Systematic stylometric differences in men and women authors: a corpus-based study

George K. Mikros, Athens

Abstract. The aim of this paper is to explore the differences existing in written texts between male and female authors. In particular we will investigate a number of well-known stylometric variables which have been used previously in authorship attribution with considerable success and we will try to expand their discriminatory power in author's gender classification.

Keywords: author profiling, stylometry, MANOVA, Discriminant Function Analysis, gender, Modern Greek.

1 Introduction

The gender differences in language understanding and production is a widely researched issue which has been investigated by a number of different disciplines, from neurophysiology and cognitive science to sociolinguistics. Our main research hypothesis in this article is that the author's gender influences the stylometric profile of the text in a systematic way. In order to explore this hypothesis we measure a wide variety of stylometric textual features in a balanced corpus of news texts written form both men and women. More specifically in the first part of this study we present the most important research findings of neurobiology related to cross-gender differentiation of linguistic ability. In the second part we examine whether author's gender leave a stylometric trace in the text. We use multivariate techniques in order to explore whether specific stylometric variables groups differ systematically across authors' gender.

2 Brain diversity between men and women

2.1 Anatomical differences in the brain

For a long time diverse behavior patterns between men and women were attributed to the unequal socio-cultural conditions existing in western societies.

However, recently a large number of studies have controlled the impact of social structure concluding that a significant amount of cross-gender variation can

be explained due to the biological differences existing in the brain. We know already that babies in their first months already present gender variation, reflecting the later differentiations observed in male and female brains (Moir & Jessel, 1992).

One of the most important gender-based brain anatomical differences is related to the corpus callosum, the tissue that connects the left and right cerebral hemispheres and facilitates interhemispheric communication. Different studies have connected this specific difference with female "intuition" (Gorman & Nash, 1992) and musical "talent" (Levitin, 2006). Recent research (Clarke et al., 2007) has confirmed its difference between sexes but its exact cognitive role has not been determined.

Another important brain difference can be found in the inferior-parietal lobule (Frederikse, Lu, Aylward, Barta, & Pearlson, 1999). This area is located over the ears and at the height of the temple and has been found to be significantly bigger in men than women. Also in men the right lobe is larger than the left, while in women this asymmetry is reversed. The right lobe is also connected to the temporary memory, a function that the brain needs to understand and manipulate spatial relationships and the ability to understand the relationships that exist between different parts of the body. It is also associated with the perception of our own emotions. The left lobe, on the other hand, is involved in the perception of time and speed and the ability of mental rotation of three-dimensional images.

2.2 Functional Magnetic Resonance Imaging – fMRI findings

One of the major developments in the field of diagnostic imaging is the development of Functional Magnetic Resonance Imaging (fMRI). This technique when applied to the brain can show in real time which parts of the cerebral cortex show increased activity by measuring the amount of circulating blood. The use of fMRI in experimental conditions with controlled stimuli can reveal which regions of the brain are associated with specific skills.

The main finding in language tests is the functional lateralization observed in fMRI of men, i.e. the exploitation of only the left lobe for processing linguistic data (Shaywitz et al., 1995). Instead, women seem to use both hemispheres of the cerebral cortex when they produce, as well as when they hear human speech.

The simultaneous use of both hemispheres in the female brain is partially explained by the larger corpus callosum, which has the female brain (see above in section 2.1) and is currently the most important biological interpretation of female superiority in language processing.

The ability of distributed language processing allows faster and more accurate processing of linguistic data (Kimura, 2000; Linn & Petersen, 1985). Instead, as a result of functional lateralization, men have twice the rates in dyslexia (Flannery, Liederman, Daly, & Schultz, 2000) and significantly higher rates of aphasia in stroke (McGlone, 1980).

Surveys from Kimura and her associates (Kimura, 1993; Kimura & Hampson, 1994) show that from the total number of patients who suffer some kind of damage to the left hemisphere of the brain, more men (48.5%) than women (30%) show signs of aphasia. The relationship between affected brain area and gender has now been determined more accurately and we now know that hits in the left anterior cerebral cortex affect more the linguistic ability of women, while hits in the back left portion of the frontal cortex produce more frequently symptoms of aphasia in men.

fMRI studies conducted in recent years have not produced consensus related to the existence of functional lateralization in men. One of the largest meta-analysis of fMRI data (Sommer, Aleman, Bouma, & Kahn, 2004) concluded that the hypothesis of functional lateralization cannot be accepted with certainty for the general population. However, recent research results (Harrington & Farias, 2008) show that men and women activate different regions of the brain when they face specific linguistic tests, supporting the hypothesis of biological diversity of linguistic competence between sexes.

