RECITAL 2010, Montréal, 19–23 juillet 2010

Attribution d'auteur au moyen de modèles de langue et de modèles stylométriques

Audrey Laroche
OLST, Dép. de linguistique et de traduction, Université de Montréal
audrey.laroche@umontreal.ca

Résumé. Dans une tâche consistant à trouver l'auteur (parmi 53) de chacun de 114 textes, nous analysons la performance de modèles de langue et de modèles stylométriques sous les angles du rappel et du nombre de paramètres. Le modèle de mots bigramme à lissage de Kneser-Ney modifié interpolé est le plus performant (75 % de bonnes réponses au premier rang). Parmi les modèles stylométriques, une combinaison de 7 paramètres liés aux parties du discours produit les meilleurs résultats (rappel de 25 % au premier rang). Dans les deux catégories de modèles, le rappel maximal n'est pas atteint lorsque le nombre de paramètres est le plus élevé.

Abstract. In a task consisting of attributing the proper author (among 53) of each of 114 texts, we analyze the performance of language models and stylometric models from the point of view of recall and the number of parameters. The best performance is obtained with a bigram word model using interpolated modified Kneser-Ney smoothing (first-rank recall of 75 %). The best of the stylometric models, which combines 7 parameters characterizing the proportion of the different parts of speech in a text, has a first-rank recall of 25 % only. In both types of models, the maximal recall is not reached when the number of parameters is highest.

Mots-clés : Attribution d'auteur, modèle de langue, stylométrie, n-grammes, vecteurs de traits.

Keywords: Authorship attribution, language model, stylometry, n-grams, feature vectors.

1 Introduction

L'attribution d'auteur est une tâche qui intéresse les chercheurs depuis le XIX^{e} siècle (Holmes, 1994). Elle a permis d'identifier l'auteur d'œuvres de provenance contestée, comme les *Federalist Papers* (McEnery & Oakes, 2000) ; elle a aujourd'hui des applications dans des domaines comme la détection de plagiat dans les travaux scolaires et la linguistique légale. La plupart des techniques d'identification d'auteur ont trait à la stylométrie, c'est-à-dire la mesure quantitative d'indices textuels de natures diverses qui caractérisent le style d'un auteur. Les indices stylistiques et leur quantité varient d'une étude à l'autre. Par exemple, Schaalje *et al.* (1997) tentent de déterminer quels types d'indices donnent de meilleurs résultats ; selon eux, ce sont les mots fonctionnels qui permettent de distinguer les auteurs, et non la richesse du vocabulaire (ex. ratio types/tokens). Stamatatos *et al.* (1999) forment des vecteurs de 22 indices stylistiques (ex. nombre de syntagmes nominaux par rapport au nombre total de syntagmes) qu'ils comparent statistiquement à un article de journal pour en trouver l'auteur. Pour attribuer un auteur à des courriels, Koppel & Schler (2003) combinent trois classes de traits : lexicaux (fréquence des mots fonctionnels), collocationels (fréquence

des bigrammes de parties du discours) et idiosyncratiques (ex. épellation, formatage). Les paramètres sont beaucoup plus nombreux dans Van Halteren (2004), dont la technique est basée sur des milliers de traits lexicaux et syntaxiques (mots-formes, parties du discours, bigrammes, trigrammes). D'autres études portent sur la performance de réseaux de neurones et d'algorithmes génétiques (McEnery & Oakes, 2000). Un autre type d'approche ne fait appel qu'à des modèles de langue : entre autres, Keselj *et al.* (2003) construisent des ensembles optimaux de n-grammes de lettres pour former des profils d'auteur. Les performances de toutes ces techniques sont variables (les meilleures rapportées allant de 50 % à 98 %) et dépendent fortement de la tâche effectuée, du nombre d'auteurs candidats (Luyckx & Daelemans, 2008) et de la taille des corpus d'entraînement, en plus des indices stylistiques ou des n-grammes sélectionnés.

L'objectif de la présente étude est de comparer la performance de différentes méthodes d'identification automatique de l'auteur d'un texte : certaines sont basées sur des modèles de langue et d'autres sur des modèles stylométriques. Pour déterminer l'auteur d'un texte, les modèles construits selon ces deux approches lui sont comparés tour à tour ; il s'agit en fait d'un problème de catégorisation. Les meilleurs modèles doivent combiner un rappel élevé et un petit nombre de paramètres.

