Publié dans les
Actes 8e Conférence en Recherche d'Information et Applications CORIA'11,
135-150, 2011
qui doivent être utilisés pour toute référence à ce travail

Quel est l'auteur de ce roman ?

Jacques Savoy

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

RÉSUMÉ. *Dans cet article, nous présentons le problème de l'attribution d'auteur d'une œuvre écrite. Comme représentation des textes, les études récentes s'appuient sur un ensemble restreint de mots fonctionnels ou très fréquents (50 ou 100). Sur cette base, les méthodes de l'analyse en composantes principales (ACP) ou des correspondances (AC) permettent de visualiser les affinités et différences entre les représentations des écrits. En appliquant l'approche du plus proche voisin, nous pouvons estimer l'auteur de chaque texte. Comme alternative, nous suggérons de fonder le calcul de distance entre textes sur la base de la spécificité du vocabulaire (Z score). Basée sur une évaluation de corpus en langue française et anglaise, cette solution permet d'accroître la qualité de l'attribution d'auteur.*

ABSTRACT. *In this paper, we present the authorship attribution problem. As text representation, recent studies suggest using a small set of function or very frequent words (50 or 100). On this basis, we can apply either the principal component analysis (PCA) or the correspondence analysis (CA) to visualize the relationships between text surrogates. Using the nearest neighbor approach, we can then suggest the possible author of a disputed writing. As new attribution strategy, we propose a technique based on specific vocabulary found in a text comparing to an entire corpus. Based on the nearest neighbour approach, we can derive a simple and efficient authorship attribution scheme. Using two corpora composed of excerpts taken from French and English novels, we show that the suggested classifier tends to perform better than both the PCA and the CA approach.*

MOTS-CLÉS : *Catégorisation de textes, attribution d'auteur, analyse des correspondances (AC), analyse en composantes principales (ACP).*

KEYWORDS: *Text categorization, authorship attribution, correspondence analysis (CA), principal component analysis (PCA).*

1. Introduction

L'auteur d'un écrit (œuvre littéraire, article de presse, courriel) n'est pas toujours connu avec une absolue certitude (Juola, 2006). Dans cette communication, nous traiterons l'attribution d'auteur sous l'angle de la catégorisation automatique de textes dans laquelle chaque auteur potentiel correspond à une catégorie. Le système sélectionnera l'auteur probable en fonction de la représentation du document et du classifieur utilisé (Sebastiani, 2002). Avec Internet, le problème de l'attribution

d'auteur connaît de nouveaux prolongements comme la vérification d'auteur. Dans ce cas, on souhaite déterminer si le texte a été écrit ou non par un auteur donné. Enfin, au lieu de déterminer le nom du véritable auteur, on peut estimer des informations socio-économiques concernant l'auteur (profilage) comme le sexe, l'âge, la nationalité, le niveau d'éducation, etc. (Argamon *et al*., 2009).

Contrairement à d'autres applications en catégorisation de textes, l'attribution d'auteur dispose d'une longue tradition (Love, 2002). Ainsi, dès la fin de l'Antiquité on s'est interrogé sur les épîtres de St Paul, en estimant que cet ensemble n'a pas forcément été écrit par le même auteur. Dans la littérature française, la dispute entre Molière et Corneille concerne plusieurs chefs d'œuvre de Molière (Labbé, 2009), (Marusenko & Rodionova, 2010). En langue anglaise, plusieurs débats concernent des écrits de Shakespeare (Craig & Kinney, 2009). De nos jours, l'attribution de documents à un auteur spécifique fait également parti des sciences forensiques, de débats légaux, mais surtout d'un intérêt grandissant sur le Web avec, comme variante, la crédibilité des auteurs (blogs, Twitter) ou la détection de plagiat.

Dans la suite de cet article, nous désirons présenter le contexte et les principales méthodes suggérées dans l'attribution d'auteur (section 2). La troisième section expose les grandes lignes des deux corpus utilisés dans nos expériences. La quatrième section décrit brièvement et évalue l'application de méthodes statistiques exploratoires (analyse en composantes principales et des correspondances). La cinquième section présente et évalue notre nouvelle approche. Finalement, une conclusion dresse les principales contributions de cette étude.

2. État des connaissances

Afin de proposer une méthode en catégorisation automatique de textes, nous devons d'une part représenter les documents et, d'autre part, disposer d'un modèle de classification (Sebastiani, 2002). En attribution d'auteur (Juola, 2006), les études précédentes ont cherché à définir une mesure stylométrique que l'on espère constante pour un auteur et différente d'un rédacteur à l'autre (Holmes, 1998). On reconnaît toutefois que le genre (poésie, pièce de théâtre, roman, texte en vers ou en prose) va influencer une telle mesure, de même que, dans une moindre mesure, la chronologie (le style d'un auteur pouvant se transformer avec les années).

Les premières solutions ont proposé de tenir compte de la longueur moyenne des mots ou des phrases, du nombre moyen de syllabes par mots, voire de la taille du vocabulaire V (notée |V|) par rapport à la longueur du document (Baayen, 2008) (Grieve, 2007). Dans ces diverses approches, on essaie de réduire à une seule valeur numérique le style d'un auteur, valeur devant être distincte entre auteurs différents.

Le livre de Mosteller & Wallace (1964) marque la naissance de méthodes statistiques efficaces dans l'attribution d'auteur. Dans ce cas précis, les auteurs privilégient l'analyse du vocabulaire et, plus précisément, des mots fonctionnels (déterminant, prépositions, conjonctions, et pronoms). Dans ce raisonnement, on

admet que l'usage et la fréquence d'apparition de ces formes ne sont pas sous le contrôle conscient de l'auteur et qu'ils varient d'une personne à l'autre. Durant les années suivantes, les méthodes proposées vont s'appuyer sur le vocabulaire comme élément significatif du style d'un auteur. Au niveau lexical, on s'est également tourné vers les mots apparaissant une seule fois dans le document (ou *hapax legomena*) (Morton, 1986) en tenant compte également de leur position (e.g., mot initial ou final de la phrase). Sur ces informations, la différence entre auteurs pouvant se mesurer grâce à la distance du chi-carré (Grieve, 2007). Comme variante, on peut signaler la mesure de la taille du vocabulaire connu d'un auteur (Efron & Thisted, 1976), (Thisted & Efron, 1987). Comme alternative plus simple, on a proposé la valeur $R = |V| / \text{sqrt}(n)$ (avec n la taille du corpus) de Guiraud, le rapport entre le nombre de *hapax legomena* (notée V_1) et la taille du vocabulaire (soit $|V_1| / |V|$), ou le rapport entre le nombre de *dislegomena* (noté $|V_2|$, et défini comme le nombre de mots apparaissant deux fois) et la taille du vocabulaire (Sichel, 1975). Toutefois, ces mesures ont l'inconvénient d'être assez instables (Baayen, 2008), en particulier face à des documents relativement courts ($n \leq 1\,000$).