3 Text profiling studies predicting the author's gender

Information Retrieval and Text Mining research was among the first fields that tried to profile the author of a text using stylometric features and machine learning algorithms. Author profiling falls into the standard paradigm of text classification with class labels the author's gender (Argamon, Koppel, Pennebaker, & Schler, 2007; Koppel, Argamon, & Shimoni, 2002; Schler, Koppel, Argamon, & Pennebaker, 2006), age (Argamon, Koppel, Fine, & Shimoni, 2003) or psychological type (Argamon, Dhawle, Koppel, & Pennebaker, 2005; Luyckx & Daelemans, 2008a, 2008b).

One of the first studies which tried to use stylometric features to predict author's gender was from Koppel et al. (2002). They compiled a sub corpus controlled for genre from British National Corpus (BNC) which contained 566 texts written from equal number of men and women authors. They counted a wide variety of topic-neutral stylometric features including the 405 most frequent function words and the most frequent Part of Speech n-grams. The total vector size contained 1081 features which trained a variant of the Exponential Gradient algorithm. The accuracy of the author's gender prediction ranged from 79.5% in literary texts to 82.6% in non-literary texts. One of the most interesting finding was that literary texts used different features from non-literary text to mark gender. Moreover, previous findings that women and men use more frequently different Parts of Speech (pronoun and definite article correspondingly) were confirmed.

The same research group used a large corpus from blogs (37,475 blog posts totalling 300 million words) and tried to predict both the authors' gender and age (Schler et al., 2006). The specific study used 1,502 features including specific content words, selected morphological categories, function words and blogs specific

features such as "blog words" - lol, haha, ur, etc. - and hyperlinks. The machine learning algorithm used was Multi-Class Real Winnow and the prediction accuracy for the author's gender reached 80.1%. Interestingly, the authors noted that despite the great diversity found among stereotyped word content usage between men and women, the most important gender distinctive features were semantically neutral (such as frequent functional words and Parts of Speech).

In another study Corney (2003) analyzed an email corpus and tried to predict the email sender's gender. He used a wide variety of stylometric features including the most frequent function words, the word and sentence length, etc. The prediction accuracy reached 70.1% and the most important gender predictors were the most frequent function words, the average word length and the letter frequencies.

Hota et al. (2006) studied the linguistic usage of men and women characters in 34 Shakespeare plays. The main research question posed was whether a male author could effectively approximate features of woman's language producing real characters and natural dialogues. The researchers used semantic neutral features (frequent functional words, numbers, prepositions, contracted word forms) and content words with high frequency (greater than 10) in the corpus. Prediction accuracy ranged from 60% to 75% depending on the features used. The authors interpret the somewhat less precise gender identification, in the fact that Shakespeare although intuitively approached the language of women characters, failed to deliver it in its entirety.

4 Methodology

4.1 Corpus description

A serious problem related with the corpora used in the authorship attribution studies is the lack of their homogeneity. As Rudman (1997) states the most striking deficiencies are:

- The improper selection, unavailability or fragmentation of the texts.
- The text normalization that often applies from the editor or the publisher causing serious distortion in the writer's style.
- The differences observed in many cases between training and cross-validated texts in terms of genre, topic, date and medium.

Linguistic variation extends across text genre, topic and medium. The linguistic boundaries between these categories are obscure and the linguistic structures exhibit frequencies which co-vary with topic and genre, or medium and topic. Even the most abstract stylometric variables exhibit significant correlation with text metadata such as topic and genre. Examples of this correlation can be found in Mikros & Argiri (2007) who examined the effect of the topic in the authorship information

carried by several stylometric variables widely used in authorship research. The study demonstrated that many stylometric variables can be used with success in topic classification. This characteristic is highly undesirable especially in cases where the researcher attempts authorship attribution in corpora where topic and other textual metadata have not been taken into account.

For the needs of our research we developed a corpus which was controlled simultaneously for the author's gender, text topic, genre and medium. More specifically the corpus design was based on the following premises:

- Equal number of texts (50) written by male and female authors.
- Each text from a male author with specific topic and genre should be matched by a text in the same topic and genre from a female author.
- All texts should be published in the same newspaper (Eleftherotypia) in a brief time span (1 year).
- The collected texts should belong to many different and distinct topics and genres in order to represent a wide socio-pragmatic space of language usage.

The resulting corpus contains 700 texts equally divided in 7 male and 7 female authors. Although, there are some small differences between specific topic and genre categories in spite of the strict sampling restrictions described above, the corpus should be considered balanced. Its size in words (W) and number of texts (N) is displayed in Table 1.

