Automatic Authorship Attribution

E. Stamatatos, N. Fakotakis and G. Kokkinakis
Dept. of Electrical and Computer Engineering
University of Patras
26500 – Patras
GREECE
stamatatos@wcl.ee.upatras.gr

Abstract

In this paper we present an approach to automatic authorship attribution dealing with real-world (or unrestricted) text. Our method is based on the computational analysis of the input text using a text-processing tool. Besides the style markers relevant to the output of this tool we also use analysis-dependent style markers, that is, measures that represent the way in which the text has been processed. No word frequency counts, nor other lexically-based measures are taken into account. We show that the proposed set of style markers is able to distinguish texts of various authors of a weekly newspaper using multiple regression. All the experiments we present were performed using real-world text downloaded from the World Wide Web. Our approach is easily trainable and fully-automated requiring no manual text preprocessing nor sampling.

1 Introduction

The vast majority of the attempts to computer-assisted authorship attribution has been focused on literary texts. In particular, a lot of attention has been paid to the establishment of the authorship of anonymous or doubtful texts. A typical paradigm is the case of the *Federalist papers* twelve of which are of disputed authorship (Mosteller and Wallace, 1984; Holmes and Forsyth, 1995). Moreover, the lack of a generic and formal definition of the idiosyncratic style of an author has led to the employment of statistical methods (e.g., discriminant analysis, principal components, etc.). Nowadays, the wealth of text available in the World Wide Web in electronic form for a wide variety of genres and languages, as well as the development of reliable text-processing tools open the way for the solution of the authorship attribution problem as regards real-world text.

The most important approaches to authorship attribution involve lexically based measures. A lot of style markers have been proposed for measuring the richness of the vocabulary used by the author. For example, the type-token ratio, the *hapax legomena* (i.e., once-occurring words), the *hapax dislegomena* (i.e., twice-occurring words), etc. There are also functions that make use of these measures such as Yule's *K* (Yule, 1944), Honore's *R* (Honore, 1979), etc. A review of this metrics can be found in (Holmes, 1994). In (Holmes and Forsyth, 1994) five vocabulary richness functions were used in the framework of a multivariate statistical analysis of the *Federalist papers* and a principal components analysis was performed. All the disputed papers lie in the side of James Madison (rather than Alexander Hamilton) in the space of the first two principal components. However, such measures require the development of large lexicons with specialized information in order to detect the various forms of the lexical units that constitute an author's vocabulary. For languages with a rich morphology, i.e. Modern Greek, this is an important shortcoming.

Instead of counting how many words occur certain number of times, Burrows (1987) proposed the use of a set of common function (or context-free) word frequencies in the sample text. This method combined with a principal components analysis achieved remarkable results when applied to a wide variety of authors (Burrows, 1992). On the other hand, a lot of effort is required regarding the selection of the most appropriate set of words that best distinguish a given set of authors (Holmes and Forsyth, 1995). Moreover, all the lexically-based style markers are highly author and language dependent. The results of a work using such measures, therefore, can not be applied to a different group of authors nor another language.

In order to avoid the problems of lexically-based measures, (Baayen, *et al.*, 1996) proposed the use of syntax-based ones. This approach is based on the frequencies of the rewrite rules as they appear in a syntactically annotated corpus. Both high-frequent and low-frequent rewrite rules give accuracy results comparable to lexically-based methods. However, the computational analysis is considered as a significant limitation of this method since the required syntactic annotation scheme is very complicated and current text-processing tools are not capable of providing automatically such information, especially in the case of unrestricted text.