Le plan de l'article est le suivant. Notre corpus est décrit à la section 2. La section 3 présente les deux types d'approches testés dans la tâche d'attribution d'auteur. Nous discutons dans la section 4 des résultats obtenus lors des expériences. La section 5 conclut l'article en suggérant des pistes d'amélioration.

2 Corpus

Le corpus utilisé dans les expériences est constitué de 167 textes de 53 auteurs différents (minimum de 2 textes par auteur ; moyenne de 3,2). Ces textes, écrits en français (pas de traductions), proviennent de diverses régions de la francophonie (France, Québec, etc.) et sont antérieurs à la seconde moitié du XX^{e} siècle. Ils s'inscrivent dans des genres littéraires hétérogènes : romans (73 textes), essais (21), poèmes (14), pièces de théâtre (14), mémoires (12), nouvelles (9), biographies (8), lettres (6), journaux (6) et dialogues (4). De plus, les textes de 21 des auteurs ne sont pas du même genre. Les textes sont généralement assez longs : le plus court fait 2100 mots, le plus long 261 700 mots, et un texte compte en moyenne 53 200 mots.

Cent six textes de 31 auteurs sont extraits de la Bibliothèque Universelle (ABU)[1] ; les 61 textes des 22 auteurs restants sont tirés du Projet Gutenberg (PG)[2]. Pour les expériences, l'ensemble du corpus est divisé arbitrairement en sous-corpus d'entraînement et de test. Le corpus d'entraînement, qui sert à modéliser chaque auteur, est formé d'un texte par auteur. Le corpus de test est donc constitué de 114 textes dont l'auteur est inconnu, mais fait partie des 53 auteurs modélisés.

À notre connaissance, aucun corpus français ayant servi dans les études antérieures sur l'attribution d'auteur n'est disponible. Les corpus grec de Stamatatos *et al.* (1999) et chinois de Fuchun *et al.* (2003) ont bien été repris dans Keselj *et al.* (2003), mais ces derniers ont été contraints, comme nous, d'assembler pour l'anglais un corpus constitué de textes classiques (ex. Shakespeare, Dickens) libres de droits. À des fins de reproductibilité et de comparaison, l'ensemble du corpus que nous avons constitué (sous ses formes originale et prétraitée) est disponible à l'adresse `http://olst.ling.umontreal.ca/~audrey/recital2010/`.

[1] `http://abu.cnam.fr/BIB/auteurs/`

[2] `http://www.gutenberg.org/browse/languages/fr`

2.1 Prétraitement

Les textes des corpus d'entraînement et de test passent par une série de transformations avant d'être modélisés. D'abord, au moyen de scripts, les licences d'ABU et du PG sont enlevées, de même que, pour anonymiser les œuvres, les 10 premières lignes des textes d'ABU et les 20 premières des textes du PG. Les textes sont segmentés finement en mots à l'aide d'un script écrit par Tanguy & Hathout (2003). Ce script de segmentation est adapté au français : une liste d'exceptions permet de ne pas séparer les constituants des mots complexes (*au fur et à mesure*) et des mots comprenant un signe de ponctuation (*R.-de-ch.*). Par la suite, une étiquette morphosyntaxique et un lemme sont attribués à chaque mot à l'aide de TreeTagger[3]. La lemmatisation de TreeTagger est désambiguïsée à l'aide d'un autre script de Tanguy & Hathout (2003) qui sélectionne le lemme le plus fréquent (selon un corpus de référence) dans les cas où TreeTagger propose plusieurs lemmatisations pour un mot. Enfin, pour certaines des expériences, les textes segmentés sont également découpés en syntagmes avec le *chunker* de TreeTagger.

3 Approches

Le principe général de la tâche d'attribution d'auteur consiste tout d'abord à modéliser les 53 auteurs du corpus d'entraînement selon une technique donnée (section 3.1). Ensuite, les modèles sont comparés un à un au texte de test dont nous cherchons à déterminer l'auteur (section 3.2). Cette tâche est répétée pour chacun des 114 textes du corpus de test, et la performance des différents modèles, qui fait l'objet de notre étude, est évaluée à l'aide des métriques présentées à la section 3.3.