Table 1
Corpus size breakdown by author's gender, text topic and genre

	Topic / **Genre**	**Science**		**Society**		**Economy**		**Art**		**Total**	
		N	W	N	W	N	W	N	W	N	W
Female	Opinion	7	3,169	84	60,489	14	5,748	17	14,568	**122**	**83,974**
	News	11	6,308	111	77,811	31	16,982	15	10,865	**168**	**111,966**
	Discourse	8	4,999	33	23,071	2	1,646	17	21,710	**60**	**51,426**
	Subtotal	*26*	*14,476*	*228*	*161,371*	*47*	*24,376*	*49*	*47,143*	***350***	***247,366***
Male	Opinion	8	3,847	88	60,283	14	8,453	17	11,301	**127**	**83,884**
	News	17	8,353	117	68,793	32	20,471	16	11,023	**182**	**108,640**
	Discourse			22	20,595	2	1,736	17	17,218	**41**	**39,549**
	Subtotal	*25*	*12,200*	*227*	*149,671*	*48*	*30,660*	*50*	*39,542*	***350***	***232,073***
	Total	**51**	**26,676**	**455**	**311,042**	**95**	**55,036**	**99**	**86,685**	**700**	**479,439**

The specific corpus aims to form a difficult challenge for stylometric analysis. It is highly homogeneous regarding its textual metadata and additionally contains small size texts which is untypical of most corpora used in stylometry. In particular, 84% of the texts have less than 1,000 words and this poses a further difficulty since most

stylometric variables exhibit authorship quantitative patterns in larger text sizes (Baillie, 1974; Ledger & Merriam, 1994).

The texts were obtained using "Minotauros" a tool for creating corpora from web sources (Koutsis, Kouklakis, Mikros, & Markopoulos, 2005). Tokenization and Part of Speech tagging was performed by 'Ellogon' a multi-lingual, cross-platform, general-purpose language engineering environment, developed from the Institute of Informatics and Telecommunications, NCSR "Demokritos" (Petasis, Karkaletsis, Paliouras, Androutsopoulos, & Spyropoulos, 2002). Measurements of various stylometric variables were made using "Corpus Manager" (Kouklakis, Mikros, Markopoulos, & Koutsis, 2007) as well as specialized PERL scripts.

4.2 Feature sets

In this study we measured six broad sets of stylometric features which contain both lexical and sublexical units. Each set groups a number of variables which function complementarily and all together approximate a specific textual construct. Although the listing is not exhaustive, it contains most of the variables that have been employed in modern stylometric research and we consider them as socio-linguistically neutral. All the features used in this study are the following:

1. *Lexical "richness"*
   a) Yule's K: Vocabulary richness index that exhibits stability in different text sizes (Tweedie & Baayen, 1998).
   b) Lexical Density: The ratio of functional to content words frequencies in the text, also known as Functional Density (Miranda & Calle, 2007).
   c) % of Hapax- and Dis-legomena: The percentage of words with frequency 1 and 2 in the text segment.
   d) Dis-/Hapax-legomena: The ratio of dis-legomena to hapax-legomena in the text segment, indicative of authorship style (Hoover, 2003).
   e) Relative entropy: Is defined as the ratio between the text entropy and its maximum entropy multiplied by 100. Maximum entropy for a text is calculated if we assume that every word appears with frequency 1 (Oakes, 1998, p. 62).
   f) Word rareness: Percentage of words in each text which do not belong to the 5,000 and the 10,000 most frequent words of Modern Greek.
2) *Word length*
   a) Average word length (per text) measured in letters.
   b) Word length distribution: The frequency of words of 1, 2, 3 … 14 letters long normalized in 1,000 words sample.
3) *Sentence length*
   a) The average sentence length measured in words.

b) The percentage of long sentences (>18 words) in each text.

4) *Character frequencies*
   The frequency of each letter in the text segment normalized in 1,000 words sample. We measured in total 31 letters (we calculated separately the frequencies of the stressed and the unstressed vowels since in Modern Greek spelling the stressed vowels have stress marked orthographically, thus representing different grapheme).
5) *Part of Speech frequencies*
   The frequency of each Part of Speech tag, expressed as percentage of the text size.
6) *Frequent Function Words (FFW)*
   The frequency of the 50 most frequent function words of Modern Greek normalized in 1,000 words sample.

4.3 Statistical analysis

Gender effect analysis in linguistic production requires multivariate methods. In order to examine in detail the way each of the six variable sets relates to author's gender we used Multiple Analysis of Variance (MANOVA). MANOVA is a multivariate statistical analysis which is specifically designed to analyze the effect of one or more categorical independent variables on two or more continuous dependent variables. Although the problem could be tackled with multiple univariate tests, the overall Type I error will be inflated and the probability to reject the null hypothesis when it is true is increased. MANOVA controls against Type I error and offers an omnibus test of significance that takes into account the effect of the independent variable(s) to all the dependent variables simultaneously (Weinfurth, 1995). Furthermore, MANOVA is particularly useful if the dependent variables are conceptually related and there is a moderate inter-correlation between them. Since each variable group contains stylometric variables that attempt to measure the quantitative expression of a specific textual construct, we expect a certain amount of redundancy. MANOVA takes into account this shared common information and tests the effect of the independent variable(s) in a multivariate way (i.e., taking all dependent variables at once). Another reason why a multivariate approach is preferable in our data is that it can detect differences when groups differ on a system of variables (Huberty & Morris, 1989). MANOVA finds a linear composite of the dependent variables that maximizes the separation of the categories that form the independent variable.

