Investigating topic influence in authorship attribution

George K. Mikros
Department of Italian and Spanish
Language and Literature
University of Athens
Panepistimioupoli Zografou - 15784
Athens, GREECE
+30 210 6511344
gmikros@isll.uoa.gr

Eleni K. Argiri
Department of Linguistics
University of Athens
Panepistimioupoli Zografou - 15784
Athens, GREECE
eleniargiri@hotmail.com

ABSTRACT

The aim of this paper is to explore text topic influence in authorship attribution. Specifically, we test the widely accepted belief that stylometric variables commonly used in authorship attribution are topic-neutral and can be used in multi-topic corpora. In order to investigate this hypothesis, we created a special corpus, which was controlled for topic and author simultaneously. The corpus consists of 200 Modern Greek newswire articles written by two authors in two different topics. Many commonly used stylometric variables were calculated and for each one we performed a two-way ANOVA test, in order to estimate the main effects of author, topic and the interaction between them. The results showed that most of the variables exhibit considerable correlation with the text topic and their exploitation in authorship analysis should be done with caution.

Keywords

Authorship Attribution, Stylometry, Topic-neutral features.

1. Introduction

Authorship attribution research is based on the "authorship fingerprint" notion. According to this view, each person possesses an idiosyncratic way to utilize their linguistic means, which are unique and their quantitative description can discriminate him/her among every other possible author. In order to find which parts of the human linguistic behavior reflect authorship, researchers have investigated a large number of text characteristics in many linguistic levels. We now know that there are at least 1000 textual attributes relevant to authorship [24]. The selection of these variables is based on their ability to reveal subconscious mechanisms of language variation, which are unique to each author. Therefore, authorship analysis is based on detecting and counting unconscious linguistic habits that are directly related to the text author.

2. Related work

2.1 Corpora controlled for topic in authorship attribution studies

Recently, text metadata influence has been acknowledged as a serious bias in authorship attribution studies. Rudman [24] provides a systematic exposition of the various pitfalls of authorship research and cites specifically that the corpora used for authorship analysis should be matched for genre and time period. Since then, many studies appeared, systematically using corpora that are controlled for topic, genre, medium etc. Baayen et al. [3] created a balanced corpus of written essays in 3 different genres and in 3 topics for each genre. Graham [8] used the Risks corpus, a one-topic corpus, which consists of nearly 1 million words of postings on the Forum on Risks to the Public in Computers and Related Systems (comp.risks). Koppel & Schler [14] used an e-mail discussion group concerning automatic information extraction. It included 480 e-mails written by 11 different authors, during a period of one year. All posts were about the same subject, forming a highly homogeneous corpus with regard to topic. Luyckx & Daelemans [16], in order to isolate the effects of topic and genre, collected 300 texts on the same topic and genre, distributed in 3 author categories (2 separate authors and 1 author category named "Others" with texts of 10 different authors and some collaborative articles of the previous two authors). Argamon et al. [1] developed a benchmark collection of electronic messages for experimentation on author attribution. The collection was based on three Usenet groups with different topics (books, computer theory, programming language C). In each topic, four subcorpora were created, based on different numbers of authors for attribution. In Mikros [22], authorship attribution was attempted in a highly homogeneous newswire corpus, controlled for topic, genre and medium. In total, 1200 texts were collected, written by four different authors in the same topic (Politics).

2.2 Topic independent features

Stylometric variables used in authorship attribution should be independent of any metalinguistic entity, that is genre, topic, medium, chronological era etc. At the same time, they should have a reasonable frequency of occurrence, in order to facilitate their statistical analysis. The above characteristics are fulfilled in the lexical level by the well-known class of function words.

Mosteller & Wallace [23] were among the first to search for text attributes that were systematically topic-neutral. They ended up using specific function words, which have high frequency of occurrence and at the same time remain corpus independent. Recently, Koppel et al. [15], using experimental methodology, found that function words are indeed the best candidates for a universal, corpus-independent feature set for authorship attribution. They used the measure of "stability", which represents quantitatively the degree of available synonymy of a specific linguistic item. Function words are unstable, in the sense that they can be substituted easily in a passage, without affecting the meaning of the text.

