Evaluation of text representation schemes and distance measures for authorship linking

Mirco Kocher and Jacques Savoy
University of Neuchatel, Switzerland

Abstract

Based on n text excerpts, the authorship linking task is to determine a way to link pairs of documents written by the same person together. This problem is closely related to authorship attribution questions, and its solution can be used in the author clustering task. However, no training information is provided and the solution must be unsupervised. To achieve this, various text representation strategies can be applied, such as characters, punctuation symbols, or letter n-grams as well as words, lemmas, Part-Of-Speech (POS) tags, and sequences of them. To estimate the stylistic distance (or similarity) between two text excerpts, different measures have been suggested based on the L^1 norm (e.g. Manhattan, Tanimoto), the L^2 norm (e.g. Matusita), the inner product (e.g. Cosine), or the entropy paradigm (e.g. Jeffrey divergence). From those possible implementations, it is not clear which text representation and distance functions produce the best performance, and this study provides an answer to this question. Three corpora, extracted from French and English literature, have been evaluated using standard methodology. Moreover, we suggest an additional performance measure called high precision (HPrec) capable of judging the quality of a ranked list of links to provide only correct answers. No systematic difference can be found between token- or lemma-based text representations. Simple POS tags do not provide an effective solution but short sequences of them form a good text representation. Letter n-grams (with $n = 4$–6) give high HPrec rates. As distance measures, this study found that the Tanimoto, Matusita, and Clark distance measures perform better than the often-used Cosine function. Finally, applying a pruning procedure (e.g. culling terms appearing once or twice or limiting the vocabulary to the 500 most frequent words) reduces the representation complexity and might even improve the effectiveness of the attribution scheme.

Correspondence:
Jacques Savoy, University of Neuchatel, rue Emile Argand, 112000 Neuchatel, Switzerland.
A preliminary study of this research appears in Joint Conference on Digital Libraries, Toronto, June 2017.
E-mail:
Jacques.Savoy@unine.ch

1 Introduction

Due to the presence of numerous pseudonymous posts, chats, threatening e-mails, and anonymous messages on the Web, the authorship attribution domain has stimulated an increasing level of interest (Olsson, 2008). To accurately determine the true author of a text, various approaches have been proposed and evaluated (Juola, 2006; Stamatatos, 2009). This field can mainly be subdivided into four distinct questions. First, the closed-class attribution problem assumes that the real author is one of the specified candidates. Second, in the open-set situation, the real author could be one of the proposed authors or another unknown one. Third, the verification question provides a binary response as

doi:10.1093/llc/fqy013

to whether a given author did in fact write a given text (Koppel *et al.*, 2007). Finally, authorship attribution can be limited to determining demographic (e.g., gender, age class, native language) or psychological traits of the author (Argamon *et al.*, 2009; Pennebaker, 2011; Rangel and Rosso, 2016).

In all these cases, the proposed methods assume that a set of documents written by the different possible authors (or categories of authors, such as men and women) is available. The current study focuses on a radically different perspective where the presence of such labeled data is not provided. The targeted question, called authorship linking, is defined as follows. Having a set of n documents (or text excerpts) written by several distinct authors, determine the pairs of documents written by the same person. In the related task called author clustering, the objective is similar and usually builds upon this task. In the clustering case, the number k of distinct authors must be determined to form k distinct author clusters based on a preset threshold for the ranked list of authorship links. As possible applications for both problems, a set of proclamations written by different terrorist groups can be regrouped, a collection of reviews written by the same author can be gathered (Almishari and Tsudik, 2012), or a set of poems (or excerpts of literary works) can be assembled. To solve this task, an unsupervised approach must be designed and evaluated.

In this context, the first challenge is to represent the text in an effective way, and it is not clear which text representation proposes the highest linking effectiveness. Past studies indicate that very frequent word-types or functional words can closely reflect the personal style of each writer, while other researchers prefer to examine the entire vocabulary. As a third view, other experiments propose to ignore terms having a low occurrence frequency (e.g. appearing once or twice). Finally, the Part-Of-Speech (POS) distribution can be used to reflect the stylistic characteristics of the different authors. Further concerns can be found when choosing the most appropriate distance (or similarity) measure between two text extracts. For example, in the information retrieval (Manning *et al.*, 2008) or deep learning community (Goodfellow *et al.*, 2016), Cosine corresponds to the most popular measure. However, many other distance measures (Duda *et al.*, 2001) do exist and their success in the authorship linking problem is largely unknown.

To provide an answer to these questions, and to determine the most effective text representation, the rest of this article is organized as follows. The next section gives an overview of recent research in authorship attribution. The third section describes the three test collections used in our experiments, while the fourth exposes the evaluation methodology. The fifth section evaluates various word-based text representations and distance measures for the authorship linking task. In the sixth section, an evaluation of different word-based representations is described, while in the seventh, various POS text representations are presented and evaluated. Different letter n-gram text surrogates are built and evaluated in the eighth section, while some efficiency questions and their effectiveness as a tool are described in the ninth section. The main findings of this study are outlined in the conclusion.

2 Related Work

To achieve an effective solution for the authorship linking task, two main challenges must be solved. First, a text representation must be defined reflecting the stylistic aspects of the author, without specifically taking account of the text genre or the topics. Second, an effective distance measure between two text representations must be determined. Such a function must return a low value when the two documents are written by the same author and a higher one otherwise. Instead of applying a distance measure, a similarity measure can be used to state that two texts were written by the same person when the similarity value is high enough.

The choice of the text representation and the distance measure are related to classical challenges in authorship attribution, but we must solve them in an unsupervised perspective. In the current context, training data are not available, and thus, author profiles cannot be derived from a sample of documents for which the authorship is known.

To capture the stylistic aspects of an author, a first set of methods suggests defining an invariant stylistic measure (Holmes, 1998) reflecting the particular style of a given author and varying from one person to another. As possible solutions, different lexical richness measures or word distribution indicators have been proposed such as Yule's K measure, statistics related to the type-token ratio (e.g. Herdan's C, Guiraud's R or Honoré's H), the proportion of word-types occurring once or twice (e.g. Sichel's S) as well as the average word length, or the mean sentence length. None of these measures has proven very satisfactory owing, in part, to word distributions ruled by a large number of very low probability elements (Large Number of Rare Events) (Baayen, 2008).

As a second framework, a multivariate method can be applied to project each document representation into a reduced space under the assumption that texts written by the same author should appear close together. Some of the main approaches applicable here are principal component analysis (Burrows, 1992; Binongo and Smith, 1999; Craig and Kinney, 2009), hierarchical clustering (Labbé and Labbé, 2006; Cortelazzo *et al.*, 2016; Tuzzi and Cortelazzo, 2018), or discriminant analysis (Ledger and Merriam, 1994; Jockers and Witten, 2010). As stylistic features, these approaches tend to employ the top 50–200 most frequent word (MFW)-types, as well as some POS information.

As a third useful paradigm, and based on various word selection schemes, different distance-based measures have been suggested. As well-known strategies, one can mention Burrows' Delta (2002) using the top m MFW (with $m = 40$–$1{,}000$), the Kullback–Leibler divergence (Zhao and Zobel, 2007) using a predefined set of 363 English words, or Labbé's method (2007) using the entire vocabulary and opting for a variant of the Manhattan distance from the L^1 norm distance measure.

Such distance measures can also be applied with less frequent words. For example, Burrows (2007) proposed two distinct but complementary tests. The first one is based on words used regularly by one author but sporadically by the others, while the second is grounded on words used infrequently by one author and ignored by the others. The remaining question is to know whether restricting the representation to the top MFW is effective or using the entire range of vocabulary would produce better results for the authorship linking problem. This question will be discussed later in this article.

