Investigating topic influence in authorship attribution

George K. Mikros
Department of Italian and Spanish Language and Literature
University of Athens

Eleni K. Argiri
Department of Linguistics,
University of Athens

Aim of the research

- 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.

Overview

- Introduction
- Related work
  - The effects of stylistic choices in topic categorization
  - The effects of topic in authorship attribution
- Methodology
  - Topic-controlled corpus
  - Stylometric variables
- Results

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.
- 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 measuring linguistic habits that are directly related to the text author.

Stylometric variables in authorship attribution

- Stylometric variables used in authorship attribution should be independent of any extralinguistic 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.
- Authorship attribution studies have widely utilize sub-word level variables such as frequency of characters in a text. At this 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.

The effects of stylistic choices in topic categorization

- Relevant studies in stylistic analysis used for text categorisation have shown that stylistic markers play at least an auxiliary role in discriminating topic:
  - Karlgren/Karlgren & Cutting: Biber's feature set and DFA for genre classification
  - Kessler et al: cue words for genre classification
  - Mikros: DFA for newswire article topic classification/average word length and frequency of punctuation marks produced accurate results
  - Mikros & Carayannis: mainly non lexical features/performance reached 81%
  - Michos et al.: syntactic and verbal identifiers, focusing on functional style
  - Stamatatos et al: text categorisation in terms of genre and author
  - Argiri (dissertation): use of authorship attribution style variables in topic categorisation produced quite accurate results in topic classification
- Almost all research has shown that stylistic markers may have subject-revealing power and has set the foundations for investigating further their reliability as topic discriminators in text categorisation.

The effects of topic in authorship attribution

- Increasing number of topic-controlled corpora in authorship studies shows awareness of topic bias in author discrimination accuracy.
- Few relevant studies, focusing mainly on e-mail messages and blogs, report contradicting results:
  - Corney: authorship attribution unaffacted by e-mail topic/function words were the best feature set which was independent of topic
  - Madigan et al.: topic interacts with author/bag of words performed poorly
  - De Vel et al.: inter- and intra-topic authorship attribution is possible but authorship precision is not stable across all authors
  - Finn & Kushmerick: topic and genre overlapping

- The problem of topic effect in authorship attribution, combined with the problem of selecting the best style attributes for topic categorisation should be further investigated in order to pinpoint the exact nature of stylistic variables in terms of their discriminatory role, in either author or topic identification.

Corpus design

- Two authors/two thematic categories (Politics & Culture) were selected
- Articles were taken from Greek newspapers (electronic editions)
- Articles come from the same column and section – they are less affected by post-editing
- Overlapping style is used between authors and between author and topic

The corpus in numbers

Corpus Statistics

Words

Author	Topic	N	Sum	Mean	Std. Deviation	Minimum	Maximum
Boukalas	Culture	50	41107	822,14	382,607	362	1399
	Politics	50	21561	431,22	179,787	362	1179
	Total	100	62668	626,68	356,432	362	1399
Maronitis	Culture	50	30645	612,90	57,673	479	739
	Politics	50	28850	577,00	50,238	471	666
	Total	100	59495	594,95	56,753	471	739
Total	Culture	100	71752	717,52	291,817	362	1399
	Politics	100	50411	504,11	150,380	362	1179
	Total	200	122163	610,82	255,065	362	1399

Stylometric variables used in the study

- **Lexical "richness" variables (8 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]
- **Sentence level measures (2 variables):**
  - Average length of sentences measured in words [SL],
  - Standard deviation of sentence length per text [SLstdev]
- **10 most Frequent Function Words of Modern Greek (10 variables).**
- **Word level measures (16 variables):**
  - 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).
- **Character level measures: Frequency of the letters normalized to 1000 word fixed text length (32 variables).**

Classification experiments

- All classification experiments were performed using Discriminant Function Analysis (DFA).
- Popular technique in authorship attribution
- Cross-validation using U-method based on “leave-one-out” principle.