A non-significant MANOVA result means that the specific set of dependent variables examined simultaneously do not differ across the categories of the independent variable and no further analysis should be made. A significant MANOVA however indicates that at least one dependent variable differs significantly across the

categories of the independent variable. In the relevant literature most researchers perform univariate tests (t-tests in our case) with adjusted alpha level (Bonferroni correction) for each of the dependent variables in order to detect which variable is different between the categories of the independent variable (Hair Jr, Anderson, Tatham, & Black, 1995; Stevens, 2002). This procedure however has been criticized (Bray & Maxwell, 1982; Huberty & Morris, 1989) among others for confusing the univariate with the multivariate research questions. Since gender and language structure interact in complex and multilevel ways we chose to further explore significant MANOVAs with Discriminant Function Analysis (DFA). Conducting DFA following a significant multivariate effect allows the researcher to investigate in detail the linear composites of the dependent variables and to determine their structure as well as the weights of each dependent variable (Meyers, Gamst, & Guarino, 2006).

5 Results

5.1 Feature group importance

In our data we performed separated MANOVAs for each one of the six stylometric groups with independent variable the author's gender. The multivariate statistic we calculated was Hotelling T^2 which is the multivariate counterpart of the univariate t statistic. Furthermore, partial η^2 was calculated indicating the percentage of the variance explained by the combined dependent variables.
Table **2** summarizes the MANOVA results in the six variable groups.

Table 2
Ranking of the feature groups based on their explanatory power (Partial η^2) in the author's gender

Feature Groups	**Hotelling T^2**	**p**	**Partial η^2**
Frequent Function Words	416.008	0.000	0.374
Character Frequencies	252.676	0.000	0.266
Word Length	101.908	0.000	0.128
Part of Speech frequencies	59.33	0.000	0.078
Lexical "richness"	17.45	0.032	0.024
Sentence length	2.792	0.211	0.004

As can be seen all stylometric groups had a multivariate statistically significant effect in author's gender except sentence length. The group that accounts for the biggest amount of variance is Frequent Function Words (37%) followed by

Character Frequencies (27%) and Word Length (13%). Small (< 10%) but statistically significant amount of explained variance in author's gender have the Part of Speech frequencies and the Lexical "richness". For each of the five variable groups that Hotelling T^2 was found statistically significant we performed DFA in order to further explore the linear composite structure and to assess each dependent's variable contribution to author's gender discrimination.

5.2 Frequent function words DFA analysis

The DFA with Frequent function words variables as independent and author's gender as dependent showed that 28 variables differ significantly between the male and the female authors.

Table 3 and all subsequent tables in next sections summarize the DFA results. In particular each table presents the Wilk's λ, of each statistically significant predictor, mean (M) and standard deviation (SD) in male ($_M$) and female ($_F$) authors as well as the within-groups correlations between the predictors and the discriminant function. Furthermore, standardized weights for each variable are reported in order to assess their relative importance in author's gender discrimination.

In Table 3 we see that male authors use more frequently the words (with decreasing importance in the author's gender discrimination) όμως [instead], αλλά [but], στην [in], σ'[contracted form of 'in'], ο [male singular article], απ' [contracted form of 'from'] , τη [female singular article], της [female singular article in genitive case], την [female singular article or female personal pronoun], με [with], που [where], η [female singular article] while female authors present higher percentages in the use of μας [us], των [article in genitive case], το [neutral singular article], δεν [not], σε [in], οι [plural article], μόνο [just], μέσα [inside], πως [how], σου [personal pronoun in genitive], τους [them], γιατί [why], τα [neutral plural article], πάνω [on], στα [in], από [from]. Among the words that males use more frequently we can group two distinct categories: a) coordinated conjunctions (αλλά, όμως) and b) contracted forms of prepositions (σ', απ'). The former category characterizes the syntactic structure of the text and previous research has revealed that can be used as a potential gender discriminator (Mulac, Bradac, & Mann, 1985; Mulac, Studley, & Blau, 1990). The latter grouping (contracted forms) has also been described by many linguists as a typical male marker in text production (Baron, 2004).

