Travaux neuchâtelois de linguistique, 2011, 55, 59-75

Who Wrote this Novel? Authorship Attribution across Three Languages

Jacques SAVOY
Institut d'informatique, Université de Neuchâtel

Based on different writing style definitions, various authorship attribution schemes have been proposed to identify the real author of a given text or text excerpt. In this article we analyze the relative performance of word types or lemmas assigned to represent styles and texts. As a second objective we compare two authorship attribution approaches, one based on principal component analysis (PCA), and a new authorship attribution method involving specific vocabulary (Z score classification scheme). As a third goal we carry out our experiments on data from three corpora written in three different languages (English, French, and German). In the first we categorize 52 text excerpts (taken from 19th century English novels) written by nine authors. In the second we work with 44 segments taken from French novels (mainly 19th century) written by eleven authors. In the third we extract 59 German text excerpts written by 15 authors and covering the 19th and early 20th centuries. Based on these collections and two specific features (word types or lemmas) we demonstrate that the Z score method performs better than the PCA, while demonstrating that lemmas tend to produce slightly better performance than word types.

1. Introduction

The problem of identifying the real author behind a text excerpt has a long history, going back perhaps to St Jerome (347-420 AD) and his well-known commentaries on the Bible (Love, 2002). Applying statistics and computer technology, more recent approaches aim at automatically determining the correct author of a given document, based on various text samples written by known authors (Juola, 2006). Generally this question has been analyzed from a variety of perspectives. First we face with the *closed class* attribution method, in which the real author may be one of several known candidates. Second, when limiting ourselves to two possible authors, we focus on a *binary* or two-case classification method, a classic example being the *Federalist Papers* (Mosteller & Wallace, 1964). Third *verification* method is applied to determine whether or not a given author did in fact write a document (Koppel *et al.*, 2009). Finally we may simply want to discover certain demographic or psychological information about the author (*profiling*) (Argamon *et al.*, 2009).

In solving text categorization problems such as these, we first represent the documents by a numerical vector comprising its relevant features (word types or lemmas in this study) (Sebastiani, 2002). This process involves selecting the most pertinent features or generating new synthetic features (PCA) useful for identifying differences between several authors (or catego-

ries, genres, etc.). In a second stage we weight them according to their discriminative capability and importance in the textual representation. Finally, through applying a classification scheme, the system automatically assigns the most appropriate author to a given input text.

The rest of this paper is organized as follows. Section 2 presents a brief overview of related authorship attribution work while Section 3 provides an overview of our three corpora. Section 4 describes the principal component analysis (PCA) technique, and the Z score-based approach we use as authorship attribution schemes. Finally, the last section presents the main conclusions that can be drawn from this study.

2. Related Work

Various interesting surveys on authorship attribution have recently been published (Love, 2002; Juola, 2006; Zheng *et al.*, 2006; Koppel *et al.*, 2009). They promote authorship attribution approaches based on statistics, with the first paradigm proposed being based on a unitary invariant value reflecting the particular style of a given author and varying from one to another (Holmes, 1998). Studies on this principle have suggested different statistics related to the type-token ratio (e.g., Herdan's *C*, Guiraud's *R* or Honoré's *H*), lexical richness measures, average word length, certain letter occurrence frequencies or mean sentence length. None of these attempts has however proven satisfactory (Grieve, 2007), due in part to the way word distributions are ruled by a large number of very low probability elements (*Large Number of Rare Events* or LNRE) (Baayen, 2008). Unlike many other domains, when analyzing larger samples of text, we must always face with numerous new instances of *hapax legomena* (words occurring only once).

Instead of applying only a single value to capture each author's discriminative stylistic features, various researchers have suggested applying multivariate techniques (Holmes & Crofts, 2010). A well-known approach in this case is the principal component analysis (PCA) (Binonga & Smith, 1999; Craig & Kinney, 2009) where new composite features are generated as a linear combination of input terms, which are then applied to represent documents as points within a new space. To determine who might be the possible author of a new text, we then simply search for the closest document, assuming that the author of this nearest document probably is the author of the text in question.

In an approach such as this, the major issue is to determine which important stylistic features are most capable of discriminating between possible authors. We have identified three main sources as possible solutions. First at the lexical level we can use the word occurrence frequency of selected terms or punctuation symbols (Grieve, 2007). Mosteller & Wallace (1964) for example found that the writer Hamilton used the word type *while* more

frequently yet in Madison's texts the conjunction *whilst* appears more frequently. As an alternative method we consider more topic-independent features, hopefully those which more precisely reflecting an author's style. In this vein we focus on function words (determiners (*the*, *an, ...*), prepositions (*in*, *of, ...*), conjunctions (*or*, *but, ...*), pronouns (*she*, *our*), as well as certain verbal forms (*is*, *was*, *would, ...*). Given the difficulty in defining precisely such a list, a wide variety of them have been suggested by researchers. Burrows (2002) for example suggests the top *n* most frequent word types (with $n = 40$ to 150), Zhao & Zobel (2007) propose 363 words, while Hoover (2007) recommends more than 4,000 frequently occurring words.

Secondly, at the syntactic level we account for part-of-speech (POS) information through measuring either distribution or frequency, or various combinations thereof. Thirdly, some authors advise considering structural and layout features, including the number of lines per sentence or per paragraph, paragraph indentation, or the presence of greetings. Additional features could also be considered, such as particular orthographic conventions (e.g., British vs. US spelling) or the occurrence of certain specific spelling errors. The resulting number of potential features that might be considered gets rather large, such as the 270 possible features compiled by Zheng *et al*. (2006).