Although the frequency of function words has been proved a reliable author discriminator feature in many studies, there are

SIGIR '07 Amsterdam. Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection

many other stylometric variables which have been used extensively and at least in theory are topic-neutral. Many of them are smaller than the word units, such as characters. At this sub-word level we can safely assume that it is very difficult to trace conscious linguistic usage. Other variables attempt to capture the vocabulary size used in a text, such as Yule's K and Language Density. These measures should also be topic independent, and since vocabulary "richness" is an author's characteristic it should not correlate with topic information. Readability measures, such as word length and sentence length, are also some of the oldest and most common features used in authorship attribution studies and are used extensively as topic-neutral variables.

2.3 The effects of stylistic choices in topic categorization

Although most stylometric features used in authorship attribution studies are considered to be topic independent, recent advances in text topic categorization have shown that topic categorization accuracy can be further improved, if we add stylistic information to the classifier models. Relevant research of stylistic analysis in text categorization has shown that stylistic markers, utilized notably for authorship attribution studies, play at least an auxiliary role in topic classification. The first important attempts to construct text classification systems for recognizing text genres and thus set the foundations for further research were the works of Kalgren & Cutting [11] and Karlgren [12], who used Biber's [4] feature set and Discriminant Function Analysis (DFA) to classify documents according to genre. Kessler, et al.[13] used cue words for the same purpose.

The reliability of style markers as topic discriminators was investigated by Argiri [2] in experiments involving the categorization of Internet articles into predefined thematic categories, with the use of machine learning schemes. The results proved that stylistic features may have subject-revealing power and significantly enhance topic classification.

Mikros & Carayannis [20] used exclusively non lexical features in order to classify 1200 texts in four topic categories. The feature set used was based exclusively on stylometric variables such as lexical "richness" and various sentence and word level measures including specific sociolinguistic attributes. Overall topic classification accuracy reached 81%, providing evidence that these features carry content information.

Mikros [19] used DFA and compared various features, lexical and non lexical in topic categorization using a corpus of 900 newswire articles. Each variable's contribution was measured using Wilks'L and the results showed that stylometric variables like the Average Word Length and frequency of the Punctuation Marks were among the most influential variables in the analysis.

Tambouratzis, et al. [28] carried out style-based text classification tests for the Greek language, focusing on polysemy and grammatically equivalent word forms. They counted morphological, as well as structural features of the texts and deployed cluster analysis on three categories (Fiction, History, Politics), with high accuracy results.

Another study was effected by Michos, et al. [18], focusing on functional rather than literary style. In their automatic text categorization experiments, they used syntactic and verbal identifiers, such as adjective/noun and adverb/verb ratios, and studied the positive/negative effects of linguistic features in real-life texts.

Overall, more and more text categorization studies seem to focus on the discriminatory role of stylistic attributes within various topics, producing interesting results, that should be further explored.

2.4 The effects of topic in authorship attribution

The increasing number of topic-controlled corpora used in authorship attribution studies, described in 2.1, reveals an awareness of topic bias in author discrimination accuracy. However, a small number of studies that have directly researched this issue report contradicting results.

Corney [5] investigated the effect of e-mail topics in authorship classification. The corpus used in this study consisted of e-mails written by a small closed group of authors on a specific set of topics. To measure the topic effect, classifier models were built for each of these authors, using the e-mails of one of the topics. Other topics' e-mails were then used as the test data for the classifier learning models from the original topic. The obtained results showed that authorship attribution accuracy was unaffected by e-mail topic and that function words were consistently the best individual feature set independent of topic.

Madigan et al. [17] also underlined the need to research topic effect in authorship attribution using cross-topic corpora. In order to test the effect of topic in authorship attribution, they used a corpus of Usenet postings compiled from two users, who systematically post many messages in discussion groups of different topic. Results showed that topic interacts with authorship and the Bag of Words (BoW) representation, which was the most successful feature set in data sets of multitopic authorship attribution, performed poorly on this experiment.

De Vel et al. [6] used a corpus of 1259 Usenet postings in four topics written by four authors. Results showed that inter- and intra- topic authorship attribution is possible but authorship categorization precision is not stable across all authors. In specific cases, the categorization obtained was biased towards the e-mail document topic content rather than on its author.

Finn & Kushmerick [7] investigated genre classification corpora controlled for topic. They evaluated their classifiers using two text collections. The first experiment calculated the accuracy of the classifier in a single subject domain. The second experiment measured the classifier accuracy, when trained on one subject domain, but tested on another. This specific task was used as a measure of the performance of a genre classifier across multiple subject domains and gave an indication of the classifier's ability to generalize to new domains. The results showed that topic and genre besides their theoretical distinctiveness, in practice, they partially overlap. The standard stylometric features used in this study were able to discriminate genres but the models built were partially topic dependent.

