English Studies

ISSN: 0013-838X (Print) 1744-4217 (Online) Journal homepage: http://www.tandfonline.com/loi/nest20

Cross-Genre Authorship Verification Using Unmasking

Mike Kestemont , Kim Luyckx , Walter Daelemans & Thomas Crombez

To cite this article: Mike Kestemont , Kim Luyckx , Walter Daelemans & Thomas Crombez (2012) Cross-Genre Authorship Verification Using Unmasking, English Studies, 93:3, 340-356, DOI: 10.1080/0013838X.2012.668793

To link to this article: http://dx.doi.org/10.1080/0013838X.2012.668793

Published online: 22 May 2012.

English Studies
Vol. 93, No. 3, May 2012, 340–356

Routledge
Taylor & Francis Group

Cross-Genre Authorship Verification Using Unmasking

Mike Kestemont, Kim Luyckx, Walter Daelemans and Thomas Crombez

In this paper we will stress-test a recently proposed technique for computational authorship verification, ''unmasking'', which has been well received in the literature. The technique envisages an experimental set-up commonly referred to as ''authorship verification'', a task generally deemed more difficult than so-called ''authorship attribution''. We will apply the technique to authorship verification across genres, an extremely complex text categorization problem that so far has remained unexplored. We focus on five representative contemporary English-language authors. For each of them, the corpus under scrutiny contains several texts in two genres (literary prose and theatre plays). Our research confirms that unmasking is an interesting technique for computational authorship verification, especially yielding reliable results within the genre of (larger) prose works in our corpus. Authorship verification, however, proves much more difficult in the theatrical part of the corpus.

1. Introduction

Over a decade ago, Joseph Rudman claimed that non-traditional authorship studies ''have had enough time to pass through any 'shake-down' phase and enter one marked by solid, scientific, and steadily progressing studies. But after 30 years and 300 publications, they have not''.[1] Rudman particularly criticized the wild proliferation of approaches: he estimated that, for instance, nearly a thousand different stylometric features had been proposed, but at the same time, he deplored the absence of a scientific consensus on a best practice in computational authorship studies.[2] When looking back on his paper in 2010, he concluded that an additional six hundred new publications had not significantly changed this situation.[3] Rudman's view is sobering, but correctly stresses the importance of the thorough evaluation of existing techniques. In this paper we will stress-test the recently proposed

Mike Kestemont, Kim Luyckx, Walter Daelemans and Thomas Crombez are affiliated with the University of Antwerp, Belgium. Email: Mike.Kestemont@ua.ac.be

[1]Rudman, ''The State of Authorship,'' 351.

[2]Ibid., 360; cf. Stamatatos, 553.

[3]Rudman, ''The State of Non-Traditional Authorship.''

ISSN 0013-838X (print)/ISSN 1744-4217 (online)
http://dx.doi.org/10.1080/0013838X.2012.668793

"unmasking" technique,[4] which has been well received in the literature.[5] The technique envisages an experimental set-up commonly referred to as "authorship verification", a task generally deemed more difficult than so-called "authorship attribution".[6] We will apply the technique to authorship verification across genres, an extremely complex text categorization problem that so far has remained unexplored. We focus on five representative contemporary English-language authors. For each of them, the corpus under scrutiny contains several texts in two genres (literary prose and theatre plays).

2. Background: Cross-Genre Authorship Verification

The task of authorship verification is commonly distinguished from that of authorship attribution.[7] In both text classification approaches, the task is to decide whether a given text has been written by a candidate author. In authorship attribution, the actual author is known to be included in the set of candidates (*closed case*). In authorship verification, however, this assumption cannot be made: the given text might have been written by one of the candidate authors, but could also be written by none of them (*open case*).[8] Note that this scenario is typical of forensic applications where it cannot be presupposed that the author of, for example, a letter bomb is among the suspect candidate authors. Forensic applications (such as extortion or suicide letters) show the potential of computational authorship studies.[9] In the case of a suicide letter (possibly faked by a murderer), however, it is highly likely that this is the only suicide letter the victim ever wrote. In absence of similar material, it is difficult to extract reliable style markers from pre-existing writings to determine the authorship of the suicide letter.

This brings us to an issue that is being paid all too little attention in present-day research: authorship *across genres*. Most studies only consider corpora that are restricted to a single text variety, such as student essays, newspaper articles, blog posts or entire novels.[10] Although the corpora in question may show a good deal of topic variation, we hardly find any studies that deal with authorship attribution across text varieties explicitly.[11] The few remarks that have been made on this issue agree that authorship attribution is difficult within a single textual genre, even more difficult when several topics are involved, and likely to be extremely difficult with several

[4]Koppel and Schler; Koppel, Schler, and Bonchek-Dokow; Koppel, Schler, and Argamon, "Computational Methods."
[5]Kacmarcik and Gamon; Stein, Lipka, and Prettenhofer.
[6]Koppel, Schler, and Argamon, "Computational Methods," 18.
[7]Koppel, Schler, and Argamon, "Authorship Attribution in the Wild," 83–4.
[8]Ibid.
[9]Chaski; Lambers and Veenman.
[10]For the categories listed, consult respectively: van Halteren et al.; Luyckx and Daelemans, "The Effect"; Stamatatos, Fakotakis, and Kokkinakis; Luyckx and Daelemans, "Shallow Text Analysis"; Sanderson and Guenter; Koppel, Schler, and Argamon, "Authorship Attribution in the Wild"; Koppel, Schler, and Bonchek-Dokow.
[11]Stamatatos, 553–4.

genres involved.[12] Consequently, cross-genre authorship verification deserves much more attention than it has attracted so far. Although it is generally assumed that an author will display stable style characteristics throughout his oeuvre, irrespective of genre, this remains speculative in the absence of systematic empirical investigation.