3.1 Phase d'entraînement

Nous avons implémenté deux catégories de méthodes pour modéliser des auteurs à partir d'un de leurs textes. La première est constituée de modèles de langue d'ordres, de lissages et d'unités distincts. La seconde catégorie regroupe des modèles stylométriques simples et complexes.

3.1.1 Acquisition des modèles de langue

Un modèle de langue d'ordre n, ou modèle n-gramme, dans son acception la plus courante, est un modèle statistique qui calcule la probabilité d'un mot étant donnés les n-1 mots qui le précèdent (Goodman, 2001). Un tel modèle est habituellement lissé afin d'accorder une probabilité non nulle à des séquences non observées dans le corpus d'entraînement. L'utilisation de modèles de langue dans l'attribution d'auteur est assez répandue ; elle est souvent conjuguée à des indices stylistiques, comme dans Koppel & Schler (2003) et Van Halteren (2004). Keselj *et al.* (2003) en font une étude plus détaillée en tentant de construire de petits modèles de caractères d'ordres différents. Nous voulons vérifier si les modèles les plus petits sont effectivement meilleurs et étudier l'influence des propriétés d'un modèle de langue, soit son type de lissage, son ordre et son unité de base. Dans notre première série d'expériences, 16 modèles de langue sont construits pour chacun des 53 auteurs à l'aide de la boîte à outils SRILM (Stolcke, 2002). Le lissage de ces modèles est soit celui de Kneser-Ney modifié interpolé (KN), soit celui de repli de Witten-Bell (WB).

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Pour chaque type de lissage, des modèles d'ordre 2, 3, 4 et 5 sont créés. Les modèles à lissage KN sont des modèles de mots. Pour les modèles à lissage WB, trois unités correspondant à des niveaux d'analyse plus ou moins abstraits sont utilisées : modèles de mots, de lemmes et de parties du discours (les deux derniers sont créés à partir du corpus étiqueté par TreeTagger). Le vocabulaire des modèles de lemmes et de parties du discours dans le corpus étant restreint, le lissage KN (bien qu'il soit généralement le meilleur selon Chen & Goodman (1998)) n'est pas approprié pour construire des modèles de langue basés sur ces unités[4] ; Stolcke *et al.* (2010) recommandent plutôt d'utiliser le lissage WB dans ces cas.

3.1.2 Acquisition des modèles stylométriques

L'approche employée dans la deuxième série d'expériences consiste à acquérir des modèles stylométriques pour représenter les auteurs. Les indices stylistiques étudiés formant ce type de modèle sont pour la plupart tirés de l'état de l'art de Holmes (1994) et ont trait aux proportions de parties du discours. Les modèles stylométriques sont construits à partir des textes analysés par TreeTagger. Chaque auteur est d'abord modélisé neuf fois à partir d'un trait stylistique unique parmi les suivants : 1) nombre de noms par rapport au nombre de verbes, 2) nombre de noms par rapport au nombre total de mots dans le texte (N), 3) nombre de verbes par rapport à N, 4) nombre d'adverbes par rapport à N, 5) nombre de mots fonctionnels[5] par rapport à N, 6) nombre de signes de ponctuation par rapport à N, 7) nombre d'adjectifs par rapport au nombre de noms, 8) nombre d'adjectifs par rapport à N, et 9) longueur moyenne des syntagmes nominaux (SN). Des modèles stylométriques complexes sont ensuite construits de façon vorace : selon leur performance respective dans la tâche d'attribution d'auteur, les indices stylistiques simples sont combinés de façon incrémentale en un vecteur de deux à neuf traits pour former des modèles d'auteur complexes. Notons que les indices stylistiques que nous étudions sont beaucoup moins nombreux que ceux dans Koppel & Schler (2003), Van Halteren (2004), Luyckx & Daelemans (2008), etc. Ceux-ci utilisent des techniques d'apprentissage machine pour pondérer les indices. Comme nous avons pour objectif d'utiliser le moins de paramètres possibles, nous n'avons pas encore exploré cette voie qui donne de bons résultats dans la littérature, bien qu'elle nécessite beaucoup de ressources.