Dans une troisième étape, Burrows (2002) propose de définir un ensemble de mots très fréquents pouvant refléter le style d'un auteur et qui soit indépendant du thème traité. La taille du vocabulaire à étudier comprend entre 50 et 150 formes. Mais la définition précise des termes à inclure dans l'analyse reste floue. On peut définir un tel ensemble comme les k formes les plus fréquentes d'un corpus, d'une langue ou inclure uniquement les mots appartenant aux parties du discours fermées (déterminants, prépositions, conjonctions, pronoms). Afin de confronter les auteurs potentiels d'un ouvrage, on utilisera des outils de statistiques multivariées comme l'analyse en composantes principales (ACP) (Burrows, 1992), (Binonga & Smith, 1999) ou, mieux connue dans l'école française, l'analyse des correspondances (AC) (Dixon & Mannion, 1993) (Lebart *et al*., 1998). D'autres approches ont également été proposées comme la classification automatique (*clustering*), parfois en complément à l'ACP (Holmes, 1992) ou en combinant quelques mesures de la richesse du vocabulaire comme les mesures H (Honoré), S (Sichel) ou K (Yule). Finalement, le recours à l'analyse discriminante a également été suggéré, avec la fréquence des lettres comme évidence, permettant une distinction entre auteurs (Ledger & Merriam, 1992). Dans ces études, la distinction entre auteurs n'est pas parfaite et les attributions peuvent se compliquer, en particulier face à des auteurs ayant une culture commune (e.g., Goldsmith, Kelly & Murphy et leurs racines anglo-irlandaises (Dixon & Mannion, 1993)) ou lorsque le pouvoir discriminant s'estompe (e.g., l'emploi de la fréquence des lettres pour distinguer entre les pièces écrites par Shakespeare, Fletcher, ou Dekker (Ledger & Merriam, 1992)).

Le recours à des méthodes tirées de l'apprentissage automatique (*machine learning*) marque une quatrième étape dans les problèmes liés à l'attribution d'auteur (Stamatatos, 2009). Les études faites par Zhao & Zobel (2005) ou par Zhao & Zobel (2007) sont des exemples caractéristiques. Sur la base des 365 formes les plus fréquentes, Zhao & Zobel (2005) cherchent à déterminer l'auteur d'articles de presse.

Comme stratégie de classification, ils démontrent qu'une approche Naïve Bayes tend à fournir une meilleure performance que les arbres de décision. Laroche (2010) indique qu'une représentation par bigrammes de mots et le recours à des modèles de langues apportent de meilleurs résultats qu'une représentation par lemmes ou parties du discours. Zheng *et al*. (2006) étudient l'application d'arbres de décision, de réseaux neuronaux et de machine à vecteurs de support (SVM) pour l'attribution de courts messages électroniques écrits en langue anglaise ou chinoise. Dans ces expériences, non seulement les éléments lexicaux sont pris en compte, mais les auteurs étudient l'efficacité de tenir compte d'éléments syntaxiques, de contenu ou de présentation.

3. Corpus d'évaluation

Contrairement à la recherche d'information, les études en attribution d'auteur disposent d'un nombre restreint de corpus permettant de vérifier et de comparer les performances des diverses démarches. De plus, les corpus disponibles sont souvent écrits uniquement en langue anglaise. Désirant fonder nos conclusions sur une base plus solide, nous avons décidé d'évaluer nos propositions sur deux corpus écrits dans deux langues différentes. Afin de répondre à ces critères, nous avons sélectionné un corpus en langue française et un second en anglais.

Le premier corpus retenu a été créé par E. Brunet et utilisé par D. Labbé. Ecrit en langue française, il regroupe 44 extraits composés d'environ 10 000 formes et numérotés de “01” à “44”. Ces fragments sont tirés de romans ou d'essais écrits principalement durant le XIXe siècle (l'annexe en donne une description plus précise). La composition de ce corpus est assez systématique dans le sens que nous retrouvons onze auteurs distincts, avec deux œuvres pour chacun et exactement deux extraits par œuvre (11 x 2 x 2 = 44).

Cette collection comprend 439 662 mots ou formes (en comptant la ponctuation, les chiffres et les nombres). Le lemme (ou entrée dans le dictionnaire) le plus fréquent est le déterminant “le” (38 270 occurrences), suivi par la virgule (30 782 occurrences) et par la préposition “de” (27 382 occurrences). Sur l'ensemble du corpus, on retrouve 13 915 lemmes distincts (taille du vocabulaire). Un extrait typique comprendra 9 992 lemmes en moyenne (médiane : 10 024, écart-type : 145,7). La différence entre le fragment le plus long (n^{o} 29 comprenant 10 239 lemmes) et le plus court (n^{o} 23, avec 9 612 lemmes) demeure faible (soit 627 formes).

Le second corpus, nommé *Oxquarry,* comprend 52 extraits de romans écrits à la fin du XIXe et au début du XXe siècle. Créée par G. Ledger, cette collection comprend des œuvres dont *apriori* la distinction entre auteurs s'avère difficile sur la base du vocabulaire (Labbé, 2007). Chaque extrait comprend environ 10 000 mots (ou formes). Chaque extrait est identifié par une étiquette allant de “1A” à “1Z” et de “2A” à “2Z”. Comme indiqué dans l'annexe, on y retrouve des écrits de neuf

auteurs provenant de 16 nouvelles. Trois auteurs (Chesteron, Forster et Tressel) sont représentés par trois fragments, tandis que douze textes ont été sélectionnés de quatre œuvres de Hardy (*Jude*, *Madding*, *Well Beloved* et *Wessex Tales*).

Si l'on analyse ce corpus, on compte un total de 530 813 lemmes, pour 21 065 lemmes distincts (taille du vocabulaire). Le déterminant “the” correspond au lemme le plus fréquent (30 048 occurrences), suivi par le verbe “be” (19 919 occurrences), et le pronom “he” (17 752 occurrences). Pour un fragment typique, on retrouve, en moyenne, 10 207 lemmes (médiane : 10 201, écart-type : 55,6).