In summary, it seems reasonable to suggest that we should make use of vocabulary features. The presence or absence of words and their occurrence frequencies might reveal the underlying and unknown 'fingerprint' of a particular author during a specific period and relative to a particular genre. Similarly, it is known that word frequencies could change over time and use, as could genres have an impact on vocabulary usage (e.g., poetry or romance, drama or comedy, prose or verse) (Labbé, 2007).

3. Evaluation Corpora

There is an empirical tradition within the authorship attribution domain whereby any text classifiers proposed must be evaluated using a corpus. Although such evaluations were usually limited to the English language, we decided to compensate by carrying out our own experiments using corpora in three languages, namely English, French, and German. The French language is characterized by greater inflectional variability than English, while German clearly makes use of a more complex morphology (e.g., compound construction). The three test corpora chosen for this testing were extracted from the Gutenberg Project (www.gutenberg.org).

The English *Oxquarry* corpus comprises 52 segments, each about 10,000 tokens in length. Created by G. Ledger, this corpus was drawn from 16 novels written during the end of the 19th century and beginning of the 20th century, and had been used in previous authorship attribution experiments

(Labbé, 2007). As shown in the Appendix, this corpus consists of more or less contemporaneous novels written by nine distinct authors, with each of the 52 segments being coded with a series of alphanumeric tags such as 1A – 1Z followed by 2A – 2Z.

For each text, we replaced certain system punctuation marks in UTF-8 coding with their corresponding ASCII symbols (e.g., “” by "), and also removed a few diacritics found in certain English words (e.g., *résumé*), although we kept the diacritics for the French and German texts. In the German excerpts however, we replaced the ß with a double s, and to standardize the spelling, during this process we expanded contracted forms for both the English language (e.g., *don't* into *do not*), and German (e.g., *im* into *in dem*, *aufs* into *auf das*).

As the basis for our experiments we selected either word types or lemmas (dictionary entry). In the first case, each distinct word form had its own entry (e.g., *house* and *houses*) while for lemmas we conflated all inflected forms under the same entry (e.g., *writes*, *wrote*, *written* were regrouped under the lemma *write*). To determine the correct lemmas, we used the part-of-speech (POS) tagger developed by Toutanova & Manning (2000) for the English language, Labbé's system (Labbé, 2001) for French, and the Tree-Tagger system (Schmid, 1995) for German.

The precise definition of lemmas is however not always clear. We do not consider the two pronouns *I* and *me* as dissimilar for example, and thus we merge them under the common headword *I* (we do the same with the lemmas *we* and *you*, and also for the two other languages). This conflation approach can be viewed as a step towards more abstract lexical information representation.

Table 1 provides an overall picture of our three corpora. For the English corpus only, we did not account for punctuation symbols due to the fact that they were missing in a few English text excerpts.

	English	French	German
Number of text excerpts	52	44	59
Number of distinct authors	9	11	15
Total number of lemmas	517,123	439,532	594,513
Most frequent lemma	the (30,048)	le (38,270)	, (50,176)
Number of distinct lemmas	20,400	13,919	31,725
Number of distinct word types	23,872	25,841	45,752
Mean number of lemmas / text	9,948	9,989	10,076
Min number of lemmas / text	9,795 (1T)	9,611 (T #23)	9,999 (T #11)
Max number of lemmas / text	10,118 (2C)	10,239 (T #29)	10,149 (T #37)

Table 1. English, French and German corpora statistics.

For the French language we used a corpus comprising novels written by eleven distinct authors, mainly during the 19th century. This corpus provided by D. Labbé consists of 44 texts, each around 10,000 word tokens in length. To identify each text we simply used numbers from 01 to 44 (see the Appendix for information on the works' authors and titles).

For the German language we build a corpus by extracting text excerpts taken from novels written mainly during the 19th century and the beginning of the 20th century, and made up of 59 texts around 10,000 word tokens each in length. To identify each German text, we simply used numbers from 01 to 59 (see the Appendix for more detailed information).

4. Text Classification Models

To design an authorship attribution system we need to choose a text representation scheme as well as a classifier model. Section 4.1 describes the representation employed in our experiments. As a classifier scheme, we selected principal component analysis (PCA) coupled with the nearest neighbour approach (Section 4.2). The Z score used to measure term specificity is described in Section 4.3, while the last section defines a measure based on the Z score method to define the distance between text pairs and evaluate it as a new authorship attribution method.

4.1 Text Representation and Feature Selection

Two distinct textual representations were used in our experiments. In the first surface forms represent the distinct features taken into account (e.g., *go*, *goes*, *gone*). With the second lemma-based representation, we want to freeze one source of possible variation between the three languages (inflectional morphology) when comparing our results.

As with all text categorization problems, we are faced with a term space characterized by a huge number of dimensions. But not all terms (word types or lemmas) examined prove very useful in distinguishing between different authors, and thus as a dimension reduction scheme we remove any terms (word types or lemmas) having a document frequency (or *df*, the number of texts in which they appear) equal to one or two. Using document frequency as a selection criterion has proven effective in various text categorization problems (Yang & Pedersen, 1997).

By ignoring low-frequency lemmas in the English corpus, we can reduce the feature space from 20,400 to 7,087 lemmas (a reduction of around 65.3%), and when analyzing word types, there are similar reductions made, from 23,872 to 8,765 types. With the French and German corpus we detect similar patterns. For example in the French corpus we start with 13,919 lemmas and obtain a reduced set of 5,761 (a relative reduction of 58.6%), while

we diminish the German corpus from 31,725 lemmas and finally obtain a feature space of 8,078 lemmas (a reduction of 58.8%).