3. Methodology

3.1 The topic-controlled authorship corpus

In order to study the topic effect in authorship attribution we compiled a small-scale corpus consisting of 200 newspaper articles written by two authors (Dimitris Maronitis, who is actually a scholar and Pantelis Boukalas, who is a philologist) for the electronic editions of two major Greek newspapers, TO VIMA and KATHIMERINI, during the period 1997-2006. All articles were downloaded from the websites of the newspapers in question.

We collected articles from two topic categories, Culture and Politics, keeping in mind the authors' similar writing style. A special criterion for the selection of the specific articles was the authors' natural register, as well as their overlapping in terms of writing within the same genre, but also each one's similar style when writing for different topics. Another interesting aspect of the texts is that their authors mix various topics while analysing certain political aspects of these topics and vice versa. For instance, they may write about a political subject and use historic or cultural examples to illustrate their point, or they may write about a cultural event or review a book and discuss them in a political context. The latter case is more frequent in the articles written by Pantelis Boukalas. Moreover, each text per author comes from the same column and section in each newspaper, as included in the newspaper supplements consisting of essays and articles regarding culture, history, science, social and political issues etc. In principle, this means that such texts undergo some low-level post-editing, as opposed to editorial or reportage articles, which are subject to a stricter editing, so that they conform to the overall style of the newspaper. Therefore, the style of the specific authors is more personal and independent of outer influences. Similar texts have also been used in a corpus compiled by Stamatatos [27] in his study on ensemble-based author identification.

The corpus size distribution per author and topic is shown in the table below (Table 1):

Table 1: Distribution of words and texts across Topic and Author categories.

	Topics					
	Culture		**Politics**		*Total*	
Authors	Texts	Words	Texts	Words	Texts	Words
Boukalas	50	41,107	50	21,561	***100***	***62,668***
Maronitis	50	30,645	50	28,850	***100***	***59,495***
Total	***100***	***71,752***	***100***	***50,411***	***200***	***122,163***

3.2 Stylometric variables

We used different categories of stylometric variables all of which are in theory topic-neutral:

1) Lexical "richness" variables: Yule's K [Yule's K], Standardized Type Token Ratio [stTTR], Lexical Density (ratio of content to function words) [LexDens], Percentage of hapax-legomena [HapaxL], Percentage of dis-legomena [DisL], Ratio of Dis- to Hapax legomena [Dis_Hap], Relative Entropy [RelEntr], Percentage of numbers in the text [Numbers] - 8 variables

2) Sentence level measures: Average length of sentences measured in words [SL], Standard deviation of sentence length per text [SLstdev] - 2 variables

3) 10 most Frequent Function Words of Modern Greek – 10 variables

4) Word level measures: Average word length per text measured in letters [AWL], Standard deviation of word length per text [AWLstdev], Word length distribution containing frequency of 1 letter word to frequency of 14 letters word [1LW, 2LW… 14LW), - 16 variables

5) Character level measures: Frequency of the letters normalized to 1000 word fixed text length – 32 variables

4. RESULTS

4.1 Classification accuracy in author and topic discrimination

In order to test the discriminatory power of the above-mentioned features, we used DFA, a well documented classification function, which has been used extensively in authorship attribution research (e.g. [3], [25], [29], [22]).

DFA involves deriving a variate, the linear combination of two (or more) independent variables that will discriminate best between a priori defined groups. Discrimination is achieved by setting the variate's weight for each variable to maximize the between-group variance, relative to the within-group variance [9].

If the dependent variables have more than two categories, DFA will calculate k-1 discriminant functions, where k is the number of categories. Each function allows us to compute discriminant scores for each case for each category, by applying the following equation:

$$\mathbf{D_{jk}= a + W_1X_{1k} + W_2X_{2k} + ... + W_nX_n}$$

where

D_{jk}= Discriminant score of discriminant function j for object k

a= intercept

W_i= Discriminant weight for the independent variable i

X_{ik}= Independent variable i for object k

For the validation of the DFA results, we used the U-method, a cross-validation procedure based on the "leave-one-out" principle [10]. Using this method, the discriminant function is fitted to repeatedly drawn samples of the original sample. This procedure estimates k-1 samples, eliminating one observation at a time from a sample of k cases.

