

## An experiment in authorship attribution

Harald Baayen<sup>1</sup>, Hans van Halteren<sup>1</sup>, Anneke Neijt<sup>1</sup>, Fiona Tweedie<sup>2</sup>

<sup>1</sup> University of Nijmegen – P.O. Box 9102 – 6500 HC – Nijmegen – The Netherlands

<sup>2</sup> University of Edinburgh – James Clerk Maxwell Building – King's Buildings – Mayfield Road – Edinburgh – EH9 3JZ – U.K.

### Abstract

This paper reports an experiment in authorship attribution that reveals considerable authorial structure in texts written by authors with very similar background and training, with genre and topic being strictly controlled for. We interpret our results as supporting the hypothesis that authors have 'textual fingerprints', at least for texts produced by authors who are not consciously changing their style of writing across texts. What this study has also taught us is that discriminant analysis is a more appropriate technique to use than principal components analysis when predicting the authorship of an unknown (held-out) text on the basis of known (training) texts of which the authorial provenance is available. Finally, standard discriminant analysis can be enhanced considerably by using an entropy-based weighting scheme of the kind used in latent semantic analysis (Landauer et al., 1998).

**Keywords:** authorship attribution, principal components analysis, discriminant analysis, latent semantic analysis

### 1. Introduction

Stylometric attempts to trace the authorship of texts by unknown or contested authors have a long history. They have been applied to influential texts such as the Bible, the works of Shakespeare, and the Federalist Papers. A wide variety of techniques from many disciplines have been considered, from multivariate statistical analysis to neural networks and machine learning. Many different facets of texts have been analysed, from sentence and word length to the most common or the rarest words, or linguistic features. (Holmes, 1998) provides a chronological review of methods used in the pursuit of the authorial "fingerprint".

A key issue in the study of authorship studies is whether authorial "fingerprints" do in fact exist. Is it truly the case that any two authors can always be distinguished on the basis of their style, so that stylometry can provide unique stylistic fingerprints for any author, given sufficient data?

Despite the long history of authorship attribution, almost all stylometric studies have been carried out on the assumption that stylometric fingerprinting is possible. However, often control texts are inappropriately chosen or not available. In addition, the imposition of editorial or publisher's style can distort the original words of the author. (Rudman, 1998) provides an excellent overview of the problems and pitfalls that characterize quantitative approaches to authorship attribution. A further complication is that in traditional problems of authorship in literary studies, the authors to which texts of doubtful or unknown provenance have to be attributed tend to be major writers, and more often than not they differ substantially with respect to background and training. What we do not know is to what extent naive writers unwittingly imprint their texts with their own specific authorial fingerprint. Does success in quantitative authorship attribution

depend on authors having deliberately developed and cultivated their own style? Or is style a marker that can be used to trace authorship even for non-specialist writers with a very similar training and background?

The aim of this study is to address this question by means of a strictly controlled experiment of authorship attribution, with texts of known authorship about strictly controlled topics being analysed between and within genres as well as between and within authors. The goal of this experiment is to gauge to what extent non-professional authors with a very similar background and training can be distinguished on the basis of their written production, under conditions in which topic, age, and incentive are strictly controlled.

In what follows, we first describe the design of the experiment. We then proceed with the statistical analyses of the experimental data. Finally, we summarize our conclusions.

## **2. Experimental design**

Our experiment in authorship attribution was carried out in Dutch. Eight students of Dutch literature at the University of Nijmegen participated in the study. All the students were native speakers of Dutch, four were in their first year of study, and four were in their fourth year. The students were asked to write texts of around 1000 words. They were paid for their participation. To encourage serious participation, the best text in each genre was awarded a prize of Hfl 125.

