|V|/n$). 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 instables lors de comparaisons entre 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). Remarquons que l'emploi de ces mots s'effectue très souvent inconsciemment et diffère d'une personne à l'autre. 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) (Savoy, 2015a), 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 reflétant 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 d'œuvres littéraires ainsi que leur évolution sur une période d'environ 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*), (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.) (Zheng *et al.*, 2006), (Abbaci & Chen, 2008) possédant le meilleur pouvoir discriminant, puis entraîner un classifieur. Le recours aux fréquences des lettres (Kjell, 1994) ou des n-grammes de lettres (pour $n = 2, 3, \ldots, 6$) constitue souvent un choix efficient d'attributs (Kešelj *et al.*, 2003), (Juola, 2006). Par contre, l'explication de la décision proposée soulève plus de difficultés. Toute stratégie basée sur l'apprentissage par machine requiert un ensemble d'entrainement, données qui ne sont pas toujours disponibles. De plus, une forte corrélation doit exister entre le jeu de données de l'entraînement et celui de la production (ou du test).

3 Profilage d'auteur

Abordons en premier la question du profilage de l'auteur. On ne cherche pas à déterminer son véritable nom mais à identifier certaines de ses caractéristiques démographiques comme son âge approximatif, son sexe, sa langue maternelle, son origine sociale, voire ses traits psychologiques. Dans cette perspective, on fait l'hypothèse que ces caractéristiques ont une influence sur le style et que ces traces peuvent être détectées. Débutons par un cas simple : Est-ce que les hommes et les femmes écrivent dans un style distinct dont les différences sont perceptibles par une machine.

Prenons un exemple extrait d'un blog et posons la question : Est-ce que ce passage a été écrit par un homme ou une femme ?

> « Yesterday we had our second jazz competition. Thank God we weren't competing. We were sooo bad. Like, I was so ashamed, I didn't even want to talk to anyone after. I felt so rotton, and I wanted to cry, but...it's ok. »

Le choix devant se faire entre deux possibilités, nous avons 50 % de chance de trouver la bonne réponse par hasard. Pour augmenter notre taux de réussite, la lecture du texte indique des émotions (*to cry*, pleurer) qui nous fera plutôt penser à une fille. Cette décision s'avère correcte. En fait, la machine (ou l'être humain) peut détecter le sexe de l'auteur en recourant aux informations suivantes. Les femmes ont tendance à employer plus de pronoms (je, elle, moi, nous, ...) ou de négations. Elles décrivent plus facilement leurs émotions (heureux, triste, ...) et les relations sociales (ami, père, ...) (Pennebaker, 2011). Ceci n'implique pas que les hommes ne discutent pas des émotions ou des relations sociales. Toutefois la fréquence d'occurrence de tels vocables s'avère plus faible auprès de la gente masculine.

À ces éléments rattachés plus au style, nous pouvons inclure les thèmes de la discussion ou du texte. 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. Pour être plus précis, Argamon *et al.* (2009) ont analysé des blogs (rédigés en anglais) et ont dressé une liste des termes plus abondants selon le sexe de l'auteur (voit tableau 1). Ainsi, recourant conjointement aux données stylistiques et thématiques, le profilage permet d'atteindre des taux de réussite entre 65 % et 80 % face à des textes relativement brefs (de 100 à 1 000 mots).

Le profilage ne s'arrête pas à cette première information. On peut estimer l'âge approximatif de l'auteur, sa nationalité, ses traits psychologiques, etc. (Kocher & Savoy, 2017). Depuis 2009 et dans le but d'encourager la recherche et de partager l'état des connaissances dans ce domaine, une campagne d'évaluation a lieu chaque année dans le cadre du forum CLEF-PAN (voir le site web pan.webis.de), (Stamatatos *et al.*, 2015).

Tableau 1. *Fréquence des termes (pour 10 000) dans les blogs selon le sexe.*

Vocables	Homme	Femme
job	68.1 ± 0.6	56.5 ± 0.5
money	43.6 ± 0.4	37.1 ± 0.4
sports	31.2 ± 0.4	20.4 ± 0.2
TV	21.1 ± 0.3	15.9 ± 0.2
sex	32.4 ± 0.4	43.2 ± 0.5
family	27.5 ± 0.3	40.6 ± 0.4
eating	23.9 ± 0.3	30.4 ± 0.3
friends	20.5 ± 0.2	25.9 ± 0.3
sleep	18.4 ± 0.2	23.5 ± 0.2
pos-	248.2 ± 1.9	265.1 ± 2
neg-	159.5 ± 1.3	178 ± 1.4