We first applied DFA using Author as dependent variable and obtained the cross-validated classification results. In the second phase, we applied DFA again using the same stylometric variables, but we used Topic as dependent variable. Both DFA's were computed using the stepwise method. The confusion matrix of both DFAs is presented below (Table 2):

Table 2: Cross-validated classification results in Author and Topic categorization.

Overall Author classification accuracy = 96%	***Predicted author***	
Author	Boukalas (%)	Maronitis (%)
Boukalas	97	3
Maronitis	5	95
Overall Topic classification accuracy = 79.5%	***Predicted topic***	

Topic	Culture (%)	Politics (%)
Culture	76	24
Politics	17	83

The authorship attribution achieved an overall 96% accuracy, showing that the selected feature set was indeed useful in capturing authorship information. However, the topic categorization accuracy was also very high (79.5%), especially if we consider that we used only stylometric variables and no content words at all. This result indicates that the features used, at least some of them, correlate with topic information and are not topic-neutral.

4.2 Testing the topic-neutral hypothesis of common stylometric variables

In order to explore further which features are truly topic independent, we performed a series of two-way ANOVA with dependent variable each time a specific stylometric variable and factors, the Author and the Topic of the text. Two-way ANOVA can reveal not only the main effects of Author and Topic in the dependent variable, but also the interaction effect between them. We examined the distribution of all the variables using Kolmogorov-Smirnov test and we found 30 variables that were not normally distributed. In these variables we used additionally the non-parametric Mann-Whitney U test in order to validate the p values of the ANOVA. In all these cases ANOVA results were confirmed although the normality assumption was violated.

The ANOVA results are reported in the following tables organized by feature sets. Grey cells are statistically significant ($p<0.05$):

Table 3: ANOVA significance in main and interaction effects with dependent variables Lexical "richness" features.

Lexical "richness" variables	***Author***	***Topic***	***Author~Topic***
Yule's K	0.00	0.02	0.08
stTTR	0.00	0.2	0.00
LexDens	0.00	0.31	0.21
DisL	0.07	0.00	0.23
RelEntr	0.57	0.00	0.05
HapaxL	0.7	0.00	0.57
Dis_Hap	0.12	0.27	0.4
Numbers	0.67	0.01	0.00

The lexical "richness" variables displayed above (Table 3), exhibit considerable variation regarding their correlation with topic. Lexical Density seems to be the only variable that discriminates authorship exclusively. All the others have some interaction with topic. In particular, four of them, appear to discriminate only topic (Hapax Legomena, Dis Legomena, Relative Entropy, Numbers). Yule's K, one of the most widely used stylometric variables in authorship attribution, relates both to authorship and topic. Standardized TTR discriminates authors, but at the same time exhibits author~topic interaction effect.

Table 4: ANOVA significance in main and interaction effects with dependent variables Sentence level features.

Sentence level variables	***Author***	***Topic***	***Author~Topic***
SL	0.00	0.84	0.03
SLstdev	0.00	0.92	0.04

The two sentence level variables have similar behavior as can be seen in the above table (Table 4). They discriminate authors and not topics, but they present statistical significance in author~topic interaction, as can be seen in Figure 1:

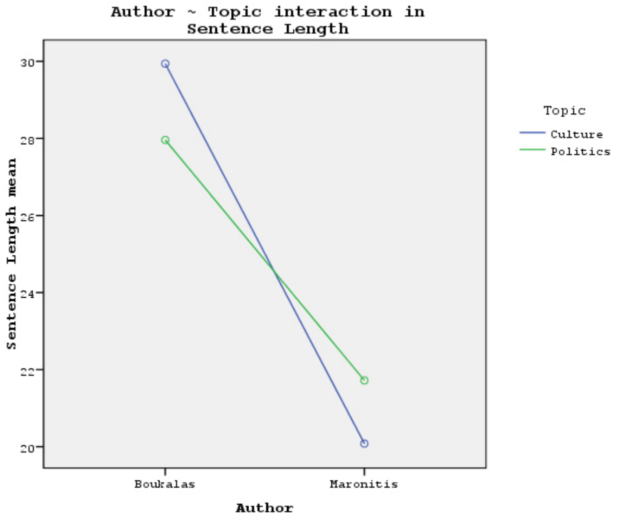

Figure 1: Author ~ Topic interaction in Sentence Length.