3. Unmasking

Unmasking is a fairly complex meta-learning approach to authorship verification.[13] In the original paper, unmasking is contrasted with the following naive approach. Consider text A, written by a known author, and text X of unknown signature. One could divide each work into chunks and build a model of the stylistic difference between both chunk collections. Subsequently, a cross-validation experiment could be carried out to assess the magnitude of the differences between A and X. High generalization accuracy in distinguishing between A and X could be considered indicative of non-identical authorship, whereas low accuracy could imply identical authorship. This method nevertheless fails to resolve a real-world verification example consisting of four works by three authors (Melville, Cooper, and Hawthorne). One of the four novels—*The House of the Seven Gables*—could be distinguished with high accuracy (>98 per cent) from the three other novels in the experiment. The conclusion that none of these authors wrote the novel, however, would be wrong, since Hawthorne wrote it.

An error analysis of the model that could distinguish between both Hawthorne texts showed that a surprisingly small number of features was doing all the work. It is indeed common for authors to use ''a small number of features in a consistently different way between works''.[14] Such features often relate to topic differences (e.g., names of *dramatis personae*), narrative differences (e.g., third-person versus first-person narratives), or thematic differences (e.g., love story versus detective story). Note that even the high-frequency items typically used in computational authorship studies are often affected by these differences; several attempts have for instance been made to diminish the effect of the narrative perspective on the frequency of personal pronouns in text.[15] As such, a limited number of features can wrongfully maximize the differences in writing style between two works of identical authorship.

The unmasking approach does not test *whether* a stylistic model can be built, distinguishing between two texts, since this is often all too easy. Rather, it tests the *robustness* of this model by deliberately impairing it over a number of iterations, each time removing those features that are most discriminative between the two texts. Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow noted that the ''degradation curves'' (in cross-validation accuracy) resulting from this process could

[12]Ibid.; Luyckx, 2–3.
[13]E.g., Koppel, Schler, and Bonchek-Dokow.
[14]Ibid., 1263.
[15]For example, pronoun culling, see Hoover.

be indicative of whether or not two texts were written by the same author.[16] When a degradation curve is built for two texts by the same author and only a few features are doing all the work, the curve displays many sudden drops in accuracy. When the most telling features are removed during each iteration, it becomes increasingly difficult to differentiate between two texts. In the case of two texts of non-identical authorship, however, a far larger number of features is discriminative, causing less dramatic drops in accuracy during degradation. Unmasking exploits this apparent regularity. Using training material in the form of a series of same-author and different-author degradation curves, they try to classify previously unseen degradation curves involving two works. It then becomes possible to accept or reject (non-) identical authorship for two texts based on the degradation curve for them.

The unmasking technique has been well received in the secondary literature, and has been successfully applied, for example, in intrinsic plagiarism detection.[17] In a paper on style obfuscation, Gary Kacmarcik and Michael Gamon even concluded that a number of simple methods for computational authorship studies "may appear to work well, but are far less resilient to obfuscation attempts than Koppel and Schler's unmasking approach".[18] Conrad Sanderson and Simon Guenter confirmed the positive effect of unmasking for longer texts, but demonstrated that it is less reliable for shorter texts, below 10,000 words in size.[19] In this paper, we will apply unmasking to authorship verification across genres. Unmasking is especially attractive for this task, because of the interference between genre markers and authorial style markers. It can be anticipated that an author's texts in different genres will display a number of superficial differences in style. Theatrical texts, for instance, can be expected to contain many more lexical features relating to direct speech or stage indications than reflective essays. Overestimating the importance of such shallow characteristics could create the impression of an artificially large distance between two same-author works, written in two genres. Interestingly, the unmasking technique might help remedy these genre-related artefacts: superficial genre-related differences between same-author texts in different genres will be filtered out easily and removed from the model early in the degradation process. After the removal of these non-essential stylistic features, one could hypothesize that only features more relevant to authorial identity will be preserved.

4. Methodology and Evaluation

Our unmasking implementation closely adheres to the original description of the procedure.[20] Consider an example corpus of four texts written by two authors. Let A1

[16]See Koppel, Schler, and Bonchek-Dokow, or Koppel, Schler, and Argamon, "Computational Methods."
[17]Stein, Lipka, and Prettenhofer.
[18]Kacmarcik and Gamon, 451.
[19]Sanderson and Guenter. Their observation is acknowledged in Koppel, Schler, and Argamon, "Computational Methods," 21.
[20]See Koppel, Schler, and Bonchek-Dokow.

and A2 be written by author A and B1 and B2 by author B. Each text is divided into tokens by splitting it along white space. All non-alphanumeric characters are deleted from the tokens. Subsequently, these texts are divided into equal-sized chunks. For each combination of two texts (e.g., <A2, B1>), we generate a degradation curve (leading to six text pair curves in total). From the chunks in both texts, we select the n tokens with the highest cumulative frequency (with a weighted average for both texts). We represent all chunks under a term-frequency vector space model using the relative frequencies of these n highly frequent tokens. During m iterations, we train a *Support Vector Machine* with a linear kernel—the *Sequential Minimal Optimization* (SMO) implementation under its default algorithmic settings in the *Weka* machine learning software package—on this chunk collection and assess its performance through x-fold cross validation.[21] At the end of each iteration, we remove the k tokens that the SMO classifier (when trained on the entire chunk collection) indicates as the ''most strongly weighted positive features'', as well as the k ''most strongly weighted negative features''. After each iteration, the feature set will contain 2^*k very discriminating tokens less than in the previous iteration. The validation accuracies collected for each of the m iterations will constitute degradation vectors, which can be depicted as curves for each of our six text pairs. In this visualization, classification accuracy is plotted as a function of the iteration index. Naturally, these curves typically display a descending slope.