4.2 Principal Component Analysis (PCA)

As a first authorship attribution approach we suggest applying principal component analysis (PCA) in order to obtain a graphical view of affinities between the various text representations (Binonga & Smith, 1999; Craig & Kinney, 2009). The input takes the form of a contingency table, in which columns correspond to the texts and rows to the terms. Table 2 illustrates a brief example, revealing the difficulty of manually detecting possible similarities between texts.

Lemma	1A Hardy	1B Butler	1C Morris	1D Stevenson	1E Butler	1F Stevenson	1G Conrad
the	709	557	606	421	592	540	502
be	346	420	411	482	467	444	376
he	484	536	102	146	464	390	464
and	323	273	322	339	291	374	260
of	321	250	432	275	222	310	319
I	114	212	204	815	279	480	541
an	292	204	230	271	141	290	347
to	267	267	281	309	307	271	238
have	176	240	136	171	203	156	168
in	201	124	158	166	114	161	150

Table 2. Ten most frequent lemmas extracted from the English corpus.

When applied to display similarities between texts, the PCA method generates a new space having fewer dimensions (than the number of terms), that are ordered and orthogonal to each other (also called principal components). This method does not select some of the possible terms but generates new dimensions as linear combinations of input variables (mainly by considering the correlation coefficients).

After this transformation, the PCA computes projections of each point (text) to hyperplanes, resulting in fewer dimensions. During this process the system accounts for a decreasing proportion of the underlying variability (or variance). The first coordinate reflects the best distance but is limited to one dimension (a line) for displaying the respective distances between texts. When displaying the first two principal components, a two-dimensional graphic view (or plane) is obtained, showing the location of the various texts (see Fig. 1).

Taking the English corpus, we select the 50 most frequent lemmas before applying the PCA method. Fig. 1 shows the resulting relative position of each text according to the two most important dimensions. As indicated, the first principal coordinate corresponds to 20.5% of the total variance

while the second represents 11.5%. To identify each text excerpt, we add the first letter of its author to each text identifier. At the top right corner for example the label "F 2V" corresponds to Forster's *Room with a View* while "B 2A" corresponds to Butler's *Erewhon*.

In Fig. 1, an excerpt located near the origin corresponds to a text very similar to the mean profile generated by all documents. As shown in our study's analysis, this mean characteristic can be attributed to "C 1G" (Conrad's *Lord Jim*) or "T 1Y" (Trestel's *Ragged Trousered Philanthropists*). On the other hand those texts located far from the origin tend to have very distinct frequency profiles, as evidenced by "M 2B" (Morris's *Dream of John Ball*) shown in the bottom as well as a second fragment of this work corresponding to the point labelled "M 1J".

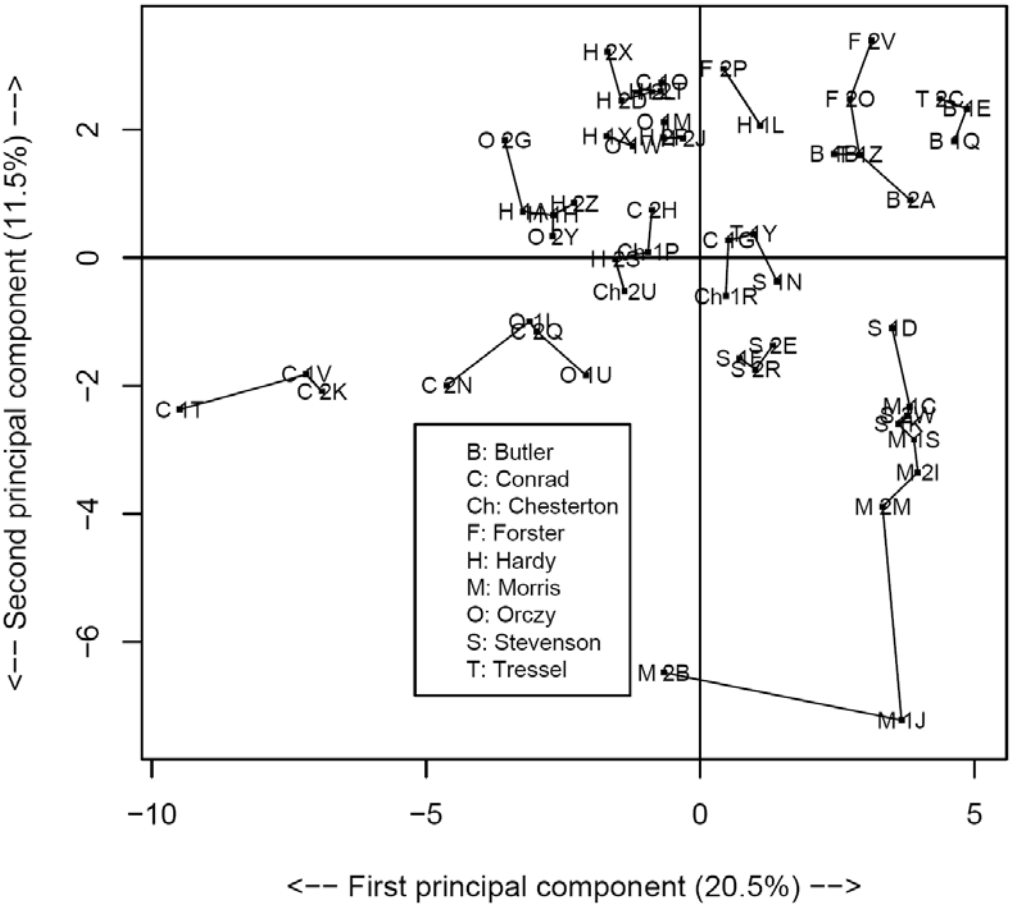

Fig. 1. Graphical representation of distances between 52 English texts based on first two axes of principal component analysis (50 lemmas).