3.2 Phase de test

Dans la phase de test, notre système doit identifier l'auteur d'un texte du corpus de test parmi les auteurs modélisés à l'aide des deux méthodes d'acquisition présentées à la section 3.1. Pour ce faire, il attribue un score à chaque modèle d'auteur pour indiquer à quel point celui-ci ressemble au texte d'auteur inconnu. Les modèles sont ordonnés selon leur score ; le modèle en tête de liste correspond à l'auteur attribué au texte par le système. La façon de calculer ce score dépend du type de modèle employé.

3.2.1 Attribution à l'aide de modèles de langue

Le score indiquant à quel point un modèle de langue modélise bien le texte d'auteur inconnu est donné par la perplexité. La perplexité d'un modèle de langue est la moyenne géométrique de la probabilité inverse

[4]En effet, SRILM ne parvient pas à calculer ces modèles.

[5]Les mots considérés comme fonctionnels sont les déterminants, les conjonctions, les pronoms et les prépositions.

des mots du corpus de test (Goodman, 2001, p. 4) :

$$score_{perplexité}(x_1...x_n) = \sqrt[n]{\prod_{i=1}^{n} \frac{1}{P(x_i|x_{1...i-1})}}$$

La perplexité des modèles de langue par rapport au texte de test est calculée à l'aide de la boîte à outils SRILM. Plus la perplexité est basse, plus le modèle de langue réussit bien à prédire le texte. L'auteur présumé d'un texte est par conséquent celui qui obtient la plus petite perplexité parmi tous les auteurs modélisés.

3.2.2 Attribution à l'aide de modèles stylométriques

Tel que décrit à la section 3.1.2, chaque auteur est représenté à l'aide de modèles stylométriques simples et complexes. Les modèles stylométriques simples sont constitués d'un paramètre unique. Pour attribuer l'auteur d'un texte de test, ce dernier doit être représenté par le même paramètre que le type de modèle stylométrique à l'étude. Le score indiquant à quel point un modèle d'auteur représente bien le texte d'auteur inconnu est égal à la différence, en valeur absolue, entre la valeur du paramètre pour le modèle (*m*) et celle du même paramètre pour le texte (*t*) :

$$score_{MS\ simple}(m,t) = |m - t|$$

Le score est calculé pour tous les auteurs modélisés ; le plus petit des scores obtenus indique lequel des modèles d'auteur correspond le mieux au texte.

Pour comparer des modèles stylométriques complexes, constitués d'un vecteur d'indices stylistiques, à un texte d'auteur inconnu, ce texte est d'abord caractérisé par un vecteur formé des mêmes paramètres que les modèles. Le vecteur de chaque modèle d'auteur ($\vec{m}$) est ensuite comparé au vecteur du texte ($\vec{t}$) à l'aide de la mesure du cosinus :

$$score_{cosinus}(\vec{m},\vec{t}) = \frac{\sum_{i=1}^{n} m_i \times t_i}{\sqrt{\sum_{i=1}^{n} m_i^2}\sqrt{\sum_{i=1}^{n} t_i^2}}$$

Plus le cosinus entre deux vecteurs est élevé, plus ces vecteurs sont similaires. L'auteur attribué au texte est donc celui qui donne la plus grande valeur de cosinus.

3.3 Métriques d'évaluation

Dans la phase de test, les modèles d'auteur sont ordonnés selon un score indiquant leur degré de ressemblance au texte dont nous cherchons à déterminer l'auteur. Pendant les expériences, cette procédure est répétée pour chaque type de modèle de langue et de modèle stylométrique parmi ceux décrits à la section 3.1. La comparaison des performances des différentes approches est basée sur deux critères : le rappel et le nombre de paramètres que renferme le modèle. Le rappel au rang *n* est donné par :

$$rappel_n = \frac{nombre\ de\ bonnes\ réponses\ au\ rang\ n}{nombre\ de\ textes\ de\ test} \times 100$$

Ainsi, le rappel d'un modèle au rang 1 correspond au pourcentage des 114 textes de test pour lesquels le système trouve le bon auteur en première position. Plus le rappel est élevé, meilleure est la performance du modèle. Quant au nombre de paramètres, il équivaut à la quantité de n-grammes distincts caractérisant

les modèles de langue ou à la taille des vecteurs d'indices stylistiques dans les modèles stylométriques simples et complexes. Étant donné qu'une quantité réduite de paramètres requiert moins de ressources (temps de calcul, mémoire), la performance d'un modèle du point de vue du nombre de paramètres est meilleure si ce modèle implique moins de paramètres.