To the best of our knowledge, there is no computational system for the automatic detection of authorship dealing with real-world text. In this paper, we present an approach to this problem. In particular, our aim is the discrimination between the texts of various authors of a Modern Greek weekly newspaper. We use an already existing text processing tool able to detect sentence and chunk boundaries in unrestricted text for the extraction of style markers. Instead of trying to minimize the computational analysis of the text, we attempt to take advantage of this procedure. In particular, we use a set of analysis-level style markers, i.e., measures that represent the way in which the text has been processed by the tool. For example, a useful measure is the percentage of the sample text remaining unanalyzed after the automatic processing. In other words, we attempt to adapt the set of the style markers to the method used by the sentence and chunk detector in order to analyze the sample text. The statistical technique of multiple regression is, then, used for extracting a linear combination of the values of the style markers that manages to distinguish the different authors. The experiments we present, for both author identification and author verification tasks, were performed using real-world text downloaded from the World Wide Web. Our approach is easily trainable and fully automated requiring no manual text preprocessing nor sampling.

A brief description of the extraction of the style markers is given in section 2. Section 3 describes the composition of the corpus of real-world text used in the experiments. The training procedure is given in section 4 while section 5 comprises analytical experimental results. Finally, in section 6 some conclusions are drawn and future work directions are given.

2 Extraction of Style Markers

As aforementioned, an already existing tool is used for the extraction of the style markers. This tool is a Sentence and Chunk Boundaries Detector (SCBD) able to deal with unrestricted Modern Greek text (Stamatatos, *et al.*, forthcoming). Initially, SCBD segments the input text into sentences using a set of disambiguation rules, and then detects the boundaries of intrasentential phrases (i.e., chunks) such as noun phrases, prepositional phrases, etc. It has to be noted that SCBD makes use of no complicated resources (e.g., large lexicons). Rather, it is based on common word suffixes and a set of keywords in order to detect the chunk boundaries using empirically derived rules. A sample of its output is given below:

> VP[Δεν θέλω να ρίξω] NP[λάδι] PP[στη φωτιά] CON[αλλά] VP[πιστεύω] CON[ότι] NP[η επιβάρυνση] PP[στον προϋπολογισμό] PP[από τους βουλευτές] VP[δεν μπορεί να προσμετρείται] μόνο PP[με τα '5 *δισ. *δρχ. των αναδρομικών] που NP[πήραν τελευταία] VP[προκαλώντας] NP[τη δυσφορία της κοινής γνώμης] .

Based on the output of this tool, the following measures are provided:

- **Token-level:** sentence count, word count, punctuation mark count, etc.
- **Phrase-level:** noun phrase count, word included in noun phrases count prepositional phrase count, word included in prepositional phrases count etc.

In addition, we use measures relevant to the computational analysis of the input text:

Table 1. The Corpus Consisting of Texts Taken from the Weekly Newspaper *TO BHMA*.

Code	Author name	Texts	Total words	Thematic area
A01	D. Maronitis	20	11,771	Culture, society
A02	M. Ploritis	20	22,947	Culture, history
A03	K. Tsoukalas	20	30,316	International affairs
A04	C. Kiosse	20	34,822	Archeology
A05	S. Alachiotis	20	19,162	Biology
A06	G. Babiniotis	20	25,453	Linguistics
A07	T. Tasios	20	20,973	Technology, society
A08	G. Dertilis	20	18,315	History, society
A09	A. Liakos	20	25,826	History, society
A10	G. Vokos	20	20,049	Philosophy

- **Analysis-level:** unanalyzed word count after each pass, keyword count, non-matching word count, and assigned morphological descriptions for both words and chunks.

The latter measures can be calculated only when this particular computational tool is utilized. In more detail, SCBD performs multiple pass parsing (i.e., 5 passes). Each parsing pass analyzes a part of the sentence, based on the results of the previous passes, and the remaining part is kept for the subsequent passes. The first passes try to detect the simplest cases of the chunk boundaries which are easily recognizable while the last ones deal with more complicated cases using the findings of the previous passes. The percentage of the words remaining unanalyzed after each parsing pass, therefore, is an important stylistic factor that represents the syntactic complexity of the text. Additionally, the measure of the detected keywords and the detected words that do not match any of the stored suffixes include crucial stylistic information.

The vast majority of the natural language processing tools can provide analysis-level style markers. However, the manner of capturing the stylistic information may differ since it depends on the method of analysis.