To determine the author of a given text we can look at its nearest neighbour. Although limited to the first two most important dimensions, the left part of Fig. 1 does suggest that the author of "C 1T" (Conrad's *Almayer's Folly*) could be the same as that of "C 1V" (Conrad's *Lord Jim*), or perhaps "C 2K" (Conrad's *Almayer's Folly*), a classification strategy corresponding to the nearest neighbour method (or *k*-NN, with *k*=1). As such we

could determine the closest neighbour of each text and then consider this closest neighbour as the probable author. To facilitate the visualization of this assignment, for each text we add a straight line to its closest neighbour. Fig. 2 displays the same information for the French collection and in Fig. 3 for the German corpus.

Table 3 reports the PCA's performance when applying the nearest neighbour approach for the top 50, 100 and 150 most frequent terms (word types or lemmas) and considering the first two or five principal components (under the label "2 axes" or "5 axes"). Under 5 axes we considered all those dimensions reflecting more than 5% of the total variability. This limit of five is therefore not selected on an arbitrary basis, but rather capable of selecting only those axes which are truly pertinent in reflecting the affinities and dissimilarities.

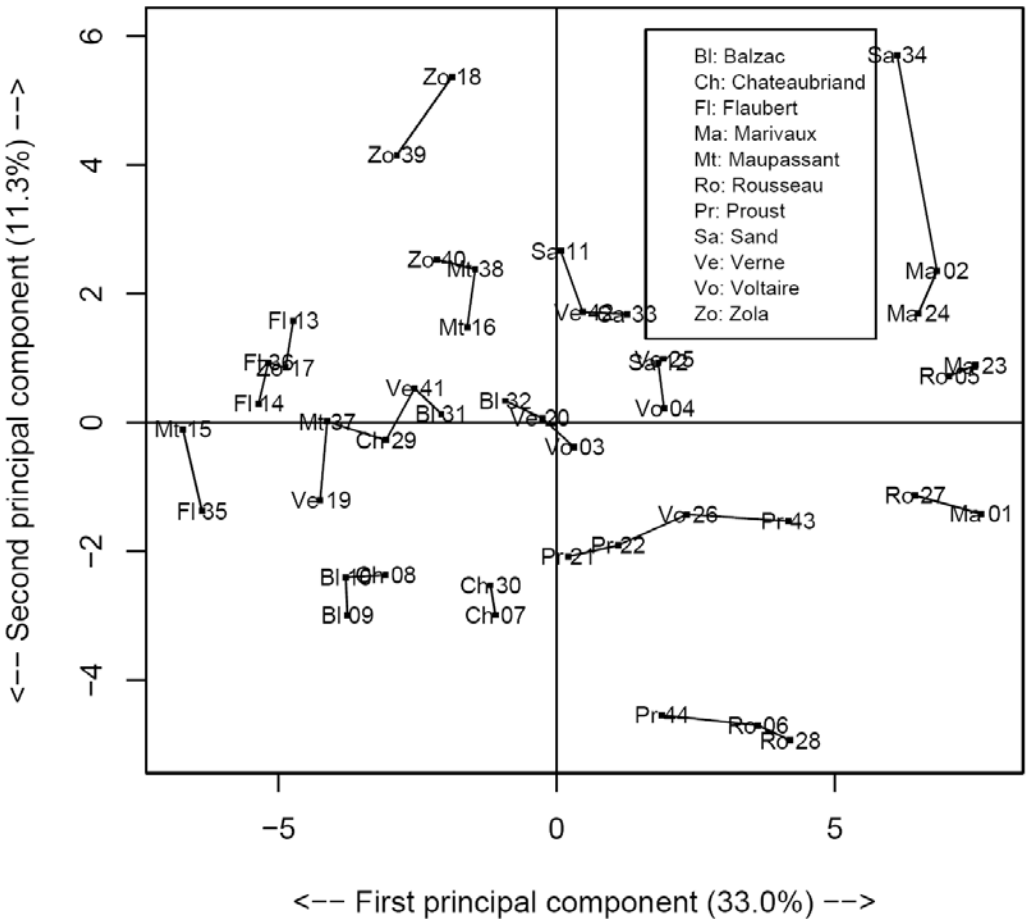

Fig. 2. Nearest neighbour representation for each of 44 texts, based on principal component analysis (50 lemmas, French corpus).

		English		French		German	
		Types	Lemmas	Types	Lemmas	Types	Lemmas
50	2 axes	48.1%	36.5%	34.1%	31.8%	22.0%	30.5%
50	5 axes	88.5%	86.5%	68.2%	68.2%	62.7%	62.7%
100	2 axes	61.5%	57.7%	45.4%	54.6%	32.2%	39.0%
100	5 axes	**92.3%**	**92.3%**	68.2%	**70.4%**	52.5%	66.1%
150	2 axes	63.5%	57.7%	43.2%	54.6%	35.6%	23.7%
150	5 axes	80.8%	84.6%	68.2%	68.2%	61.0%	**69.5%**

Table 3. PCA evaluation using 2 or 5 principal components with 50 to 150 terms.