La mise au point de techniques plus perfectionnées dans le profilage d'auteur possède des applications diverses. Ainsi, une surveillance des blogs et réseaux sociaux permettrait de détecter la volonté de suicide chez les adolescents, les signes avant-coureurs d'une dépression, ou la présence d'un prédateur sexuel dans un chat entre jeunes adolescents. Plus récemment, la détection entre vrais et faux commentaires sur un produit ou un service (restaurant, hôtel, film, ...) a enregistré

quelques progrès. La propagande étant souvent uniquement positive (ou négative) et elle ne s'accompagne pas (ou très peu) d'information factuelle ou très précise. Le dépistage des fausses nouvelles (en politique, média) connaît également un attrait grandissant.

4 Qui se cache derrière Elena Ferrante ?

Le droit d'auteur tel que nous le comprenons correspond mal à la vision des auteurs du XVIIe ou XVIIIe siècle. De plus, de nombreux écrits (*e.g.*, livres ou pamphlets politiques) sont parus sous des pseudonymes pour éviter la censure ou la police du roi. En littérature, le recours à un pseudonyme s'explique par d'autres raisons comme d'être trop jeune (le cas de E. Poe (Schöberlein, 2017)), être une femme (*e.g.*, les sœurs Brontë), voire le souhait de renaître dans un style nouveau (*e.g.*, le cas Gary-Ajar en France, J.K Rowling en Angleterre (Juola, 2016)).

Examinons un cas récent. De nos jours, l'Italie vit dans la fièvre Ferrante, nom de plume sous lequel est paru *L'amica geniale* (2011) (*L'amie prodigieuse*, 2016), un succès mondial traduit dans de nombreuses langues. Mais personne ne répond à ce nom. Aucune étude scientifique n'a abordé la question de déterminer le nom du véritable auteur ou d'établir son profil.

Afin de lever le voile sur cette énigme, une équipe pluridisciplinaire dirigée par le prof. A. Tuzzi (Université de Padoue) a entrepris de générer un corpus de romans italiens contemporains. Sur support électronique, cet ensemble regroupe 150 œuvres rédigées principalement entre 1987 et 2016 par 40 auteurs différents (voir tableau en annexe). Tous les noms probables derrière le pseudonyme Ferrante ont été inclus y compris sept romans d'Elena Ferrante. Au total, ce corpus compte environ 10 million de mots.

Durant l'été 2017, ce corpus a été traité par des chercheurs en statistique lexicale et informatique venant de France, Grèce, Italie, Pologne et Suisse. Afin de partager leurs conclusions, une conférence s'est déroulée le 7 septembre à Padoue. Tous les résultats convergent vers un unique auteur, à savoir Domenico Starnone (né en 1943 à Saviano, près de Naples) qui vit en couple avec Anita Raja (dont le nom a été évoqué comme auteur possible (Gatti, 2016)). Comme l'action de *L'amie prodigieuse* se déroule à Naples dans les années 50, les lecteurs de ce roman (et de sa suite) y verront un lien direct avec les origines de Starnone. Mais cet élément demeure un indice extratextuel et Starnone n'est pas le seul auteur napolitain du corpus. Alors comment peut-on être plus ou moins certain de cette identification ?

Afin de déterminer si deux documents sont rédigés par le même auteur, nous comparons leur vocabulaire et analysons les fréquences des mots (Labbé, 2007). Si les termes employés et leur fréquence s'avèrent proches, la distance intertextuelle sera faible. Dans le cas contraire, elle s'élèvera jusqu'à un maximum de 1.0 lorsque deux textes ne possèdent rien en commun comme une nouvelle écrite en français et une autre en finnois. Si les deux documents sont identiques, la distance sera nulle.

Plus précisément, la distance intertextuelle entre le texte A et B (notée D(A,B)) est indiqué dans l'équation 1 dans laquelle n_A signale le nombre de mots du texte A et tf_{iA} la fréquence absolue du terme i (pour $i = 1, 2, \ldots, m$) dans le texte A. La taille du vocabulaire est indiquée par m. Si l'on admet que le texte B est plus long que le texte A, nous devons réduire les fréquences des termes appartenant à B. Ces dernières (notées tf_{iB}) sont multipliées par le rapport des tailles comme présenté à droite la formule 1.