Testing topic-independence in 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.

Authorship and Topic classification accuracy (all variables)

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

Lexical “Richness” variables (1)

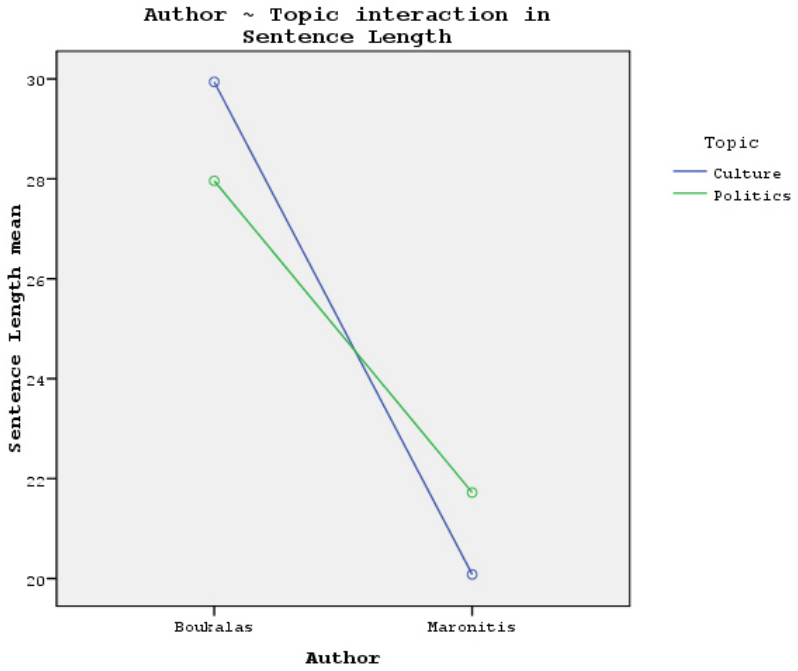
Lexical “richness” variables	*Author*	*Topic*	*Author~Topic*
Yule’s K	0.00	0.02	0.08
stTTR	0.00	0.2	0.00
LexDens	0.00	0.31	0.21
DisL	0.07	0.00	0.23
RelEntr	0.57	0.00	0.05
HapaxL	0.7	0.00	0.57
Dis_Hap	0.12	0.27	0.4
Numbers	0.67	0.01	0.00

Lexical “Richness” variables (2)

- Lexical Density seems to be the only variable that discriminates authorship exclusively.
- All the others have some interaction with topic. 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.

Sentence level variables (1)

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



Sentence level variables (2)

- The two sentence level variables have similar behavior since they discriminate authors and not topics, but at the same time they present statistical significance in author~topic interaction.
- 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.

Function words variables (1)

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

Function words variables (2)

- From the ten most frequent function words of Modern Greek, 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 contribute to phrase complexity and build cohesion patterns, which can indirectly be linked with topic information.

Character level variables (1)

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

Character level variables (2)

- Character frequencies are not a 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 particularly 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 character 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.