Pour les deux corpus, nous avons retenu les entrées dans le dictionnaire ou lemmes (e.g., “être”, “(to) be”) et non les formes de surface (e.g., “est”, “était”, “été” pour le français ou “is”, “was”, “been” pour la langue anglaise). En effet, il existe des homographies, particulièrement en français, que nous pouvons distinguer par l'emploi de lemmes. Ainsi, il convient de distinguer entre le verbe être et la saison dans la forme “été”, ou avec “est” entre la direction et ce même verbe.

Notons que si nous avons retenu uniquement l'entrée dans le dictionnaire, les ambigüités liées à certaines homographies ne sont pas toujours levées. Ainsi, nous ne distinguons pas entre le “livre” (écrit) et la “livre” (mesure de poids) ou, pour la langue anglaise entre le lieu “desert” (même forme pour le nom et l'adjectif) et le verbe “desert” (déserter) (ou entre “to” préposition et “to” infinitif, distinction faite manuellement dans (Burrows, 2002)). Ce processus de lemmatisation a été réalisé grâce au logiciel de Labbé (2001) pour la langue française et par celui de Toutanova *et al*., (2003) pour la langue anglaise.

4. Méthodes statistiques multivariées et le plus proche voisin

Afin de représenter les extraits d'œuvres littéraires, nous avons repris les k lemmes les plus fréquents de notre corpus (avec $k = 50$ ou 100). Ce choix a été suggéré par plusieurs auteurs (Burrows, 1992), (Binonga & Smith, 1999) (Zhao & Zobel, 2007). Afin de visualiser ces textes et leurs affinités, nous avons retenu deux méthodes statistiques multivariées, soit l'analyse en composantes principales (ACP) et l'analyse des correspondances (AC).

La base de notre analyse repose sur un tableau lexical, soit un ensemble de termes et leurs fréquences en fonction des œuvres. Le tableau 1 présente un exemple simplifié (dix lemmes) avec les romans *Le cousin Pons* (Balzac), *La mare au diable* (Sand) et *Madame Bovary* (Flaubert). Les fréquences d'occurrences provenant d'extraits de taille similaire peuvent donc être comparées directement. Cependant, attribuer un auteur sur la base d'un seul ou de quelques éléments s'avère périlleux. Par exemple, dans le tableau 1, le texte n° 10 laisse entrevoir un sous-emploi marqué du pronom “je” (une seule occurrence). Ce sous-emploi pourrait nous conduire à le rapprocher du texte n° 35 présentant également un sous-emploi (deux occurrences). Or, ces deux œuvres ont été écrites par des auteurs différents, tandis que le texte n° 32, possédant 102 occurrences du pronom “je”, provient du

même roman que le texte n° 10. Un raisonnement similaire peut être fait sur la base des fréquences d'occurrence du déterminant "le", sur-employé dans les fragments n° 10 et n° 35 comparés aux deux autres. Parfois on peut distinguer une œuvre des autres, comme en analysant l'abondance de virgules dans l'extrait de *Madame Bovary* ou la faible fréquence du lemme "de" dans *La mare au diable*. Face à uniquement quatre extraits et dix termes, ce travail demeure relativement aisé.

	n° 10 *Le cousin Pons*	n° 32 *Le cousin Pons*	n° 34 *La mare au diable*	n° 35 *Mme Bovary*
le	1 096	862	629	1 114
,	770	714	704	939
de	716	694	381	736
.	288	375	415	268
à	258	246	201	327
il	126	168	221	254
et	148	156	227	200
être	152	195	236	103
que	110	135	223	97
je	1	102	263	2

Tableau 1. *Fréquence absolue des dix lemmes les plus fréquents dans des extraits de trois romans écrits par trois auteurs différents*

Le tableau 2 présente une situation similaire avec quatre extraits de notre corpus en langue anglaise. Ainsi, le premier extrait (n° 1D) possède un nombre restreint d'occurrences du déterminent "the" (421) tandis que dans *Almayer* (n° 1T) Conrad utilise abondamment ce terme. Par contre, le pronom "you" s'avère fréquent dans l'extrait n° 1D de *Catriona* (280 occurrences). Cette fréquence revient à une valeur plus proche de la moyenne avec le second extrait (n° 2R, 127 occurrences), et à une valeur similaire au fragment de *Jude* (135 occurrences). Une abondance du pronom "you" ne peut donc caractériser le style de *Stevenson*.

	1D *Catriona*	2R *Catriona*	1T *Almayer*	1L *Jude*
the	421	548	838	540
a	3	5	7	12
you	280	127	54	135
which	27	39	14	41
would	42	48	31	28
my	18	4	1	2
not	73	57	70	135
that	149	115	82	128

Tableau 2. *Fréquence d'occurrence de huit lemmes parmi les plus fréquents dans des extraits de* Catriona *(Stevenson),* Almayer *(Conrad) et* Jude *(Hardy)*

Afin de visualiser *au mieux* et en deux dimensions ces 44 extraits en fonction des 50 lemmes, nous pouvons appliquer l'analyse en composantes principales (ACP), une technique connue en attribution d'auteur (Craig & Kinney, 2009) (Hoover &

Hess, 2009). On comprend aisément que sur la base d'un tableau comprenant 44 colonnes (ou individus représentant les 44 textes) et 50 lignes (ou caractères correspondant aux 50 lemmes), l'œil humain a de la peine à distinguer rapidement les fragments similaires et ceux qui sont dissemblables.

Sur la base de ces informations, la technique de l'ACP va construire de nouveaux axes orthonormés, combinaisons linéaires des caractères spécifiés en entrée. On ne procède donc pas à une sélection d'un nombre restreint de caractères en fonction, par exemple, de leur pouvoir discriminant. L'ACP fait l'hypothèse que tous les caractères jouent le même rôle et qu'ils correspondent à des critères numériques réels. Or, notre tableau lexical ne correspond pas à cette seconde attente; ce sont des entiers. Nous devons donc soit centrer et réduire les données, soit fournir la matrice de corrélation. Nous avons opté pour la seconde stratégie. L'étude de Holmes (1992) indique que les deux possibilités fournissent des résultats similaires.