If words seem a natural way to generate a text surrogate, other studies have suggested using the letter occurrence frequencies (Kjell, 1994; Ledger and Merriam, 1994) or the distribution of short sequences of letters (character n-grams) (Juola, 2006). As demonstrated by Kešelj *et al.* (2003) such a representation can produce much better results. This approach can be justified, for example, by considering that an author employing the continuous present tense more frequently can be detected by a high frequency of the tri-gram 'ing' and verbal forms related to the verb 'to be' (e.g. 'am', 'is', 'are'). As another example, one can identify more adverbial forms with a word ending in 'ly'. However, it is not clear which n value for the character n-gram is needed to achieve the highest performance level, and this value may depend on the collection, language, as well as other factors (e.g. text genre, OCR text) (McNamee and Mayfield, 2004).

Finally, the fingerprint of an author can be identified by the POS distribution. For example, one writer prefers using noun phrases more frequently than verb phrases implying more nouns and adjectives. For example, when comparing Presidents Kennedy's and Obama's speeches, one can clearly see this difference, with Obama adopting more verbal constructions, meaning a style oriented toward action ('yes, we can') (Savoy, 2017). Such text representations do not usually produce very high performance levels, but instead of considering only the distribution of single POS tags, short sequences of POS tags can be a more effective way of detecting some discriminative stylistic aspects of different authors.

3 Test Collections and Evaluation Methodology

As test collections for evaluating authorship linking algorithms, the PAN CLEF evaluation campaigns (Stamatatos *et al.*, 2016) have generated some

Table 1 Selected statistics about the test collections

Name	Language	# Texts	# Authors	Mean length	# Links
Oxquarry	EN	52	9	10,377	160
Brunet	FR	44	11	8,231	66
St Jean	FR	100	18	9,410	464

corpora written in the English, Dutch, and Greek languages. However, only the training texts are currently available, not the full collection. Besides, all those texts are rather short (e.g. from 126 to 1,086 words on average in each text) corresponding to newspaper articles or online reviews.

To gain a better understanding of the advantages and drawbacks of various approaches, a test collection containing longer texts is required. To this end, the Oxquarry corpus has been selected (Labbé, 2007). This corpus regroups fifty-two excerpts from novels written by nine distinct authors (e.g., eight excerpts written by Conrad, seven by Stevenson, six each by Morris and Orczy, etc.). A condition when generating this corpus was that each author should provide at least two texts. The mean size per document (in number of word-tokens) is 10,377. Similarly, the French corpus (Labbé and Labbé, 2006), called Brunet, contains forty-four texts of novels written by eleven different well-known writers (e.g., Marivaux, Voltaire, Sand, Balzac, Zola, Proust, etc.). In this corpus, each author provides exactly four text passages taken from two of their novels. Table 1 provides some statistics about these corpora, and a more complete description can be found in the Appendix.

As a new test collection, the St Jean (Series A) corpus will be used. The entire corpus (Series A + Series B) will contain 200 text excerpts, but only the first part is used in our experiments. Like the Brunet corpus, it contains passages of novels written in French and published during the 19th century. In this corpus, one can find thirteen excerpts from novels written by Balzac, eleven by Flaubert, ten by Maupassant and Zola, and six by Dumas, Sand, Stendhal, and V. Hugo. As shown in Table 1, this last corpus contains more authors and documents than any previous test collections. Moreover, to select each text excerpt, the author must be identified without any doubt, and the text must not contain any modifications or alterations. For the St Jean corpus, there are an abundance of indices (such as correspondences, notebooks, drafts, proofs) which make it possible to affirm that the given author is known. As a counterexample, one can mention the difficulty encountered with several of Shakespeare's works (Ledger and Merriam, 1994; Michell, 1996; Craig and Kinney, 2009; Tassinari, 2009). For some works, the original text may have been modified, such as *Le Secret de Wilhelm Storitz* published in 1910 after the death of the author (Jules Verne in 1898), in a version modified by his son.

In the last column of Table 1, the number of correct links is indicated. In this context, a link establishes a relationship between two texts written by the same author. For example, with the Brunet corpus, each author provided four texts. To regroup those four texts into a cluster, we can create $(4 \times 3)/2 = 6$ links. Having eleven authors, the number of correct links to resolve this problem is $6 \times 11 = 66$ links.

4 Evaluation Methodology

As proposed in the PAN CLEF campaigns, an authorship linking algorithm is evaluated with the AP (average precision), a measure well-known in different NLP domains (Manning *et al.*, 2008). The usual output is a ranked list (denoted L) of links between two texts. Each link indicates that the same author wrote the two texts. Preferably, each link also contains a numerical value indicating a degree of belief (or a probability) that the pair of texts was written by the same author. With a test collection, the entire set of true links (denoted R) is known. A passage of such a ranked list is depicted in Table 2. For example, the first row indicates that the system correctly establishes a link between Texts #3 and #48 (both written by Stevenson, author name added with a posteriori knowledge) with a distance of 0.431 (computed by the Manhattan function).

Based on this notation, one can verify whether the link at the ith position (denoted l_i) belongs to the set R. If this is the case, the link is relevant, otherwise it is not (see Equation (1)). Based on

Table 2 Excerpt of an output based on the Oxquarry corpus, Manhattan distance, token-based representation

Rank	Distance	ID 1	Author	ID 2	Author
1	0.431	3	Stevenson	48	Stevenson
2	0.455	5	Stevenson	30	Stevenson
3	0.458	10	Stevenson	48	Stevenson
4	0.470	13	Stevenson	30	Stevenson
5	0.473	3	Stevenson	10	Stevenson
6	0.479	18	Morris	38	Morris
7	0.493	2	Morris	34	Morris
8	0.497	12	Orczy	50	Orczy
9	0.502	4	Butler	16	Butler
10	0.503	34	Morris	38	Morris
...	...	...	...	...	...
58	0.621	16	Butler	29	Hardy

this indicator function, one can define the accuracy up to a fixed rank. Equation (2) defines this performance value, and normally, the performance is provided at rank 10 (denoted Prec@10) or 20 (Prec@20). These two limits are used frequently in the information retrieval domain because they correspond to the first two pages of results returned by a commercial search engine. As one can see in Table 2, the Prec@10 = 1.0; all links up to the 10th rank are correct.

$$\text{relevant}(i) = \begin{cases} 1, \textit{ if } l_i \in R \\ 0, \textit{ otherwise} \end{cases} \quad (1)$$

$$\text{precision}(k) = \sum_{j=1}^{k} \text{relevant}(j) \Big/ k. \quad (2)$$

Another interesting limit is defined by $|R|$, the number of true links in the test collection. This limit varies from one test collection to another, and it is denoted as R-precision (or RPrec). These values are indicated in the last column of Table 1 for our three corpora. One can compute this value by using Equation (2), with $k = |R|$.

For all these measurements, the best value is 1.0, which is achieved when all links are relevant, while the lowest value is 0.0 when no relevant link is found. Such a performance measure provides a direct and simple interpretation. For example, when the accuracy after ten links is 0.8, the final user knows that in the ten first results, 80% are correct (or eight of ten links). As a main drawback of this measure one can mention that the rank is not considered. In this example, the two incorrect answers can appear in the first two positions or in the last two. In both cases, the performance measure Prec@10 is the same and equal to 0.8, although users would certainly prefer having the incorrect results in the bottom part of the ranked list instead of in the top positions.