We obtain the most effective performance with the English corpus when considering the 100 most frequently occurring terms and the first five principal components, with the same performance levels using word types or lemmas (48 correct attributions out of 52). For this language, it seems also that using word types (Column 4) tends to be more effective than the lemma-based representation (Column 5). Note that using more terms (e.g., 150 instead of 100) tends to result in decreased performance.

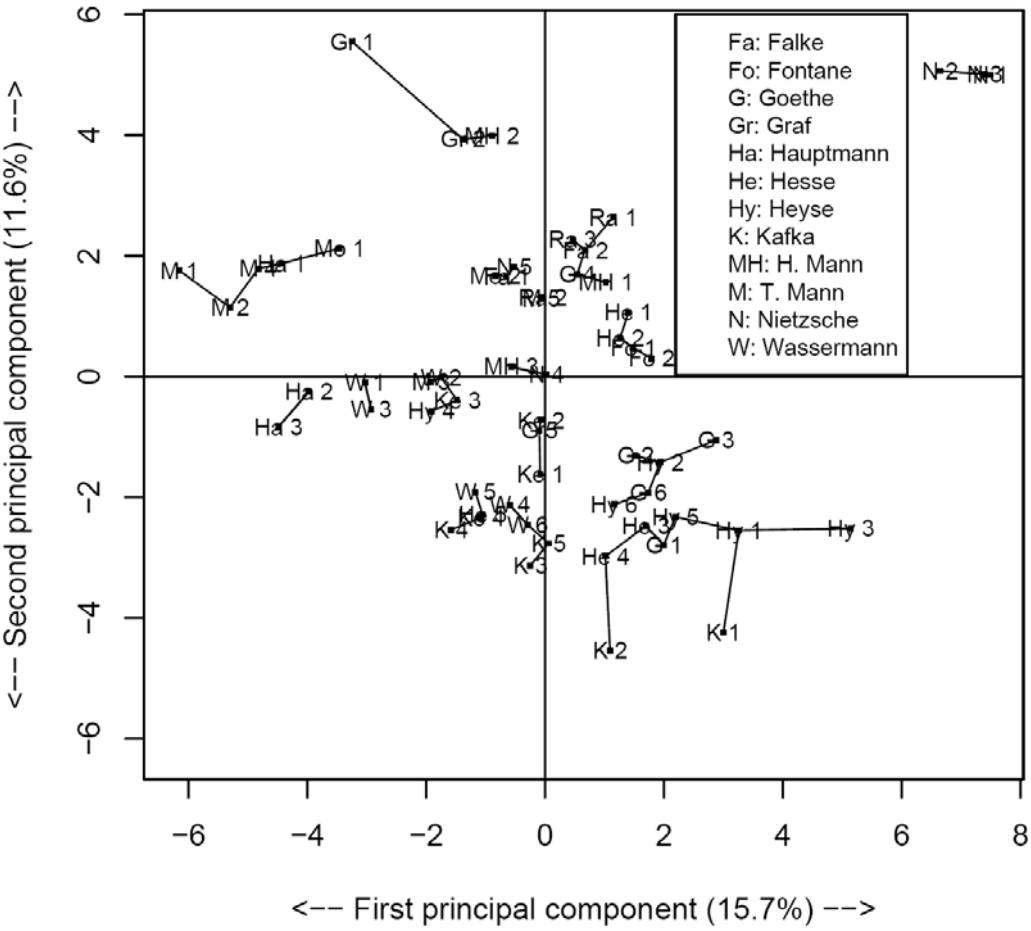

Fig. 3. Nearest neighbour representation for each of 59 texts, based on principal component analysis (50 lemmas, German corpus).

With the French collection we achieve the best performance by considering the 100 most frequently occurring lemmas and the first five principal com-

ponents. This level corresponds to 31 correct attributions out of 44. We can thus conclude that using 50, 100 or 150 word types or lemmas with the first five dimensions results in similar performance levels. Unlike the English corpus, representations based on lemmas tend to be more effective than those based on words when using only the first two dimensions. With the German collection, the best results are obtained with the 150 most frequent lemmas and the first five principal components (41 correct attributions out of 59). Finally, lemma-based representation tends to produce better performances.

4.3 Z Score and Specific Vocabulary

As a new authorship attribution approach, we suggest representing each text based on terms having a document frequency larger than two and corresponding to the text's specific vocabulary, as proposed by Muller (1992) and developed by Savoy (2010). To define and measure a term's specificity, we need to split the entire corpus into two disjoint parts denoted P_0 and P_1. For a given term t_i, we compute its occurrence frequency in the set P_0 (denoted tf_{i0}) and its occurrence frequency in the second part P_1 (or tf_{i1}). In the current context, the set P_0 corresponds to the disputed text, while P_1 represents all other texts, and thus for the entire corpus the occurrence frequency for the term t_i becomes $tf_{i0}+tf_{i1}$. The total number of tokens in part P_0 is denoted n_0, similarly with P_1 and n_1, and the size of the corpus is defined by $n = n_0+n_1$.