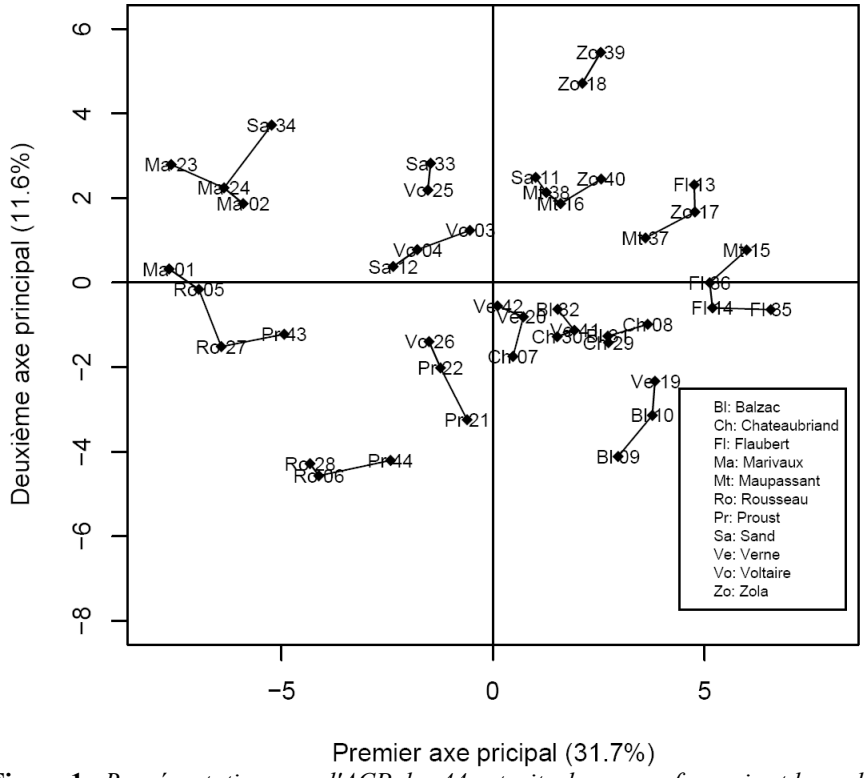

Figure 1. *Représentation avec l'ACP des 44 extraits du corpus français et leur plus proche voisin (basée sur 50 lemmes les plus fréquents)*

En résultat, nous avons un nouveau système d'axes perpendiculaires reflétant une portion décroissante de la variance sous-jacente. Ainsi, en sélectionnant les q premiers axes, nous obtenons une représentation la plus fidèle possible tout en limitant l'espace à un nombre retreint q de facteurs. Dans le cas où $q = 2$, une

visualisation sur le plan est possible comme l'illustre la figure 1. Dans ce cas, le premier axe représente 31,7 % de la variance totale, le deuxième 11,6 % et le troisième 10,1 % (le 4e 7,4 %, le 5e 5,7 %, le 6e 4,4 %, …, le 22e 0,6 %, …, le 36e 0,0 %, …). Prendre en compte l'ensemble des axes principaux n'apporte pas un intérêt particulier et on peut se limiter aux axes représentant 5 % ou plus de la variance totale, soit dans notre exemple les cinq premiers.

Dans la figure 1, basée sur 43,3 % (31,7 % + 11,6 %) de la variance, nous avons fait précéder le numéro de chaque fragment par les deux premières lettres du nom de l'auteur, facilitant ainsi le repérage des textes écrits par la même personne. Cette figure permet de visualiser les textes partageant des profils lexicaux similaires comme, dans la partie supérieure gauche, entre les textes "Ma 23" (*La vie de Marianne*) et "Ma 24" (*Le paysan parvenu*) ou, dans la partie supérieure droite, entre les documents "Zo 39" (*Thérèse Raquin*) et "Zo 18" (*La bête humaine*).

Comme l'ACP permet de visualiser les distances entre les différentes représentations de nos romans, nous pouvons utiliser cette mesure pour identifier le *plus proche voisin* de chaque extrait (*nearest neighbor*). Afin de déterminer l'auteur d'un passage, nous lui attribuons le nom de l'auteur du texte le plus proche. Afin de déterminer clairement le plus proche voisin de chaque texte, nous avons dessiné, dans la figure 1, un trait entre chaque paire de points. On voit ainsi que, dans la partie inférieure droite, le plus proche voisin du point "Bl 09" (*Les Chouans*) est le point "Bl 10" (*Le Cousin Pons*). Cette stratégie d'appariement n'est pas exempte d'erreur comme le démontre le lien établi entre le texte "Bl 10" (*Le Cousin Pons*) et le document "Ve 19" (*De la terre à la lune*).

		ACP 50	ACP 100	AC 50	AC 100
original	2 axes	40,9 %	50,0 %	40,9 %	36,4 %
	5 axes	72,7 %	72,7 %	72,7 %	72,7 %
7 500	2 axes	36,4 %	54,6 %	36,4 %	38,6 %
	5 axes	**79,6 %**	72,7 %	75,0 %	72,7 %
5 000	2 axes	36,4 %	45,5 %	36,4 %	27,3 %
	5 axes	68,2 %	65,9 %	63,6 %	63,6 %
2 500	2 axes	38,6 %	31,8 %	27,3 %	34,1 %
	5 axes	45,5 %	63,6 %	56,8 %	54,6 %
1 000	2 axes	29,6 %	34,1 %	27,3 %	31,8 %
	5 axes	29,6 %	36,4 %	20,5 %	36,4 %

Tableau 3. *Évaluation de l'attribution d'auteur par analyse en composantes principales (ACP) ou analyse des correspondances (AC) (corpus français) avec 50 ou 100 lemmes et en tenant compte des deux ou cinq premiers axes*

En appliquant cette stratégie du *plus proche voisin* basée sur les distances obtenues en tenant compte des deux premiers axes principaux de l'ACP (50

lemmes), nous obtenons un taux de succès de 40,9 % (18 attributions correctes sur 44). Par contre, nous pouvons déterminer le *plus proche voisin* en considérant les cinq premiers axes principaux. Dans ce cas, le taux de succès passe à 72,7 % (ou 32 attributions correctes sur 44). De plus, au lieu de considérer uniquement les $k = 50$ lemmes les plus fréquents, nous pouvons retenir les 100 lemmes. Dans ce cas, la stratégie du *plus proche voisin* permet de déterminer correctement 50 % des textes avec les deux premiers axes et 72,7 % avec les cinq premiers axes. Ces résultats sont repris dans le tableau 3 sous la ligne "original" et les colonnes "ACP 50" (pour 50 lemmes) et "ACP 100" (100 lemmes). Si l'on tient compte de tous les axes possibles, la performance moyenne se situe aux environs de 8 %, indiquant bien que de nombreux axes ajoutent plus de bruit que d'information pertinente.