4 Résultats et discussion

Les deux types d'approches — modèles de langue et modèles stylométriques — ont fait l'objet d'expériences dans lesquelles un auteur parmi les 53 modélisés doit être attribué à chacun des 114 textes d'auteur inconnu. De façon générale, les meilleurs résultats en termes de rappel sont obtenus avec des modèles de langue. Dans les deux approches, un nombre de paramètres plus élevé n'est pas directement lié à un rappel supérieur. Les paragraphes qui suivent présentent les résultats obtenus lors des expériences d'attribution d'auteur dans lesquelles le type de modélisation varie (sections 4.1 et 4.2) ; la section 4.3 montre les résultats d'expériences sur l'influence de la taille des corpus d'entraînement et de test sur la performance.

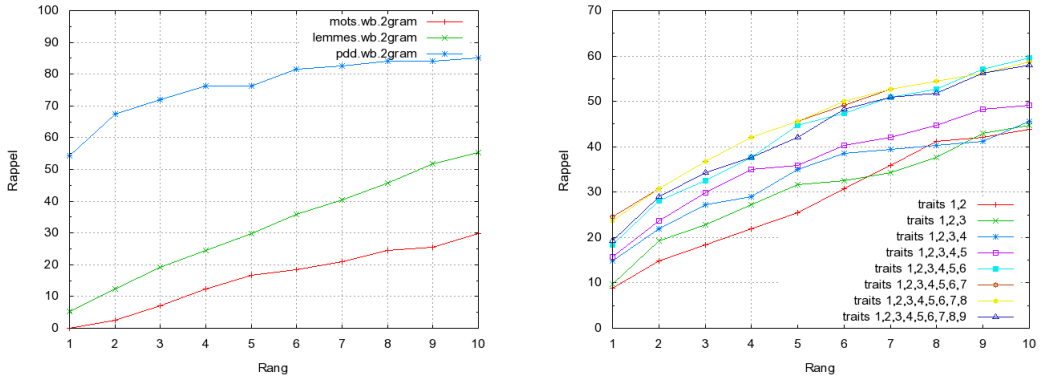
4.1 Performance des modèles de langue

Dans une première série d'expériences, les auteurs sont modélisés avec des modèles de langue dont le lissage, l'ordre et l'unité varient. Le tableau 1 permet de comparer les différents types de modèles de

TAB. 1 – Rappel au rang 1 des modèles de langue

	Mots		Lemmes	Parties du discours
	KN	WB	WB	WB
2-gramme	74,56	0,00	5,26	54,39
3-gramme	71,93	0,00	7,02	50,00
4-gramme	72,81	0,00	9,65	47,37
5-gramme	72,81	0,00	10,53	49,12

langue. Il montre que le choix de la méthode de lissage est très important : le rappel au rang 1 pour les modèles de mots à lissage de Kneser-Ney modifié interpolé est supérieur à 70 %, tandis que les modèles de mots à lissage de Witten-Bell ne donnent jamais la bonne réponse au rang 1. L'ordre des modèles de langue a lui aussi une influence (irrégulière) sur la performance. L'ordre 2 donne le meilleure rappel dans le cas des modèles de mots KN (74,56 %) et des modèles de parties du discours WB (54,39 %), mais c'est l'ordre le plus élevé (5) qui produit le meilleur rappel parmi les modèles de lemmes WB (10,53 %) ; l'ordre n'influence pas le rappel au rang 1 des modèles de mots WB, qui demeure nul. Par ailleurs, plus le niveau d'analyse du texte est abstrait (du plus concret au plus abstrait : mot, lemme, partie du discours), plus le rappel des modèles WB augmente. Par exemple, le rappel au rang 1 des modèles WB bigrammes est décuplé lorsque l'unité passe du lemme (5,26 %) à la partie du discours (54,39 %). La figure 1(a) (page suivante) montre l'importante amélioration du rappel aux rangs 1 à 10 selon l'unité du modèle de langue WB. Tout se passe comme si la variation au niveau des suites de mots entre les textes d'un même auteur est plus marquée que celle au niveau des suites de parties du discours. En construisant des modèles d'auteur encore plus abstraits, des modèles syntaxiques, Baayen *et al.* (1996) ont en effet noté que les structures syntaxiques constituent de meilleurs indices stylistiques que les indices portant sur le lexique. Les modèles KN sont impossibles à construire lorsque la taille de vocabulaire est réduite (Stolcke *et al.*, 2010), comme



(a) Modèles de langue Witten-Bell bi-grammes

(b) Modèles stylométriques complexes

FIG. 1 – Rappel en fonction du rang

c'est le cas des modèles de lemmes et des modèles de parties du discours. S'il existait un type de lissage plus performant que WB (comme KN) et adapté aux vocabulaires restreints, la performance dans la tâche d'attribution d'auteur serait probablement améliorée.