The distribution of a term t_i is assumed to be binomial with parameters n_0 and Prob[t_i], representing the probability of randomly selecting the term t_i within the entire corpus. This probability is estimated by Eq. 1, based on the maximum likelihood principle (MLE).

$$\text{Prob}[t_i] = \frac{tf_{i0} + tf_{i1}}{n} \quad (1)$$

This first approach causes certain problems, especially concerning terms never occurring in the corpus, which are assigned a probability of 0. Word distribution does however tend to follow a LNRE (*Large Number of Rare Events* (Baayen, 2008)) pattern, and we therefore suggest smoothing the estimations Prob[t_i] as $(tf_{i0}+tf_{i1}+\lambda) / (n+\lambda\cdot|V|)$, where λ is a parameter and $|V|$ the vocabulary size (Lidstone's law (Manning & Schütze, 2000)). This modification would slightly shift the probability density function's mass towards rare and unseen words (or words that do not yet occur). In our experiments we set $\lambda = 0.1$, and choose to do so because we do not want to assign a large probability to rare words (e.g., large λ value). This smoothing technique is rather easy to implement, and in certain circumstances the maximum likelihood estimation (e.g., Eq. 1) provides a better estimate, thus justifying a smaller value for λ.

Repeating this draw n_0 times allows us to estimate the expected number of occurrences in part P_0 by $n_0 \cdot \text{Prob}[t_i]$. We then compare this expected number to the observed number (namely tf_{i0}), and any *large* differences between these two values would indicate a deviation from the expected behaviour. To obtain a more precise definition of *large* we could account for the variance of the binomial process (defined as $n_0 \cdot \text{Prob}[t_i] \cdot (1 - \text{Prob}[t_i])$). Eq. 2 defines the final standardized Z score for term t_i using the partition P_0 and P_1.

$$\text{Z score}(t_{i0}) = \frac{tf_{i0} - n_0 \cdot \text{Prob}[t_i]}{\sqrt{n_0 \cdot \text{Prob}[t_i] \cdot (1 - \text{Prob}[t_i])}} \quad (2)$$

This Z score value can be used to verify whether the underlying term is used proportionally with roughly the same frequency in both parts (Z score value close to 0). On the other hand, when a term has a positive Z score greater than a given threshold δ (e.g., 2), we could consider it as being over-used or belonging to the specific vocabulary found in part P_0. A large negative Z score (less than $-\delta$) indicates that the corresponding term is under-used in P_0. Knowing that the Z score is assumed to follow a Normal N(0,1) distribution within the limits of $\delta = \pm 1$, we might theoretically find that 68.26% of the terms belong to the common vocabulary, while 15.87% form part of the specific vocabulary (over-used terms).

For the German corpus we analyze the 15 most significant Z scores on a per-author basis. To derive an author profile, we simply average the Z scores calculated for all texts written by the same person. In this set, we could first find terms used more or less exclusively by certain writers in one of their works. These lemmas (or word types) might correspond to main character names (e.g., *Wilhelm* in Goethe's *Die Leiden des jungen Werther*, *K.* in Kafka's novel *Der Prozeß*, or *Tonio* in Mann's *Tonio Kröger*), geographical names or locations (e.g., *Venedig* and *Hotel* in Mann's *Der Tod in Venedig*), or words related to main characters or actions (e.g., *Advokat* and *Prokurist* in *Der Prozeß*). Within the higher Z scores we might also find certain frequent words, such as *wir* (we), *;* (semicolon), *ich* (I), *sie* (she/they) in Goethe's profile, *allerdings* (however) or *d* (lemma corresponding to *der*, *dem*, *die*, ... (the)) in Kafka's profile, *und* (and), *,* (comma), *Meer* (sea) in T. Mann, or *:* (colon), *oh* (interjection), *mein* (my), *reden* (to talk), and *ich* (I) in Nietzsche's most significant terms. As we can see, some lemmas (e.g., *ich* (I)) may appear in two distinct profiles having high Z score values.

	Lemma	Goethe	Nietzsche	Kafka	Hesse	Mann T.
1	,	2.63	-2.37	-1.78	4.80	4.62
2	d	-3.66	-0.75	3.39	-5.80	3.31
3	.	-4.20	-4.66	-2.76	0.54	-0.44
4	und	-2.79	-0.57	-5.51	2.42	4.91
5	sein	-1.13	0.72	-0.01	4.14	1.58
6	er	-4.70	-9.52	4.89	6.30	3.11
7	ein	-1.01	-2.68	-3.45	-0.20	1.20
8	ich	4.76	7.51	-4.66	1.55	-8.07
9	"	2.18	-1.82	3.26	4.90	-0.80
10	in	-1.83	-1.96	-1.31	-0.33	1.97
11	zu	2.02	0.67	3.07	-1.98	0.30
12	sie	3.76	-4.30	1.81	-5.00	-0.15
13	haben	-0.78	-2.93	2.59	4.03	-3.48
14	sich	1.42	-3.17	3.29	-3.21	1.56
15	nicht	0.67	0.40	3.60	1.23	-2.60

Table 5. Top 15 most frequent lemmas and their corresponding Z scores, according to five author profiles (German corpus).

Another interesting approach is to list the most frequent terms (lemmas in this case) along with their Z scores in each author's profile. Table 5 reports the 15 most frequent lemmas extracted from the German corpus. The first column indicates the lemmas in decreasing order of occurrence frequency (within the entire corpus), while the following five columns show the mean Z scores for these lemmas in the corresponding author's profile. From the first row we could infer that the comma (,) is used more significantly in the works of Hesse (Z score 4.8) and T. Mann (4.62), compared to the rest of the works in our German corpus. Goethe also tended to follow a similar pattern (2.63), while Nietzsche (-2.37) or Kafka (-1.78) used this punctuation symbol significantly less. An analysis of the eighth row reveals that Goethe (based on the three novels included in our corpus) employs *ich* (I) (Z score 4.76) more often although for the pronoun *ich* (I) Nietzsche (7.51) appears to be the real champion. On the other hand this personal pronoun was rejected by T. Mann (Z score -8.07) or Kafka (-4.66), who clearly preferred the pronoun *er* (he) (Z score respectively of 3.11 and 4.89), while Hesse used this pronoun significantly more than the other writers (Z score of 6.3).