Cette première évaluation se fondait sur l'ensemble du fragment mis à notre disposition, soit environ 10 000 mots. Afin de connaître la performance de cette stratégie face à ces documents plus courts, nous avons extrait un sous-ensemble possédant une longueur variant de 7 500 à 1 000 mots. Comme on peut s'y attendre, la performance a tendance à diminuer face à des textes de longueur plus faible, sans que cette diminution soit systématique (voir tableau 3). Ainsi, la meilleure performance s'obtient avec 50 lemmes, cinq axes principaux mais en ayant des extraits comprenant 7 500 mots.

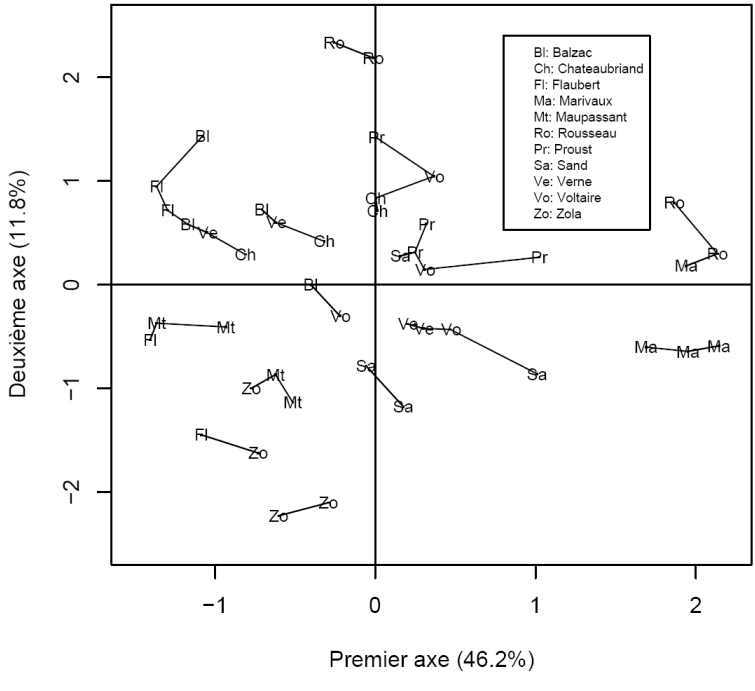

Figure 2. *Représentation des 44 extraits du corpus français et de leur plus proche voisin (AC, 50 lemmes)*

Comme méthode alternative, notre choix s'est porté sur l'analyse des correspondances (AC) (Lebart *et al*., 1998), (Greenacre, 2007), une approche similaire à l'ACP. En entrée toutefois, nous pouvons directement introduire un tableau lexical. Le calcul de distance entre textes ne se fonde pas sur une mesure euclidienne basée sur les fréquences. En effet, ce choix aurait pour effet d'accorder beaucoup d'importance aux grandes différences. Par exemple, sur la base du tableau 1, les variations de fréquences liées au déterminent “le” sont, en valeurs absolues, souvent plus importantes que celles liées au pronom “je”. Pour corriger quelque peu ce phénomène, la distance du chi-carré sous-jacente à l'analyse des correspondances équivaut à une distance euclidienne pondérée par le profil moyen.

Comme pour l'ACP, le système retourne une représentation basée sur une série ordonnée d'axes orthonormés. Basé sur 50 lemmes, le premier axe représente 46,2 % de la variance totale (ou de l'inertie), le deuxième 11,8 % et le troisième 9,1 % (le 4e 6,2 %, le 5e 5,0 %, le 6e 3,2 %, …, le 22e 0,3 %, …, le 43e 0,0 %). Dans la figure 2 basée sur 58 % (46,2 % + 11,8 %) de l'inertie totale, le croissement des axes marque l'individu composé de la moyenne des caractères. Plus les points s'éloignent de ce centre, plus ils se différencient. Ainsi, sur la partie droite, on retrouve les points “Ma 23”, “Ma 02” ou “Ro 05”, indiquant que Marivaux et Rousseau se distinguent du centre, mais également des œuvres situées à l'autre extrême de l'axe, soit Flaubert (“Fl 35”) ou Maupassant (“Mt 15”). Le premier groupe se caractérise par une fréquence forte sur les lemmes “je”, “vous”, “moi” ou “mon”, tandis que le second groupe comprend plus souvent les lemmes “sur”, “se”, “son” ou “ils”.

		ACP 50	ACP 100	AC 50	AC 100
original	2 axes	61,5 %	63,5 %	42,3 %	40,4 %
	5 axes	84,6 %	**96,2 %**	84,6 %	90,4 %
7 500	2 axes	44,2 %	44,2 %	38,5 %	36,5 %
	5 axes	86,5 %	82,7 %	69,2 %	80,8 %
5 000	2 axes	42,3 %	50,0 %	32,7 %	38,4 %
	5 axes	82,7 %	78,9 %	75,0 %	76,9 %
2 500	2 axes	28,9 %	36,5 %	26,9 %	28,9 %
	5 axes	59,6 %	50,0 %	53,9 %	46,2 %
1 000	2 axes	28,9 %	28,9 %	26,9 %	30,8 %
	5 axes	32,7 %	32,7 %	30,8 %	30,8 %

Tableau 4. *Évaluation de l'attribution d'auteur par ACP ou AC (corpus anglais) avec 50 ou 100 lemmes et en tenant compte des deux ou cinq premiers axes*

Afin de connaître la qualité de cette approche en attribution d'auteur, nous avons repris les deux ou cinq axes principaux de même que les 50 ou 100 lemmes les plus fréquents. En appliquant la stratégie du *plus proche voisin*, nous pouvons obtenir les

performances indiquées dans le tableau 3, sous les colonnes “AC 50” (pour 50 lemmes) et “AC 100” (avec 100 lemmes). Sur la base de 50 lemmes, les performances sont assez similaires à celles obtenues avec l'ACP, tandis que pour 100 lemmes, la performance avec l'AC s'avère moins bonne lorsque les extraits sont relativement longs (entre 10 000 et 5 000 mots). Pour des tailles plus réduites, les performances avec l'AC et l'ACP demeurent proches.