In order to normalize the calculated style markers we make use of ratios of them (e.g., words / sentences, noun phrases / total detected chunks, words remaining unanalyzed after parsing pass 1 / words, etc.). The total set of style markers comprises 22 markers, namely: 3 token-level, 10 phrase-level, and 9 analysis-level ones.

3 Corpus

The corpus used for this study consists of texts downloaded from the World Wide Web-site of the Modern Greek weekly newspaper *TO BHMA* (Dolnet, 1998). This newspaper comprises several supplements. We chose to deal with authors of the supplement B, entitled *ΝΕΕΣ ΕΠΟΧΕΣ* (i.e., new ages), which comprises essays on science, culture, history, etc. since in such writings the indiosyncratic style of the author is not likely to be overshadowed by the characteristics of the corresponding text-genre. In general, the texts included in the supplement B are written by scholars, writers, etc., rather than journalists. Moreover, there is a closed set of authors that regularly publish their writings in the pages of this supplement. The collection of a considerable amount of texts by an author was, therefore, possible.

Initially, we selected 10 authors whose writings are frequently published in this supplement. No special criteria have been taken into account. Then, 20 texts of each author were downloaded from the Web-site of the newspaper. No manual text preprocessing nor text sampling was performed aside from removing unnecessary headings. All the downloaded texts were taken from issues published during 1998 in order to minimize the potential change of the personal style of an author over time. Some statistics of the downloaded corpus are shown in table 1. The last column of this table refers to the thematic area of the majority of the writings of each author. Notice that this information was not taken into account during the construction of the corpus.

4 Training

The corpus described in the previous section was divided into a training and a test corpus. As it is shown by Biber (1990; 1993), it is possible to represent the distributions of many core linguistic features of a stylistic category based on relatively few texts from each category (i.e., as few as ten texts). Thus, for each author 10 texts were used for training and 10 for testing. All the texts were analyzed using SCBD which provided a vector of 22 style markers for each text. Then, the statistical methodology of multivariate linear multiple regression was applied to the training corpus. Multiple regression provides predicting values of a group of *response* (dependent) variables from a collection of *predictor* (independent) variable values. The response is expressed as a linear combination of the predictor variables, namely:

$$y_i = b_0 + z_1 b_{1i} + z_2 b_{2i} + \ldots + z_r b_{ri} + e_i$$

where y_i is the response for the i-th author, z_1, z_2,..and z_r are the predictor variables (i.e., in our case r=22), b_0, b_{1i}, b_{2i},..., and b_{ri}, are the unknown coefficients, and e_i is the random error. During the training procedure the unknown coefficients for each author are determined using binary values for the response variable (i.e., 1 for the texts written by the author in question, 0 for the others). Thus, the greater the response variable of a certain author, the more likely to be the author of the text.

Some statistics measuring the degree to which the regression functions fit the training data are presented in table 2. Notice that R^2 is the *coefficient of determination* defined as follows:

$$R^2 = \frac{\sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2}$$

where n is the total number of training data (texts), $\bar{y}$ is the mean response, $\hat{y}_j$ and y_j are the estimated response and the training response value of the j-th author respectively. Additionally, a significant F-value implies that a statistically significant proportion of the total variation in the dependent variable is explained.

Table 2. Statistics of the Regression Functions.

Code	R^2	F Value
A01	0.40	2.32
A02	0.72	9.12
A03	0.44	2.80
A04	0.44	2.80
A05	0.32	1.61
A06	0.51	3.57
A07	0.59	5.13
A08	0.35	1.87
A09	0.53	4.00
A10	0.63	5.90

It has to be noted that we use this particular discrimination method due to the facility offered in the computation of the unknown coefficients as well as the computationally simple calculation of the predictor values. However, we believe that any other methodology for discrimination-classification can be applied (e.g., discriminant analysis, neural networks, etc.).