For evaluation purposes, a ''leave-one-text-out validation'' is carried out on this set of curves. For each text, we iteratively collect as test vectors the degradation curves that involve that specific text and have all other degradation curves make up the training vectors. If the current test text were A2, for instance, the initial six curves would be divided as follows:

Training curves	**Test curves**
<A1, B1> = *different author*	<A1, A2> = *same author*
<A1, B2> = *different author*	<A2, B1> = *different author*
<B1, B2> = *same author*	<A2, B2> = *different author*

These curves are represented as vectors using the following features for $i = 0, \ldots, m$: the accuracy after i elimination rounds, the accuracy difference between round i and $i+1$ and the accuracy difference between round i and $i+2$. Additionally, the highest accuracy drop in one iteration and highest accuracy drop in two iterations are included as a feature. Next, we train the SMO classifier on this representation of the training curves and use it to classify each of the test curves as a *same-author* or *different-author* curve (i.e., binary classification). When all predictions have been collected, one can report on the overall classification accuracy, the macro-averaged *F1*-score as well as the number of correctly classified *same-author* curves and the number of correctly classified *different-author* curves (*true positives*). In our example

[21]The software etc. are described in Platt; and Hall et al.

of four texts, twelve test predictions (three predictions per test text) would be included in the evaluation. Two quite distinct baselines can be used for this classification task. The *chance*-baseline stresses the fact that one is dealing with a two-class classification problem: curves are either labelled as *same-author* curves or as *different-author* curves. One baseline for such problems could be the chance-level of a 50 per cent classification accuracy. The *majority* baseline is more demanding and stresses the fact that in these experiments the probability of a text pair by the same author is much lower than that of a text pair by different authors, so that a high overall accuracy can be reached by simply labelling each curve as a *different-author* curve.

For the experiments in this paper, we will not venture into advanced parameter optimization and throughout this paper we will use the same generic parameter settings as tentatively adopted by Koppel, Schler, and Bonchek-Dokow: a *chunk size* of 500 tokens, $n = 250$, $m = 10$ and $k = 3$.[22] An important deviation, however, is that they used an x-fold cross validation with x invariably set to 10 to assess the classifier's performance in each iteration, whereas we have used a leave-one-out validation scheme (i.e. an x-fold cross validation with x dynamically set to the cumulative number of chunks in both texts). Another difference is that we did not normalize the chunk length of texts using a random sampling procedure. Preliminary experimentation (not reported for the sake of brevity) showed that this set-up generally yielded more nuanced degradation curves—an effect especially notable with shorter texts. Class imbalance (e.g., for the combination of a longer and shorter text) moreover did not seem an impediment: rather, the procedure seemed to benefit from including as many of the available chunks as possible.

5. Corpus and Selection of Texts

The corpus we collected for the experiments in cross-genre authorship verification consists of published texts by five contemporary authors: Edward Bond (1934), David Mamet (1947), Harold Pinter (1930–2008), Sam Shepard (1943), and Arnold Wesker (1932). The main criterion for selecting an author was the availability of texts in more than one literary genre. Theatre and prose were the genres these five authors were most productive in, so these were chosen for the experiments. For an individual text to be included in the experiments, it needed to have a minimum length of 10,000 words—or twenty chunks of 500 tokens. Since Sanderson and Guenter demonstrated that the unmasking technique was less reliable for texts below that length and the original unmasking papers all worked with long texts, this seemed a reliable threshold.[23] Table 1 shows the list of works used in our experiments. Digitization of the contemporary material—still under copyright protection—involved three steps. After the books were scanned, they were analyzed using software for Optical

[22]Koppel, Schler, and Bonchek-Dokow.

[23]Sanderson and Guenter; cf. Koppel, Schler, and Argamon, 21.

Table 1 List of Authors, Genres, and Works Used in Our Experiments, with Text Length Information (expressed in number of 500-token chunks)

Author	Genre	Title	Chunks (500 tokens)
Edward Bond	Theatre	*Bingo*	29
		The Fool	37
		Summer	33
	Prose	*Fables*	20
David Mamet	Theatre	*American Buffalo*	33
		Lakeboat	22
		Glengarry Glen Ross	28
		Sexual Perversity in Chicago	20
		The Woods	24
	Prose	*The Old Religion*	71
		The Village	104
Harold Pinter	Theatre	*The Caretaker*	37
		The Homecoming	32
		The Hothouse	34
		No Man's Land	23
	Prose	*The Dwarfs*	90
Sam Shepard	Theatre	*Fool for Love*	27
		The Unseen Hand	23
		Melodrama Play	20
		La Turista	29
	Prose	*Great Dream of Heaven*	73
		Cruising Paradise	125
Arnold Wesker	Theatre	*Annie Wobbler*	21
		I'm Talking about Jerusalem	39
		The Journalists	32
		Roots	40
		Shylock	45
		Chicken Soup with Barley	38
		The Wedding Feast	35
	Prose	*The Man Who Became Afraid*	24
		Love Letters on Blue Paper	31
		Said the Old Man to the Young Man	31
		Six Sundays in January	24
		The Visit	63

Character Recognition (OCR), in this case the *Abbyy FineReader* software package, and saved in the UTF-8 format. The final step was a manual correction—focusing mainly on typical confusables (e.g., *rn* vs. *m*)—of the OCRed texts to ensure the quality of the digitized material. We removed any text that was added by the editor (e.g., lists of *dramatis personae* or actors), not written by the author (e.g., a foreword), or written in a different genre (e.g., a foreword to a theatre play by the author himself). Our decision not to remove stage directions (e.g., *walks around the store a bit in silence*) or names of characters speaking was inspired by the unmasking technique. Since it iteratively removes those features that are most discriminative between texts—often, these will be names of *dramatis personae*, topic markers

etc.[24]—we hypothesized that removing these from the input material would be a redundant pre-processing step.