Si l'on considère le corpus anglais et ses 52 extraits d'œuvres littéraires, l'application de l'ACP ou de l'AC (50 ou 100 lemmes) sur l'ensemble des fragments (soit 10 000 lemmes en moyenne), nous obtenons les performances indiquées dans le tableau 4. Comme pour la langue française, les niveaux de performance restent bons si l'on réduit un peu la taille des fragments (e.g., 7 500 lemmes). Par contre, avec des volumes entre 1 000 et 2 500 mots par texte, l'attribution devient plus difficile, et particulièrement lorsque l'on se limite aux deux premiers axes.

Quelques difficultés peuvent survenir dans l'application de ces méthodes comme, par exemple, la présence d'un lemme n'apparaissant dans aucun texte (e.g., “him” dans les extraits de taille 1 000 mots). Dans de tels cas, le calcul ne peut s'effectuer sans élimination, au préalable, de ces caractères.

5. Vocabulaire spécifique et attribution d'auteur

Comme nouvelle stratégie d'attribution d'auteur ou, plus généralement, de catégorisation de textes, nous proposons de tenir compte de la spécificité des termes (lemme, vocable, ponctuation, etc.) du vocabulaire appartenant à une catégorie (ou à un auteur) en comparaison des autres. La définition d'un terme spécifique a été développée par Muller (1992). Afin de mesurer cette spécificité, on subdivise notre corpus en deux parties disjointes, soit entre un extrait *e* donné (par exemple le n° “02”) et le reste du corpus (noté *r*). Pour mesurer la spécificité d'un lemme ω (par exemple le verbe “être” repris dans le tableau 5a ou le terme “son” dans le tableau 5b), on compte le nombre d'occurrences dans la partie *e* (valeur notée tf_e) et sa fréquence dans *r* (notée tf_r). Pour l'ensemble du corpus, nous retrouverons tf_e+tf_r occurrences de ce terme. La taille de l'ensemble *e* s'élève à n_e, tandis que le volume du corpus sera n (= $n_e + n_r$).

	extrait 02	**reste**	corpus
lemme “être”	297	8 411	8 708
autres lemmes	5 779	236 676	242 455
	6 076	245 087	251 163

Tableau 5a. *Table de contingence pour le lemme* **“***être***”** *dans l'extrait 02 (Marivaux,* Le paysan parvenu*) et le reste du corpus*

En faisant l'hypothèse que la fréquence d'apparition du lemme ω suit une loi binomiale, nous devons estimer Prob[ω] sa probabilité d'apparition lorsque nous tirons un terme au hasard dans le corpus. Cette probabilité s'estime par (tf_e+tf_r) / n.

Si l'on répète n_e fois ce tirage aléatoire, on peut estimer le nombre de termes ω tiré par Prob[ω]·n_e. Ce nombre correspond au nombre attendu que nous comparons avec le nombre observé, soit tf_e. De manière précise, nous calculons un score Z pour chaque vocable ω selon l'équation 1 dans laquelle n_e·Prob[ω] représente la moyenne et n_e·Prob[ω]·(1-Prob[ω]) la variance de la distribution binomiale.

$$\text{Z score}(\omega) = \frac{tf_e - n_e \cdot \text{Prob}[\omega]}{\sqrt{n_e \cdot \text{Prob}[\omega] \cdot (1 - \text{Prob}[\omega])}} \quad (1)$$

	extrait 02	**reste**	corpus
lemme “son”	89	4 953	5 042
autres lemmes	5 987	240 134	246 121
	6 076	245 087	251 163

Tableau 5b. *Table de contingence pour la préposition* **“***son***”** *dans l'extrait 02 (Marivaux,* Le paysan parvenu*) et le reste du corpus*

Sur la base des valeurs indiquées dans la tableau 5a, on peut estimer Prob[être] = 8708 / 251163 = 0,0347 et le score Z associé sera de (297 – 6076 · 0,0347) / √(6076 · 0,0347 · (1-0,0347)) = 6,05. Pour le lemme “son”, les valeurs du tableau 5b nous permettent d'obtenir la valeur du score Z s'élevant à -3,02. Une valeur de score Z positive et supérieure à un seuil δ donné (e.g., δ = 2) indique un sur-emploi dans l'extrait considéré. Un sous-emploi sera indiqué par un score Z inférieur à un seuil - δ donné. Dans nos exemples, le lemme “être” correspond à un sur-emploi dans l'extrait du *Le paysan parvenu* et le terme “son” à un sous-emploi.

La valeur du score Z est attachée à chaque terme (lemme, vocable, bigramme de mots, etc.) et elle va nous servir de base au calcul d'une distance entre deux textes. Si nous avons deux documents D_j et D_k, et un ensemble de terme t_i, pour i = 1, 2, … m, nous pouvons définir une distance basée sur le score Z en appliquant l'équation 2 dans laquelle t_{ki} indique le terme t_i dans le document D_k.

$$\text{Distance Z}(D_j, D_k) = \frac{1}{m} \cdot \sum_{i=1}^{m} \left(Zscore(t_{ji}) - Zscore(t_{ki}) \right)^2 \quad (2)$$

Avec cette mesure, lorsque deux documents possèdent des scores Z proches, la distance sera faible, indiquant une proximité du vocabulaire spécifique. De fortes différences entre deux scores Z vont entraîner une plus grande distance. Enfin, mettre au carré la différence entre les scores Z correspondant au même terme permet souvent de réduire l'impact du vocabulaire commun, c'est-à-dire des termes ayant un score Z compris entre - δ et + δ. Avec une définition d'une distance, nous pouvons recourir au *plus proche voisin* afin de déterminer l'auteur probable d'un écrit.

Afin d'éviter de traiter l'ensemble du vocabulaire, nous pouvons ignorer les lemmes apparaissant une fois (*hapax*) ou deux (*dislegomena*). Nous avons également éliminé les lemmes dont 90 % des occurrences apparaissent dans un seul texte. Pour le corpus français (fragments de 10 000 mots), nous disposons de 13 916 lemmes distincts avec 5 240 *hapax*, 1 985 *dislegomena* et 273 lemmes trop fréquents

dans un seul fragment. Notre stratégie œuvrera donc sur les 6 418 lemmes restants. Pour le corpus anglais, le vocabulaire possède une taille de 21 065 auquel on retire 9 006 *hapax*, 3 188 *dislegomena* et 409 lemmes très fréquents dans un seul écrit. L'attribution se fondera donc sur 8 462 lemmes. En se limitant à des textes de 1 000 mots, l'attribution pour le corpus français se basera sur 1 735 termes, tandis que pour la langue anglaise, nous aurons recours à 1 418 lemmes.