Each student wrote in three genres: fiction, argument and description. Three texts were written in each genre, on the following topics. For the fiction, we asked our participants to write a retelling of the fairy tale of Little Red Riding-Hood, to write a detective story about a murder in the university, and to compose a romance of chivalry. For the argumentative texts, their task was to defend a position about the television program 'Big Brother', to write about the unification of Europe, and to take a position about the health risks of smoking. The descriptive texts concerned football, the (then) upcoming new millennium, and a book-review of the book read most recently by the participant. The order of writing the texts was randomised so that practice effects were reduced as much as possible. We thus have nine texts from each participant, making a total of seventy-two texts in the analysis. All texts were produced in the same week. There was no contact between the participants. Participants were not allowed to consult encyclopedias or dictionaries.

The main question that will concern us here is whether it will be possible to group texts by their authors using the state-of-the-art methods of stylometry. A positive answer would support the hypothesis that stylistic fingerprints exist, even for authors with a very similar background and training. A negative answer would argue against the hypothesis that each author has her/his unique stylistic fingerprint.

## **3. Statistical analysis**

The average text length for our 72 elicited texts is 908 words. The shortest text has 628 words and the longest 1342. A principal components analysis (PCA) along the lines of (Burrows, 1992) of the most frequent function words in the texts shows no authorial structure. This is illustrated shown in upper panels of Figure 1. The upper left panel plots authors in the plane spanned by the first two principal components, the upper right panel shows their location in the plane of the second and third principal component. Higher principal components reveal similar random patterns. By contrast, a principal components analysis does reveal some structure for education

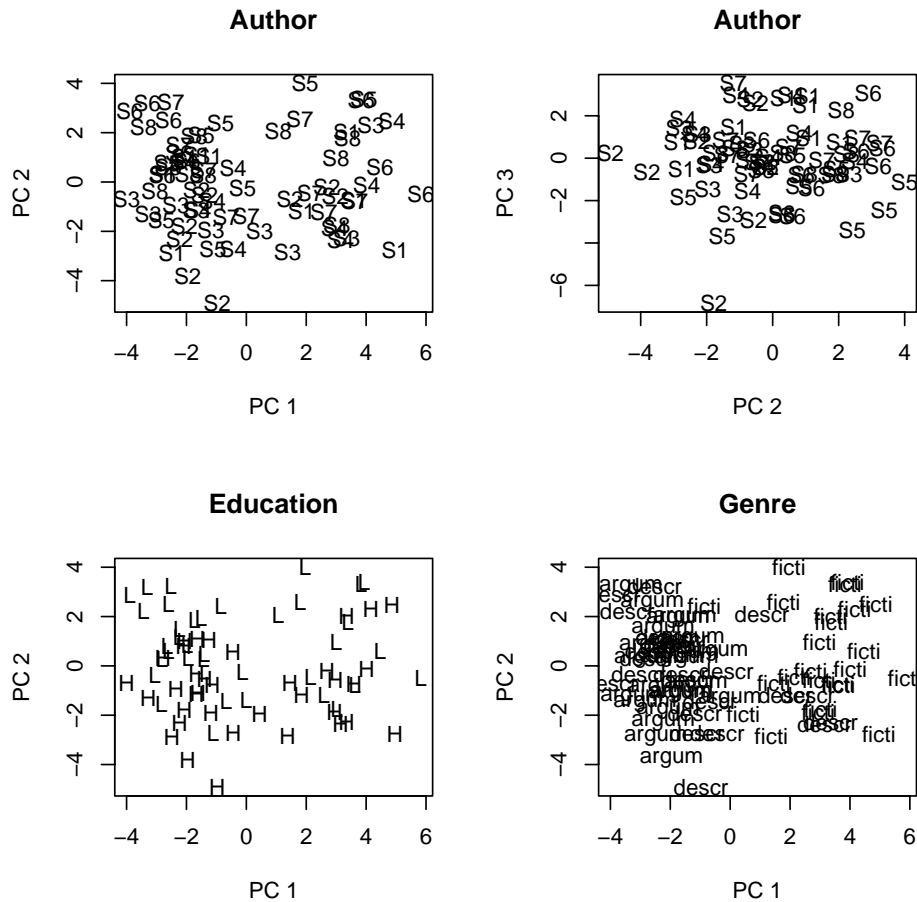


Figure 1: *Principal components analysis of the 50 most frequent function words in the texts produced in the experiment. The upper two panels plot authors in the planes spanned by principal components 1 and 2 (left) and 2 and 3 (right). The lower left panel plots education level (High versus Low) for the first two principal components. The lower right panel plots genre for the same principal components.*