Unlike previous research, we did not discard any texts, even if they were the only prose or theatre work by a particular author in the corpus. The reason for this was twofold. Koppel, Schler, and Bonchek-Dokow stated "the pair Emily Bronte/ *Wuthering Heights* can't be tested since it is the author's only work".[25] This decision can be explained from a Machine Learning perspective, since including the work in the training set implies that a correct *same-author* label for Emily Brontë is impossible, whereas including it in the test set implies that there is no training material for Emily Brontë. We decided to include, for example, Bond's only prose work *Fables*—the only one in our corpus that was sufficiently long—in order to mimic a realistic forensic situation. Consider the case of the (potentially faked) suicide letter: it is likely that the author is not in the training material, since the text is compared to a number of suspect authors, possibly not including the actual author. Even in those cases, the unmasking technique should be able to predict a *different-author* provenance. A second reason for not removing texts was of a more practical nature. By keeping all materials, we were able to work on a complete matrix of authors and genres, allowing both intra-genre and cross-genre experiments for all authors.

6. Intra-Genre Experiments

A first experiment has been carried out on the five authors' eleven prose works in the corpus. The degradation curves for this experiment are depicted in Figure 1. Solid lines represent *same-author* curves, whereas dotted lines represent *different-author* curves. Figure 1 is highly illustrative of the potential of the unmasking approach. All curves tend to display downward slopes, with decreasing cross-validation accuracies, as more predictive features are eliminated in each iteration. For *same-author* curves, however, it is clearly visible that the effect of degradation generally sets off sooner and more dramatically. *Different-author* curves are more robust to the impairment attempts and tend to yield higher cross-validation accuracies, even when a large number of strongly discriminative features is deleted. Consequently, the *different-author* curves dominate the upper regions of the plot in Figure 1, well distinguishable from most of the *same-author* curves in the lower part of the plot. Intersections between both curve types are minimal. A leave-one-text-out validation test on this set of curves confirms the success of the approach: the overall accuracy amounts to 96.36 per cent, which is only just over the F1 score of 94.67 per cent. Overall, 22 *same-author* curves and 84 *different-author* curves have been correctly classified as such (*true positives*). Our result thus confirms the large

[24]Luyckx.
[25]Koppel, Schler, and Bonchek-Dokow, 1265.

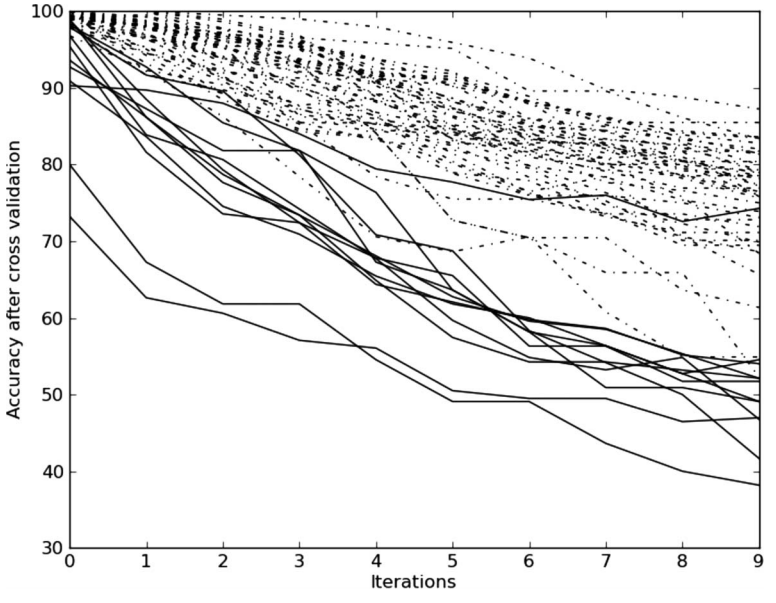

Figure 1 Unmasking Applied to Prose Texts by Five Authors.

potential of this state-of-the-art approach for authorship verification in prose work collections.

A second experiment has been carried out on the twenty-three theatrical works in the corpus. Note that the corpus contains many more theatrical works than prose works but the former tend to be shorter in length (cf. Table 1). Figure 2 displays the degradation curves resulting from the unmasking of this collection of plays. Figure 2 displays a much less clear-cut differentiation of the *same-author* curves and their *different-author* counterparts. Again, we see that a large number of *same-author* curves is concentrated in the lower region of the figure, displaying the anticipated effect that *same-author* curves are less robust to the unmasking's feature impairment. Similarly, a large number of *different-author* curves is situated in the upper part of the plot. Nevertheless, Figure 2 suggests that the unmasking approach with its default settings is less effective for the theatrical section of the corpus. The leave-one-text-out validation confirms this, yielding an overall accuracy of 83.99 per cent, an F1 score of 61.98 per cent and 20 and 405 true positive predictions for the *same-author* and *different-author* class respectively.

7. Cross-Genre Experiments

The unmasking procedure assumes that the degradation curves of *same-author* and *different-author* text pairs display different characteristics, the former being much more susceptible to feature elimination. So far, the procedure has been mainly investigated for text pairs within the same text variety, although Koppel, Schler, and

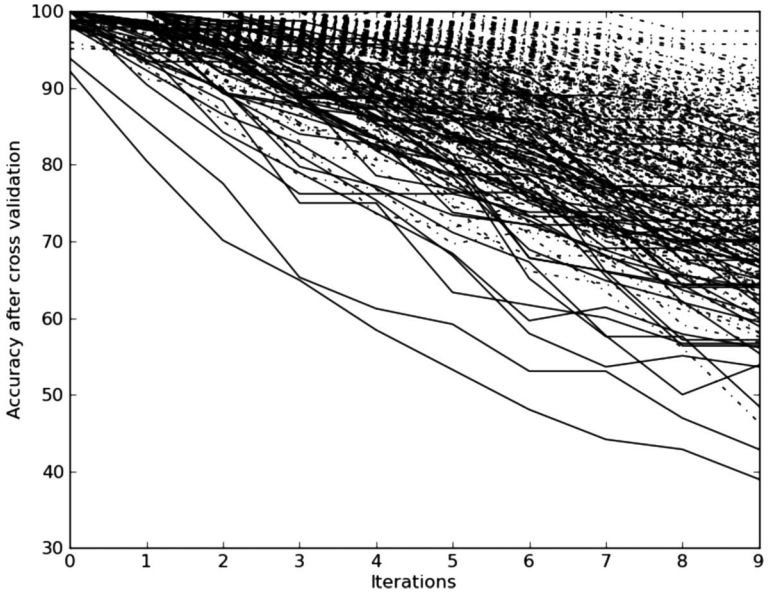

Figure 2 Unmasking Applied to Theatre Plays by Five Authors.