Cette stratégie a été appliquée sur nos deux corpus et les performances obtenues sont indiquées dans le tableau 6. Comparés à l'approche ACP ou AC, les résultats indiqués s'avèrent meilleurs avec, en moyenne, une progression de 7,9 % par rapport à la meilleure approche ACP pour le corpus français et de 15,4 % pour le corpus anglais. Pour un seul cas, la technique de l'ACP génère un meilleur résultat. Lorsque l'on considère 2 500 mots par extraits pour le corpus français, la méthode de l'ACP (100 lemmes, 5 axes) offre une performance de 63,6 %, soit une valeur supérieure à celle indiquée dans le tableau 6 (56,8 %).

	Français	Anglais
original	**100 %**	**100 %**
7 500	90,9 %	98,1 %
5 000	72,7 %	86,5 %
2 500	56,8 %	69,2 %
1 000	38,6 %	57,7 %

Tableau 6. *Évaluation de l'attribution d'auteur par la méthode du vocabulaire spécifique*

6. Conclusion

Déterminer précisément l'auteur d'un texte repose sur différentes approches comme l'attribution faite par les contemporains, l'étude des faits historiques, l'analyse de la correspondance des auteurs potentiels, les affinités politiques, idéologiques ou religieuses, de même que les études antérieures concernant l'attribution de l'œuvre en question. Dans le cadre de cette communication, nous avons discuté d'un ensemble d'approches provenant des techniques en statistiques multivariées et de l'apprentissage automatique (catégorisation de textes).

Les méthodes étudiées sont fondées sur le vocabulaire et la fréquence d'emploi des lemmes. Nous pensons que le recours aux mots ou vocables à la place des lemmes aurait certainement donné des résultats assez similaires. Par contre, le choix des lemmes s'explique par la richesse morphologique du français (les vocables “est”, “es”, “avais” sont regroupés sous la même entrée) que nous souhaitons maîtriser quelque peu et par notre souhait d'éliminer quelques ambigüités lexicales (“été” comme verbe ou nom). Ainsi, les diverses formes verbales (e.g., pouvoir, faire) ou liées à la morphologie flexionnelle ne vont pas influencer nos calculs.

Basé sur un corpus de 44 extraits d'œuvres littéraires écrites en français et de 52 autres écrites en anglais, les approches basées sur l'ACP ou l'AC, conjuguées avec la stratégie du *plus proche voisin*, offrent des résultats relativement moyens lorsque l'on se limite aux deux premiers axes principaux. La prise en compte de tous les axes représentant au moins 5 % de la variance permet d'améliorer ces premiers résultats.

La stratégie proposée repose sur une distance intertextuelle basée sur le vocabulaire spécifique (score Z). Associée au *plus proche voisin*, cette démarche fournit d'excellents résultats, supérieurs aux approches basées sur l'ACP ou l'AC.

Remerciements

L'auteur tient à remercier D. Labbé pour nous avoir donné accès aux deux corpus utilisés, ainsi que la version lemmatisée du corpus français. Cette recherche a été financée en partie par le Fonds national suisse pour la recherche scientifique (subside FN 200021-124389/1)

7. Bibliographie

Argamon S., Koppel M., Pennebaker J.W., Schler J. « Automatically profiling the author of an anonymous text », *Communications of the ACM*, vol. 52, n° 2, 2009, p. 119-123.

Baayen R.H., *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*. Cambridge, Cambridge University Press, Cambridge, 2008.

Binonga J.N.G., Smith M.W., « The application of principal component analysis to stylometry », *Literary and Linguistic Computing*, vol. 14, n° 4, 1999, p. 445-465.

Burrows, J.F., « Delta: A measure of stylistic difference and a guide to likely authorship », *Literary and Linguistic Computing*, vol. 17, n° 3, 2002, p. 267-287

Burrows J.F., « Not unless you ask nicely: The interpretative nexus between analysis and information », *Literary and Linguistic Computing*, vol. 7, n° 1, 1992, p. 91-109.

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

Dixon P., Mannion D., « Goldsmith's periodical essays: A statistical analysis », *Literary and Linguistic Computing*, vol. 8, n° 1, 1993, p. 1-19.

Efron B., Thisted R. « Estimating the number of unseen species: How many words did Shakespeare know? », *Biomerika*, vol. 63, n° 3, 1976, p. 435-447.

Greenacre M. *Correspondence Analysis in Practice*, 2nd Ed. Chapman & Hall/CRC, Boca Raton, 2007.

Grieve J., « Quantitative authorship attribution: An evaluation of techniques », *Literary and Linguistic Computing*, vol. 22, n° 3, 2007, p. 251-270.

Holmes D.I., « A stylometric analysis of Mormon scripture and related texts », *Journal of the Royal Statistical Society A*, vol. 155, n° 1, 1992, p. 91-120.

Holmes D.I., « The evolution of stylometry in humanities scholarship », *Literary and Linguistic Computing*, vol. 13, n° 3, 1998, p. 111-117.

Hoover D.L., Hess S., « An exercise in non-ideal authorship attribution: The mysterious Maria Ward », *Literary and Linguistic Computing*, vol. 24, n° 4, 2009, p. 467-489.

Juola P., « Authorship attribution », *Foundations and Trends in Information Retrieval*, vol. 1, n° 3, 2006.

Labbé, D. « Normalisation et lemmatisation d'une question ouverte », *Journal de la Société Française de Statistique*, vol. 142, n° 4, 2001, p. 37-57.

Labbé D. « Experiments on authorship attribution by intertextual distance in English », *Journal of Quantitative Linguistics*, vol. 14, n° 1, 2007, p. 33-80.

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

Laroche A., « Attribution d'auteur au moyen de modèles de langue et de modèles stylométriques », *Actes RECITAL*, 2010.

Lebart L., Salem A., Berry L. *Exploring Textual Data*. Dordrecht (NL), Kluwer, 1998.

Ledger G., Merriam R., « Shakespeare, Fletcher, and the *Two Noble Kinsmen* », *Literary and Linguistic Computing*, vol. 9, n° 3, 1994, p. 235-248.