Sentence length mean is not statistically different between the two topics. However, Boukalas is using statistically significant larger sentences than Maronitis in Culture texts and smaller sentences than Maronitis in Politics texts. This kind of interaction reveals that each author manipulates this variable in a different way, according to the topic of the text. In general, an author~topic statistically significant interaction in a stylometric variable falsifies its topic-neutral character.

Table 5: ANOVA significance in main and interaction effects with dependent variables Frequent Function Words features. In parenthesis a rough translation in English.

Frequent Function Words variables	***Author***	***Topic***	***Author~Topic***
kai (and)	0.00	0.61	0.13
na (to)	0.00	0.00	0.64
tha (will)	0.00	0.01	0.25
den (don't)	0.00	0.00	0.06
oti (that)	0.00	0.00	0.03
apo (from)	0.06	0.43	0.83
pou (where ~ who/m)	0.12	0.97	0.63
gia (for)	0.37	0.09	0.24
se (in)	0.5	0.45	0.93
me (with)	0.73	0.05	0.68

From the ten most frequent function words of Modern Greek displayed in the above table (Table 5), half of them do not have any discriminatory power over author or topic (apo, pou, gia, se, me). From the remaining five, only “kai” discriminates exclusively authorship, while the others distinguish both author and topic. These results show that, although function words are indeed semantically free, they do however contribute indirectly to the meaning of the text. This is happening probably through syntax and discourse level, since many function words construct phrase complexity and build cohesion patterns, which can indirectly be linked with topic information.

Table 6: ANOVA significance in main and interaction effects with Word level features as dependent variables.

Word level variables	*Author*	*Topic*	*Author~Topic*
AWL	0.00	0.00	0.38
2LW	0.00	0.6	0.93
7LW	0.00	0.00	0.51
8LW	0.00	0.03	0.72
9LW	0.00	0.05	0.38
10LW	0.00	0.5	0.86
11LW	0.00	0.08	0.97
12LW	0.00	0.18	0.72
14LW	0.11	0.00	0.34
3LW	0.13	0.07	0.77
4LW	0.22	0.24	0.14
AWLstdev	0.36	0.00	0.31
1LW	0.4	0.23	0.71
13LW	0.55	0.04	0.44
6LW	0.75	0.03	0.13
5LW	0.9	0.82	0.5

The word level variables discriminate both author and topic, as shown in the above table (Table 6). Authorship is exclusively distinguished by 2, 9, 10, 11, 12 letters words and topic by 6, 13, 14 letters words plus Average Word Length standard deviation. Discrimination of both author and topic is observed by Average Word Length and 7 and 8 letters words. The influence of topic on word level variables is important. A possible explanation could be that long words tend to be terms with specific topic meaning. Furthermore, average word length standard deviation is higher in texts with many long words, which make this variable topic-dependent.

Table 7: ANOVA significance in main and interaction effects with Character level features as dependent variables. In parentheses, the character in Modern Greek.

Character level variables	*Author*	*Topic*	*Author~Topic*
gh (γ)	0.00	0.00	0.00
f (φ)	0.00	0.00	0.00
s (σ)	0.00	0.00	0.05
k (κ)	0.00	0.00	0.09
dh (δ)	0.00	0.00	0.16
u (υ)	0.00	0.06	0.14
n (ν)	0.00	0.1	0.73
i_st (ί)	0.00	0.14	0.63
r (ρ)	0.00	0.2	0.13
h (η)	0.00	0.3	0.15
sfin (ς)	0.00	0.41	0.98
e (ε)	0.00	0.5	0.26
ks (ξ)	0.00	0.62	0.9
h_st (ή)	0.00	0.69	0.26
th (θ)	0.00	0.75	0.46
m (μ)	0.00	0.84	0.03
a (α)	0.00	0.9	0.87
bh (β)	0.00	0.99	0.08
l (λ)	0.02	0.07	0.67
omg (ω)	0.03	0.34	0.34
e_st (έ)	0.04	0.18	0.05
a_s (ά)	0.07	0.19	0.38
x (χ)	0.07	0.9	0.00
t (τ)	0.25	0.04	0.77
ps (ψ)	0.31	0.02	0.51
u_st (ύ)	0.33	0.12	0.02
z (ζ)	0.6	0.15	0.7
o_st (ό)	0.68	0.82	0.17
i (ι)	0.78	0.02	0.07
p (π)	0.83	0.00	0.06
omg_st (ώ)	0.94	0.92	0.88
o (ο)	0.95	0.18	0.55