5 Performance

Before proceeding to the presentation of the analytical results of our disambiguation method, a representation of the test corpus into a dimensional space would illustrate the main differences and similarities between the authors. Towards this end, we performed a principal components analysis and the representation of the 100 texts of the test corpus in the space defined by the first and the second principal components (i.e., accounting for the 43% of the total variation) is depicted in figure 1. As can be seen, the majority of the texts written by the same author tend to cluster. Nevertheless, these clusters cannot be clearly separated.

According to our approach, the criterion for identifying the author of a text is the value of the response linear function. Hence, a text is classified to the author whose response value is the greatest. The confusion matrix derived from the application of the disambiguation procedure to the test corpus is presented in table 3, where each row contains the responses for the ten test texts of the corresponding author. The last column refers to the identification error (i.e.,

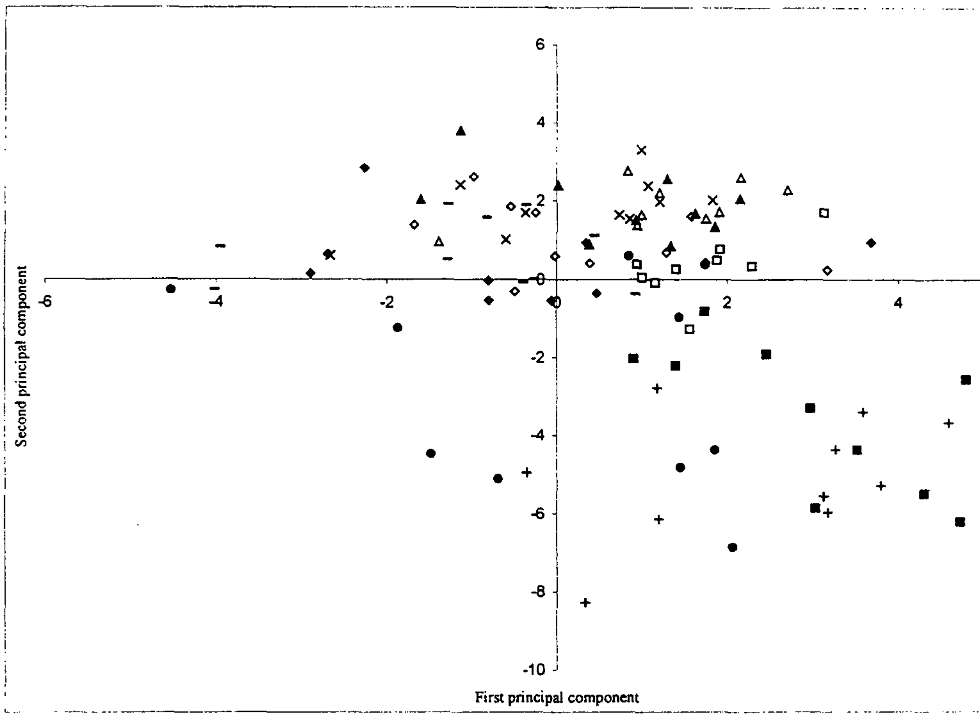

Figure 1. The Test Corpus in the Space of the First Two Principal Components.

Table 3. Confusion Matrix of the Author Identification Experiment.

Actual	Guess										
	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	Error
A01	3	2	0	0	2	0	0	2	0	1	0.7
A02	0	10	0	0	0	0	0	0	0	0	0.0
A03	0	0	8	0	0	0	0	1	0	1	0.2
A04	0	0	0	9	0	0	0	0	0	1	0.1
A05	0	0	0	3	3	1	0	0	3	0	0.7
A06	2	1	0	0	0	7	0	0	0	0	0.3
A07	0	0	0	0	0	0	10	0	0	0	0.0
A08	1	2	0	1	0	2	0	4	0	0	0.6
A09	0	0	0	0	0	0	0	1	9	0	0.1
A10	0	0	2	1	1	0	0	0	0	6	0.4
									Average		0.31

erroneously classified texts / total texts) for each author. Approximately 65% of the average identification error corresponds to three authors, namely: A01, A05, and A08. Notice that these are the authors with an average text-size smaller than 1,000 words (see table 1). It appears, therefore, that a text sample of relatively short size (i.e., less than 1,000 words) is not adequate for the representation of the stylistic characteristics of an author's style. Notice that similar conclusions are drawn by Biber (1990; 1993).