Le nombre de paramètres, notre autre critère de performance, ne varie pas en fonction du type de lissage (seul le poids des n-grammes varie), mais selon l'ordre et l'unité des modèles de langue. Plus l'ordre du modèle est élevé, plus le nombre de paramètres (nombre de n-grammes distincts dans le texte modélisé) est élevé. Par exemple, pour un texte de 48 500 mots (taille médiane des textes du corpus d'entraînement), le modèle bigramme contient 41 400 paramètres, celui d'ordre 3, 45 900 paramètres, d'ordre 4, 47 600 paramètres et d'ordre 5, 48 200 paramètres. L'abstraction des niveaux d'analyse textuelle est inversement proportionnelle à la quantité de paramètres. À titre d'exemple, dans le texte de 48 500 mots, un modèle de mots bigramme contient 41 400 paramètres, un modèle de lemmes 24 700 paramètres et un modèle de parties du discours, 300.

Comme le montre le tableau 1 ci-dessus, un nombre de paramètres plus élevé n'est pas une condition nécessaire pour obtenir un meilleur rappel. Au contraire, parmi tous les modèles WB, c'est celui qui compte le moins de paramètres (bigramme, parties du discours) qui donne le meilleur rappel au rang 1 (54,39 %). Parmi les modèles KN, c'est également celui qui contient le moins de paramètres, le modèle bigramme, qui donne le meilleur rappel (74,56 %). Les modèles d'ordre plus élevé souffrent donc d'un problème de surapprentissage.

4.2 Performance des modèles stylométriques

L'attribution d'auteur à l'aide de modèles stylométriques simples donne des rappels au rang 1 (tableau 2, page suivante) de beaucoup inférieurs à ceux obtenus avec les modèles de mots à lissage KN et de parties du discours à lissage WB. Les modèles stylométriques simples, qui comptent un seul paramètre, ont cependant un meilleur rappel au rang 1 que les modèles WB de mots et de lemmes, qui contiennent des milliers de paramètres.

Il est intéressant de constater que les meilleurs indices stylistiques sont ceux qui sont liés aux proportions

TAB. 2 – Rappel des modèles stylométriques simples aux rangs 1 et 10

Id	Trait	R1	R10	Id	Trait	R1	R10
1	‖Noms‖/‖Verbes‖	9,65	42,98	6	‖Ponctuations‖/‖Mots‖	6,14	41,23
2	‖Noms‖/‖Mots‖	9,65	40,35	7	‖Adjectifs‖/‖Noms‖	5,26	36,84
3	‖Verbes‖/‖Mots‖	7,02	43,86	8	‖Adjectifs‖/‖Mots‖	3,51	40,35
4	‖Adverbes‖/‖Mots‖	7,02	40,35	9	Longueur moyenne SN	2,63	26,32
5	‖Mots fonctionnels‖/‖Mots‖	6,14	47,37				

des mots lexicaux (traits 1 à 4). En effet, plusieurs études stylométriques utilisent des indices liés aux mots fonctionnels et à la ponctuation (McEnery & Oakes, 2000), mais nos résultats montrent que ces indices sont moins performants que ceux liés aux noms, aux verbes et aux adverbes. Notons toutefois qu'au rang 10, le meilleur rappel est celui donné par la proportion de mots fonctionnels dans les textes (47,37 %). Les indices liés aux proportions d'adjectifs et à la longueur moyenne des syntagmes nominaux (indices 7 à 9) ont une performance médiocre.

Pour former les modèles stylométriques complexes, nous utilisons une approche vorace : les indices simples sont graduellement combinés en des vecteurs de traits par ordre de leur rappel au rang 1 tel que noté dans le tableau 2. La figure 1(b) montre le rappel aux rangs 1 à 10 de ces modèles de vecteurs de traits stylométriques.