Table 3

Ranking of frequent function words based on their overall usefulness (absolute value of the standardized coefficient) in the authors' gender differentiation

Predictors	Wilk's λ	p	MM	SDM	MF	SDF	Correlation Coeff.	Stand. Coeff.
μας	0.953	0	0.143	0.247	0.306	0.459	0.286	0.443
όμως	0.908	0	0.282	0.271	0.136	0.178	-0.411	-0.427
των	0.973	0	1.002	0.689	1.274	0.917	0.217	0.381
το	0.975	0	2.055	0.88	2.335	0.87	0.208	0.37
δεν	0.986	0.002	0.775	0.526	0.912	0.615	0.156	0.347
σε	0.973	0	0.731	0.47	0.89	0.48	0.216	0.346
οι	0.977	0	0.809	0.525	0.995	0.68	0.198	0.291
αλλά	0.993	0.03	0.293	0.264	0.251	0.244	-0.106	-0.214
στην	0.968	0	0.887	0.521	0.708	0.462	-0.235	-0.192
μόνο	0.986	0.002	0.104	0.15	0.142	0.172	0.152	0.161
μέσα	0.985	0.001	0.076	0.131	0.112	0.158	0.159	0.159
σ'	0.994	0.045	0.074	0.136	0.049	0.179	-0.099	-0.155
ο	0.988	0.004	1.372	0.774	1.206	0.743	-0.142	-0.13
πως	0.991	0.013	0.081	0.179	0.122	0.248	0.122	0.107
σου	0.994	0.041	0.014	0.059	0.027	0.112	0.1	0.096
απ'	0.992	0.021	0.041	0.101	0.025	0.074	-0.113	-0.082
τους	0.976	0	0.678	0.532	0.854	0.594	0.202	0.075
τη	0.991	0.011	0.824	0.439	0.74	0.441	-0.125	-0.075
της	0.956	0	2.204	0.977	1.794	0.945	-0.276	-0.059
γιατί	0.994	0.045	0.105	0.178	0.132	0.178	0.098	0.052
τα	0.981	0	0.908	0.575	1.092	0.737	0.18	0.046
πάνω	0.99	0.008	0.043	0.099	0.067	0.136	0.129	0.046
την	0.972	0	1.778	0.763	1.534	0.676	-0.219	-0.044
με	0.991	0.01	1.54	0.764	1.407	0.591	-0.127	-0.039
που	0.982	0	1.768	0.61	1.603	0.604	-0.176	-0.033
στα	0.992	0.015	0.285	0.289	0.345	0.351	0.119	0.024
η	0.958	0	1.792	0.882	1.459	0.703	-0.271	0.021
από	0.989	0.006	1.313	0.605	1.444	0.645	0.136	0.004

Instead, in the words that characterize women authors we can distinguish the presence of personal pronouns (μας, σε, σου). The preference of personal pronoun usage has been confirmed by previous corpus-based studies (Argamon et al., 2007; Holmes, 1990; Preisler, 1986; Rayson, Leech, & Hodges, 1997) and is related to the fact that female discourse is characterized by interpersonal involvement. This has also been described by Tannen (1991) as the "report vs. rapport" distinction, i.e. the female speaker/author's tendency to produce texts that concentrate on interaction with her readers/listeners and maintain their relationship while males focus on the information transmission.

5.3 Character frequencies DFA analysis

The DFA with character frequencies as independent variables and author's gender as dependent showed that 12 variables differ statistically significant between the male and the female authors. Table 4 presents the analysis findings.

Table 4
Ranking of character frequencies based on their overall usefulness (standardized coefficient) in the authors' gender differentiation

Predictors	Wilk's λ	p	M_M	SD_M	M_F	SD_F	Correlation Coeff.	Stand. Coeff.
η	0.961	0.000	3.937	0.731	3.643	0.741	0.333	1.241
κ	0.967	0.000	4.038	0.556	3.833	0.563	0.305	1.043
ρ	0.938	0.000	4.473	0.508	4.197	0.567	0.426	1.033
λ	0.979	0.000	2.765	0.482	2.627	0.473	0.240	0.870
ί	0.987	0.003	2.501	0.387	2.419	0.346	0.187	0.816
ό	0.993	0.028	2.091	0.402	2.029	0.348	0.138	0.777
ο	0.977	0.000	7.871	0.859	8.139	0.922	-0.250	0.749
α	0.983	0.000	9.246	0.871	9.476	0.887	-0.218	0.558
ζ	0.988	0.004	0.358	0.197	0.318	0.168	0.181	0.475
δ	0.990	0.011	1.712	0.364	1.788	0.423	-0.159	0.359
γ	0.973	0.000	1.703	0.402	1.829	0.369	-0.273	0.320
ω	0.992	0.020	1.430	0.338	1.492	0.366	-0.146	0.274
φ	0.978	0.000	0.801	0.238	0.882	0.307	-0.245	0.144

In Table 4 we can see that male authors use more frequently (with decreasing importance in the author's gender discrimination) the characters η, κ, ρ, λ, ί, ό, ζ. On the other hand female authors use more frequently the characters ο, α, δ, γ, ω, φ. The relative importance of each character was determined by its standardized coefficient.