From the above table (Table 7), we conclude that letter frequencies are not topic-neutral feature. From the 32 measured characters, 12 correlate with topic either as a main effect (gh, f, s, k, dh, t, ps, i, p) or as interaction with the Author variable (m, x, u_st). This result is particular interesting since the letters, which present statistically significant main effects in topic, are among the most frequent consonants in Modern Greek. A partial explanation of this could be found if we inspect more closely the distribution of the specific consonants at the word level. Mikros et al. [21], found that dh, p, k, t, gh, f, s are the most frequent letters in the beginning of a word. This could reveal a covert relation to the topic of a text, since specific topics contain terms, which begin with specific characters. If this is true, then letter frequencies should not be used as topic-neutral authorship attribution variables, since different topics will change dynamically the correlation with specific characters. As a result, each authorship attribution corpus will present different character~topic correlations in an unpredictable way.

We repeated author and topic classification with 22 features that have been found to be really topic-neutral (that is, features that present statistically significant main effect to Author). The confusion matrix of both DFA's is presented below:

Table 8: Cross-validated classification results in Author and Topic categorization using only topic-neutral features.

Overall Author classification accuracy = 93%	***Predicted author***	
Author	Boukalas (%)	Maronitis (%)
Boukalas	93	7
Maronitis	7	93
Overall Topic classification accuracy = 49%	***Predicted topic***	
Topic	Culture (%)	Politics (%)
Culture	50	50
Politics	52	48

The results reported in the above table (Table 8), show that authorship attribution accuracy remained high (93%), while topic categorization dropped to baseline percentage (49%). Although accuracy in authorship attribution dropped 3% relating to the stepwise DFA reported in Table 2, the feature set that obtained this attribution is far more robust and can be used reliably in measuring author's style, excluding text topic influence.

5. Conclusions and future work

This study investigated the topic-neutral character of some widely used stylometric variables in authorship attribution studies. In order to research the influence of topic in author discrimination, we compiled a balanced corpus of two authors, whose articles are equally divided in two distinctive topics, culture and politics. In this corpus, we measured five feature sets that in theory are topic independent. Using DFA, we showed that the same feature set could provide author and topic classification with high accuracy. A more detailed study, using a series of two-way ANOVA, revealed that many stylometric variables are actually discriminating topic rather than author. Among them, we found Frequent Function Words, specific characters, word lengths, and commonly used lexical "richness" measures, such as Yule's K. The main conclusion is that, when we apply these stylometric variables for authorship attribution to multitopic corpora, we should be extremely cautious. Authorship attribution could become a by-product of the correlation of authors with specific topics. Although this could be a useful parameter, when the set of possible authors is large, or have specific aims [17], it should be avoided in authorship attribution problems with a limited number of authors, where the analysis is focused in identifying the real person behind a text. The reported results are based on a limited corpus in both author and topic categories but they are indicative of the complex interaction between an author's style and the text topic he writes.

Future research will be directed in other languages than Greek, as well as testing other variables, such as bigrams, trigrams, Part of Speech tags, Part of Speech bigrams etc. In addition, a larger experiment is under preparation, containing more author and topic categories.

6. REFERENCES

[1] Argamon, S., Šarić, M., and Stein, S. Style mining of electronic messages for multiple author discrimination. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining*, 2003.

[2] Argiri E., *Style-based topic categorisation with the use of machine learning techniques*. MSc Dissertation, University of Athens/National Technical University, Greece, 2006.

[3] Baayen, H., van Halteren, H., Neijt, A., Tweedie, F. An experiment in authorship attribution. In *Proceedings of JADT 2002* (St. Malo 2002). 2002, 29-37.

[4] Biber D. *Variation across speech and writing*. Cambridge: Cambridge University Press, 1988.

[5] Corney, Malcolm. *Analysing E-mail Text Authorship for Forensic Purposes.* MA thesis, Queensland University of Technology, 2003.

[6] de Vel, O., Anderson, A., Corney, M., Mohay, G. Multi-Topic E-mail Authorship Attribution Forensics. In *Proceedings of ACM Conference on Computer Security - Workshop on Data Mining for Security Applications*, (November 8, 2001), Philadelphia, PA, USA.

[7] Finn, A. & Kushmerick, N. Learning to classify documents according to genre. In S. Argamon, (ed.), *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico, 2003, 35-45.

[8] Graham, Neil. *Automatic detection of authorship changes within single documents.* MSc thesis, Graduate Department of Computer Science, University of Toronto, 2003.