Comme dans le cas des modèles de langue, les modèles stylométriques complexes ayant le plus grand nombre de paramètres (équivalant à la taille du vecteur) ne sont pas ceux qui ont le meilleur rappel. En effet, le meilleur ensemble de paramètres combine les indices stylistiques 1 à 7 et a un rappel de 24,56 % au rang 1 ; le plus gros des modèles stylométriques (9 paramètres) a un rappel de 19,30 % au premier rang. La figure 1(b) montre que l'introduction des traits 4, 6 et 7 a une plus grande influence sur le rappel au rang 1 que celle des traits 5, 8 et 9 (ces deux derniers diminuant même le rappel). Notre étude, dans son état actuel, étant donné qu'elle ne fait intervenir que 9 indices, ne permet pas de conclure qu'un nombre inférieur de paramètres stylistiques donne nécessairement de meilleurs résultats. Il pourrait au contraire être profitable de sélectionner une grande quantité d'indices comme le font Luyckx & Daelemans (2008) puis de les pondérer automatiquement.

4.3 L'influence de la taille du corpus

Les textes d'entraînement et de test utilisés dans les expériences décrites ci-dessus sont longs (53 200 mots en moyenne) et leur taille varie beaucoup : cela peut influencer les résultats. Nous avons vérifié la performance du meilleur modèle de chaque approche — modèle de mots bigramme KN et modèle stylométrique complexe à 7 paramètres — sur des textes d'entraînement de tailles égales (2000, 2500 et 3000 mots), puis sur des textes de test de 500, 750 et 1000 mots[6]. Lorsque la taille des textes d'entraînement est réduite par rapport aux textes originaux, les textes de test conservent leur taille originale, et inversement.

Comme l'indique le tableau 3, plus la taille des textes d'entraînement est importante, meilleur est le rappel au rang 1 (ce qui est conforme aux résultats de Luyckx & Daelemans (2008)). Il n'en est pas ainsi pour les textes de test (ceux dont on cherche à déterminer l'auteur) : le rappel maximal est atteint avec la

[6]Ces tailles sont déterminées en fonction de la taille des plus petits textes d'entraînement (6000 mots) et de test (2000 mots). La tâche a été effectuée deux ou trois fois pour chaque taille ; les résultats rapportés correspondent à la moyenne des rappels obtenus.

TAB. 3 – Rappel au rang 1 des meilleurs modèles en fonction de la taille des corpus

Modèle	Taille du corpus d'entraînement			Taille du corpus de test		
	2000 mots	2500 mots	3000 mots	500 mots	750 mots	1000 mots
Modèle de langue	38,89	42,11	44,30	46,49	48,25	47,81
Modèle stylométrique	6,14	9,21	10,53	5,85	6,14	5,26

taille médiane. Ces constats sont valides pour les deux approches de modélisation. Dans tous les cas, cependant, le rappel dans cette série d'expériences (dans laquelle les textes sont coupés) est inférieur à celui obtenu lorsque tous les textes ont leur taille originale. Cette différence est plus marquée lorsque les textes d'entraînement sont raccourcis dans le cas des modèles de langue et, dans le cas des modèles stylométriques, lorsque c'est la taille des textes de test qui est réduite.

5 Conclusion

Nous avons présenté une tâche d'attribution d'auteur dans laquelle un texte doit être catégorisé en étant comparé à des modèles d'auteurs. Plusieurs approches de modélisation sont possibles ; nous avons testé la performance de modèles de langue et de modèles stylométriques, ces derniers étant très répandus dans la littérature. Nos résultats montrent que les modèles de langue atteignent de meilleurs rappels que les modèles stylométriques, bien que des disparités de performance importantes existent entre les différents types de modèles dans chaque approche. Le rappel maximal obtenu est de 75 % (avec un modèle de mots bigramme à lissage de Kneser-Ney modifié interpolé) ; ce résultat est difficilement comparable à ceux mentionnés dans les travaux antérieurs, ceux-ci traitant en général moins d'auteurs (entre 2 et 10) et leurs corpus étant variés. Soulignons que Luyckx & Daelemans (2008), qui se sont penchés sur l'influence du nombre d'auteurs, obtiennent un rappel semblable au nôtre (76 %) pour 20 auteurs (avec une approche basée sur l'apprentissage machine), alors que notre étude porte sur 53 auteurs. Nous constatons par ailleurs qu'un grand nombre de paramètres dans les modèles n'est pas gage de meilleurs résultats ; dans le cas des modèles stylométriques, ces résultats suggèrent qu'il serait intéressant de pondérer les indices stylistiques, comme le font plusieurs auteurs. Des expériences futures porteront aussi sur des modèles de langue basés sur les caractères et sur les structures syntaxiques.