$$D(A,B) = \sum_{i=1}^{m} \left| tf_{iA} - \widehat{tf}_{iB} \right| / (2 \cdot n_A) \quad \text{avec } \widehat{tf}_{iB} = tf_{iB} \cdot n_A / n_B \qquad (1)$$

La machine calcule la distance entre toutes les paires de romans présents dans notre corpus (soit 150 x 149 / 2 = 11 175) et les classe par ordre croissant (voir tableau 2). Au premier rang, on retrouve deux œuvres signées Ferrante (ID 51 *Storia di chi fugge e di chi resta* et ID 52 *Storia Della Bambina Perduta*) avec la distance la plus faible (0,111) comme l'indique le tableau 1. Avec l'accroissement des rangs, les distances augmentent et notre certitude que le même auteur apparaît conjointement sur les deux colonnes décroît.

Arrivé au rang 33, on rencontre le premier cas avec deux noms distincts. L'ordinateur indique une distance très faible (0,193) entre *Storia Della Bambina Perduta* (ID 52, Ferrante) et *Lacci* (ID 132, Starnone) ou de 0,195 entre *Storia di chi fugge e di chi resta* (ID 51, Ferrante) et *Autobiografia erotica di Aristide Gambia* (ID 131, Starnone) (voir annexe). Un scénario identique se rencontre au rang 41, puis 42. En descendant jusqu'au rang 84, on retrouve une distance relativement faible (0,216) entre deux romans écrits par deux autres auteurs (Carofiglio et Veronesi). Arrivé à ce rang, nous avons rencontré soit des œuvres écrites par le même auteur, soit, dans dix cas, un appariement entre Ferrante et Starnone, jamais d'autres noms.

Tableau 2. *Liste ordonnée des paires de romans pairs avec la distance de Labbé.*

Ra	Dist	ID	Auteur	ID	Auteur
1	0,11	51	Ferrant	52	Ferrante
2	0,12	50	Ferrant	51	Ferrante
3	0,12	49	Ferrant	50	Ferrante
4	0,13	50	Ferrant	52	Ferrante
5	0,14	145	Verone	147	Veronesi
6	0,14	42	Faletti	44	Faletti
...	...	...	...	...	...
33	0,19	52	Ferrant	132	Starnone
38	0,19	51	Ferrant	131	Starnone
41	0,19	51	Ferrant	132	Starnone
42	0,19	47	Ferrant	127	Starnone
...	...	...	...	...	...
84	0,21	25	Carofigl	147	Veronesi

En modélisant les distances entre romans écrits soit par le même auteur, soit par deux écrivains distincts (Savoy, 2016), nous pouvons estimer que la probabilité qu'une distance de 0,2 se retrouve entre deux œuvres du même auteur s'élève à environ 99 %. Domenico Stanone est-il vraiment l'homme de plume derrière le pseudonyme Elena Ferrante ?

Le vocabulaire usité par une personne permet donc de l'identifier car le choix des mots n'est pas le simple fruit du hasard. Chacun possède ses habitudes lexicales. Par exemple, le président Jacques Chirac utilisait volontiers le terme *abracadabrantesque* tandis que Emmanuel Macron recourt aux termes *in petto*, *chicayas*, ou *bovarysme*. Dans ses romans, E. Ferrante utilise les mots *sfottente* (railleur), *risatella* (gloussement) ou *malodore* (mal odorant) mais pas avec les variantes orthographiques *mal odore* ou *maleodore*. Ces mots se rencontrent également chez D. Starnone mais jamais chez les autres 38 écrivains.

Le plus souvent, les termes apparaissent sous la plume de plusieurs auteurs, mais la fréquence d'occurrence s'avère nettement plus élevée chez Ferrante et

Starnone. Par exemple, Ferrante utilise onze fois *alunna* (élève), Starnone 28 fois et les 38 autres écrivains six fois seulement. Autre exemple, avec *tassare* (taxer) que l'on retrouve 22 fois chez Ferrante, dix fois dans les romans de Starnone mais seulement trois fois ailleurs. On peut aussi citer *minutamente* (minutieusement), *reattività* (réactivité), ou *fiaccamente* (avec lassitude). Parfois le dialecte apparaît comme dans les insultes avec 19 fois *strunz* (variante de *stronzo*) dans les écrits de Ferrante, 68 fois chez Starnone, et seulement quatre fois ailleurs.