level, as can be seen in the lower left panel of Figure 1. Texts written by students with less than 1 year of university education (L) tend to have higher scores on the second principal component than texts written by students with at least three years of university education (H). The lower right panel of Figure 1 shows that there is some separation by genre as well. The fiction texts tend to have higher loadings on the first principal component than the argumentative texts and the descriptive texts. In hindsight, our instructions for these two genres were probably not specific enough to allow a reasonable separation to emerge.

Analysis of letter frequencies gives similar results, while measures of vocabulary richness show some indication of structure with respect to the education level of the writer. First year of studies appear to have lower values of Yule's K (Yule, 1944), and hence a lower repeat-rate. In addition, higher values of Orlov's Z (Orlov, 1983) are the province of first-year students also, indicating a greater richness of vocabulary.

While authorial structure is not visible in a principal components analysis, it does emerge from a series of linear discriminant analyses (LDA). In contrast to PCA, LDA allows us to bring into

the analysis our knowledge of the authorship of our texts. To obtain a measure of discrimination accuracy, we carried out pairwise leave-one-out cross-validation using linear discriminant prediction of the authorship of a held-out text on the basis of a training set of texts with known authorship. For each of the  $8 \times 7 / 2 = 28$  pairs of authors, the authorship of each of the 18 texts contributed by these two authors was estimated on the basis of the 17 remaining texts. The discriminability of a pair of authors is the proportion of correctly attributed texts. The overall discrimination score is the average of the pairwise discriminability scores. In order to avoid problems with collinear predictors, we orthogonalized our data matrix before carrying out the linear discriminant classification.

The first column labeled LDA in Table 1 reports the resulting accuracy scores averaged over the 28 pairwise comparisons involved in one standard cross-validation sequence, for 3 differently-sized sets of highest-frequency function words. Note that this approach is not successful, with an average accuracy at chance level. Thus far, the LDA analysis and the principal components analysis both suggest that our authors are too similar in background and training to allow them to be distinguished on the basis of their texts.

However, the cross-validation procedure described above has the disadvantage that there is an imbalance in the coverage of topics and genres between the texts of two authors. For instance, if the fiction text about Little Red Riding-Hood happens to be the left-out text for one author, then the presence of the corresponding text for the other author implies that the number of training texts for the two authors differ with respect to the support for fiction texts (2 for the one and 3 for the other author), and with respect to the topics covered (one author's version of the fairy tale is missing). To counterbalance for this asymmetry, we modified the cross-validation procedure by taking the corresponding text of the second author out of the training set. Thus, if the held-out text is the fairy tale, then the fairy tale of the second author is excluded from the training set as well. The success scores for standard linear discriminant analysis under this modified cross-validation regime are shown in the third column of Table 1, see also Figure 2. Removal of the text and genre imbalance leads to an improvement in classification accuracy of roughly 10%.

Additional improvements are obtained by modifying the input to the linear discriminant analysis. Landauer and his colleagues ((Landauer and Dumais, 1997; Landauer et al., 1998)) have pointed out in the context of corpus-based semantic analysis that linear discriminant analysis can be substantially enhanced by weighting the vectors describing the frequency distributions of words across texts for their by-text entropy. Consider the data matrix,  $M = (f_{ij})$ , with  $f_{ij}$  the frequency of word  $i$  in text  $j$ . The intuition underlying entropy-weighted linear discriminant analysis (ELDA) is that words that have a more or less uniform frequency distribution across the texts are less useful and less informative than words that have a decidedly non-uniform distribution. For ELDA, we use as data matrix  $M' = (f'_{ij})$  instead of  $M$ , with