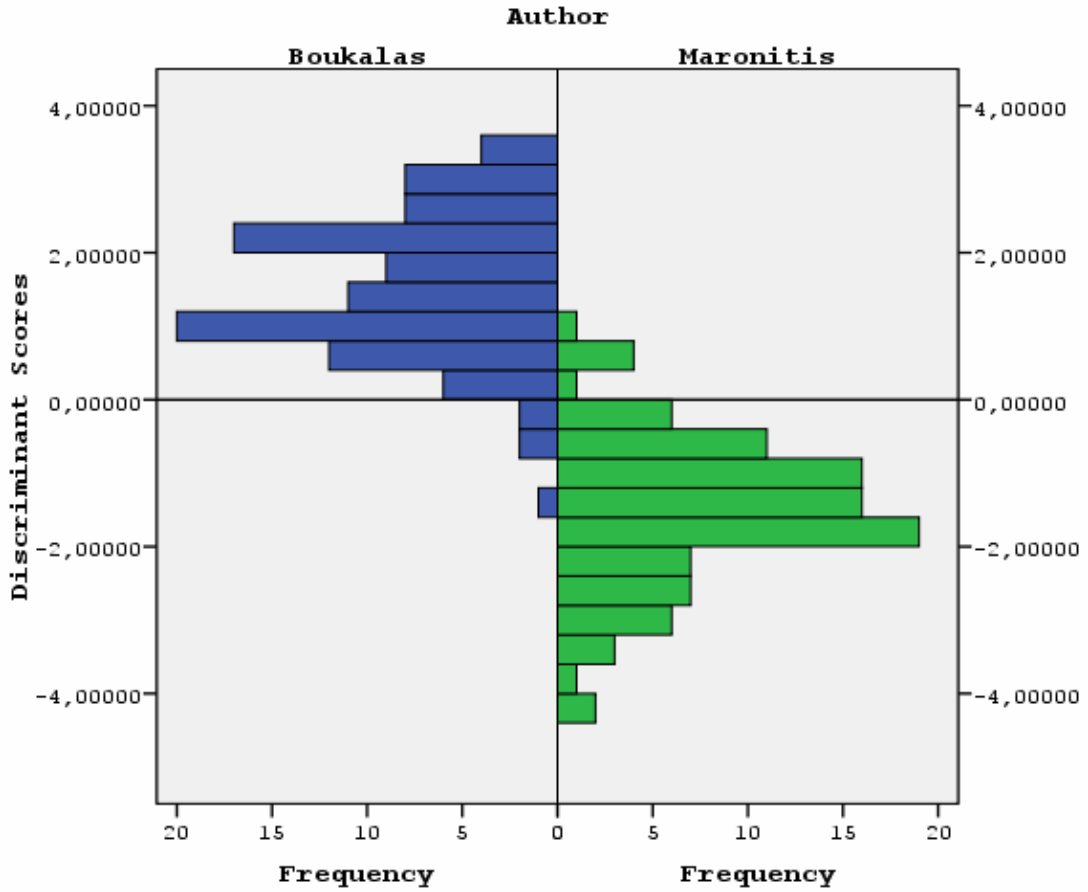
“Pure” authorship variables

Variables	*Author*	*Topic*	*Author~Topic*	Variables	*Author*	*Topic*	*Author~Topic*
LexDens	0.00	0.31	0.21	***h (η)***	0.00	0.3	0.15
kai (and)	0.00	0.61	0.13	***sfin (ς)***	0.00	0.41	0.98
2LW	0.00	0.6	0.93	***e (ε)***	0.00	0.5	0.26
9LW	0.00	0.05	0.38	***ks (ξ)***	0.00	0.62	0.9
10LW	0.00	0.5	0.86	***h_st (ή)***	0.00	0.69	0.26
11LW	0.00	0.08	0.97	***th (θ)***	0.00	0.75	0.46
12LW	0.00	0.18	0.72	***a (α)***	0.00	0.9	0.87
u (υ)	0.00	0.06	0.14	***bh (β)***	0.00	0.99	0.08
n (ν)	0.00	0.1	0.73	***l (λ)***	0.02	0.07	0.67
i_st (ί)	0.00	0.14	0.63	***omg (ω)***	0.03	0.34	0.34
r (ρ)	0.00	0.2	0.13	***e_st (έ)***	0.04	0.18	0.05