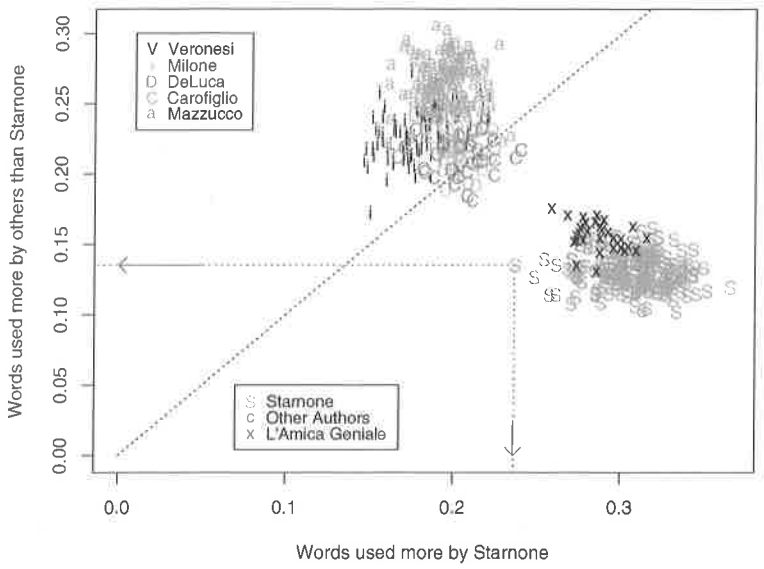

Figure 1. *Graphiques des mots sur-employés par Starnone comparés à ceux plus abondants chez d'autres auteurs italiens.*

Pour mettre de visualiser cette proximité lexicale, Craig & Kinney (2009) proposent une méthode permettant d'extraire le vocabulaire plus usité par un auteur (soit Starnone dans notre étude) et les autres (représentés par Carofiglio, De Luca, Mazzucco, Milone, et Veronesi). À l'aide de ces deux listes de mots, la machine calcule, pour chaque tranche de 4 000 mots, le pourcentage de termes apparaissant dans la liste de Stanone (donnant la valeur sur l'axe horizontal) et dans celle de autres écrivains (axe vertical). Par exemple, un des fragments écrit par Starnone possède 23,7 % des vocables représentatifs de cet auteur et 13,5 % des autres écrivains (voir figure 1).

En reprenant les romans de ces six auteurs on obtient la figure 1. Dans cette dernière, les romans de Starnone apparaissent à droite et en bas, tandis que les autres se regroupent en haut et à gauche. Enfin, selon le même processus, nous ajoutons un roman écrit par Ferrante (*L'amie prodigieuse*). Les points correspondants à cette œuvre se regroupent avec les romans signés Starnone. Le résultat indique la même personne derrière Ferrante, constatation que D. Starnone a réfuté (Fontana, 2017).

5 Analyse des discours des présidents américains

L'application des techniques issues de l'attribution d'auteur nous permet de suivre l'évolution stylistique ou thématique à travers les décennies. Par exemple, dans l'étude de la présidence américaine, les interventions les plus importantes sont le discours inaugural du président, celui des adieux et les messages sur l'État de l'Union (*State of the Union* ou SOTU). Les deux premiers ne se présentent que tous les quatre ans tandis que les derniers possèdent une fréquence quasi-annuelle.

Afin de créer notre corpus, nous avons téléchargé ces discours depuis le site *www.presidency.ucsb.edu*. Ce corpus comprend 58 discours inauguraux, et 228 messages sur l'Etat de l'Union écrits sous 45 présidences. Dans notre corpus, le premier discours a été prononcé par Washington (30 avril 1789) et le dernier par Trump (28 février 2017). Nous avons écarté les discours d'adieu correspondant plus à une forme de testament politique plutôt qu'à une allocution sur les actions futures du gouvernement.

De plus, nous avons également écarté deux présidents (W. H. Harrison (1841) et J. A. Garfield (1881)) qui n'ont pas écrit de discours sur l'État de l'Union car leur mandat a été trop bref. Enfin, on notera que Cleveland apparaît deux fois (1885-1888 et 1893-1896) ce qui correspond à ses deux mandats séparés par la présidence de Harrison.

Cette analyse des allocutions devrait faire ressortir les affinités entre présidents. De tels rapprochements devraient exister entre présidents appartenant au même parti, donnant ainsi naissance à deux identités linguistiques distinctes, l'une républicaine et l'autre démocrate. En effet, nous pourrions nous attendre à ce que les présidences démocrates s'expriment plutôt sur l'éducation, la famille et la santé tandis que les républicains devraient axer leurs interventions sur la libre entreprise, une réduction de l'État et un soutien au secteur militaire. Est-ce qu'une analyse lexicale peut confirmer ou infirmer cette hypothèse ? Quels outils sont les plus aptes à décrire les diverses tendances, affinités ou oppositions entre présidents ?