$$f'_{ij} = \frac{\log(f_{ij} + 2)}{-\sum_{k=1}^J \left( \frac{f_{ik} + 1}{\sum_{k=1}^J f_{ik} + 1} \cdot \log\left(\frac{f_{ik} + 1}{\sum_{k=1}^J f_{ik} + 1}\right) \right)}, \quad (1)$$

with  $J$  the number of texts. Note that we used the transposed versions of  $M$  and  $M'$  as input to the discriminant analysis. Table 1 and Figure 2 show that the use of ELDA leads to a substantial improvement in classification accuracy. For a data matrix with 60 function words, ELDA with the modified cross-validation scheme yields an accuracy of 81.5%.

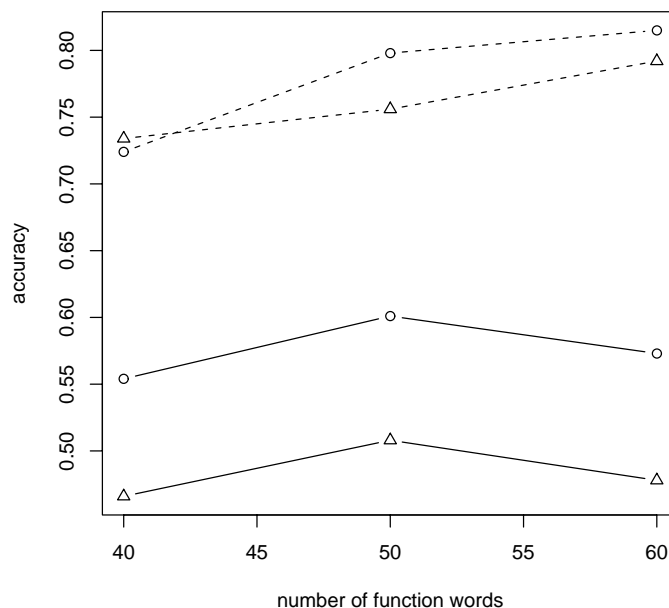


Figure 2: Cross-validation accuracy scores for pairwise linear discriminant analyses. Solid lines represent standard linear discriminant analysis, dashed lines represent entropy-enhanced linear discriminant analysis. The triangles represent standard leave-one out estimates, the circles represent estimates in which the training set for pairwise comparisons does not contain the text from the other author on the topic covered by the held-out text.

$n$	standard CV		modified CV	
	LDA	ELDA	LDA	ELDA
40	0.466	0.734	0.554	0.724
50	0.508	0.756	0.601	0.798
60	0.478	0.792	0.573	0.815

Table 1: Pairwise standard and modified cross-validation accuracy scores for standard linear discriminant analysis (LDA) and entropy-enhanced linear discriminant analysis (ELDA).

It turns out that enriching the data matrix with the frequencies of punctuation marks leads to a further improvement. An ELDA analysis on 42 function words and 8 punctuation marks leads to a classification accuracy of 83.5%, and a data matrix with 50 function words and 8 punctuation marks yields an ELDA accuracy of 88.1%. Figure 3 shows the kind of clustering that is revealed by this last analysis. Note that the only participants that have overlapping clusters of texts are subjects 1 and 5, at the center right hand edge of the figure.

## 4. Conclusions

This experiment in authorship attribution shows that there is considerable authorial structure in written texts even when the authors of these texts come from very similar backgrounds. This surprising result provides support for the hypothesis that authors may have 'textual fingerprints', at least for texts produced by writers who are not consciously changing their style of writing across texts.