The definition of AP given by Equation (3) provides a solution to this issue (Manning *et al.*, 2008). With this measure, the ranks are considered. Suppose that the output list computed by System A and B contains four links. With System A, all links are relevant, and therefore, the AP is 1.0. With System B, only the link indicated in the first position is incorrect. Therefore, the AP is $(0.0 + 0.5 + 0.666 + 0.75)/4 = 0.479$ indicating a relative change of more than 100% between the two rankings.

$$\text{AP} = \left(\sum_{j=1}^{|L|} \text{precision}(j) \times \text{relevant}(j)\right) \Big/ |L|. \quad (3)$$

With the AP, a simple interpretation is not possible. Even if this measure takes account of the ranks, it is sensitive to the first rank(s) as shown in our example. On the other hand, AP does not punish verbosity, i.e. every true link counts even when appearing near the end of the ranked list. Therefore, by providing all possible authorship links, one can attempt to maximize AP, without penalizing the Prec@10.

Overall, the AP and Prec@10 (Prec@20 and RPrec as well) are useful for comparing two (or more) linking strategies. However, in some cases, it is important to return only good results and to specify 'I don't know' when a link between two texts is not fully ascertained. Returning an answer that appears to be wrong creates a lack of trust in the system for the final user leading to a lack of confidence, or engendering a perception that the computer is making mistakes. It is known in the PR domain (Public Relations) that a happy customer will talk to only four to six friends, but a dissatisfied

user will tell nine to fifteen people about their bad experience (Blackshaw, 2008). This phenomenon is relatively unknown in the academic world where the traditional performance measures tend to underestimate the real 'cost' of incorrect classifications. As a counterexample, one can cite the robust track at TREC (Voorhees and Harman, 2005) in which the focus is to penalize more severely the retrieval of irrelevant items from a search engine.

To measure the capability of a system to return only good results (or links in our context), one can measure its high precision (denoted HPrec) by indicating the Rank-1 of the first incorrect answer appearing on the top of the returned list. For example, HPrec = 57 indicates that the first fifty-seven results are correct before the first incorrect answer appears at Rank 58, as it is the case in our example in Table 2.

5 Text Representations and Distance Measures

To solve the authorship linking problem, each text (or excerpt) must be represented in such a way as to closely reflect its stylistic aspects instead of the topics. In this perspective, language style is present as pervasive and frequent forms used by an author for mainly aesthetical value (Love, 2002; Biber and Conrad, 2009). Previous studies have found that the top m MFW (with $m = 50$–500) tends to produce a high level of effectiveness (Burrows, 2002; Savoy, 2015). This set may or may not include punctuation symbols. Moreover, the distinction between uppercase and lowercase letters is ignored, meaning all uppercase letters are transformed into their lowercase equivalent.

As a possible variant, one can consider only functional words, namely, determiners, prepositions, pronouns, conjunctions, and modal verbs (or all closed POS categories). Moreover, to define those frequent words, a stemmer can be applied to remove inflectional suffixes (e.g. related to a variation in number, gender, or grammatical case). For the English language, the S-stemmer (Harman, 1991) applies three ordered rules to replace the plural form of a word with the corresponding singular form (e.g. the last rule is to remove the ending '-s' unless the word ends in '-ss' or '-us').

Instead of restricting the vocabulary to very frequent word-types, Labbé (2007) suggests considering the entire vocabulary. This solution is also adopted by Burrows (2007) who proposes to subdivide the vocabulary into three strata based on the term occurrence frequency.

In addition, an effective text representation can be generated in relation to the letter distribution or letter n-gram (Kešelj *et al.*, 2003). Typical values of n vary from 1 to 5, but higher values can also be considered (McNamee and Mayfield, 2004). Moreover, the POS distribution or a short sequence of such POS tags will be analyzed as other possible stylistic representations.

On the other hand, considering more words or character n-grams increases the complexity of the system and requires more processing time. Therefore, the words (or n-gram of characters) appearing only once (*hapax legomenon*) or twice (*dis legomenon*) can be ignored. This filtering decision can be justified to prevent overfitting to single occurrences. Moreover, due to the Zipf distribution of term occurrence frequencies, removing words appearing once or twice tends to reduce the vocabulary size by half.

The numerous distance measures can be regrouped under different families (Duda *et al.*, 2001; Manning *et al.*, 2008) where the most frequent one is the L^p family (or L^p norm). In this paradigm, the value of the parameter p determines different groups. To define the distance measure, uppercase letters will denote vectors (or points), while lowercase letters with a subscript indicate the values inside a vector. Thus, A or B specify vectors, while a_i indicates the element in the ith position of vector A, and m is the length of the vector.

To limit the investigations on the distance functions, a reduced set of functions has been selected due to their usefulness in a related task (Kocher and Savoy, 2017). First, with fixing $p = 1$, the Manhattan distance is obtained as defined in Equation (4). The underlying assumption is that the distance must be computed in proportion to

the sum of the absolute differences for all dimensions. The distance value can be broken down into contributions made by each dimension (or stylistic feature).

$$dist_{Manhattan}(A, B) = \sum_{i=1}^{m} |a_i - b_i|. \quad (4)$$

Based on the L^1 norm (absolute difference), several variants of this distance measure have been proposed, such as the Tanimoto formula depicted in Equation (5). The value returned by the Manhattan distance is not normalized, and it is sometimes difficult to figure out when a given distance is small or large. With the Tanimoto distance, a normalization factor is used corresponding to the sum of the maximum values of the coefficients.

$$dist_{\text{Tanimoto}}(A, B) = \sum_{i=1}^{m} |a_i - b_i| \Big/ \sum_{i=1}^{m} max(a_i, b_i). \quad (5)$$

Changing the value of p to 2, the Euclidean (L^2 norm) distance is obtained and represents a straight line between two points. This approach usually does not perform well, and thus variants, of the Euclidian distance have been suggested such as the Matusita formulation shown in Equation (6) or the Clark distance given by Equation (7).

$$dist_{Matusita}(A, B) = \sqrt{\sum_{i=1}^{m} \left(\sqrt{a_i} - \sqrt{b_i}\right)^2}. \quad (6)$$

$$dist_{Clark}(A, B) = \sqrt{\sum_{i=1}^{m} \left(\frac{|a_i - b_i|}{a_i + b_i}\right)^2}. \quad (7)$$

As another well-known family, different variants based on the inner product (or dot product) have been suggested. The main drawback of the inner product is the absence of normalization. It is not clear when a distance value must be interpreted as large or small. Therefore, different variants have been proposed, and the most popular is certainly the Cosine similarity (Equation (8)) which can be transformed into a distance value between 0 and 1 (Equation (9)) (Manning *et al.*, 2008).

$$sim_{Cosine}(A, B) = \sum_{i=1}^{m} a_i b_i \Big/ \sqrt{\sum_{i=1}^{m} a_i^2} \sqrt{\sum_{i=1}^{m} b_i^2}. \quad (8)$$

$$dist_{Cosine}(A, B) = \cos^{-1}(sim_{Cosine}(A, B)) \Big/ \pi. \quad (9)$$

Shannon's concept of entropy (Manning *et al.*, 2008) is also a main source of a family of distance measures. The Jeffrey divergence (denoted JDivergence) computes the difference between two probability distributions (see Equation (10)). In this case, all values a_i of each vector must be non-negative, and they must sum up to 1. Moreover, the basis of the logarithm is fixed to two in Shannon's entropy measure. However, in the author profiling context, or when only the ranking of the different categories is relevant, changing the basis of the logarithm does not affect the ordering of the answers. As for other distance measures, a larger distance value indicates a larger difference between the writing style of the two authors (or points).

$$dist_{JDivergence}(A, B) = \sum_{i=1}^{m} (a_i - b_i) \log\left({}^{a_i}/_{b_i}\right). \quad (10)$$

Finally, we must recall that a distance measure must respect three properties, namely, the identity, the symmetry, and the triangle inequality. On this set, one can add the characteristic that a distance must be always positive or null. As shown in (Kocher and Savoy, 2017), the Manhattan, Tanimoto, Matusita, and Clark functions respect these criteria. On the other hand, the Cosine can return null even if the vectors are different, and the JDivergence function does not respect the triangle inequality.