Dans notre analyse, tous les discours d'un président sont rassemblés pour former son profil. Toutefois seuls les mots fonctionnels (articles, pronoms, préposition, conjonctions, verbes auxiliaires) ont été retenus. Cette représentation correspond au style des présidents et non au contenu sémantique (fourni par les noms, adjectifs, verbes et adverbes). Pour chaque paire de profils présidentiels, une distance intertextuelle (Labbé, 2007) est calculée. Redonner simplement toutes ces valeurs (43 x 43 = 1 859) ne présente pas un grand intérêt. Par contre, nous pouvons utiliser cette information pour opérer une classification automatique (Kaufman & Rousseeuw, 1990). Le résultat final nous permettra de découvrir les divers groupes formés de textes les plus similaires.

Pour proposer une visualisation alternative (Savoy, 2015b), (Savoy, 2017), nous avons opté pour une représentation arborée, une technique classique (Baayen, 2008) en particulier en génomique (Paradis, 2011). La figure 2 illustre le résultat obtenu. Dans ce cas, la distance séparant chaque président sera mieux visualisée. Dans cette figure, la longueur des lignes permettant de rejoindre deux présidents indique la distance entre eux. Ainsi, pour aller de Trump à Kennedy, nous devons parcourir une distance moins importante qu'entre Trump et Washington. Enfin, être situé à gauche ou à droite de l'axe central n'a pas d'importance. Ce placement latéral permet simplement une meilleure visualisation.

En partant du bas à droite, on peut remonter le temps. On remarquera que le style de Trump se rapproche légèrement de celui de Bush (fils), tandis que Clinton et Obama possède des profils stylistiques similaires. De même, la présidence de Bush (père) (notée H Bush) se rapproche de celle de Reagan (dont il a été le vice-

président lors du second mandat). Ensuite, les présidences précédentes possèdent chacune son propre style qui les démarque les unes des autres, à l'exception du duo Nixon–Ford. Jusqu'à la présidence de F.D. Roosevelt, l'ordre chronologique est *grosso modo* respecté.

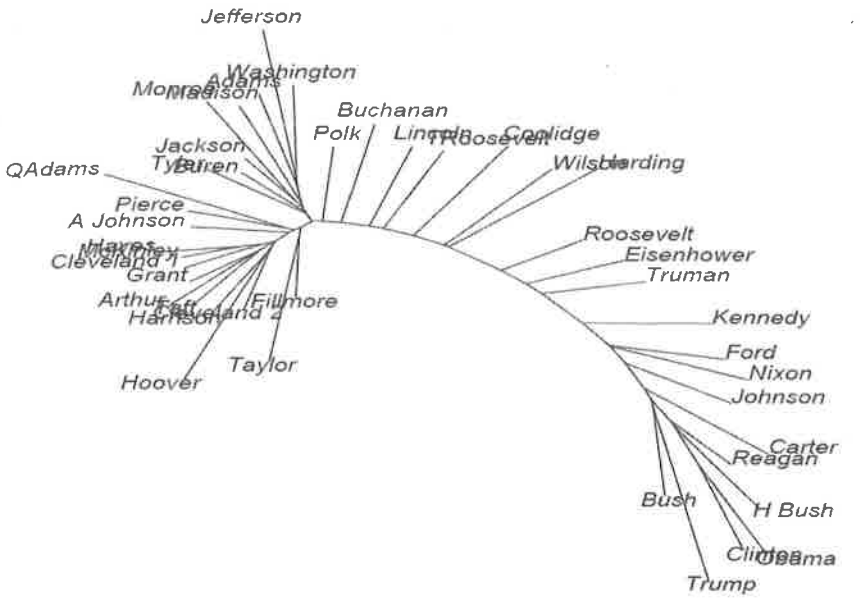

Figure 2. *Représentation arborée des distances entre les profils stylistiques des présidents.*

Pour les années précédant la deuxième guerre mondiale, on peut remarquer quelques présidences qui s'éloignent nettement de leurs contemporains. Ainsi Wilson (président de 1913–1920) a marqué une rupture stylistique importante (son prédécesseur, Taft, se retrouve avec tout à gauche avec Arthur (1881–1884) et Harrison (1889–1892)). Cette différence a certainement été rendue possible par T. Roosevelt (1901–1908) qui se démarque des présidences de la fin du XIXe siècle. La position de Lincoln (1861–1865) mérite aussi d'être soulignée comme distincte des autres présidences de sa décennie (Pierce (1853–1856) et A. Johnson (1865–1868) qui se trouve tout à gauche). Enfin on notera un regroupement des pères fondateurs (Washington à Monroe), un groupe proche du trio Jackson (1829–1836) – van Buren (1837–1840) – Tyler (1841–1844).