5.4 Word length DFA analysis

The DFA with Word length as independent variable and author's gender as dependent showed that seven variables differ significantly between the male and the female authors. Table 5 summarized the analysis findings.

Table 5
Ranking of word lengths based on their overall usefulness (absolute value of the standardized coefficient) in the authors' gender differentiation

Predictors	Wilk's λ	p	M_M	SD_M	M_F	SD_F	Correlation Coeff.	Stand. Coeff.
2 letter words	0.967	0.000	11.361	2.079	12.131	2.073	0.487	0.869
3 letter words	0.994	0.046	23.156	2.208	23.498	2.331	0.198	0.720
13 letter words	0.983	0.000	1.100	0.613	1.270	0.667	0.349	0.442
14 letter words	0.991	0.011	0.704	0.471	0.799	0.517	0.252	0.367
8 letter words	0.976	0.000	7.402	1.488	6.932	1.488	-0.414	-0.086
4 letter words	0.984	0.000	10.585	2.045	10.050	2.163	-0.334	-0.040
9 letter words	0.987	0.003	6.188	1.562	5.842	1.513	-0.295	0.009
10 letter words	0.991	0.012	5.317	1.400	5.052	1.408	-0.247	0.003

In Table 5 we see that female authors use greater percentage of 2, 3, 13 and 14 letter words while male authors present higher percentages in the use of 4, 8, 9 and 10 letter words. The examination of the standardized coefficients shows that the most useful markers for the detection of female writing is the percentage of 2 and 3 letter words followed by the percentage of 13 and 14 letter words. Correspondingly, the most useful markers for the detection of the male writing is the percentage of 8 and 4 letter words followed by the percentage of 9 and 10 letter words. These results give us a relative clear picture regarding female writing in news and word length. Female authors use more than males the lower and upper boundary of the word length spectrum. They use smaller words (2-3 letter words) which in their majority in Modern Greek belong to the group of function words. They use also many words which have many letters (13 and 14 letter words) and they are related inversely to function word usage. Since the 13 and 14 letter words have relatively small effect on gender discrimination compared to the 2 and 3 letter words, we can hypothesize that they reflect inversely the major trend of the small words to characterize women's writing. This hypothesis is further supported by examining the correlations of 2, 3, 13 and 14 letter words with Lexical density in the female data. 2 and 3 letter words appear to be in a statistically significant negative correlation (r_{2lw}= -0.354, r_{3lw} = -0.385) with the lexical density, meaning that increase in lexical density (i.e. more content words) relates inversely to the percentage of 2 and 3 letter words. On the other hand, 13 and 14 letter words have smaller but statistically significant positive correlation (r_{13lw} = 0.134, r_{14lw} = 0.144) with lexical density, meaning that as lexical density increases the percentage of longer words increases also but with smaller pace.

5.5 Part of Speech frequencies DFA analysis

The DFA with Part of Speech frequencies as independent variables and author's gender as dependent showed that only the usage of Adverbs (Wilk's $\lambda = 0.987$, $p = 0.003$) and the usage of Adjectives (Wilk's $\lambda = 0.995$, $p = 0.049$) present statistically significant differences between male and female authors. More specifically male authors use increased percentage of adverbs ($M = 8.2$, $SD = 1.9$) compared to female authors ($M = 7.8$, $SD = 1.7$) and female authors use increased percentage of adjectives ($M = 8.2$, $SD = 2.2$) compared to male authors ($M = 7.9$, $SD = 1.9$).

Adverb usage demonstrated strong relationship with the discriminant function with correlation coefficient -0.403 and standardized coefficient -0.435 whereas adjective usage exhibited weaker association with correlation coefficient 0.256 and standardized coefficient 0.464.

5.6 Lexical "richness" DFA analysis

The DFA with Lexical "richness" as independent variables and author's gender as dependent showed that only the Percentage of hapax legomena (Wilk's $\lambda = 0.993$, $p = 0.025$) present statistically significant differences between male and female authors. More specifically, male authors have higher percentage of hapax legomena ($M = 41.2$, $SD = 7.6$) compared to female authors ($M = 39.9$, $SD = 7.2$).

Percentage of hapax legomena demonstrated strong correlation with the discriminant function with correlation coefficient 0.542 and standardized coefficient 0.875.