6 Evaluation of Word-Based Text Representations

To compare different text representations, our experiments start by using all word-types with the six

Table 3 Evaluation over two word-based text representations and six distance measures

Distance and text representation	Oxquarry			Brunet			St Jean		
	AP	RPrec	HPrec	AP	RPrec	HPrec	AP	RPrec	HPrec
T-Manhattan	0.588	0.525	57	0.648	0.561	**26**	**0.666**	**0.585**	64
T-Tanimoto	0.620	0.556	59	0.653	0.561	**26**	0.663	0.573	**65**
T-Matusita	0.561	0.519	51	0.569	0.485	16	0.504	0.470	44
T-Clark	**0.731**	**0.650**	**70**	0.603	0.515	*23*	0.533	0.512	47
T-Cosine	0.511	0.500	28	0.590	0.561	15	0.648	0.570	52
T-JDivergence	0.595	0.544	67	**0.659**	**0.636**	19	0.608	0.542	56
L-Manhattan	0.643	0.556	63	0.662	**0.606**	**25**	**0.652**	**0.576**	**68**
L-Tanimoto	0.685	0.563	66	**0.675**	**0.606**	**25**	0.651	0.573	**68**
L-Matusita	0.611	0.538	53	0.565	0.530	15	0.489	0.461	53
L-Clark	**0.737**	**0.644**	68	0.558	0.500	*15*	0.452	0.421	48
L-Cosine	0.553	0.500	38	0.568	0.545	15	0.589	0.542	30
L-JDivergence	0.613	0.538	**71**	0.656	0.636	20	0.603	0.536	62

Note: Best performances in bold.

distance measures described previously. In the top part of Table 3, the word-tokens have been used for the three corpora (e.g. the label 'T-Manhattan' indicates a text surrogate generated with word-tokens and a distance computed with the Manhattan measure).

In the bottom part, the text representations are built based on the lemmas (or the dictionary entries, denoted 'L-Manhattan'). As a general rule, in the English language, the difference between these two forms can be small (e.g. houses versus house, running versus run). For the French language, however, one can expect a larger difference due to a richer inflectional morphology (e.g. aimerais versus aimer (to love), blanches versus blanc (white)).

As performance measures, the AP and RPrec have been reported, with the higher the value, the better the effectiveness. To reflect the quality of a text representation and distance measure to return only good answers, the HPrec value is also reported.

From data shown in Table 3 under the English corpus (Oxquarry), it can be seen that the AP values are on average 8% higher for the lemma-based representation than for the tokens. The highest values (always depicted in bold) are however similar in both cases. The situation is similar for the French Brunet collection, with a mean AP difference of 3.3% in favor of the lemmas. The last corpus (St Jean) indicates a better AP performance when applying tokens (on average, 5.3%). When considering the HPrec values, usually the lemma-based representations tend to produce better answers, but for the Brunet corpus, the value 23 achieved with the Clark function using tokens is clearly an exception, compared to 15 obtained with lemmas (both values shown in italics in Table 3).

When analyzing the variations related to the distance measures, one can see that none of them performs the best in all cases. For the Oxquarry corpus, the Clark measure (L^2 family) produces the best effectiveness, while for the St Jean collection, the highest level of accuracy is obtained with the Manhattan distance (L^1 family). For the Brunet corpus, the Jeffrey divergence offers the best precision values for one text surrogate (token-based), while Tanimoto is a better choice for the second (lemma-based). However, in all these experiments, the Cosine distance never produces the best answer. In mean, and compared to the best AP solution, the performance of the Cosine function is 9.5% lower with the token-based representation and 16% worse with the lemmas. Overall, and considering the two text representations, the Matusita distance offers the lowest AP values. Finally, one can see that the results achieved by Manhattan and Tanimoto distance are correlated.

When analyzing the ranked lists for the English corpus, we found that correctly linked texts appearing on the top are novels written by Stevenson (*Catriona*, *The Master of Ballantrae*), Morris (*News from Nowhere*), or Hardy (*Well-beloved, Jude the*

Table 4 Evaluation over three text representations and six distance measures

Distance and text representation	Oxquarry			Brunet			St Jean		
	AP	RPrec	HPrec	AP	RPrec	HPrec	AP	RPrec	HPrec
P-Manhattan	0.494	0.425	26	0.482	0.470	10	0.563	**0.555**	**53**
P-Tanimoto	0.505	0.444	24	0.494	0.485	10	0.508	0.494	26
P-Matusita	**0.517**	**0.463**	**36**	**0.497**	0.485	**12**	0.448	0.448	20
P-Clark	0.419	0.431	4	0.301	0.318	3	0.456	0.455	24
P-Cosine	0.465	0.406	25	0.474	0.485	11	0.448	0.448	20
P-JDivergence	0.513	**0.463**	**36**	0.495	**0.500**	**12**	**0.555**	0.491	41
G-Manhattan	0.470	0.406	29	0.435	**0.424**	5	**0.406**	**0.418**	**14**
G-Tanimoto	0.471	0.406	**30**	0.438	**0.424**	5	0.360	0.367	6
G-Matusita	**0.489**	**0.456**	27	**0.462**	0.439	11	**0.406**	0.415	13
G-Clark	0.410	0.431	4	0.216	0.182	4	0.248	0.291	9
G-Cosine	0.442	0.388	23	0.440	**0.424**	7	0.360	0.367	8
G-JDivergence	0.482	**0.456**	**30**	**0.462**	**0.424**	**12**	0.401	0.406	**14**
N-Manhattan	0.308	0.306	**6**	0.230	0.318	0	0.237	0.285	0
N-Tanimoto	0.316	0.338	**6**	0.233	0.333	0	0.238	0.285	0
N-Matusita	0.324	**0.344**	2	0.281	0.364	0	0.239	0.306	0
N-Clark	0.189	0.194	0	0.120	0.136	0	0.121	0.139	0
N-Cosine	0.314	0.319	5	0.230	0.273	0	0.226	0.270	0
N-JDivergence	**0.329**	**0.344**	4	**0.286**	**0.379**	0	**0.244**	**0.312**	**5**

Note: Best performances in bold.

Obscure). Determining the specific functional terms of those authors (Savoy, 2016), we found that Stevenson uses more frequently the words *I, my, me, ye, myself*, and the comma. With Morris, the most specific words are *thou, shall, we, three, and, our*, and the comma, while Hardy's characteristic terms are *her, she, had, till, being*, and the quote. The other writers tend to share more specific terms in common such as the full stop between Conrad, Orczy, and Butler, the determiner *the* appearing with both Chesterton and Conrad, or the pronoun *it* belonging to the favorite terms of Tressel and Chesterton.