[9] Hair, J., Anderson, R., Tatham, R., and Black, W. *Multivariate data analysis*. New Jersey: Prentice Hall, 1995.

[10] Huberty, C., Wisenbaker, J. and Smith, J. Assesing predictive accuracy in discriminant analysis. *Multivariate Behavioural Research,* 1987, 22:307-329.

[11] Karlgren, J, Cutting, D. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. *Proceedings of the 15th International Conference on Computational Linguistics*, NJ: ACM Press, 1994, 1071-1075.

[12] Karlgren, J. Stylistic Experiments for Information Retrieval Experiment. In: T. Strzalkowski (ed.), *Natural Language Information Retrieval*. Norwell: Kluwer Academic Publishers, 1999, 147-166.

[13] Kessler, B., Nunberg, G., Schutze, H. Automatic detection of text genre. *Proceedings of the 35th Annual Meeting of the ACL and the 8th Meeting of the European Chapter of the ACL*. San Francisco: Morgan Kaufmann, 1997, 32-38.

[14] Koppel, M., and Schler, J. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico, 2003, 69-72.

[15] Koppel, M. Akiva, N. and Dagan, I. A Corpus-Independent Feature Set for Style-Based Text Categorization. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[16] Luyckx, K. and Daelemans, W. Shallow Text Analysis and Machine Learning for Authorship Attribution. In

Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands (CLIN 2004), 2005, 149-160.

[17] Madigan, D., Genkin, A., Lewis, D., Argamon, S., Fradkin, D., and Ye, L. Author identification on the large scale. In *Proceedings of Joint Annual Meeting of the Interface and the Classification Society of North America, 2005*.

[18] Michos, S.E., Stamatatos, E., Fakotakis, N., and Kokkinakis, G. An empirical text categorizing computational model based on stylistic aspects. *Proceedings of the 8th International Conference on Tools with Artificial Intelligence*. Washington: IEEE Computer Society, 1996, 71-77.

[19] Mikros, G. Statistical approaches to automatic text categorisation in modern Greek: A pilot study for evaluating stylistic markers and statistical methods. Paper for the 6th International Conference on Greek Linguistics, in CD-ROM, 2003.

[20] Mikros, G., and Carayannis, G. Modern Greek corpus taxonomy. In: *Proceedings of the Second International Conference on Language Resources and Evaluation.* Athens: National Technical University of Athens, 2000, 129-134.

[21] Mikros, G., Hatzigeorgiu, N., and Carayannis, G. Basic quantitative characteristics of the Modern Greek language using the Hellenic National Corpus. *Journal of Quantitative Linguistics,* 2005*,* 12: 167-184.

[22] Mikros, G. Authorship attribution in Modern Greek newswire corpora. In Uzuner, O., Argamon, S. & Karlgren, J. (eds), *Proceedings of the SIGIR 2006 Workshop on Directions in Computational Analysis of Stylistics in Text Retrieval*, Seattle, USA, August 10, 2006, 43-47.

[23] Mosteller, F. and Wallace, D. *Applied bayesian and classical inference. The case of The Federalist Papers.* New York: Springer – Verlag, 1964.

[24] Rudman, J. The State of Authorship Attribution Studies: Some Problems and Solutions. *Computer and the Humanities, 31,* 1998, 351–365.

[25] Stamatatos, E., Fakotakis, N, and Kokkinakis, G. Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics,* 2000*,* 26: 471-495.

[26] Stamatatos, E., Fakotakis, N., and Kokkinakis, G. Computer-Based Authorship Attribution Without Lexical Measures. *Computers and the Humanities,* 2001*,* 35: 193-214.

[27] Stamatatos, E. Ensemble-based author identification using character N-grams. In: *Proceedings of the 3rd International Workshop on Text-based Information Retrieval* (TIR'06), 2006, 41-46.

[28] Tambouratzis, G., Markantonatou, S., Hairetakis, N., and Carayannis, G. Automatic Style Categorisation of Corpora in the Greek Language. In: *Proceedings of the Second International Conference on Language Resources and Evaluation.* Athens: National Technical University of Athens, 2000, 135-140.

[29] Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Carayannis, G., and Tambouratzis, D. Discriminating the Registers and Styles in the Modern Greek Language – Part 2: Extending the feature Vector to Optimize Author Discrimination. *Literary & Linguistic Computing,* 2004*,* 19: 221-242.