Les distances stylistiques les plus courtes se rencontrent souvent entre le président et son successeur (Jackson – Van Buren (0,057)) ou lorsque le même président se retrouve une seconde fois à la Maison Blanche (Cleveland (1885–1888, 1893–1996) (0,057)). Les styles les plus éloignés se situent entre Obama et Quincy Adams (1825–1828) (0,32) ou entre Clinton et Quincy Adams (0,32). Chaque époque a ses préférences esthétiques et le style évolue clairement avec les années.

Notre première hypothèse impliquait un rapprochement entre présidents issus du même parti. Les regroupements Clinton – Obama, Bush – Trump, Nixon – Ford, ou Jackson – Van Buren – Tyler correspondraient à cet effet. Cependant, on constate également des présidences clairement distinctes les unes des autres (Carter, Johnson, Kennedy) ou des rapprochements entre des visions politiques opposées (par exemple, entre Washington et Jefferson).

L'effet temporel semble jouer un rôle plus important dans la formation des groupes présidentiels. En effet, les groupes que l'on observe sont formés de présidences relativement proches dans le temps. Si l'on considère le contenu sémantique (les noms, adjectifs, verbes et adverbes), est-ce que notre vision va changer ? En répétant les mêmes calculs avec des profils présidentiels sans leurs éléments stylistiques, nous obtenons la figure 3.

Cette deuxième illustration indique aussi que le temps explique également la tendance générale sous-jacente. Ainsi, les dernières présidences se situent vers le bas, tandis que Washington et les autres pères fondateurs se placent en haut. Dans la période récente, les thèmes regroupent le trio Ford – Carter – Nixon (en bas à droite), puis les duos Reagan, H Bush (père) et d'autre part Clinton – Obama. La présidence de Bush fils rejoint le dernier duo, mais avec un léger décalage. Trump dispose d'une position éloignée de ses contemporains directs avec des thèmes distincts. Toutefois, cette présidence dispose d'un seul discours d'investiture (relativement bref) et d'un seul message sur l'État de l'Union.

En remontant, nous rencontrons souvent des groupes formés de deux ou trois présidents qui se suivent dans l'ordre chronologique. Par exemple, on voit le duo Truman-Eisenhower (provenant de deux partis différents), Hoover – Coolidge (même parti), T. Roosevelt – Taft (même parti), ou Lincoln – A. Johnson (deux partis différents). Dans ces deux derniers cas, la figure 2 indiquait une grande différence de style tandis que les sujets abordés rapprochent fortement ces présidences.

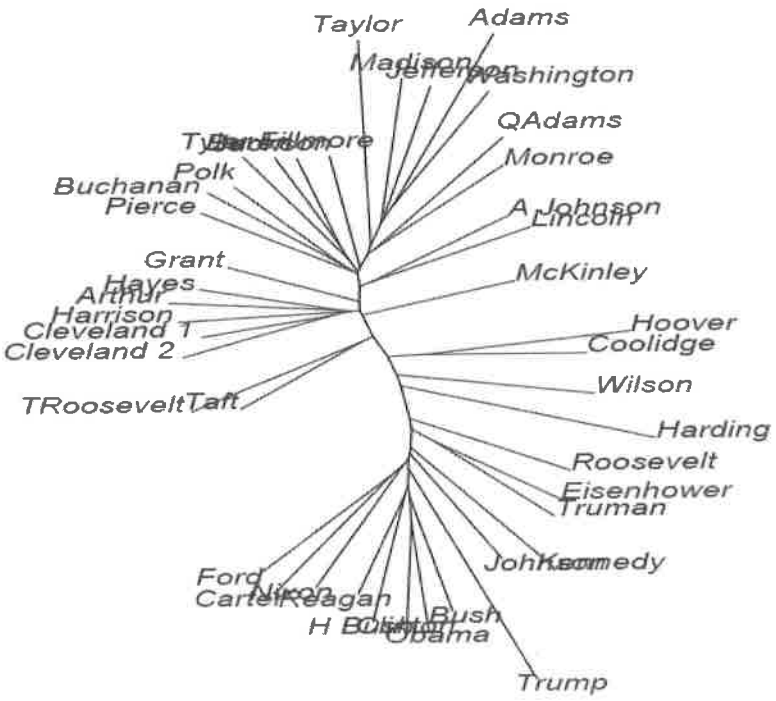

Figure 3. *Représentation arborée des distances entre les profils des présidents selon le contenu de leurs discours.*