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

Marusenko M., Rodionova E., « Mathematical methods for attributing literary works when solving the "Corneille-Molière" problem », *Journal of Quantitative Linguistics*, vol. 17, n° 1, 2010, p. 30-54.

Morton A.Q., « Once. A test of authorship based on words which are not repeated in the sample », *Literary and Linguistic Computing*, vol. 1, n° 1, 1986, p. 1-8.

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

Muller C. *Principes et Méthodes de Statistique Lexicale*. Paris, Honoré Champion, 1992.

Sebastiani F., « Machine learning in automatic text categorization », *ACM Computing Survey*, vol. 14, n° 1, 2002, p. 1-27.

Sichel H.S., « On a distribution law for word frequencies », *Journal of the American Statistical Association*, vol. 70, n° 351, 1975, p. 542-547.

Stamatatos E., « A survey of modern authorship attribution methods », *Journal of the American Society for Information Science and Technology*, vol. 60, n° 3, 2009, p. 433-214.

Thisted R., Efron B. « Did Shapespeare write a newly-discovedred poem? », *Biomerika*, vol. 74, n° 3, 1987, p. 445-455.

Toutanova K., Klein, D. Manning C., Singer Y. « Feature-rich part-of-speech tagging with a cyclic dependency network », *Proceedings of HLT-NAACL*, 2003, p. 252-259.

Zhao Y., Zobel J. « Effective and scalable authorship attribution using function words », *Proceedings of the Second AIRS Asian Information Retrieval Symposium*, 2005, Berlin, Springer-Verlag, p. 174-189.

Zhao Y., Zobel J. « Searching with style: Authorship attribution in classic literature », *Proceedings of the Thirtieth Australasian Computer Science Conference (ACSC2007)*, 2007, Ballarat, p. 59-68.

Zheng R., Li J., Chen H., Huang, Z. « A framework for authorship identification of online messages: Writing-style features and classification techniques », *Journal of the American Society for Information Science & Technology*, vol. 57, n° 3, 2006, p. 378-393.

Zubaryeva O., Savoy J. « Evaluation de modèles de classification appliqués à la détection d'opinions », *Actes 7ième COnférence en Recherche d'Information et Applications CORIA'2010*, mars 2010, Sousse, p. 271-286.

Ids	Année	Auteur	Titre
1, 23	1731	Marivaux	*La vie de Marianne (L.1)*
2, 24	1734	Marivaux	*Le paysan parvenu (L.1)*
3, 25	1747	Voltaire	*Zadig*
4, 26	1759	Voltaire	*Candide*
5, 27	1761	Rousseau	*La nouvelle Héloïse (L.1)*
6, 28	1762	Rousseau	*Emile (L.5)*
7, 29	1801	Chateaubriand	*Atala*
8, 30	1844	Chateaubriand	*La vie de Rancé*
9, 31	1799	Balzac	*Les Chouans*
10, 32	1847	Balzac	*Le cousin Pons*
11, 33	1831	Sand	*Indiana*
12, 34	1846	Sand	*La mare au diable*
13, 35	1857	Flaubert	*Madame Bovary*
14, 36	1881	Flaubert	*Bouvard et Pécuchet*
15, 37	1883	Maupassant	*Une vie*
16, 38	1888	Maupassant	*Pierre et Jean*
17, 39	1867	Zola	*Thérèse Raquin*
18, 40	1890	Zola	*La bête humaine*
19, 41	1865	Verne	*De la terre à la lune*
20, 42	1910	Verne	*Le secret de Wilhelm Storitz*
21, 43	1913	Proust	*Du côté de chez Swann*
22, 44	1927	Proust	*Le temps retrouvé*

Tableau A.1. *Description de notre corpus de test en langue française*

Id	Auteur	Titre abrégé	Chap.	Id	Auteur	Titre abrégé	Chap.
1A	Hardy	*Jude*	I	**2A**	Butler	*Erewhon revisit.*	XIV
1B	Butler	*Erewhon revisit.*	II	**2B**	Morris	*Dream of JB*	
1C	Morris	*News*	XIII	**2C**	Tressel	*Ragged TP*	
1D	Stevenson	*Catriona*	V	**2D**	Hardy	*Jude*	
1E	Butler	*Erewhon revisit.*	XVIII	**2E**	Stevenson	*Ballantrae*	IV
1F	Stevenson	*Ballantrae*	II	**2F**	Hardy	*Wessex Tales*	
1G	Conrad	*Lord Jim*	XIV	**2G**	Orczy	*Elusive P*	VII
1H	Hardy	*Madding*	III	**2H**	Conrad	*Lord Jim*	XXI
1I	Orczy	*Scarlet P*	I	**2I**	Morris	*News*	VIII
1J	Morris	*Dream of JB*	VII	**2J**	Hardy	*Well beloved*	I
1K	Stevenson	*Catriona*	X	**2K**	Conrad	*Almayer*	VI
1L	Hardy	*Jude*	VII	**2L**	Hardy	*Well beloved*	XII
1M	Orczy	*Scarlet P*	XIV	**2M**	Morris	*News*	XIX
1N	Stevenson	*Ballantrae*	V	**2N**	Conrad	*Almayer*	XI
1O	Conrad	*Lord Jim*	VII	**2O**	Forster	*Room with view*	I
1P	Chesterton	*Man who was*	I	**2P**	Forster	*Room with view*	IV
1Q	Butler	*Erewhon revisit.*	VII	**2Q**	Conrad	*Almayer*	IX
1R	Chesterton	*Man who was*	VII	**2R**	Stevenson	*Catriona*	XVI
1S	Morris	*News*	I	**2S**	Hardy	*Madding*	X
1T	Conrad	*Almayer*	II	**2T**	Hardy	*Well beloved*	VI
1U	Orczy	*Elusive P*	I	**2U**	Chesterton	*Man who was*	III
1V	Conrad	*Lord Jim*	II	**2V**	Forster	*Room with view*	VIII
1W	Orczy	*Elusive P*	XIV	**2W**	Stevenson	*Catriona*	I
1X	Hardy	*Wessex Tales*		**2X**	Hardy	*Well beloved*	VIII
1Y	Tressel	*Ragged TP*		**2Y**	Orczy	*Scarlet P*	VII
1Z	Tressel	*Ragged TP*		**2Z**	Hardy	*Madding*	XVIII

Tableau A.2. *Description de notre corpus de test en langue anglaise*