Bonchek-Dokow report on a successful application of the technique to Hebrew-Aramaic texts across different topics.[26] It is an interesting question whether the degradation differences between *same-author* and *different-author* curves would also hold for pairs of texts that do not belong to the same genre. To assess this question, we have carried out an experiment on the entire corpus, but considering only degradation curves for pairs of texts that belong to different genres (i.e. combination of a prose work and a play). The restriction to different-genre curves is essential in this set-up, since the curves for text pairs within and across both genres are not directly comparable.

Figure 3 shows the result for the unmasking procedure in this cross-genre set-up. Already at first sight, the results seem hardly better than for the theatrical texts. Although a number of *same-author* curves are situated in the extreme lower part of the plot, most curves are indistinguishably concentrated in the central regions of the plot, with no clear difference between *same-author* and *different-author* curves. A leave-one-text-out validation confirms the poor performance of unmasking (with its default settings) in this experiment, with an overall accuracy of 77.27 per cent and a macro-averaged F1 of 55.72 per cent. Interestingly, the number of correctly classified same-author curves is fairly high (19) when compared to the number of correctly classified different-author curves (372). Unmasking, however, when applied under its ''default'' settings, generally does not seem able to capture the overall difference between *same-author* and *different-author* text pairs across two genres in our corpus.

[26]Ibid., 1271–2.

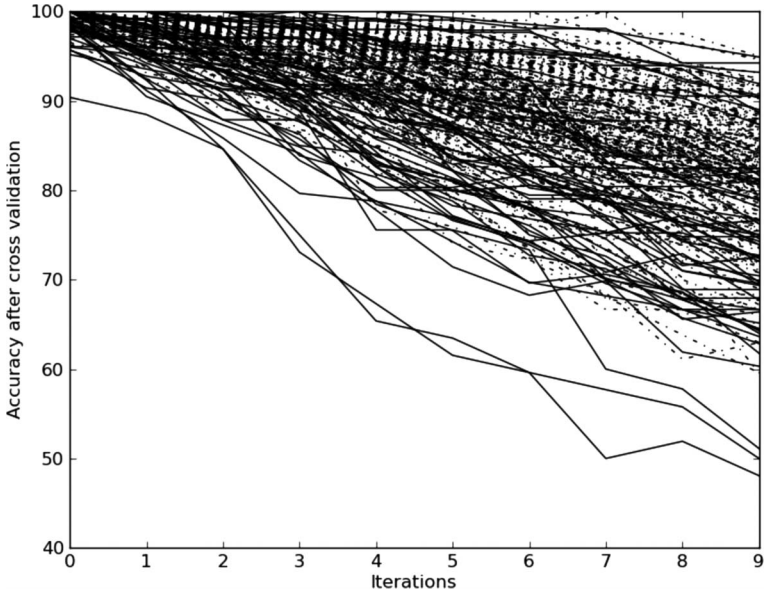

Figure 3 Unmasking Applied in a Cross-Genre Set-Up, Using Prose and Theatre Works by Five Authors.

8. Interpretation

It is interesting to have a closer look at a number of exemplary degradation curves from the experiments in the previous paragraph. The relationship between Pinter and Mamet is interesting, for instance, because these authors were personal friends and Mamet acknowledged Pinter as a key influence on his work.[27] The degradation curve (Example 1) between, for example, Pinter's play *The Caretaker* and Mamet's prose text *The Old Religion* yields the following accuracies during the degradation: 1: 100.00 per cent; 2: 100.00 per cent; 3: 100.00 per cent; 4: 100.00 per cent; 5: 99.07 per cent; 6: 98.15 per cent; 7: 97.22 per cent; 8: 95.37 per cent; 9: 93.52 per cent; 10: 91.67 per cent. Even when a large number of highly discriminative features are deleted, the classifier continues to succeed in building a high-quality model of the stylistic differences in both texts. These Mamet and Pinter texts thus appear to adopt well-distinguishable styles, notwithstanding the close stylistic affinities Mamet acknowledged between their works. However, if we confront two of Pinter's texts—Example 2: again, *The Caretaker* and his prose work *The Dwarfs*—quite another series of accuracies is obtained: 1: 100.00 per cent; 2: 97.64 per cent; 3: 96.85 per cent; 4: 93.70 per cent; 5: 89.76 per cent; 6: 83.46 per cent; 7: 81.10 per cent; 8: 77.16 per cent; 9: 74.80 per cent; 10: 74.80 per cent. Clearly, the stylistic difference between these texts—although they also belong to distinct genres—is much less robust to the

[27]Mamet.

elimination of discriminative features. Arguably, when most of the superficial differences are eliminated, the remaining features predominantly capture a similarity in authorial style between both works, making it difficult to uphold high validation accuracies.

These are of course positive examples while the cross-genre experiment also counts multiple less successful examples. Let us, for instance, consider two Mamet works: the play *Sexual Perversity in Chicago* and the prose text *The Village* (Example 3). Even though we would expect the *same-author* curve for this text pair to display clear drops in accuracy, the returned values suggest otherwise: 1: 100.00 per cent; 2: 100.00 per cent; 3: 99.19 per cent; 4: 98.39 per cent; 5: 96.77 per cent; 6: 95.97 per cent; 7: 95.16 per cent; 8: 95.16 per cent; 9: 93.55 per cent; 10: 88.71 per cent. Clearly, the classifier's performance is hardly harmed by the feature elimination. Another less successful example (for *different-author* curves, Example 4) is the text pair of Bond's play *Summer* and Wesker's prose collection *Love Letters on Blue Paper*. Although the resulting curve is of the *different-author* type, it shows pronounced drops in accuracy from iteration 3 onwards: 1: 100.00 per cent; 2: 100.00 per cent; 3: 96.87 per cent; 4: 90.62 per cent; 5: 85.94 per cent; 6: 87.50 per cent; 7: 84.37 per cent; 8: 79.69 per cent; 9: 75.00 per cent; 10: 71.87 per cent.