4.4 Z Score Distance and Evaluation

To estimate the distance between a disputed text D_j and an author profile A_k we apply Eq. 3 in which the author profile A_k is simply the average Z scores calculated for all texts written by that person (see examples given in Table 5). Based on a set of terms t_i, for $i = 1, 2, \ldots m$, and a set of possible author A_k, $k = 1, 2, \ldots, r$ we simply select the lowest distance to determine the most probable writer.

$$\text{Dist}(D_j, A_k) = \frac{1}{m}\sum_{i=1}^{m}\left(Z\,score(t_{ij}) - Z\,score(t_{ik})\right)^2 \quad (3)$$

When both Z scores are very similar for all terms the resulting distance is small, meaning that the same author probably wrote both texts. Moreover, the power of two in this computation tends to reduce the impact of any differences less than 1.0, mainly occurring in the common vocabulary. On the other hand, large differences could also occur for a given term, when both Z scores are large and have opposite signs. In this case one author tends to use the underlying term more frequently than the mean (term specific to this author) while for the other this term is under-used.

English		French		German	
Types	Lemmas	Types	Lemmas	Types	Lemmas
100%	100%	100%	100%	83.1%	84.7%

Table 6. Evaluation of Z score method.

With the English and French corpora, the Z score distance makes it possible to correctly classify all the texts, while with the German collection, the authorship attribution approach is able to correctly identify around 85% of the cases (50 out of 59). Unlike the two other methods, this suggested approach was parameter free, so Table 6 contains just one row.

5. Conclusion

Authorship attribution problems involve a number of interesting challenges. In this study, we investigate these problems linked to literary works written in three languages (English, French, and German), and process around ten distinct authors for each corpus. Unlike previous studies limited to a single corpus, usually written in English, we are able to base our findings on a broader and more solid foundation. In empirical studies the use of more than one collection should be the norm and in our opinion analyzing corpora comprising works in more than one language allows us to obtain a better overview of the relative merits of the various methods.

In our case we first apply the principal component analysis (PCA) method in which word-frequency (or lemma-based) information is used to visualize similarities and dissimilarities between text excerpts. As a data visualization tool, PCA defines a new ordered set of orthogonal dimensions in which we can place the text points, and which constitutes a real advantage over other approaches. By working with a reduced space and applying a distance measure, we are able to apply a nearest-neighbour learning scheme. However the resulting success rates with the three corpora are not perfect (see Table 3).

In order to improve categorization performance, we describe a new classification scheme based on term Z score defined by Muller (1992). These Z score values prove to be useful in identifying the specific vocabulary in a given text, and in this study, revealing the author's style characteristics. Instead of considering the *n* most frequent words (e.g, with $n = 40$ to 150 (Burrows, 2002)), the model we propose accounts for all selected terms. In our opinion however weighting features according to their occurrence frequencies constitutes a reasonable strategy. As for the need to design a robust classifier capable of generalization, we believe that an appropriate method for pruning the feature space's high dimensionality could consist of ignoring terms having a small occurrence or document frequencies, and therefore we applied this feature selection procedure in our study. As shown in Table 5 with our German corpus, the proposed approach may provide some useful information about the authors' similarities and differences. This approach is based on a distance measure and thus we suggest that it should be used as an authorship attribution method. Evaluations made on the English, French, and German corpora lead to high success rates, clearly superior to those obtained when applying the PCA approach (see Table 6).

Upon comparing the two approaches with respect to word type and lemma text representations, we find that both tend to provide similar results. At the limit slight improvement is evident when applying a lemma-based text representation. There are of course certain questions that must be addressed in future studies. An evaluation of the Z score's reliability in authorship attribution and its impact on text representation quality when using word bigrams or trigrams (or lemmas) in a complementary manner are just two examples.

Acknowledgments

The author would like to thank Dominique Labbé for providing us the English and French corpora, together with a version of the French lemmatized corpus. This work was supported by the Swiss NSF (#200021-124389).

References

Argamon, S., Koppel, M., Pennebaker, J.W., & Schler, J. (2009): Automatically profiling the author of an anonymous text. Communications of the ACM, 52(2), 119-123.

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

Binonga, J.N.G., & Smith, M.W. (1999): The application of principal component analysis to stylometry. Literacy and Linguistic Computing, 14(4), 445-465.

Burrows, J.F. (2002): Delta: A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17(3), 267-287.

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

Grieve, J. (2007): Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing, 22(3), 251-270.

Holmes, D.I. (1998): The evolution of stylometry in humanities scholarship. Literary and Linguistic Computing, 13(3), 111-117.

Holmes, D.I., & Crofts, D.W. (2010): The Diary of a Public Man: A case study in traditional and non-traditional authorship attribution. Literary and Linguistic Computing, 25(2), 179-197.

Hoover, D.L. (2007): Corpus stylistics, stylometry and the styles of Henry James. Style, 41(2), 160-189.

Juola, P. (2006): Authorship attribution. Foundations and Trends in Information Retrieval, 1(3).