For the French corpus St Jean, correct links appearing in the top of the ranked lists connect novels written by Zola (*L'assomoir, La Fortune des Rougon*), Flaubert (*Mme Bovary, Bouvard et Péruchet*), or Maupassant (*Mont-Oriol, Bel-Ami*). In this case, the specific terms associated with Zola are *ça*, *avait*, *elle*, *aurait*, and *était* (it, had, she, he, would have, was), while Flaubert uses more frequently *des* (of), *ils* (they), *les* (the), the exclamation mark, and the semicolon. Finally, Maupassant can be distinguished from the others with his use of the following terms: the colon, *il*, *puis*, *elle*, and *et* (he, then, she, and).

7 Evaluating POS-Based Text Representations

As another text representation, one can consider the POS distribution. A closer look reveals that the information returned from the tagger (Stanford POS tagger for the English language (Toutanova *et al.*, 2003) and Labbé's POS tagger for French (Labbé, 2007)) contains not only the POS category (e.g. verb, noun, pronoun) but also some morphological information (e.g. personal pronoun, third-person, plural). The punctuation symbols are also included as additional tags. In Table 4, the performance with text surrogate generated with those POS tags are presented in the top part (e.g. 'P-Manhattan'), while in the middle part ('G-Manhattan') only the grammatical categories (e.g. verb, pronoun, adverb) and punctuations have been used to build text representations.

Taking account of the POS tags (with the associated morphological information) produces a better text representation than would be the case if only the grammatical categories were used. Comparing the two models, the AP measure reflects, on average, a 5% difference with the Oxquarry

Table 5 Evaluation of short sequences of POS tags

Distance and text representation	Oxquarry			Brunet			St Jean		
	AP	RPrec	HPrec	AP	RPrec	HPrec	AP	RPrec	HPrec
P-Tanimoto	0.505	0.444	24	0.494	0.485	10	0.508	0.494	26
P2-Tanimoto	0.554	0.475	37	0.612	0.576	14	0.671	0.618	56
P3-Tanimoto	0.555	0.500	30	0.663	0.621	18	0.712	0.645	59
P4-Tanimoto	0.524	0.450	39	0.616	0.545	9	0.619	0.567	36
P-Matusita	0.517	0.463	36	0.497	0.485	12	0.448	0.448	20
P2-Matusita	0.553	0.481	55	0.631	0.561	22	0.691	0.630	80
P3-Matusita	0.529	0.481	38	0.661	0.621	18	0.699	0.609	68
P4-Matusita	0.490	0.419	37	0.561	0.515	13	0.531	0.482	37

corpus, 11.7% with the Brunet, and 26.6% with the St Jean corpus. Based on the HPrec values, this relative change is higher, up to 62.3% with the St Jean corpus. Finally, the POS text representation produces lower effectiveness levels than either the lemma- or the token-based models (see Table 3). Compared to the average token-based performance shown in Table 3, the mean relative change in AP is 18% for the Oxquarry, 26.2% with the Brunet, and 17.3% with the St Jean collection, and always in favor of the token-based models.

Concerning the distance measures, Table 4 indicates that the Matusita distance usually produces the best AP results with the three corpora with the single exception being the performance of the Manhattan function for the St Jean collection (0.563 versus 0.448).

Finally, in the bottom part of Table 4, the text representation is based on the distribution of the token length (e.g. 'N-Manhattan'). This surrogate is not limited to a single value, i.e. the mean token size, but presents all possible token lengths with their occurrence frequency. The performance reported in Table 4 clearly indicates that such an approach is not a pertinent representation. Moreover, the HPrec value is often 0, indicating that even the first link is wrong.

As the number of distinct POS tags (42 for the English language, 29 for the French) is rather limited compared to the vocabulary size, a text representation can be built using short sequences of such tags. Considering only two distance measures, Table 5 reports the evaluations of these text surrogates generated form sequences of two to four POS tags. Compared to the baselines ('P-Tanimoto' or 'P-Matusita' repeated from Table 4) corresponding to single POS tags, sequences of two or three tags improve the result significantly. For example, with the Brunet corpus and using the Matusita function, the AP increases from 0.497 to 0.661 (+33%). The best performance depicted in Table 5 is usually below those achieved based on word-based representation (see Table 3). In some cases, however, the difference is rather small, i.e. with the Brunet corpus and Tanimoto function, 0.653 for token-based versus 0.663 for sequences of three POS tags, corresponding to a relative change of +1.5%.

8 Letter *N*-Gram Evaluation

As another text representation, one can select short sequences of letters, denoted *n*-grams, extracted from the text. In this generation process, a few variants are possible. Each word boundary may stop the creation of the *n*-grams. The distinction between uppercase and lowercase could be preserved, and the adjacent *n*-grams could overlap. In our experiments, the word boundary does not stop the *n*-grams generation. All punctuation symbols are replaced by a space, and the uppercase letters are replaced by their corresponding lowercase. As an example, based on the sentence 'Paul's book is red.', the following overlapping 4-grams are extracted: '_pau', 'paul', 'aul_', 'ul_s', 'l_s_', '_s_b', ..., 's_red', '_red', 'red_', where '_' indicates a space.

As possible values for *n*, one can consider any value between 1 and 10. However, after $n = 5$, 6,

Table 6 Evaluation over six different *n*-gram text representations

n-gram length	Oxquarry			Brunet			St Jean		
	AP	RPrec	HPrec	AP	RPrec	HPrec	AP	RPrec	HPrec
Token-Tanimoto	0.620	0.556	59	0.653	0.561	**26**	**0.663**	**0.573**	65
1/2-Tanimoto	0.654	0.613	52	0.614	0.561	23	0.549	0.491	52
3 Tanimoto	0.817	0.738	61	0.641	0.576	20	0.641	0.576	20
4 Tanimoto	0.854	0.788	82	0.655	0.606	20	0.609	0.545	65
5 Tanimoto	0.872	0.806	99	0.670	0.606	21	0.622	0.548	**73**
6 Tanimoto	0.883	0.813	**101**	0.676	0.606	16	0.631	0.545	71
7 Tanimoto	**0.888**	**0.825**	99	**0.680**	**0.621**	20	0.624	0.539	52
Token-Matusita	0.561	0.519	51	0.569	0.485	16	0.504	0.470	44
1/2-Matusita	0.587	0.538	55	0.627	0.591	18	0.532	0.479	55
3 Matusita	0.828	0.731	65	0.638	0.606	18	**0.638**	**0.606**	18
4 Matusita	0.888	**0.825**	94	0.658	**0.621**	20	0.534	0.448	**68**
5 Matusita	**0.892**	**0.825**	**96**	**0.667**	0.606	**22**	0.539	0.476	58
6 Matusita	0.883	0.819	88	0.664	0.576	21	0.556	0.494	57
7 Matusita	0.876	0.775	88	0.660	0.576	20	0.556	0.500	38

Note: Best performances in bold.

or 7, the number of generated *n*-grams becomes huge, and most them have a very low occurrence frequency. Table 6 reports the performance obtained with the Tanimoto (L^1 norm) and Matusita (L^2) distance for $n = 3$–7. Before that, the first line for each measure indicates the performance achieved with a token-based representation, and then the label '1/2' indicates a combined text representation based on both uni- and bigrams as suggested by Kjell (1994) and Goldberg (2017).

As depicted in Table 6, the best *n* value depends on the collection, but values larger than or equal to 5 tend to produce the highest performance (e.g. for the Oxquarry corpus, $n = 7$ with the Tanimoto distance, $n = 5$ with Matusita). Comparing across corpora or distance measures, slightly modifying the value *n* tends to produce similar results, e.g. Oxquarry with Tanimoto function gives an AP of 0.872 with $n = 5$ versus 0.888 with $n = 7$ (+1.8%).