For each of the four previously discussed examples 1 to 4, we have generated an insightful visualization of the feature elimination process (Figures 4–7). This plot visualizes the tokens that have been iteratively eliminated during the construction of the degradation curve: the weights assigned to these features (negative as well as positive) are plotted as a function of the index of the iteration in which they were

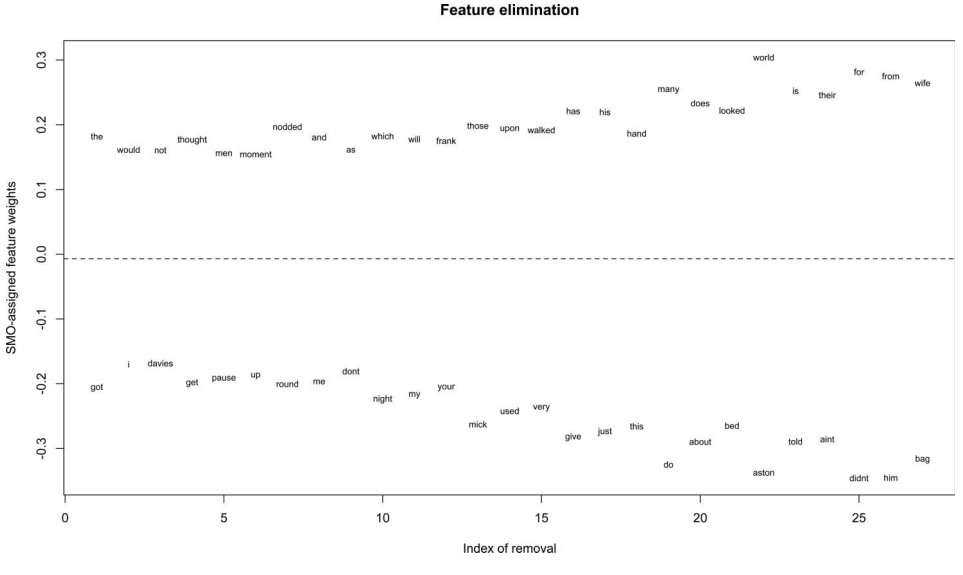

Figure 4 Visualization of the Feature Elimination Process for Pinter's Play *The Caretaker* and Mamet's Prose Text *The Old Religion* (Example 1 above).

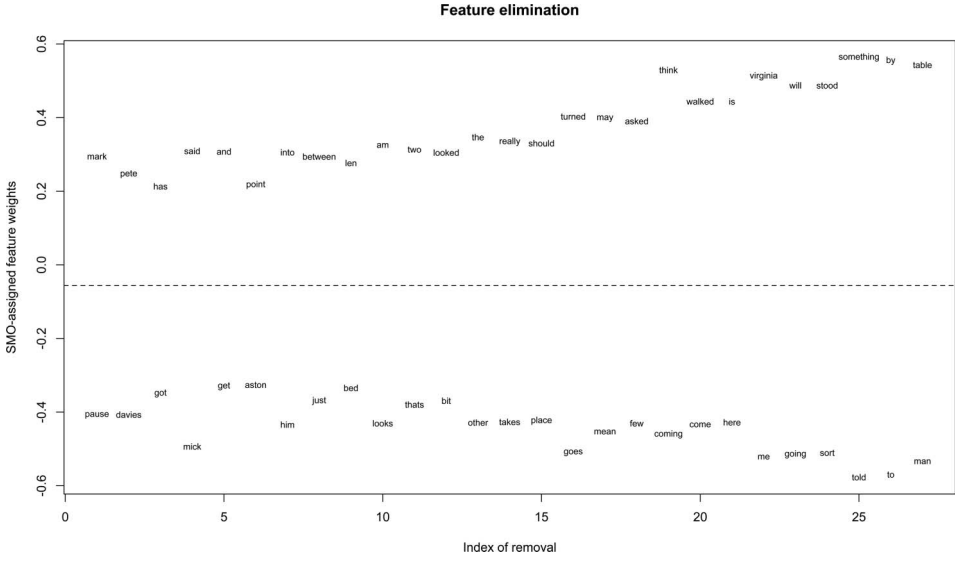

Figure 5 Visualization of the Feature Elimination Process for Pinter's *The Caretaker* (play) and his Prose Work *The Dwarfs* (prose) (Example 2 above).

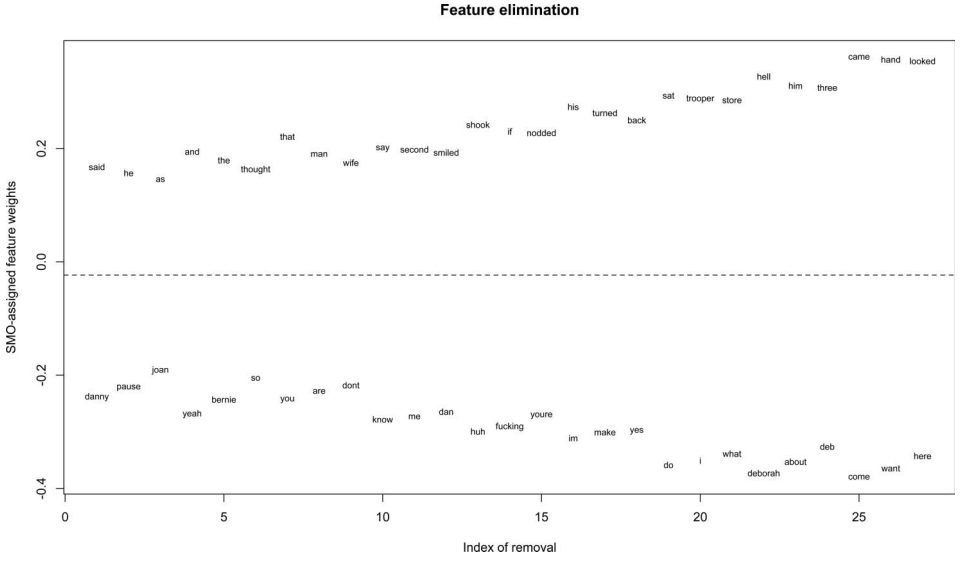

Figure 6 Visualization of the Feature Elimination Process for Two Mamet Works: The Play *Sexual Perversity in Chicago* and the Prose Text *The Village* (Example 3 above).