A closer inspection of the association of the lexical "richness" variables with the author's gender reveals a complex and heterogeneous picture that is characteristic of the complexity of the relationship between author's gender and textual stylometric profile. Although Relative entropy and the percentage of words which do not belong to the most frequent 5000 words of the corpus theoretically measure the same abstract textual property, i.e. lexical "richness", appear to be inversely related to author's gender. Women write texts with rare vocabulary while men's texts present less lexical repetition and avoidance of standardized lexical patterns.

6 Conclusions

The present study investigated the role of the author's gender in the systematic differentiation observed in the stylometric profile of texts of men and women authors. Using a corpus compiled in a way to experimentally control text topic, genre and medium we studied a wide array of stylometric features and their usage distribution in men and women's texts. Multivariate statistical analysis (MANOVA

followed by Discriminant Function Analysis) revealed that men and women use indeed differently most stylometric features, a fact that can be further exploited for the development of author's gender profiling systems.

References

Argamon, Shlomo; Dhawle, Sushant; Koppel, Moshe; Pennebaker, James (2005). Lexical predictors of personality type *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America: Theme: Clustering and Classification, 8-12 Jun 2005*. St. Louis, MO.

Argamon, Shlomo; Koppel, Moshe; Fine, Jonathan; Shimoni, Anat Rachel (2003). Gender, genre, and writing style in formal written texts. *Text, 23*(3), 321-346.

Argamon, Shlomo; Koppel, Moshe; Pennebaker, James W.; Schler, Jonathan (2007). Mining the blogosphere: Age, gender and the varieties of self–expression. *First Monday,12*(9). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index. php/fm/article/view/2003/1878

Baillie, D. W. (1974). Authorship attribution in Jacobean dramatic texts. In J. L. Mitchell (Ed.), *Computers in the humanities*. Edinburgh: Edinburgh University Press.

Baron, Naomi S. (2004). See you Online. *Journal of Language and Social Psychology, 23*(4), 397-423. doi: 10.1177/0261927x04269585

Bray, James H.; Maxwell, Scott E. (1982). Analyzing and Interpreting Significant MANOVAs. *Review of Educational Research, 52*(3), 340-367. doi: 10.3102/00346543052003340

Clarke, Dave; Wheless, James; Chacon, Monica; Breier, Joshua; Koenig, Mary-Kay; McManis, Mark; Baumgartner, James (2007). Corpus callo-sotomy: A palliative therapeutic technique may help identify resectable epileptogenic foci. *Seizure, 16*(6), 545-553.

Corney, Malcolm Walter (2003). *Analysing e-mail text authorship for forensic purposes.* (Master), Queensland University of Technology, Queensland.

Flannery, Kathleen A.; Liederman, Jacqueline; Daly, Louise; Schultz, Jennifer K. (2000). Male prevalence for reading disability is found in a large sample of black and white children free from ascertainment bias. *Journal of the International Neuropsychological Society, 6*(4), 433-442.

Frederikse, Melissa E.; Lu, Angela; Aylward, Elizabeth; Barta, Patrick; Pearlson, Godfrey (1999). Sex Differences in the Inferior Parietal Lobule. *Cereb. Cortex, 9*(8), 896-901. doi: 10.1093/cercor/9.8.896

Gorman, Christine; Nash, Madeleine (1992, 20 January 1992). Sizing up the sexes. *TIME*, 36-43.

Hair Jr. Joseph F.; Anderson, Rolph E.; Tatham, Ronald L.; Black, William C. (1995). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ, USA: Prentice-Hall.

Harrington, Greg S.; Farias, Sarah Tomaszewski (2008). Sex differences in language processing: Functional MRI methodological considerations. *Journal of Magnetic Resonance Imaging, 27*, 1221-1228.

Holmes, Janet (1990). Hedges and boosters in women's and men's speech. *Language and Communication, 10*(3), 185-205.

Hoover, David (2003). Another perspective on vocabulary richness. *Computers and the Humanities, 37*, 151-178.

Hota, Sobhan R.; Argamon, Shlomo; Koppel, Moshe; Zigdon, Iris (2006). *Performing gender: Automatic stylistic analysis of Shakespeare's characters.* Paper presented at the Proceedings of Digital Humanities 2006, Paris.

Huberty, Carl J.; Morris, John D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin, 105*(2), 302-308.

Kimura, Doreen (1993). *Neuromotor mechanisms in human communication*. Oxford: Oxford University Press.

Kimura, Doreen (2000). *Sex and cognition*. Cambridge, MA: MIT Press.

Kimura, Doreen; Hampson, Elizabeth (1994). Cognitive pattern in men and women is influenced by fluctuations in sex hormones. *Current Directions in Psychological Science, 3*(2), 57-61.

Koppel, Moshe; Argamon, Shlomo; Shimoni, Anat Rachel (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing, 17*(4), 401-412.