As depicted in Table 6, the best value of *n* depends on the collection, but the difference between the three corpora or distance functions is just ± 1 (e.g. for the Oxquarry corpus, $n = 6$ with the Tanimoto distance, $n = 5$ with the Matusita, $n = 6$ for Brunet corpus). Compared to the token-based representation, the *n*-gram approach tends to produce a higher effectiveness. With the English corpus, the improvement is significant. For example, with the Tanimoto function, the AP increases from 0.620 to 0.888 (+43.2%), and with the Matusita distance, from 0.561 to 0.892 (+59%). With the Brunet corpus and applying the Tanimoto distance, the performance difference is smaller, but still present, e.g. the AP varies from 0.653 to 0.680 (+4.1%), or from 0.569 to 0.667 (+17.2%) with the Matusita function. With the St Jean corpus, the *n*-gram approach improves the performance only with the Matusita measure.

9 Efficiency Improvement

In the previous sections, text representations were constructed considering the entire vocabulary or all possible *n*-grams. Ranking the terms (word-types or *n*-grams) in proportion to their occurrence frequency, a Zipfian distribution can be observed. If the most frequent ones cover a large proportion of all texts, the terms appearing only once or twice tend to correspond to 50% of all word-types, and usually a larger percentage when considering character *n*-grams. Moreover, assigning a text to an author based on a few words occurring only once is an unsafe decision and prone to impostors (a writer can easily pass for another).

To reduce the text representation, various pruning strategies can be applied. To assess their effects, Table 7 reports in the top part the mean number of

Table 7 Statistics of different pruning strategies

Pruning Strategy	Oxquarry		Brunet		St Jean	
	Token	$\|V\|$	Token	$\|V\|$	Token	$\|V\|$
All tokens	11,650	2,169	10,628	2,204	12,331	2,466
$tf > 1$	10,351	871	9,183	759	10,711	845
$tf > 2$	9,688	539	8,550	443	10,005	492
$tf > 3$	9,254	394	8,161	313	9,576	350
$tf > 4$	8,934	314	7,883	244	9,272	274
50 tokens	5,840	48	5,664	50	6,654	50
100 tokens	6,886	96	6,588	99	7,694	94
200 tokens	7,817	194	7,236	193	8,422	194
300 tokens	8,317	282	7,593	274	8,843	286
500 tokens	8,862	431	8,027	408	9,308	430
1,000 tokens	9,560	710	8,562	642	9,913	691
6-grams (all)	52,409	26,003	44,302	21,724	50,800	24,187
$tf > 1$	34,657	8,251	29,868	7,289	34,884	8,271
$tf > 2$	26,621	4,233	22,933	3,822	27,145	4,401
$tf > 3$	21,800	2,626	18,561	2,364	22,242	2,767
$tf > 4$	18,525	1,807	15,554	1,613	18,821	1,912

word-tokens (labeled 'Token') and the mean vocabulary size ($|V|$) per document. The first row indicates the mean values before any pruning procedure ('All tokens'). For the Brunet corpus, the averages are 10,628 tokens per document and 2,204 word-types in each document. In the next four rows, the terms appearing once ('$tf > 1$') to 4 times ('$tf > 4$') in a document representation are eliminated. The representation size decreases slowly, for example in the Oxquarry collection, from 11,650 tokens to 10,351 when ignoring terms appearing once, or to 8,934 when only keeping terms appearing at least 5 times. On the other hand, the mean vocabulary size per document decreases faster. With the Oxquarry corpus, the mean number of distinct terms begins with 2,169 and decreases to 314 when removing all terms having a term frequency smaller than or equal to 4.

In the middle of Table 7, we report the mean number of tokens per document when using only the 50–1,000 MFW-tokens when generating the text surrogates. These word lists were defined in relation to the entire corpus. Reducing the vocabulary to the top 50 MFW, the text representation size is reduced, on average, by 50%, e.g. with the Oxquarry collection, from 11,650 to 5,840 tokens. Looking at the vocabulary, the reduction is more marked. For example, with the Oxquarry collection, the mean vocabulary/document decreases from 2,169 to 48 (−97.8%).

The bottom part of Table 7 shows the statistics when considering letter n-grams with $n = 6$. For the English corpus, the most frequent 6-grams is '_that_', for the Brunet corpus it is '_vous_' (you/plural), and '_elle_' (she/singular) appears the most in the St Jean collection.

The main concern with the n-gram model is the huge number of distinct n-grams that can be generated. With the St Jean corpus, the mean number of terms in a text representation goes from 2,466 word-types to 24,187 6-grams (around 10 times more). Here too, the pruning of terms appearing less than twice is useful to reduce the complexity of the text representation, as for example, with the St Jean corpus, the size decreases from 24,187 to 4,401 6-grams appearing more than twice (−81.8% in relative value), or to 1,912 6-grams occurring at least 5 times (−92.1%).

Pruning text representations by ignoring features with a very low occurrence frequency reduces the complexity of text representation. However, such procedures may hurt the overall success. To verify this aspect, Table 8 reports the three performance measures using two distance

Table 8 Evaluation of different pruning strategies on word-based representation

Pruning strategy	Oxquarry			Brunet			St Jean		
	AP	**RPrec**	**HPrec**	**AP**	**RPrec**	**HPrec**	**AP**	**RPrec**	**HPrec**
Tok-Tanimoto	**0.620**	0.556	59	0.653	0.561	**26**	0.663	0.573	65
tf > 1 Tanimoto	0.620	0.550	63	0.661	0.606	23	0.702	0.627	58
tf > 2 Tanimoto	0.616	0.531	**64**	0.661	0.606	23	0.701	0.621	49
tf > 3 Tanimoto	0.613	0.538	62	0.657	0.636	22	0.542	0.606	53
tf > 4 Tanimoto	0.611	0.525	59	0.657	0.636	22	0.692	0.606	52
50-Tanimoto	0.533	0.533	46	0.637	0.591	19	0.711	0.639	**70**
100-Tanimoto	0.562	0.519	46	0.632	0.591	21	0.718	0.655	65
200-Tanimoto	0.580	0.519	48	0.646	0.576	20	0.725	0.667	63
300-Tanimoto	0.600	0.531	48	0.653	0.591	21	0.736	0.676	65
500-Tanimoto	0.613	**0.613**	50	0.665	0.636	23	**0.751**	**0.679**	63
1,000-Tanimoto	0.628	0.556	61	**0.676**	**0.652**	23	0.750	0.676	60
Tok-Matusita	0.561	0.519	51	0.569	0.485	16	0.504	0.470	44
tf > 1 Matusita	0.575	0.506	55	0.619	0.545	**25**	0.591	0.527	60
tf > 2 Matusita	0.571	0.488	55	0.648	0.591	24	0.604	0.542	51
tf > 3 Matusita	0.575	0.494	55	0.652	0.576	21	0.605	0.542	58
tf > 4 Matusita	0.577	**0.577**	51	0.652	0.606	22	0.595	0.530	58
50-Matusita	0.521	0.475	49	0.627	0.545	23	0.700	0.621	**84**
100-Matusita	0.554	0.500	51	0.608	0.530	17	0.692	0.621	73
200-Matusita	0.573	0.500	58	0.607	0.530	17	0.712	0.652	70
300-Matusita	0.597	0.538	54	0.626	0.545	17	**0.733**	0.667	67
500-Matusita	0.614	0.556	69	0.663	0.652	20	0.751	**0.676**	70
1,000-Matusita	**0.632**	0.563	**74**	**0.674**	**0.667**	20	0.720	0.639	77

functions and word-based text representation. In the row labeled 'Tok-Tanimoto' (and 'Tok-Matusita'), the performance obtained with all tokens are depicted as a baseline.