removed. Interestingly, words that are eliminated in a later stage tend to be assigned more extreme weights by the SMO classifier than those removed earlier—although the former are essentially less discriminative in the whole data set, since

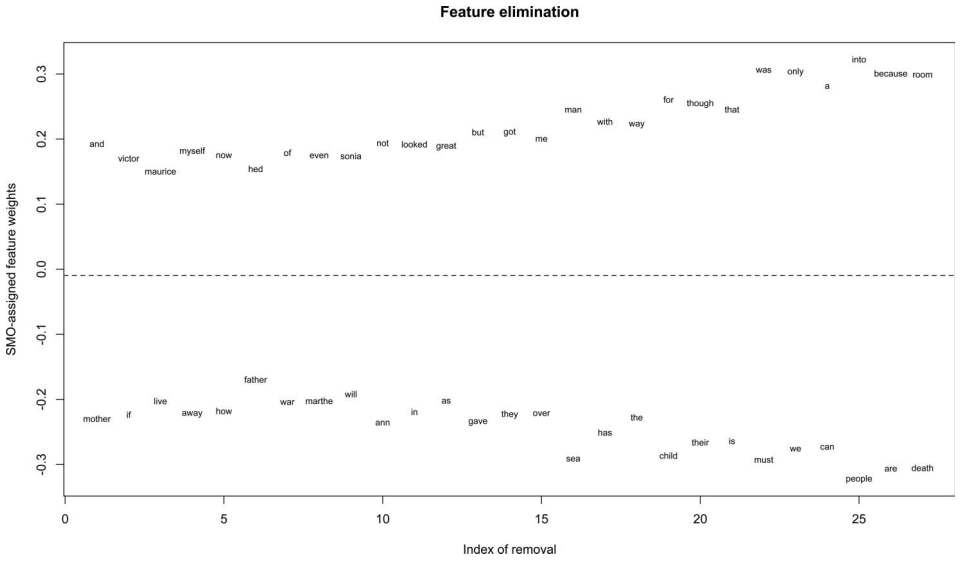

Figure 7 Visualization of the Feature Elimination Process for Bond's Play *Summer* and Wesker's Prose Collection *Love Letters on Blue Paper* (Example 4 above).

they are eliminated at later stages in the procedure. This is a fine illustration of the internal working of the unmasking procedure: the removal of highly discriminative tokens tempts the classifier into overfitting on specific, less predictive features. Precisely this effect of stylistic overfitting is likely to result in lower generalization accuracy in the case of many *same-author* text pairs, causing significant drops in the associated curve.

Note the presence of many text-specific words in the plots that are deleted at an early stage in the degradation process, such as the names of the principal characters in Figure 5 for Example 2: *davies*, *mick* and *aston* (negative features) or *mark* and *pete* (positive features) are obviously among the earliest features to be eliminated in this text pair. In the texts, such features moreover include personal pronouns that relate to a text's narrative perspective (e.g., *i* in Figure 4 for Example 1). From the point of view of cross-genre studies, it is notable that most figures moreover show multiple instances of genre-related words being deleted as anticipated above. A typical theatrical feature that is deleted concerns director's indications, such as *pause* (Figure 6 for Example 3). In general, foul and colloquial language (e.g., *fucking* or *aint*) seems to be more prominent in the direct speech of the theatre texts. Words like *nodded*, *smiled*, and *looked* are interesting prose-features that are deleted (e.g., Figure 6 for Example 3): such descriptions are obviously useless in a theatrical text where the audience will plainly *see* these actions. In prose texts, however, such behaviour needs to be explicitly described.

Interestingly, the unmasking approach seems to carry a lot of potential for (cross-genre) authorship studies but the main issue seems to be the fine-tuning of the

multiple parameters on which the procedure depends. Arguably, some *same-author* text pairs display a large number of shallow (e.g., register-related) differences that do not relate to authorial style. Consequently, in these cases, relatively many features need to be eliminated before the core of authorial style markers can be isolated. In other cases, the superficially discriminative feature can be much smaller in number for a given text pair, so that the danger exists that too many (authorship related) features are cut (yielding artificially low generalization accuracy). Especially the complex interplay between the k and n parameter seems interesting in this respect. Ideally, an additional meta-learning module should be integrated in the approach, which can automatically deduce the optimal parameter settings for a given text pair.

9. Conclusion

The experiments reported on in this paper confirm that unmasking is an interesting technique for computational authorship verification, yielding reliable results especially within the genre of (larger) prose works in our corpus. Authorship verification, however, proves much more difficult in the theatrical part of our corpus as well as in the novel cross-genre experiment reported here. The original settings for the various parameters might be genre-specific or even author-specific, so that further research is desirable into the automatic induction of the optimal experimental settings (e.g., through the use of genetic algorithms). We suspect that a dynamic parameter re-estimation for each specific text pair might be necessary.