Koppel, M., Schler, J., & Argamon, S. (2009): Computational methods in authorship attribution. Journal of the American Society for Information Science & Technology, 60(1), 9-26.

Labbé, D. (2001): Normalisation et lemmatisation d'une question ouverte. Journal de la Société Française de Statistique, 142(4), 37-57.

— (2007): Experiments on authorship attribution by intertextual distance in English. Journal of Quantitative Linguistics, 14(1), 33-80.

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

Manning, C.D., & Schütze, H. (2000): Foundations of Statistical Natural Language Processing. Cambridge (The MIT Press).

Mosteller, F., & Wallace, D.L. (1964): Inference and Disputed Authorship, The Federalist. Reading (Addison-Wesley). Reprint 2007.

Muller, C. (1992): Principes et méthodes de statistique lexicale. Paris (Honoré Champion).

Savoy, J. (2010): Lexical analysis of US political speeches. Journal of Quantitative Linguistics, 17(2), 123-141.

Schmid, H. (1995): Improvements in part-of-speech tagging with an application to German. In Proceedings of the ACL SIGDAT-Workshop.

Sebastiani, F. (2002): Machine learning in automatic text categorization. ACM Computing Survey, 14(1), 1-27.

Toutanova, K., & Manning, C. (2000): Enriching the knowledge sources used in a maximum entropy part-of-speech tagging. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), 63-70.

Yang, Y., & Pedersen, J.O. (1997): A comparative study of feature selection in text categorization. In Proceedings of the Fourteenth Conference on Machine Learning ICML, 412-420.

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

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

Appendix

In the following tables we have reported the identifiers used for each text excerpt together with the author and title of the novel from which the text segment was extracted.

Id	Author	Short Title	Id	Author	Short Title
1A	Hardy	Jude	2A	Butler	Erewhon revisit.
1B	Butler	Erewhon revisit.	2B	Morris	Dream of JB
1C	Morris	News from nowhere	2C	Tressel	Ragged TP
1D	Stevenson	Catriona	2D	Hardy	Jude
1E	Butler	Erewhon revisit.	2E	Stevenson	Ballantrae
1F	Stevenson	Ballantrae	2F	Hardy	Wessex Tales
1G	Conrad	Lord Jim	2G	Orczy	Elusive P
1H	Hardy	Madding	2H	Conrad	Lord Jim
1I	Orczy	Scarlet P	2I	Morris	News from nowhere
1J	Morris	Dream of JB	2J	Hardy	Well beloved
1K	Stevenson	Catriona	2K	Conrad	Almayer
1L	Hardy	Jude	2L	Hardy	Well beloved
1M	Orczy	Scarlet P	2M	Morris	News from nowhere
1N	Stevenson	Ballantrae	2N	Conrad	Almayer
1O	Conrad	Lord Jim	2O	Forster	Room with a view
1P	Chesterton	Man who was	2P	Forster	Room with a view
1Q	Butler	Erewhon revisit.	2Q	Conrad	Almayer
1R	Chesterton	Man who was	2R	Stevenson	Catriona
1S	Morris	News from nowhere	2S	Hardy	Madding
1T	Conrad	Almayer	2T	Hardy	Well beloved
1U	Orczy	Elusive P	2U	Chesterton	Man who was
1V	Conrad	Lord Jim	2V	Forster	Room with a view
1W	Orczy	Elusive P	2W	Stevenson	Catriona
1X	Hardy	Wessex Tales	2X	Hardy	Well beloved
1Y	Tressel	Ragged TP	2Y	Orczy	Scarlet P
1Z	Tressel	Ragged TP	2Z	Hardy	Madding

Table A.1. Detailed description of English *Oxquarry* corpus content.

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

Table A.2. Detailed description of French corpus content.

Id	Author	Title
1, 2	Goethe	Die Wahlverwandtschaften
3, 4	Goethe	Die Leiden des jungen Werther
5, 6	Goethe	Wilhelm Meisters Wanderjahre
7, 8	Morike	Mozart auf der Reise nach Prag
9, 10	Keller	Die Leute von Seldwyla - Band 1
11, 12	Keller	Die Leute von Seldwyla - Band 2
13	Heyse	L'Arrabbiata
14, 15	Heyse	Beatrice
16, 17, 18	Heyse	Der Weinhüter
19	Raabe	Deutscher Mondschein
20, 21	Raabe	Zum wilden Mann
22, 23	Fontane	Unterm Birnbaum
24, 25, 26	Nietzsche	Also sprach Zarathustra
27, 28	Nietzsche	Ecce Homo
29	Hauptmann	Bahnwärter Thiel
30, 31	Hauptmann	Der Ketzer von Soana
32, 33	Falke	Der Mann im Nebel
34, 35	Mann, H.	Flöten und Dolche
36	Mann, H.	Der Vater
37, 38	Mann, T.	Der Tod in Venedig
39, 40	Mann, T.	Tonio Kröger
41	Mann, T.	Tristan
42, 43	Kafka	Der Prozeß
44, 45	Kafka	Die Verwandlung
46	Kafka	In der Strafkolonie
47, 48	Wassermann	Caspar Hauser
49, 50	Wassermann	Der Mann von vierzig Jahren
51, 52	Wassermann	Mein Weg als Deutscher und Jude
53, 54	Hesse	Drei Geschichten aus dem Leben
55, 56, 57	Hesse	Siddhartha
58, 59	Graf	Zur Freundlichen Erinnerung

Table A.3. Detailed description of German corpus content.