Dans ce nouveau graphique, les distances les plus courtes se rencontrent entre Jackson (1829–1836) et Van Buren (1837–1840) (0,233) ou Obama et Clinton (0,241). A l'inverse, les contenus les plus dissemblables se situent entre Trump et Jackson (0,669) ou Trump et Van Buren (0,655).

On soulignera que les thèmes abordés ne dépendent pas uniquement de la seule volonté du président. Les évènements extérieurs imposent l'ordre du jour des problèmes à résoudre. Le vocabulaire tendra donc à être plus similaire entre présidents œuvrant dans les mêmes décennies qu'entre présidents plus distants dans le temps, même s'ils proviennent du même parti.

6 Conclusion

La catégorisation de textes permet de vérifier ou d'enrichir les métadonnées d'un ouvrage. Face à l'abondance des documents numériques, une demande pour de telles applications, si elles disposent d'une bonne performance, va s'accroitre au cours des prochaines années. Dans cette présentation, nous avons exposé trois exemples ayant pour socle des techniques automatiques d'attribution d'auteur. En se limitant uniquement à cet usage particulier, on constate que l'importance grandissant d'Internet s'accompagne d'un recours plus abondant à l'anonymat ou à l'emploi de pseudonyme. Cela ne doit pas nous masquer les nombreux problèmes d'attribution présents dans notre histoire comme les pamphlets politiques (par exemple, les *Federalist Papers* (Savoy, 2013)) ou les œuvres littéraires (le cas des sœurs Brontë, Edgar Poe, Shakespeare, Molière). Nous avons examiné plus en détail le cas d'*Elena Ferrante* en démontrant sa véritable identité : Domenico Starnone.

Mais parfois le nom de l'auteur ne nous intéresse pas vraiment. Dans ce cas, la machine peut dresser un profil de l'auteur en déterminant son âge approximatif, son sexe, sa nationalité, ses origines sociales ou quelques traits psychologiques (Pennebaker, 2011). Dans cette optique, la performance ne s'avère pas parfaite et les meilleurs systèmes approchent des taux de réussite avoisinant les 65 à 80 % (Stamatatos, 2016). Notre style et nos choix lexicaux ne laissent pas transparaître complètement toutes nos données personnelles. Toutefois, si le corpus d'entrainement s'éloigne du type de documents utilisés pour le test, la performance décroit de 8 % à 15 % (e.g., genre de texte différent) (Kocher & Savoy, 2017).

Enfin, l'analyse des affinités stylistiques ou thématiques nous permet d'élaborer des cartes indiquant les rapprochements ou différences entre auteurs. Une application sur les discours présidentiels américains indique que le facteur temporel constitue une source importante de variation dans des corpus couvrant de longues périodes (voir aussi (Hughes *et al.*, 2012)). L'affiliation politique ne constitue pas un facteur déterminant dans l'élaboration des rapprochements entre présidents, que ce soit au niveau stylistique ou thématique.

Remerciements

Cette étude a été réalisée en partie avec M. Kocher. Le corpus PIC a été créé par les professeurs Michele Cortelazzo et Arjuna Tuzzi de l'Université de Padoue.

Bibliographie

Abbasi, A., Chen, H. 2008. Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace. ACM – Transactions on Information Systems, 26(2).

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.

Fontana, E. 2017. Lo scrittore Domenico Starnone: "Io non sono Elena Ferrante". Il Giornale, Sept. 9th.

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

Holmes, D.I. 1998. The Evolution of Stylometry in Humanities Scholarship. Literary and Linguistic Computing, 13(3), 111-117.

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, 1(3).

Juola, P. 2016. The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions. Digital Scholarship in the Humanities, 30(1), i100-i113.

Kaufman, L. and Rousseeuw, P.J. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Interscience, Hoboken.

Kešelj, V., Peng, F., Cercone, N., Thomas, C. 2003. N-Gram-Based Author Profiles for Authorship Attribution. In: Proceedings of the Conference Pacific Association for Computational Linguistics, pp. 255-264, ACL.