As a general trend, one can observe that removing very low frequency word-types might even increase the performance. For example, comparing the baseline with the row labeled '*tf* > 2', the AP value increases for the Brunet and St Jean corpora for both distance measures. With the St Jean corpus and the Matusita distance, the performance goes from 0.504 to 0.604 (+19.8%). Conversely, with the Oxquarry and the Tanimoto distance, a slight decrease can be seen (from 0.620 to 0.616, −0.6%). This pruning strategy reduces the vocabulary from slightly more than 2,000 word-types to 443 (Brunet) or 539 (Oxquarry) as shown in Table 7.

As another example, one can analyze the row labeled '500-Matusita' where the 500 MFW are defined in relation to the whole vocabulary. As can be seen in the data depicted in Table 7, such a pruning scheme tends to reduce the mean vocabulary size per document in the range of 408 (−81.5% for Brunet) to 431 (−80.1% for the Oxquarry). The results achieved with this strategy generally indicates an improvement over the baseline performance. Considering the AP values, the increase is around +9.4% (from 0.561 to 0.614) with the Oxquarry corpus using the Matusita distance. After the pruning stage, spurious features and especially words with single occurrences have no longer an influence on the distance calculation and are therefore ignored to create the authorship links.

To obtain an overview of the time required to compute the different text representations, Table 9 reports the elapsed time in seconds (a mean based on four runs with both the Tanimoto and Matusita distance functions). The first row (labeled 'Token') corresponds to a token-based surrogate built with the entire vocabulary, while the second (labeled '1,000 MFW') signals the value when considering only the 1,000 most frequent tokens. The last three rows report the time needed when considering the *n*-gram models with different values for *n*.

When time is a critical factor, adopting a pruning scheme based on the *k* most frequent tokens should

Table 9 Elapsed time in seconds for different text representations

Text representation	Oxquarry	Brunet	St Jean
Token	297	224	1,387
1,000 MFW	49	31	106
3-grams	235	128	669
4-grams	3,381	1,594	9,603
5-grams	16,063	7,336	46,699
6-grams	41,449	19,537	127,868

be viewed as an effective approach. It can be from 7 times (Brunet corpus) to 13 times (St Jean) faster than taking account of the entire vocabulary. Moreover, such an approach is possibly more effective (see Table 8). On the other hand, adopting an effective n-gram model ($n \geq 5$ as depicted in Table 6) requires a larger processing time as indicated in the last rows of Table 9 due to the huge number of generated n-grams.

10 Conclusion

The authorship linking problem raises new challenges, and one of them is the absence of a training phase useful in determining the most effective feature set and distance measures. In this unsupervised context, our study evaluates the effectiveness of six different distance functions using three test collections. Moreover, the main findings are based on two different languages (English and French) with relatively long text excerpts (from 8,231 to 10,377 tokens/document).

None of the selected distance functions performs optimally in all cases. From the L^1 norm, the Tanimoto, strongly correlated to the Manhattan function, usually produces high-performance levels (see Table 3), at least for the two French collections. In the L^2 family, the Matusita function performs well with some text representations (see Tables 4 and 5), while the Clark distance produces better answers with token-based representation (see Table 3). In some cases, the Jeffrey divergence might produce a high performance. However, in all cases, the Cosine distance function, frequently used in various applications (Goldberg, 2017), does not perform very well.

As text representation, the word-tokens or the lemmas (dictionary entries) correspond to well-known approaches. Using lemma-based representation requires that an additional morphological analysis be performed. The results of our experiments indicate that lemmas tend to be more useful (see Table 3). While this conclusion is valid for the English corpus, the two French corpora indicate contradictory findings. With the Brunet collection, lemmas perform better than tokens, but with the St Jean corpus, it is the reverse. The performance differences are however not substantial, i.e. +3.3% with the Brunet corpus and −5.3% in the St Jean collection.

As another text representation that can extract stylistic features, POS tags (grammatical category with morphological information together with the punctuation symbols) can be applied. Compared to the word-based representation (token or lemma), the AP tends to decrease around 20%. Limiting the representation to single grammatical categories (see Table 4) lowers the results significantly more with an average decrease of 25% compared to word-based with the English collection and over 30% with the two French corpora. Thus, single POS tags do not effectively identify stylistic differences between authors. However, a short sequence of two or three POS tags is clearly more effective. Such text representations can even be more effective than a word-based model. For the AP and considering a sequence of two POS tags, the mean improvement is around 8% for the English corpus and 25% for the Brunet collection. Working with sequences of four or more POS tags is not effective in all collections.

As a third paradigm to generate a text representation, character n-grams can be used. As shown in Table 6, the best value for n depends on the collection, but values larger than or equal to 5 tend to produce the best answers. Compared to token-based models, the n-grams may perform significantly better, for example, with the Oxquarry corpus using the Tanimoto distance, the average improvement is 31.6%, and with the Matusita function is 45.4%. For the Brunet collection, this enhancement is smaller, as we can observe in mean +14.4% with the Matusita distance with a preference for the n-gram model but roughly the same precision values using the Tanimoto function.

For efficiency reasons, one can apply a pruning procedure to reduce the vocabulary size by ignoring terms appearing once or twice. This culling procedure reduces the size of the number of word-types by 50%, and around 80% for the letter 6-grams (see Table 7). Such a pruning scheme can significantly reduce the complexity of text representations based on character *n*-grams with a value of *n* larger than 4 or 5. Moreover, and as shown in Table 8, the success rate is usually higher after the pruning than before. As an alternative, reducing the word-types to the 500 MFW is still a pertinent strategy allowing better performance than considering the entire vocabulary (see Table 8).

There are various ways to extend the current study and to broaden the acquired knowledge from a ranked list of authorship links. Since the proposed methods are based on a reduced set of features, an interpretation of the results can be beneficial for the final user. We could extract information about why (and why not) the highest (and lowest) pairs of texts have a shared authorship. Furthermore, while this study is focused on authorship linking, the experience obtained can be applied to other domains and could be used to improve the performance of author clustering approaches.

Acknowledgments

The authors would thank the anonymous reviewers for their helpful suggestions and remarks.

Funding

This research was supported, in part, by the NSF under grant #200021_149665/1.

References

Almishari, M. and Tsudik, G. (2012). Exploring linkability of user reviews. In *Proceedings Computer Security ESORICS*, Pisa, September 10-12, LNCS, vol. 7459. Berlin: Springer-Verlag, pp. 307–24.

Argamon, S., Koppel, M., Pennebaker, J. W. and Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, **52**(2): 119–23.

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

Biber, D. and Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Binongo, J. N. G. and Smith, M. W. (1999). The Application of Principal Component Analysis to Stylometry. *Literary and Linguistic Computing*, **14**(4): 445–65.

Blackshaw, P. (2008). *Satisfied Customers Tell Three Friends, Angry Customers Tell 3,000*. New York: Crown Business.

Burrows, J. F. (1992). Not unless you ask nicely: the interpretative nexus between analysis and information. *Literary and Linguistic Computing*, **7**(1): 91–109.

Burrows J. F. (2002). Delta: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Burrows, J. F. (2007). All the way through: testing for authorship in different frequency strata. *Literary and Linguistic Computing*, **22**(1): 27–47.

Cortelazzo, M. A., Nadalutti, P., Ondelli, S. and Tuzzi, A. (2016). Authorship attribution and text clustering for contemporary italian novels. In *Proceedings Qualico 2016*, Trier, August 24-28, pp. 7–8.