Additionally, the technique is of theoretical interest on a number of levels that require further investigation. Whereas most comparative studies agree on the supremacy of character n-grams as the most reliable markers of authorial style,[28] unmasking so far has primarily been applied on the token level. This restriction naturally makes sense because a large number of the superficial differences between same-author text pairs are realized as individual words (e.g., names) instead of the level of word parts (e.g., inflectional endings). Nevertheless, it seems interesting to study the application of unmasking to a character n-gram text representation. Another interesting issue concerns author set size: most studies agree that experiments involving a large number of authors typically invites worse results than those that only consider a limited set of candidate authors.[29] For unmasking, however, such an effect need not necessarily be anticipated, since the classification of a curve might benefit from a larger example collection of *same-author* and *different-author* curves.

Acknowledgements

Kestemont is a PhD fellow (*aspirant*) of the Research Foundation—Flanders (FWO). The research of Luyckx and Daelemans is partially funded by the FWO project

[28]Grieve; Hirst and Feiguina; Luyckx and Daelemans, "The Effect."
[29]Koppel, Schler, and Argamon, "Authorship Attribution in the Wild"; Luyckx and Daelemans, "The Effect."

''Computational Techniques for Stylometry for Dutch''. The research by Crombez is partially funded by the FWO project ''Mass Spectacle in Flanders''. The authors would like to acknowledge Sarah Bekaert's work on the digitization of the corpus.

References

Chaski, Carole. ''Who's at the Keyboard? Authorship Attribution in Digital Evidence Investigations.'' *International Journal of Digital Evidence* 4, no. 1 (2005): 1–13.

Grieve, Jack. ''Quantitative Authorship Attribution: An Evaluation of Techniques.'' *Literary and Linguistic Computing* 22, no. 3 (2007): 251–70.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. ''The WEKA Data Mining Software: An Update.'' *SIGKDD Explorations* 11, no. 1 (2009): 10–18.

Hirst, Graeme, and Olga Feiguina. ''Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts.'' *Literary and Linguistic Computing* 22, no. 4 (2007): 405–17.

Hoover, David. ''Multivariate Analysis and the Study of Style Variation.'' *Literary and Linguistic Computing* 18, no. 4 (2003): 341–60.

Kacmarcik, Gary, and Michael Gamon. ''Obfuscating Document Stylometry to Preserve Author Anonymity.'' In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, 444–51. Sydney, Australia: International Committee on Computational Linguistics and Association for Computational Linguistics, 2006.

Koppel, Moshe, and Jonathan Schler. ''Authorship Verification as a One-class Classification Problem.'' In *Proceedings of the 21st International Conference on Machine Learning*, 489–95. Banff, Canada: Association for Computing Machinery, 2004.

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. ''Authorship Attribution in the Wild.'' *Language Resources and Evaluation* 45, no. 1 (2011): 83–94.

———. ''Computational Methods in Authorship Attribution.'' *Journal of the American Society for Information Science and Technology* 60, no. 1 (2009): 9–26.

Koppel, Moshe, Jonathan Schler, and Elisheva Bonchek-Dokow. ''Measuring Differentiability: Unmasking Pseudonymous Authors.'' *Journal of Machine Learning Research* 8 (2007): 1261–76.

Koppel, Moshe, Jonathan Schler, Shlomo Argamon, and Eran Messeri. ''Authorship Attribution with Thousands of Candidate Authors.'' In *Proceedings of the 29th International Conference of the Special Interest Group on Information Retrieval (SIGIR)*, 659–60. Seattle, USA: Association for Computing Machinery, 2006.

Lambers, Maarten, and Cor Veenman. ''Forensic Authorship Attribution Using Compression Distances to Prototypes.'' In *Proceedings of the 3rd International Workshop on Computational Forensics, Lecture Notes in Computer Science 5718*, 13–24. Berlin: Springer-Verlag, 2009.

Luyckx, Kim. *Scalability Issues in Authorship Attribution.* Brussels: University Press Antwerp, 2010.

Luyckx, Kim, and Walter Daelemans. ''Shallow Text Analysis and Machine Learning for Authorship Attribution.'' In *Computational Linguistics in the Netherlands 2004: Selected Papers from the 15th CLIN Meeting*, 149–60. Utrecht, The Netherlands: Landelijke Onderzoekschool Taalkunde, 2005.

———. ''The Effect of Author Set Size and Data Size in Authorship Attribution.'' *Literary and Linguistic Computing* 26, no. 1 (2011): 35–55.

Mamet, David. Interview. ''Landmarks,'' *Night Waves.* BBC Radio, 3 March 2005.

Platt, John. ''Fast Training of Support Vector Machines Using Sequential Minimal Optimization.'' In *Advances in Kernel Methods—Support Vector Learning,* edited by B. Schoelkopf, C. Burges, and A. Smola. Cambridge, Mass.: MIT Press, 1998.

Rudman, Joseph. "The State of Authorship Attribution Studies: Some Problems and Solutions." *Computers and the Humanities* 31, no. 4 (1997): 351–65.

———. "The State of Non-traditional Authorship Studies—2010: Some Problems and Solutions." In *Proceedings of the Digital Humanities 2010*, *London*, *UK*, 217–19. London: Alliance of Digital Humanities Organizations, 2010.

Sanderson, Conrad, and Simon Guenter. "Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation." In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sydney, Australia, 482–91. Sydney, Australia: SIGDAT/ACL, 2006.

Stamatatos, Efstathios. "A Survey of Modern Authorship Attribution Methods." *Journal of the American Society for Information Science and Technology* 60, no. 3 (2009): 538–56.

Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis. "Automatic Text Categorization in Terms of Genre and Author." *Computational Linguistics* 26, no. 4 (2000): 471–95.

Stein, Benno, Nedim Lipka, and Peter Prettenhofer. "Intrinsic Plagiarism Analysis." *Language Resources and Evaluation* 45, no. 1 (2011): 63–82.

Van Halteren, Hans, Harald Baayen, Fiona Tweedie, Marco Haverkort, and Anneke Neijt. "New Machine Learning Methods Demonstrate the Existence of a Human Stylome." *Journal of Quantitative Linguistics* 12, no. 1 (2005): 65–77.