Instead of trying to identify who the author of a text is, some applications require the verification of the hypothesis that a given person is the author of the text. In such a case, only the response function of the author in question is involved. Towards this end, a threshold value has to be defined for each response function. Thus, if the response value for the given author is greater than the threshold then the author is accepted.

Additionally, for measuring the accuracy of the author verification method as regards a

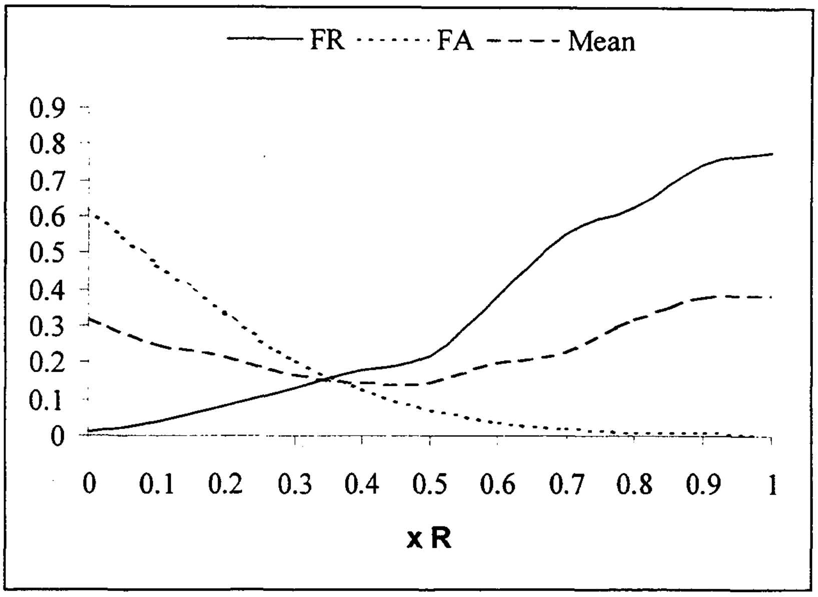

Figure 2. FR, FA, and Mean Error as Functions of Subdivisions of *R*.

certain author, we defined False Rejection (FR) and False Acceptance (FA) as follows:

$$FR = \frac{\text{rejected texts of the author}}{\text{total texts of the author}}$$

$$FA = \frac{\text{accepted texts of the author}}{\text{total text of other authors}}$$

Similar measures are widely utilized in the area of speaker recognition in speech processing (Fakotakis, *et al.*, 1993).

The *multiple correlation coefficient* $R = +\sqrt{R^2}$ of a regression function (see table 2) equals 1 if the fitted equation passes through all the data points. At the other extreme, it equals 0. The fluctuation of average FR, FA, and mean error (i.e., (FR+FA)/2) for the entire test corpus using subdivisions of *R* as threshold (x-axis) is shown in figure 2, and the minimum mean error corresponds to *R/2*. Notice that by choosing the threshold based on the minimal mean error the majority of applications is covered. On the other hand, some applications require either minimal FR or FA, and this fact has to be taken into account during the selection of the threshold.

The results of the author verification experiment using *R/2* as threshold are presented in table 4. Approximately 70% of the total false rejection corresponds to the authors A01, A05, A08 as in the case of author identification. On the other hand, false acceptance seems to be highly relevant to the threshold value. The smaller the threshold value, the greater the false acceptance. Thus, the authors A03, A04, A05, and A08 are responsible for 72% of the total false acceptance error.

Table 4. Author Verification Results (threshold=*R/2*).