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

Duda, R. O., Hart, P. E. and Stork, D. G. (2001). *Pattern Classification*. New York, NY: Addison-Wesley.

Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. San Rafael, CA: Morgan & Claypool Publ.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. Cambridge: The MIT University Press.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science*, **42**(1): 7–15.

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, **13**(3): 111–17.

Jockers, M. L. and Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, **25**(2): 215–23.

Juola, P. (2006). Authorship attribution. In Foundations and Trends in Information Retrieval, vol. **1**.

Kešelj, V., Peng, F., Cercone, N. and Thomas C. (2003). N-gram-based author profiles for authorship

attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03*, Halifax, pp. 255–64.

Kjell, B. (1994). Authorship determination using letter pair frequency features with neural network classifier. *Literary and Linguistics Computing*, **9**(2): 119–24.

Kocher, M. and Savoy, J. (2017). Distance measures in author profiling. *Information Processing and Management*, **53**(5): 1103–19.

Koppel, M., Schler, J. and Bonchek-Dokow, E. (2007). Measuring differentiability: unmasking pseudonymous authors. *Journal of Machine Learning Research*, **8**(6): 1261–76.

Labbé, D. and Labbé, C. (2006). A tool for literary studies. *Literary and Linguistic Computing*, **21**(3): 311–26.

Labbé, D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics*, **14**(1): 33–80.

Ledger, G. and Merriam, R. (1994). Shakespeare, Fletcher, and the *Two Noble Kinsmen*. *Literary and Linguistic Computing*, **9**(3): 235–48.

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

Manning, C. D., Raghavan, P. and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

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

McNamee, P. and Mayfield, J. 2004. Character N-gram tokenization for European language text retrieval. *Information Retrieval Journal*, **7**(1/2): 73–98.

Olsson, J. (2008). *Forensic Linguistics*. London: Continuum.

Pennebaker, J. W. (2011). *The Secret Life of Pronouns. What our Words Say about us*. New York, NY: Bloomsbury Press.

Rangel, F. and Rosso, P. (2016). On the impact of emotions on author profiling. *Information Processing and Management*, **52**(1): 73–92.

Savoy, J. (2015). Comparative evaluation of term selection functions for authorship attribution. *Digital Scholarship in the Humanities*, **30**(2): 246–61.

Savoy, J. (2016). Text representation strategies: an example with the *State of the Union* addresses. *Journal of the American Society for Information Science and Technology*, **67**(8): 1858–70.

Savoy, J. (2017). Analysis of the style and the Rhetoric of the American Presidents over two centuries. *Glottometrics*, **38**: 55–76.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, **60**(3): 433–214.

Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., and Potthast, M. (2016). Clustering by authorship within and across documents. In *Notebook Papers of CLEF 2016 Labs and Workshop*. Aachen: CEUR.

Tassinari, L. (2009). *John Florio, the Man who was Shakespeare*. New York, NY: Giano Books.

Toutanova, K., Klein, D., Manning, C. and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclid dependency network. In *Proceedings of HLT-NAACL 2003*, pp. 252–9.

Tuzzi, A. and Cortelazzo, M. (2018). What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer. *Digital Scholarship in the Humanities*, in press. DOI: https://doi.org/10.1093/llc/fqx066.

Voorhees, H. and Harman, D. (2005). *The TREC Experiment and Evaluation in Information Retrieval*. Cambridge: The MIT University Press.

Zhao, Y. and Zobel, J. (2007). Searching with style: authorship attribution in classic literature. In *Proceedings of the Thirtieth Australasian Computer Science Conference (ACSC2007)*, Ballarat, pp. 59–68.

Appendix

Table A1 List of fifty-two text excerpts from the Oxquarry corpus

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

Table A2 List of forty-four text excerpts from the Brunet corpus

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

Table A3 List of 100 text excerpts from the St Jean corpus

#	Author	Short title	#	Author	Short title
1	Balzac	Cousine Bette	51	Dumas	Les trois mousquetaires
2	Chateaubriand	Atala	52	Flaubert	Mme Bovary
3	Dumas	Monte Cristo	53	Gautier	Jettatura
4	Flaubert	Bouvard et Pécuchet	54	Goncourt	Germinie Lacerteux
5	Gautier	Avatar	55	Victor	Notre Dame de Paris
6	Goncourt	Mme Gervaisais	56	Maupassant	Notre cœur
7	Victor	Misérables	57	Sand	Indiana
8	Huysmans	A rebours	58	Stendhal	Rouge et Noir
9	Lamartine	Graziella	59	Verne	Tour du monde
0	Maupassant	Bel-Ami	60	Zola	L'Assommoir
11	Musset	Confession	61	Balzac	César Birotteau
12	Nerval	Aurélia	62	Dumas	Les trois mousquetaires
13	Sand	Petite Fadette	63	Flaubert	Mme Bovary
14	Stendhal	Chartreuse de Parme	64	Gautier	Spirite
15	Verne	Terre à la lune	65	Goncourt	Germinie Lacerteux
16	Vigny	Cinq-Mars	66	Victor	Notre Dame de Paris
17	Zola	L'Argent	67	Maupassant	Fort comme la mort
18	Balzac	Cousine Bette	68	Sand	La mare au diable
19	Chateaubriand	Atala	69	Stendhal	Rouge et Noir
20	Dumas	Monte Cristo	70	Vigny	Servitude et grandeur
21	Flaubert	Bouvard et Pécuchet	71	Zola	Bête humaine
22	Gautier	Avatar	72	Balzac	Colonel Chabert
23	Goncourt	Mme Gervaisais	73	Dumas	Les trois mousquetaires
24	Victor	Misérables	74	Flaubert	Un coeur simple
25	Huysmans	A rebours	75	Victor	Notre Dame de Paris
26	Lamartine	Graziella	76	Flaubert	Education sentimentale
27	Maupassant	Bel-Ami	77	Maupassant	Fort comme la mort
28	Musset	Confession	78	Sand	La mare au diable
29	Nerval	Aurélia	79	Vigny	Servitude et grandeur
30	Sand	Petite Fadette	80	Zola	Bête humaine
31	Stendhal	Chartreuse de Parme	81	Balzac	Colonel Chabert
32	Verne	Terre à la lune	82	Flaubert	Education sentimentale
33	Vigny	Cinq-Mars	83	Maupassant	Mont-Oriol
34	Zola	L'Argent	84	Zola	Fortune des Rougon
35	Balzac	Cousine Bette	85	Balzac	Le père Goriot
36	Chateaubriand	René	86	Flaubert	Hérodias
37	Dumas	Monte Cristo	87	Maupassant	Mont-Oriol
38	Flaubert	Mme Bovary	88	Zola	Fortune des Rougon
39	Gautier	Jettatura	89	Balzac	Eugénie Grande
40	Goncourt	Germinie Lacerteux	90	Flaubert	Salammbô
41	Victor	Misérables	91	Maupassant	Mont-Oriol
42	Lamartine	Graziella	92	Zola	Germinal
43	Maupassant	Notre cœur	93	Balzac	Eugénie Grandet
44	Musset	Confession	94	Flaubert	Salammbô
45	Sand	Indiana	95	Balzac	Le père Goriot
46	Stendhal	Chartreuse de Parme	96	Maupassant	Une vie
47	Verne	Le tour du monde	97	Balzac	Scènes de la vie
48	Vigny	Cinq-Mars	98	Zola	Germinal
49	Zola	L'Assommoir	99	Stendhal	Rouge et le Noir
50	Balzac	César Birotteau	100	Balzac	Scènes de la vie