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. 2017. Distance Measures in Author Profiling. Information Processing & Management, 53(5), 1103-1119.

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

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

Marusenko, M., & Rodionova, E. 2010. Mathematical Methods for Attributing Literary Works when Solving the "Corneille-Molière" Problem. Journal of Quantitative Linguistics, 17(1), 30-54.

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. Addion-Wesley, Reading.

Olsson, J. 2008. Forensic Linguistics. Continuum, London.

Paradis, E. 2011. Analysis of Phylogenetics and Evolution with R. 2nd Ed., Springer, New York

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.

Savoy, J. 2013. The Federalist Papers Revisited: A Collaborative Attribution Scheme. In Proceedings ASIST, Montreal, November.

Savoy, J. 2015a. Comparative Evaluation of Term Selection Functions for Authorship Attribution. Digital Scholarship in the Humanities, 30(2), 246-261.

Savoy, J. 2015b. Text Clustering: An Application with the State of the Union Addresses. Journal of the American Society for Information Science and Technology, 66(8), 1645-1654

Savoy, J. 2016. Estimating the Probability of an Authorship Attribution. Journal of the American Society for Information Science and Technology, 67(6), 1462-1472.

Savoy, J. 2017. Analysis of the Style and the Rhetoric of the American Presidents Over Two Centuries. Glottometrics, 38(1), 55-76

Schöberlein, S. 2017. Poe or not Poe? A Stylometric Analysis of Edgar Allan Poe's Disputed Writings. Digital Scholarship in the Humanities, 32(3), 643-659.

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., Potthast, M., Rangel, F., Rosso, P., & Stein, B. 2015. Overview of the PAN/CLEF 2015 Evaluation Lab. In Josiane Mothe et al, editors, Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings 6th International Conference of the CLEF Initiative (CLEF 15), 518-538, Berlin: Springer.

Zhao, Y., Zobel, J. 2007. Searching with Style: Authorship Attribution in Classic Literature. In: Proceedings of the Thirtieth Australasian Computer Science Conference, pp. 59-68, Ballarat.

Zheng, R., Li, J., Chen, H., & Huang, Z. 2006. A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques. Journal of the American Society for Information Science & Technology, 57(3), 378-393.

Annexe

Tableau A.1. *Nom de l'auteur, sexe et nombre de romans dans le corpus PIC (Padova Italian Corpus). En italique, les auteurs originaires de Campanie.*

Nom	Se	No	Nom	Se	No
Affinati	M	2	*Montesano*	M	4
Ammaniti	M	4	Morazzoni	F	2
Bajani	M	3	Murgia	F	5
Balzano	M	2	Nesi	M	3
Baricco	M	4	Nori	M	3
Benni	M	3	*Parrella*	F	2
Brizzi	M	3	*Piccolo*	M	7
Carofiglio	M	9	Pincio	M	3
Covacich	M	2	*Prisco*	M	2
De Luca	M	4	Raimo	M	2
De Silva	M	5	*Ramondino*	F	2
Faletti	M	5	*Rea*	M	3
Ferrante	?	7	Scarpa	M	4
Fois	M	3	Sereni	F	6
Giordano	M	3	*Starnone*	M	10
Lagioia	M	3	Tamaro	F	5
Maraini	F	5	Valerio	F	3
Mazzanti	F	4	Vasta	M	2
Mazzucc	F	5	Veronesi	M	4
Milone	F	2	Vinci	F	2

Tableau A.2. *Quelques exemples de romans inclus dans le corpus PIC.*

DocID	Auteur	Année	Taille (mots)	Titre
29	Carofiglio	2014	67 222	Regola equilibrio
...	...	...		...
46	Ferrante	1992	41 914	L'amore molesto
47	Ferrante	2002	53 546	I giorni dell abbandono
48	Ferrante	2006	36 222	La figlia oscura
49	Ferrante	2011	96 135	L'amica geniale
50	Ferrante	2012	138 622	Storia del nuovo cognome
51	Ferrante	2013	119 148	Storia di chi fugge e di chi resta
52	Ferrante	2014	136 945	Storia della bambina perduta
...	...	...		...
124	Starnone	1987	39 538	Ex cattedra
129	Starnone	2000	140 226	Via Gemito
130	Starnone	2007	40 787	Prima esecuzione
131	Starnone	2011	118 810	Autobiografia erotica di Aristide Gaml
132	Starnone	2014	36 554	Lacci
133	Starnone	2016	42 286	Scherzetto
...	...	...		...