Code	R/2	FR	FA
A01	0.32	0.3	0.022
A02	0.42	0.0	0.044
A03	0.33	0.0	0.155
A04	0.33	0.1	0.089
A05	0.28	0.6	0.144
A06	0.36	0.2	0.011
A07	0.38	0.0	0.022
A08	0.30	0.6	0.100
A09	0.36	0.0	0.055
A10	0.40	0.4	0.033
Average	0.35	0.22	0.068

Finally, the total time cost (i.e., text processing by SCBD, calculation of style markers, computation of response values) for the entire test corpus was 58.64 seconds, or 1,971 words per second, using a Pentium at 350 MHz.

6 Conclusions

We presented an approach to automatic authorship attribution of real-world texts. A computational tool was used for the automatic extraction of the style markers. In contrast to other proposed systems we took advantage of this procedure in order to extract analysis-level style markers that represent the way in which the text has been analyzed. The experiments based on texts taken from a weekly Modern Greek newspaper prove that the stylistic differences among a wide range of authors can be easily detected using the proposed set of style markers. Both author identification and author verification tasks have given encouraging results.

Moreover, no lexically-based measures, such as word frequencies, are involved. This approach can be applied to a wide-variety of authors and types of texts since any domain-dependent, genre-dependent, author-dependent style marker have not been taken into account. Although our method has been tested on Modern Greek, it requires no language-specific information. The only prerequisite of this method in order to be employed in another language is the availability of a text-processing tool of general purpose and the appropriate selection of the analysis-level measures.

The presented approach is fully-automated since it is not based on specialized text preprocessing requiring manual effort. Nevertheless, we believe that the accuracy results may be significantly improved by employing text-sampling procedures for selecting the parts of text that best illustrate the stylistic features of an author.

Regarding the amount of required training data, we proved that ten texts are adequate for representing the stylistic features of an author. Some experiments we performed using more than ten texts as training corpus for each author did not improved significantly the accuracy results. It has been also shown that a lower bound of the text-size is 1,000 words. Nevertheless, we believe that this limitation affects mainly authors with vague stylistic characteristics.

We are currently working on the application of the presented methodology to text-genre detection as well as to any stylistically homogeneous group of real-world texts. We also aim to explore the usage of a variety of computational tools for the extraction of analysis-level style markers for Modern Greek and other natural languages.

References

Baayen, H., H. Van Halteren, and F. Tweedie 1996, Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution, *Literary and Linguistic Computing*, 11(3): 121-131.

Biber, D. 1990, Methodological Issues Regarding Corpus-based Analyses of Linguistic Variations, *Literary and Linguistic Computing*, 5: 257-269.

Biber, D. 1993, Representativeness in Corpus Design, *Literary and Linguistic Computing*, 8: 1-15.

Burrows, J. 1987, Word-patterns and Story-shapes: The Statistical Analysis of Narrative Style, *Literary and Linguistic Computing*, 2(2): 61-70.

Burrows, J. 1992, Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information, *Literary and Linguistic Computing*, 7(2): 91-109.

Dolnet, 1998, *TO BHMA*, Lambrakis Publishing Corporation, http://tovima.dolnet.gr/

Fakotakis, N., A. Tsopanoglou, and G. Kokkinakis, 1993, A Text-independent Speaker Recognition System Based on Vowel Spotting, *Speech Communication*, 12: 57-68.

Holmes, D. 1994, Authorship Attribution, *Computers and the Humanities*, 28: 87-106.

Holmes, D. and R. Forsyth 1995, The Federalist Revisited: New Directions in Authorship Attribution, *Literary and Linguistic Computing*, 10(2): 111-127.

Honore, A., 1979, Some Simple Measures of Richness of Vocabulary, Association for Literary and Linguistic Computing Bulletin, 7(2): 172-177.

Mosteller, F. and D. Wallace 1984, *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, Addison-Wesley, Reading, MA.

Stamatatos, E., N. Fakotakis, and G. Kokkinakis forthcoming, On Detecting Sentence and Chunk Boundaries in Unrestricted Text Based on Minimal Resources.

Yule, G. 1944, *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Cambridge.