Il serait intéressant de tester les approches sur des corpus différents (le nôtre étant relativement long et composé de genres hétérogènes) afin de mesurer leur capacité à accomplir des tâches plus près d'applications réelles. La matrice de confusion obtenue à partir du meilleur modèle montre qu'en moyenne, 28 % des textes d'un auteur sont attribués à un autre ; 38 % de cette moyenne est dû à 5 des 53 auteurs. Si le texte d'entraînement est d'un genre différent de celui des textes de test (par ex. un poème et des romans), la catégorisation effectuée par notre système est plus mauvaise : 52 % des auteurs ayant mal été attribués au moins une fois sont représentés par des genres différents au sein de notre corpus. Par ailleurs, si les textes d'entraînement et de test portent sur le même thème (par ex. *La femme du mort, Tome I* et *La femme du mort, Tome II*), ce qui est le cas pour 9 des 53 auteurs (15 textes de test), le meilleur modèle de langue a un rappel de 100 %. Il semble donc que l'homogénéité du genre et du thème soit un facteur non négligeable pour l'attribution d'auteur ; de plus amples expériences pourraient le confirmer.

Remerciements

Nous remercions Philippe Langlais pour ses nombreuses et précieuses suggestions de même que pour la constitution d'une partie du corpus. Nous remercions aussi Patrick Drouin et les relecteurs anonymes pour leurs commentaires.

Références

BAAYEN H., VAN HALTEREN H. & TWEEDIE F. (1996). Outside the cave of shadows : using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, **11**, 121–132.

CHEN S. F. & GOODMAN J. (1998). *An Empirical Study of Smoothing Techniques for Language Modeling*. Rapport interne.

FUCHUN P., SCHUURMANS D., KESELJ V. & WANG S. (2003). Automated authorship attribution with character level language models. In *Proceedings of the tenth conference of the European Chapter of the Association for Computational Linguistics*, p. 12–17, Budapest, Hongrie.

GOODMAN J. (2001). *A Bit of Progress in Language Modeling*. Rapport interne, Redmond, WA.

HOLMES D. I. (1994). Authorship attribution. *Computers and the Humanities*, **28**(2), 87–106.

KESELJ V., FUCHUN P., CERCONE N. & THOMAS C. (2003). N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03*, p. 255–264, Halifax, Canada : Pacific Association for Computational Linguistics.

KOPPEL M. & SCHLER J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, p. 69–72.

LUYCKX K. & DAELEMANS W. (2008). Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, p. 513–520, Manchester.

MCENERY T. & OAKES M. (2000). *Authorship identification and computational stylometry*, In R. DALE, H. MOISL & H. SOMERS, Eds., *Handbook of Natural Language Processing*, p. 545–562. Marcel Dekker : New York.

SCHAALJE G. B., HILTON J. L. & ARCHER J. B. (1997). Comparative power of three author-attribution techniques for differentiating authors. *Journal of Book of Mormon Studies*, **6**(1), 47–63.

STAMATATOS E., FAKOTAKIS N. & KOKKINAKIS G. (1999). Automatic authorship attribution. In *Proceedings of the ninth conference of the European Chapter of the Association for Computational Linguistics*, p. 158–164, Morristown, NJ : Association for Computational Linguistics.

STOLCKE A. (2002). SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, p. 901–904, Menlo Park, CA : SRI International.

STOLCKE A., YURET D. & MADNANI N. (2010). SRILM-FAQ. `http://www.speech.sri.com/projects/srilm/manpages/srilm-faq.7.html`.

TANGUY L. & HATHOUT N. (2003). *Perl pour les linguistes*. Paris : Lavoisier.

VAN HALTEREN H. (2004). Linguistic profiling for author recognition and verification. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, p. 199–206, Morristown, NJ : Association for Computational Linguistics.