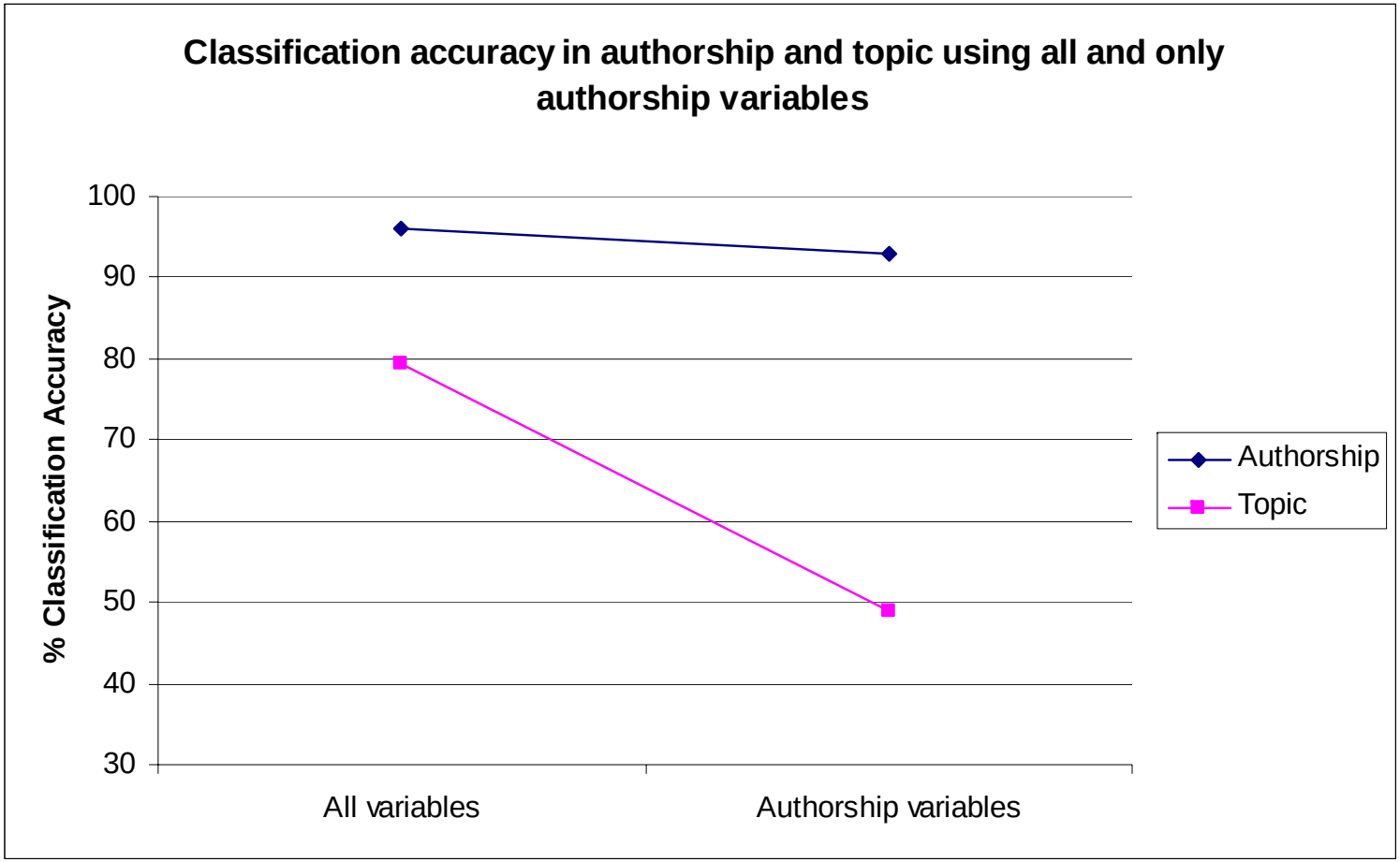
Authorship and Topic classification accuracy (authorship variables)

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

Authorship classification obtained



Comparing the accuracy of the two feature sets



Conclusions

- The main conclusion is that many widely used stylometric variables correlate systematically with the text topic rather than the author.
- The application of these features for authorship attribution to multitopic corpora, 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, 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.

THANK YOU!