Kouklakis, George; Mikros, George K.; Markopoulos, George; Koutsis, Ilias (2007). *Corpus Manager: A tool for multilingual corpus analysis*. Retrieved from http://ucrel.lancs.ac.uk/publications/CL2007/paper/244_Paper.pdf.

Koutsis, Ilias; Kouklakis, George; Mikros, George K.; Markopoulos, George (2005). *MINOTAVROS: A tool for the semi-automated creation of large corpora from the Web*(Vol.1). Retrieved from http://www.corpus.bham.ac.uk/PCLC/minotavros.doc.

Ledger, Gerard; Merriam, Thomas (1994). Shakespeare, Fletcher, and the Two Noble Kinsmen. *Literary and Linguistic Computing, 9*, 235-248.

Levitin, Daniel (2006). *This is your brain on music: The science of a human obsession*. New York: Dutton Adult.

Linn, Marcia C.; Petersen, Anne C. (1985). Emergence and characterization of sex differences in spatial ability: a meta-analysis. *Child Development, 56*, 1479-1498.

Luyckx, Kim; Daelemans, Walter (2008a). Personae: A corpus for author and personality prediction from text. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis & Daniel Tapias (Eds.), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 28-30 May 2008*. Marrakech, Morocco.

Luyckx, Kim; Daelemans, Walter (2008b). Using syntactic features to predict author personality from text *Proceedings of Digital Humanities 2008 (DH 2008)* (pp. 146-149).

McGlone, Jeanette (1980). Sex differences in human brain organization: a critical survey. *Behavioral Brain Science, 3*, 215-227.

Meyers, Lawrence S.; Gamst, Glenn; Guarino, A.J. (2006). *Applied multivariate research. Design and interpretation*. Thousand Oaks, CA: Sage.

Mikros, George K.; Argiri, Eleni K. (2007). Investigating topic influence in authorship attribution. In Benno Stein, Moshe Koppel & Efstathios Stamatatos (Eds.), *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection* (Vol. 276, pp. 29-35). Amsterdam, Netherlands: CEUR.

Miranda, García Antonio; Calle, Martín Javier (2007). Function words in authorship attribution studies. *Literary and Linguistic Computing*, *22*(1), 49-66.

Moir, Anne; Jessel, David (1992). *Brain sex: The real difference between men and women*. New York: Delta.

Mulac, Anthony; Bradac, James J.; Mann, Susan Karol (1985). Male/female language differences and attributional consequences in children's television. *Human Communication Research, 11*(4), 481-506.

Mulac, Anthony; Studley, Lisa B.; Blau, Sheridan (1990). The gender-linked language effect in primary and secondary students' impromptu essays. *Sex roles, 23*(9-10), 439-470.

Oakes, Michael P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Petasis, Georgios; Karkaletsis, Vangelis; Paliouras, Georgios; Androutsopoulos, Ion; Spyropoulos, Constantine, D. (2002, 29-31 May 2002). *Ellogon: A new text engineering platform.* Paper presented at the Proceedings of the third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain.

Preisler, Bent (1986). *Linguistic sex roles in conversation: Social variation in the expression of tentativeness in English*. Berlin: Mouton de Gruyter.

Rayson, Paul; Leech, Geoffrey; Hodges, Mary (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational com-

ponent of the British National Corpus. *International Journal of Corpus Linguistics, 2*(1), 133-152.

Rudman, Joseph (1997). The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities, 31*(4), 351-365.

Schler, Jonathan; Koppel, Moshe; Argamon, Shlomo; Pennebaker, James (2006). *Effects of age and gender on blogging.* Paper presented at the Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.

Shaywitz, Bennet A.; Shaywitz, Sally E.; Pugh, Ken R.; Constable, Todd R.; Skudlarski, Pawel; Fulbright, Robert K.; Gore, John C. (1995). Sex differences in the functional organization of the brain for language. *Nature, 373*, 607-609. doi: 10.1038/373607a0

Sommer, Iris; Aleman, André; Bouma, Anke; Kahn, René (2004). Do women really have more bilateral language representation than men? A meta-analysis of functional imaging studies. *Brain, 127*(8), 1845-1852. doi: 10.1093/brain/awh207

Stevens, James P. (2002). *Applied multivariate statistics for the social sciences* (4rth ed.). Hillsdale, NJ: Erlbaum.

Tannen, Deborah (1991). *You just don't understand: Women and men in conversation*. London: Virago Press.

Tweedie, Fiona J., & Baayen, Harald R. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, 32*(5), 323-352.

Weinfurth, Kevin P. (1995). Multivariate Analysis of Variance. In Laurence G. Grim & Paul R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 245-276). Washington, DC: American Psychological Association.