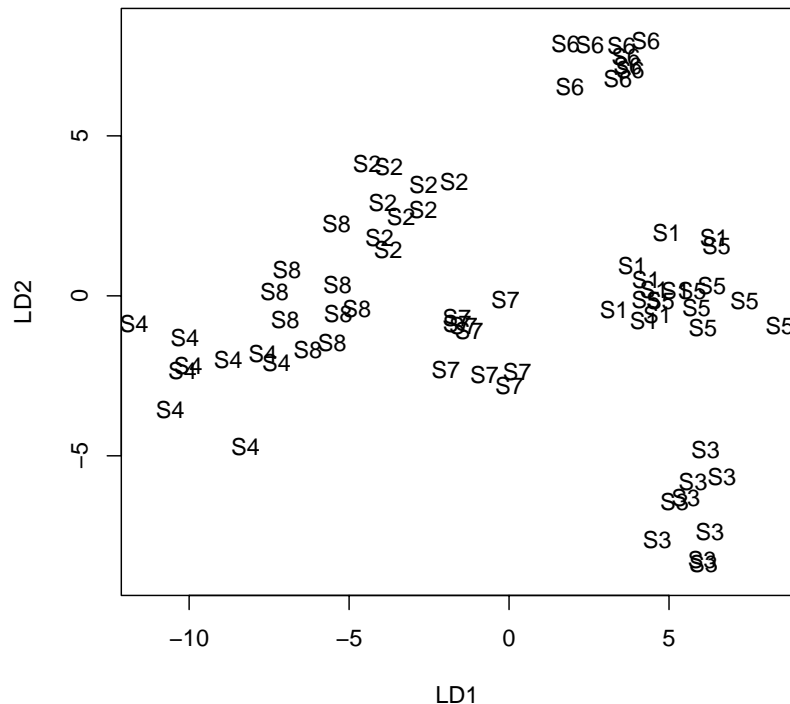


Figure 3: *Linear discriminant analysis with entropy weighting for the 50 most frequent function words in the elicited texts, combined with 8 punctuation marks (.,'?!:;).*

What this study has also taught us is that discriminant analysis is a more appropriate technique to use than principal components analysis when predicting the authorship of an unknown (held-out) text on the basis of known (training) texts of which the authorial provenance is available, and that, as claimed by Landauer and his colleagues, entropy-weighting does indeed lead to a substantial increase in classification accuracy.

Our finding that a simple principal components analysis of the highest-frequency function words in our experimental texts fails to uncover authorial structure suggests that the authors studied in literary stylometry, for which principal components analysis is reported to lead to insightful clustering (Burrows, 1992; Baayen et al., 1996), are indeed writers with very different backgrounds and training. These are authors who must have developed their own writing style far beyond the more rudimentary differences in style that we could only observe for our participants by using far more powerful analytical tools than simple principal components analysis.

The observation that in cross-validation a strict control of topic and genre leads to an increase in classificatory accuracy of some 10% provides further support for Rudman's claim that control texts in literary stylometry require rigorous matching with respect to variables such as topic and genre. Although the use of function words rather than content words makes it possible to focus on text style rather than textual content, our data suggest that style and content are intertwined to a greater extent than we had previously thought. Finally, we were surprised by the extent to which the simple inclusion of punctuation marks in the analysis enhanced classification accuracy. Punctuation marks may prove to be effective style markers, especially for texts that have not been subjected to editorial normalization.

## References

- Baayen R. H., Van Halteren H., and Tweedie F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11:121–131.
- Burrows J. F. (1992). Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7:91–109.
- Holmes D. I. (1998). Authorship attribution. *Literary and Linguistic Computing*, 13(3):111–117.
- Landauer T. K. and Dumais S. (1997). A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.
- Landauer T. K., Foltz P. W., and Laham D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.
- Orlov J. K. (1983). Ein Model der Häufigkeitsstruktur des Vokabulars. In Guiter H. and Arapov M. V. editors, *Studies on Zipf's Law*, pages 154–233. Brockmeyer, Bochum.
- Rudman J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31:351–365.
- Tweedie F. and Baayen R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32:323–352.
- Yule G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge.