SOCIOLINGUISTIC FEATURES FOR AUTHOR GENDER IDENTIFICATION, 2016
http://dx.doi.org/10.1080/09296174.2016.1226430

Routledge
Taylor & Francis Group

Sociolinguistic Features for Author Gender Identification: From Qualitative Evidence to Quantitative Analysis

Vasiliki Simaki[a,b], Christina Aravantinou[c], Iosif Mporas[d], Marianna Kondyli[e] and Vasileios Megalooikonomou[c]

[a]Department of Computer Science, Linnaeus University, Sweden; [b]Centre for Language and Literature, Lund University, Sweden; [c]Department of Computer Engineering and Informatics, University of Patras, Greece; [d]University of Hertfordshire, UK; [e]Department of Educational Science and Early Childhood Education, University of Patras, Greece

ABSTRACT

Theoretical and empirical studies prove the strong relationship between social factors and the individual linguistic attitudes. Different social categories, such as gender, age, education, profession and social status, are strongly related with the linguistic diversity of people's everyday spoken and written interaction. In this paper, sociolinguistic studies addressed to gender differentiation are overviewed in order to identify how various linguistic characteristics differ between women and men. Thereafter, it is examined if and how these qualitative features can become quantitative metrics for the task of gender identification from texts on web blogs. The evaluation results showed that the "syntactic complexity", the "tag questions", the "period length", the "adjectives" and the "vocabulary richness" characteristics seem to be significantly distinctive with respect to the author's gender.

1. Introduction

Theoretical and empirical studies prove the strong relationship between social factors and linguistic attitudes. The language loan due to the language contact, the linguistic change through time, the different social context influencing the speaker's linguistic choices, the cultural information that linguistic meanings carry, are some examples pointing out the relation between language and society. An essential principle of sociolinguistics is that social and linguistic activity and attitude are mutually dependent and influenced (Labov, 1972; Kakridi-Ferrari, 2005). The linguistic variation according to the speakers' social categories is the main object of sociolinguistics.

Since language is perceived as a social activity, reflecting and/or influencing the social reality, the relation between language and society may exist in the sense

CONTACT Vasiliki Simaki vasiliki.simaki@lnu.se

that social relations are registered in language. Sociolinguistic research examines the bidirectional and systematic relations between different linguistic systems and the social environment in which they are used (Trudgill, 1972, 2000).

The linguistic variation can be observed at different levels of language analysis (intonation, phonological, morphological, syntactic), and it is perceived as a socially different – but linguistically equal – way to say the same thing (Kobayashi, Matsumura, & Ishizuka, 2007). Various social categories related to gender, age, education, profession, and social status, do make different sociolinguistic choices in their everyday life, according to the communicative situation. It is considered that these choices behave as markers of social characteristics either of the speakers (gender, age, education, etc.) or the communicative situation (formal, informal, spontaneous, etc.) (Archakis & Kondyli, 2004).

The task of extracting distinctive language characteristics among different texts has been a challenging task in quantitative and computational linguistics, where studies primarily focused in the identification of text genre and authorship. In Stamatatos, Fakotakis, and Kokkinakis (2000) stylistic markers were extracted through classification methods in order to detect the kind of a text's genre and the identification of the text's author. Their approach is lexical-based and the researchers achieve high accuracy in an automated classification methodology. In other studies, statistical analysis is performed in order to observe the consistency and the stylistic variations of texts, as in Bagavandas and Manimannan (2008) where they reveal stylistic distinctive characteristics between different authors. Savoy (2012) evaluated an authorship attribution method in three different corpora in English, French and German, and proved that word types and lemmas features are highly efficient for the entire multilingual corpus.

The extraction of demographic information from text has also been extensively studied. Przybyła and Teisseyre (2014) reported a study on the detection of demographic features (gender, age, education, political party preferences) of the Polish parliament deputies, based on text and word characteristics, investigating several classification algorithms. Their results overcome language barriers and successfully identified the politicians' demographic profile. Moreover, the gender factor, i.e. how the linguistic behaviour of people is related to gender, has been studied.

The differences between men's and women's spoken and written language are crucial in the sociolinguistic research, since men and women are considered to be not just two biologically different entities but socially constructed as two different social groups. The participation in each social group implies different duties, privileges and, by extension, different linguistic attitudes. In modern/western societies the different linguistic choices between genders concern mostly the preference of use of specific characteristics (sex-preference differences), which can be observed both in phonetic/phonological, morphological, syntactic, lexical, semantic level and according to communicative situation.

The specific characteristics observed after empirical studies could be called "the language of women" or, from another perspective, the "feminine language". In this paper, the phenomena of differentiated linguistic choices between women and men after the relevant literature overview are listed. Then, these choices are converted into measurable characteristics and detected in a gender-annotated corpus. The aim of this study is to show that quantitative features identified in theoretical and empirical sociolinguistic studies, when converted into measurable features, can detect the differences between male and female language attitudes.

The rest of the paper is organized as follows: Section 2 presents an overview on the language and gender issue. In Sections 3 and 4 the corpus we used for our study and the sociolinguistic markers of differentiation between women and men are presented. In Section 5, we quantitatively investigate the sociolinguistic markers using statistical analysis. Finally, in Section 6 we summarize our study's results and discuss them.

2. Language and Gender

Recent empirical findings about the gender linguistic variation are indicative for the existence of the feminine sociolect. In this section, the feminine linguistic choices, on which the feature set used for the gender identification is based, are presented.

The earlier studies in the language and gender issue focused on the differences between men and women in phonological level, without any further deepening. Phonological differentiations only were observed and the researchers of the time, Wilhelm von Humboldt and Jacob Grimm (cit. in Jespersen, 1922), made a distinction of the language in terms of gender, age and educational level, but they denied the existence of a separate feminine language. They supported that women's talk has only some differentiated characteristics; moreover, they assumed that women should not have an active engagement in the elaboration and the enrichment of the language. Grimm was the first to distinguish the biological from the grammatical gender in terms of sociological criteria.

Jespersen (1922) made a more attentive study in women's language according to which women's vocabulary is smaller and more "central" (a term abandoned in current sociolinguistic terminology). Subsequently, the sociolinguistic science evolved and researchers proposed the term *gender* – instead of *sex* – in order to capture sociolinguistic variety, in which they attribute different characteristics. The vocabulary richness, though, remains an important feature in linguistics and, more specifically, in text analysis. This marker indicates not only someone's personal writing style, but also vocabulary patterns used by people belonging to different social groups. Although this characteristic was directly connected to text length, recent studies prove that the vocabulary richness can be text-length independent and more efficient in authorship attribution (Kubát & Milička, 2013).

A general opinion about women's language is that, statistically, women tend to make a more conservative use of language and they use more standard types than men (Gordon, 1997). The only occasion they evade the standard language is when they adapt to socially prestigious changes, local linguistic elements, communicative indirection, and under specific communicative situations. From another perspective, Milroy and Milroy (1985), in their social network theory, claim that gender is a non-homogeneous category in each community. They associate the women's linguistic attitude more with their social status than the gender itself.

The most important findings after a lot of research in language and gender (Lakoff, 1973, 1975, 1990; Fishman, 1983; Cameron 1998, 2005; Bucholtz, 1999; Bucholtz, Liang, & Sutton, 1999; Makri-Tsilipakou, 2010) are summarized below:

- The knowledge and use of refined colour gradations in women's talk has been examined, and compared to men's discourse; women tend to use more analytical ways to describe a specific colour tone (e.g. "cherry blossom pink", "salmon orange", "mint green", etc.), a characteristic which is more frequent in text associated to topics around fashion, makeup, etc., domains in general that attract the feminine interest.
- Another important characteristic in women's talk is the frequent use of "empty" adjectives, adjectives which carry a metaphorical sense of admiration and/or approval. Women tend to make different compliments than men by using adjectives such as "sweet", "divine", "stunning", "lovely", etc.
- Women also prefer a more "gentle" way of conversation, by using questions in place of statements. These forms lay the ground for the conversational opening and/or continuation. A statement like "This car is not a nice colour" may not open a conversation, while the interrogative phrase "Do you like the car's colour?" needs an answer at least.
- Besides specific lexical choices (use of norm types, avoiding of bad words, etc.) that women, unlike men, do, linguists observe that in many cases women try to decrease the illocutionary force of their utterances. This phenomenon is achieved by using palliative forms like tag-questions (e.g. "He is a good boy, isn't he?"), interrogative forms instead of affirmations (e.g. "I should go now?"), extension of requests (e.g. "Hey Dad, will you please drive me to the movies, if you can?"), hedges of uncertainty (e.g. "I'm not so sure", "I don't know", etc.).
- Women have different politeness strategies than men and different ways to agree/disagree. They do not express their agreement/disagreement in a sharp and curt way like men, and they use more polite phrases.
- Women also use more sentimental expressions, indirect requests and hypercorrected grammar types (grammatical construction produced by mistaken analogy, with standard usage out of a desire to be correct). Men on the other hand, tend to use more "bad" words and slang types,

in general, coarser language than women, and in case of disagreement, they use strong and explicit expressions. They insert in their vocabulary non-standard forms and neologisms (newly coined word, expression, or usage).

- Other interesting characteristics are the syntactic complexity and lexical density in male and female talk. The syntactic complexity, which characterizes the female discourse, investigates the presence of more than one clause in a sentence by the use of secondary clauses in the period. The lexical density concerns the use of content words (nouns, adjectives, verbs and adverbs). Alami, Sabbah, and Iranmanesh (2013) study the lexical density in male and female discourse, and compare its relationship to the discourse length. They observe that the lexical density does not have a statistically significant difference between male and female discourse and also, there exists a negative relationship between the lexical density of discourse and the discourse length.

Sociolinguistic studies have identified various linguistic choices related either to women or to men, but it still remains difficult to gather all empirical findings and extract a generalized profile for both genders. Eckert and McConnell-Ginet (1999) make an important effort for new generalizations and explanations in the field of the research about language and gender. The researchers emphasize the subjectivity and the contradiction in the studies, the ideologies, the methodologies used and the author's conclusions during a study. Their statement about the difficulties in that domain of search can be summarized as follows: 'Our understanding of what it means to be male or female in a particular group in the community, in society, and in the world, underlies our interpretation of gender differentiation in language use' (Eckert & McConnell, 1999, p. 188).

Recent studies about gender and language try to merge existing and more radical theories, in order to create patterns about the gender-specific variation, and tend to analyse the meaning and the social context of a given linguistic attitude (Eckert, 2012). There exists also the need to combine the total social information about a given group, in order to examine the samples in terms of more than one variable. The sociological, anthropological and stylistic information in a given communicative situation are of great importance for the explanation of the specific linguistic choice of the speaker, and various studies use this non-linguistic information in order to draw conclusions and new evidence (Bucholtz, 1998, 2002, 2003; Irvine, 2001; Bucholtz & Hall, 2005; Moore & Podesva, 2009). The interdisciplinary methods used by researchers mentioned above inspired this effort of combining sociolinguistic information with statistical and text mining techniques.

These studies, as presented, are used in the present paper, and most sociolinguistic markers of genderized discourse were collected and turned into a quantitative form. In Section 4, we present the challenging task of transforming these linguistic markers into measurable features in order to perform subsequently the statistical analysis.

3. Corpus Description

The corpus used in our study is the 'Blog author gender classification data set' (Mukherjee & Liu, 2010) which consists of a collection of 2936 blog posts from many blog hosting sites and blog search engines. For each blog post the author's gender was labelled by using the available information, i.e. the blogger's profile information, his/hers profile pictures or avatars. The collected posts are equally distributed (half male, half female). The posts may contain a unique sentence, but in most cases they contain a longer text, covering exhaustively a thematic area. The data-set covers a large thematic and stylistic range and it might be useful to extract gender-associated information according to the topic and the style of the posts. However, it should be taken into account that a blog post is a piece of written discourse displaying the main characteristics of written discourse (differences in grammatical complexity, lexical density, nominalization, explicitness). As Hadley (1995) claims 'written text conforms to rules that most successful writers unconsciously follow and native readers unconsciously expect to find'. Since most sociolinguistic researchers' findings are based on oral discourse, it was a major challenge for us to measure and confirm (or disconfirm) these indices in written texts.

For our study the Mukherjee and Liu corpus is divided into female and male texts. The female corpus contains 1390 blog posts (621.845 words, 37.225 sentences) and the male corpus contains 1546 posts (696.127 words, 37.847 sentences). In this article we denote the corpus as $C = \{D_i\}$ and its documents (i.e. blog posts) as D_i, with $1 \le i \le I$. In our case $I = 1390 + 1546 = 2936$. Each document D_i is labelled as M (male) or F (female), according to the author's gender, thus resulting in the male sub-corpus $C_M = \{D_j\}$, with $1 \le j \le J$ and the female corpus $C_F = \{D_k\}$, with $1 \le k \le K$. In our case $J = 1546$ and $K = 1390$.

We can see in Table 1 and Figure 1 the distribution of the female and male blog posts, according to the size of each post. We observe that the female and male corpus contain more blog posts of various sizes. The chart shows that the 'blog posts' text type differs from online comments, tweets, or Facebook status in terms of text length and number of sentences. Therefore, it is suggested that the blog post's size may be a feature of differentiation among social media text types.

The distribution of female and male corpora shows the wide range of the documents' length; so a further division into length-based classes is needed. As we discussed above, texts of a different size may have different characteristics (linguistic and stylistic), even when they belong to the same text type. A current trend in text mining is the classification of texts according to determined sizes. In recent studies (Chen, Jin, & Shen, 2011; Sun, 2012; Vo & Ock, 2015) researchers tend to classify short texts such as article titles, snippets, film/product/other reviews. In the present study, a corpus consisting of texts of different sizes, which are quite heterogeneous in terms of stylistic characteristics, is used. It is not possible to perform experiments so as to search the

Table 1. The number of female and male posts according to their size.

Number of Sentences	Number of posts in female corpus	Number of posts in male corpus
1	30	55
2	27	54
3	55	79
4	62	98
5	92	94
6	75	88
7	69	78
8	55	65
9	71	69
10	47	61
11	53	63
12	50	52
13	41	38
14	55	47
15	46	47
16	40	41
17	34	41
18	38	28
19	25	35
20	40	29
21–25	92	102
26–34	85	76
35–50	64	72
51–100	62	49
>100	49	55
>200	21	21
>300	12	9

same characteristics in a data-set ranging from uni-sentenced texts to texts of more than 100 sentences. For this reason, and after analysis of the statistical distribution of the number and the size of documents, a four-class division of our corpora is adopted: A, B, C, D. Class E contains the totality of the corpora documents, which is also measured. Table 2 presents the categories and their size, the documents' number distributed to each class for the male and female corpus, and shows the average blog post size for each category.

As shown in Table 2, a proportional division into classes for the female and male corpus is tried, without though ignoring the statistical particularities of each collection. It is observed that in overall corpus (Class E), women are more "chatty" per post than men, but this conclusion applies only in the blog posts category containing 1–9 sentences (Class A).

4. Turning Linguistic Characteristics Into Measurable Data

In this part, we investigate how linguistic characteristics of female and male discourse described in Section 2 can be turned into quantitative data, in a form that can be detected and measured into the Mukherjee and Liu's (2010) corpus. Each linguistic characteristic of female and male discourse have been

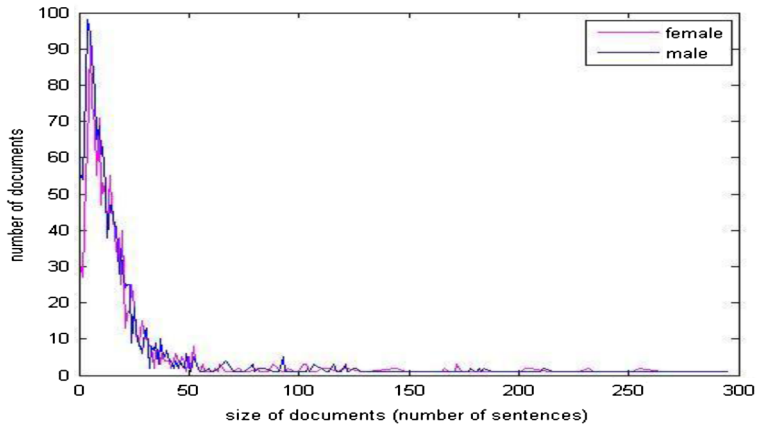

Fig. 1. The distribution of female and male corpus accordingly to the posts' size.

generalized and simplified, in order to find implicit or explicit ways to get it measured in the corpora.

In many sociolinguistic studies, an important feature is the different politeness and agreement/disagreement strategies, which can be measured by detecting some standard phrases ("thank you", "thank you very much", "you're welcome", "appreciated", "much obliged", "may I?", "please", "pardon me", "excuse me", "I'm sorry", "I'm terribly sorry", "I'm very sorry", "sorry", "I beg your pardon", "pardon me", etc.) into the corpora and compare in which of two there are more appearances. It is considered that $PLT_i = \frac{P_i}{W_i}$ is the metric for politeness and agreement/disagreement strategies for post D_i, where P_i is the number of polite, agreement/disagreement phrases in D_i and $W_i = \sum_{k=1}^{K} w_k$ for $w_k \in D_i$ is the total number of words in D_i.

The "empty" adjectives characteristic can be also captured implicitly, by counting the total number of adjectives in two corpora and compare them. It is calculated that $E_i = \frac{ADJ_i}{W_i}$ is the number of "empty" adjectives in D_i, where $ADJ_i = \sum_{l=1}^{L} w_i$ for $w_l \in D_i$, when w_l is the number of adjectives in D_i and W_i is the total number of words in D_i.

Our suggestion is that if there is a remarkable difference between female and male texts, and women use more adjectives than men, some of them will be "empty". The syntactic complexity, though, can be observed explicitly by counting the number of verbs per period in both female and male cases. The more verbs in a period the more complex syntactically this period is, containing more sentences and/or secondary clauses. It is considered that $P_j = \sum_{j=1}^{J} p_j$ is the number of periods in D_i, and $SC_j = \frac{VB_j}{W_{i_j}}$ measures the syntactic complexity for each period p_j, where $VB_j = \sum_{m=1}^{J} w_m^{i_j}$ for $w_m \in p_j$ is the number of verbs and W_{i_j} is the total number of words in $p_j \in D_i$. The syntactic complexity measurement is not complete, since conjunction and other circumstantial clause

Table 2. The female and male corpus divided into classes and the average size of every class.

Female corpus				Male corpus			
Class	Number of sentences/post	Number of posts/ class	Average # of sentences/post	Class	Number of sentences/post	Number of posts/ class	Average # of sentences/post
A^F	1–9	536	5.57	A^M	1–8	611	4.67
B^F	10–20	469	14.55	B^M	9–24	634	14.75
C^F	21–34	177	26.18	C^M	25–36	107	29.47
D^F	>35	208	109.61	D^M	>37	194	115.9
E^F	1–454	1390	26.78	E^M	1–483	1546	24.48

elements should be observed. In this effort simple measurements are made, taking into account the potentials of computational tools.

The period length of the posts is also measured, which may be combined to other characteristics, in order to lead to more secure conclusions. It is $PL_j = \frac{W_{i_j}}{W_i}$ the period length of document's period p_j, with W_{i_j} the total number of words in $p_j \in D_i$ and W_i the total number of words in D_i.

Another direct measurement is the interrogative form of utterances, which is counted as the number of simple question marks and combinations of question marks with other symbols (???, ?!?, !??, ?!!). The interrogative forms of utterances of D_i is $UT_i = \frac{QM_i}{SP_i}$, where QM_i is the number of simple question marks and combinations of question marks with other symbols in D_i, and $SP_i = \sum_{n=1}^{N} c_n$ for the characters $c_n \in \{//.//, //, //, //!//, //?//\}$ in D_i is the number of special characters in D_i. This number shows not only structural questioning punctuation, but also the use of interrogations when interaction is wanted. After the sociolinguistic data, a larger number of question marks are to be expected in women's corpus.

Concerning the lexical density, as described in Section 2, previous studies observe no difference at this level between women and men. In the present work, the lexical density is measured in both female and male corpus. It is considered that $LD_i = \frac{CN_i}{W_i}$ is the lexical density for D_i, where W_i is the total number of words in D_i and $CN_i = \sum_{q=1}^{Q} w_q$ for $w_q \in D_i$ is the number of words tagged as "adjective", "adverb", "noun" or "main verb"(content words).

Early studies claim that women have a smaller vocabulary than men. A direct way to confirm this claim is to measure the total number of different words in female and male corpus, without counting the stop words. This consists of the "pure" vocabulary of women and men's discourse or, in other words, the vocabulary richness, given that the first appearance of each word is measured. $V_i = \frac{DW_i}{W_i}$ is the metric for the pure vocabulary richness where $DW_i = \sum_{r=1}^{R} w_r$ for $w_r \in D_i$ is the number of distinctive words (excluding stop words) in D_i and W_i is the total number of words in D_i.

The "tag question" characteristic is also possible to be measured, by tracking all tag questions in female and male blog posts from an exhaustive list which was created. $T_i = \frac{TQ_i}{W_i}$ is the metric for "tag question" where TQ_i is the number of tag questions in D_i and W_i the total number of words in D_i. The non-standard types, which are more frequent in male discourse, are also measured. In order to capture this characteristic, it is assumed that the natural language processing tools used contain either a corpus-based lexicon or an electronic dictionary. Thus, the corpus types that are not recognized by these tools (types that are not part of the lexicon/dictionary) are perceived as out of the typical language types. It is considered that $NST_i = \frac{OVW_i}{W_i}$ is the metric for the "non-standard types" characteristic where OVW_i is the number of all words not recognized by the dictionary used.

In the effort to detect if women express themselves in a more sentimental way, SentiWordNet (Esuli & Sebastiani, 2006) provides the potential to discover it implicitly. The sentimentally polarized words (positive and negative meaning) of SentiWordNet can be measured into both female and male corpora, and evaluate the findings. It is calculated that $SW_i = \frac{SEN_i}{W_i}$ is the metric for the use of sentimental language for D_i where $SEN_i = \sum_{u=1} w_u$ is the number of words $w_u \in D_i$ found in SentiWordNet and W_i is the total number of words in D_i.

It is observed that the female corpus contains more polarized words, as expected after theoretical evidence, while male discourse is closer to "neutral". Finally, lists of slang types and "bad words" from the internet are used and their appearances in the female and male corpus are counted, in order to detect if the men's linguistic choice of "use of coarse language", can be confirmed. $BD_i = \frac{B_i}{W_i}$ is the metric for use of bad words and $SG_i = \frac{L}{W_i}$ the metric for use of coarse language for D_i accordingly, where $B_i = \sum_{x=1}^{X} w_x$ for $w_x \in D_i$ is the number of "bad words" and $L_i = \sum_{y=1}^{Y} w_y$ for $w_y \in D_i$ is the number of slang words elicited from our lists, and W_i is the total number of words in D_i.

Table 3 shows a summarized list of the qualitative characteristics and their quantitative counterparts.

Table 3. The linguistic markers and the corresponding quantitative features.

	Linguistic markers	Quantitative features
Related to women's language	Use of 'empty' adjectives	Number of adjectives per document/ sum of document words
	Syntactic complexity	Number of verbs per period/ sum of document words
	Interrogative forms	Number of question marks per document/ sum of document punctuation
	Tag questions	Number of tag-question-phrases per document/ sum of document words
	Use of sentimental language	Number of sentimentally polarised words per document/sum of document words
	Politeness and agreement/disagreement strategies	Number of polite, agreement, disagreement phrases per document/ sum of words
Related to men's language	Vocabulary richness	Number of different words (without the stop-words) per document/sum of document words
	Use of non-standard types	Number of unrecognized words
	Use of bad words	Number of bad words per document/sum of document words
	Coarse language and slang types	Number of slang types per document/sum of document words
Neutral	Lexical density	Number of content words per document/sum of document words
Not related to language and gender	Period length	Number of words per period/ sum of document words

5. Evaluation of Sociolinguistic Characteristics

In Section 4 we have reported ways of measuring the sociolinguistic indices in order to perform statistical analysis in female and male corpora and to evaluate their gender distinction ability. Our effort is to verify the statistical hypothesis according to which the two corpora (female and male) are different in terms of the variables proposed. The hypothesis testing is an important tool, in order to verify the proposed theory, and it offers useful conclusions after the samples' information. The mean value and the standard deviation (STD) per feature are calculated so as to quantitatively investigate their dependence to gender. In order to examine whether the two sets of data (one for men and one for women) are significantly different from each other, the *t*-statistic test (Welch, 1947) is performed. The independent-samples *t*-test compares the means between two unrelated groups of the same continuous dependent variable. The null hypothesis (H_0) in that case suggests that data from the men's, $\{X_M\}$, and women's, $\{X_W\}$, datasets are independent random samples from normal distributions with equal means, and equal but unknown variances, which means that there is no difference among the two samples' means. On the other hand, the alternative (H_1) suggests that the means are not equal. For the estimation of the value of the statistical indicator and the degrees of freedom that determine the critical areas on the table the mathematical relation below is used:

$$t = \frac{\overline{X_M} - \overline{X_W}}{\sqrt{\frac{s_M^2}{n_M} + \frac{s_W^2}{n_W}}} \tag{1}$$

where $\overline{X_M}$ and $\overline{X_W}$ are the sample means for men and women respectively, s_M and s_W are the STDs and n_M and n_W are the sample sizes of data for men and women. The critical value of the *t*-test is 1.96 and it determines whether to reject the null hypothesis and if the absolute value of the test is greater than the critical value (>1.96) statistical significance can be declared. The *t*-value is estimated with the commonly used $a = 5\%$ significance level (i.e. 95% confidence interval).The corresponding *p*-value, i.e. the probability, under the null hypothesis, of observing a value as extreme or more extreme of the *t*-statistic test was also estimated. For the cases where $p < 0.05$ the null hypothesis is rejected and, thus, the corresponding sociolinguistic features are statistically different between men and women.

In Tables 4, 5, 6, 7 and 8, the results of the mean, STD measurements and the *t*-statistic test are presented. Twelve characteristics are calculated.

In Table 4, the sum total results for the female and male corpora are presented. As observed, most of these features appear to be statistically significant. The "syntactic complexity" characteristic not only confirms the linguistic theory claiming that women tend to use more syntactically complex forms, but it appears to be one of the most informative features. The same tendency occurs

Table 4. Statistics for Classes E^F and E^M.

	Class E^F		Class E^M		Statistical test	
Feature list	Mean	STD	Mean	STD	*t*-statistic	p-value
Tag questions	0.01277	0.0104	0.009906	0.0096	7.752773	**0.00000**
Syntactic complexity	0.157124	0.0362	0.147631	0.0356	7.145601	**0.00000**
Adjectives	0.058131	0.0223	0.062444	0.0235	−5.09591	**0.00000**
Vocabulary richness	0.326558	0.0594	0.340288	0.0627	−6.08662	**0.00000**
Period length	24.35627	22.97	28.4611	25.46	−4.225	**0.00002**
Lexical density	0.553679	0.05	0.559528	0.05	−3.34042	**0.000847**
Sentimental language	0.139925	0.0339	0.13595	0.0350	3.123432	**0.001805**
Politeness strategies	0.001595	0.0037	0.001304	0.0031	2.311472	**0.020882**
Slang types	0.049287	0.0232	0.051228	0.0230	−2.27475	**0.022994**
Interrogative forms	0.293628	0.3871	0.319436	0.4191	−1.73426	0.082977
Non-standard types	0.223416	3.645	0.128795	0.0732	0.96693	0.333744
Bad words	0.000775	0.0029	0.000845	0.0027	−0.67202	0.501627

Table 5. Statistics for Classes A^F and A^M.

	Class A^F		Class A^M		Statistical test	
Feature list	Mean	STD	Mean	STD	*t*-statistic	p-value
Syntactic complexity	0.153261	0.0424	0.142193	0.0417	4.447623	**0.00000**
Vocabulary richness	0.350838	0.0629	0.365433	0.0644	−3.8733	**0.000114**
Adjectives	0.062369	0.0279	0.068144	0.0286	−3.4516	**0.000578**
Sentimental language	0.142601	0.0400	0.135429	0.0419	2.958278	**0.003158**
Period length	32.28037	33.38	37.98591	35.08	−2.73056	**0.006426**
Tag questions	0.011585	0.0116	0.01016	0.0123	2.017642	**0.043863**
Bad words	0.000559	0.0026	0.000862	0.0031	−1.79499	0.072919
Slang types	0.049645	0.0299	0.051511	0.0296	−1.05854	0.290035
Non-standard types	0.129858	0.0821	0.135082	0.0878	−1.03879	0.299121
Politeness strategies	0.001378	0.0045	0.00115	0.0037	0.932688	0.351197
Lexical density	0.554422	0.06	0.555679	0.06	−0.35889	0.719745
Interrogative forms	0.16154	0.3366	0.154416	0.3429	0.353932	0.723456

Table 6. Statistics for Classes B^F and B^M.

	Class B^F		Class B^M		Statistical test	
Feature list	Mean	STD	Mean	STD	*t*-statistic	p-value
Tag questions	0.013606	0.0110	0.00977	0.0078	6.458489	**0.00000**
Syntactic complexity	0.162341	0.0334	0.152676	0.0326	4.795546	**0.00000**
Vocabulary richness	0.32718	0.0480	0.338704	0.0484	−3.92526	**0.000009**
Period length	19.42507	6.04	21.0637	6.76	−3.88967	**0.000108**
Lexical density	0.553006	0.04	0.561075	0.04	−3.31532	**0.000949**
Interrogative forms	0.311568	0.3979	0.378016	0.4484	−2.595	**0.009589**
Politeness strategies	0.001663	0.0032	0.001245	0.0026	2.310157	**0.02111**
Adjectives	0.056966	0.0193	0.059649	0.0201	−2.23685	**0.02551**
Sentimental language	0.141895	0.0321	0.138655	0.0311	1.678391	0.093587
Slang types	0.048431	0.0175	0.049993	0.0183	−1.43559	0.151421
Non-standard types	0.119726	0.0567	0.118375	0.0542	0.397576	0.691028
Bad words	0.000831	0.0025	0.000798	0.0026	0.207884	0.83536

in the case of the "tag question" characteristic. The "adjectives" characteristic is informative enough, but it is related to men, unlike the linguistic marker of "use of empty adjectives" which is correlated to women's language. This

Table 7. Statistics for Classes C^F and C^M.

	Class C^F		Class C^M		Statistic test	
Feature list	Mean	STD	Mean	STD	*t*-statistic	p-value
Tag questions	0.014316	0.0082	0.009937	0.0073	4.672485	**0.000000**
Lexical density	0.55034	0.03	0.567346	0.04	−3.503	**0.000583**
Syntactic complexity	0.15917	0.0290	0.150423	0.0303	2.375782	**0.018398**
Vocabulary richness	0.31365	0.0419	0.327061	0.0507	−2.29671	**0.022732**
Period length	17.23316	3.94	18.58575	5.23	−1.9736	0.050825
Non-standard types	0.11758	0.041	0.13166	0.0717	−1.84212	0.067485
Sentimental language	0.139401	0.0263	0.135038	0.0286	1.273425	0.204302
Interrogative forms	0.462332	0.4199	0.523259	0.4017	−1.2122	0.226686
Slang types	0.049616	0.0181	0.051861	0.0170	−1.04842	0.29554
Adjectives	0.054898	0.0159	0.056829	0.0151	−1.02008	0.308769
Bad words	0.001052	0.0046	0.000742	0.0017	0.801856	0.42342
Politeness strategies	0.00172	0.0038	0.001659	0.0027	0.155715	0.876373

Table 8. Statistics for Classes D^F and D^M.

	Class D^F		Class D^M		Statistic test	
Feature list	Mean	STD	Mean	STD	*t*-statistic	p-value
Tag questions	0.012639	0.0061	0.009489	0.0052	5.545297	**0.00000**
Adjectives	0.05283	0.0135	0.056867	0.0144	−2.8917	**0.004046**
Period length	18.27488	6.96	20.44161	5.11	−2.89554	**0.004108**
Syntactic complexity	0.153268	0.0282	0.146454	0.0225	2.672102	**0.007855**
Slang types	0.050046	0.0180	0.053895	0.0137	−2.44756	**0.014838**
Interrogative forms	0.45323	0.3311	0.534308	0.3430	−2.39643	**0.017021**
Lexical density	0.556306	0.03	0.562165	0.03	−1.73758	0.083058
Non-standard types	0.133229	0.0681	0.141764	0.074	−1.1962	0.232327
Politeness strategies	0.001827	0.0020	0.001711	0.0021	0.569491	0.569349
Vocabulary richness	0.273332	0.0481	0.274231	0.0527	−0.17771	0.859044
Bad words	0.000977	0.0023	0.001012	0.0020	−0.16189	0.871477
Sentimental language	0.129128	0.0224	0.129213	0.0240	−0.03682	0.970648

characteristic's significance based on the statistical test emerges as a novel male differential feature.

The "vocabulary richness" marker is confirmed in measurements. This means that women have a smaller vocabulary than men, and the statistical difference between women and men makes the feature statistically significant. The "period length" characteristic, which was proposed to be measured in order to have statistical information about the length of the posts' sub-constituents, becomes an important and representative characteristic of the male linguistic choices. This characteristic does not appear in sociolinguistic bibliography; it is a new contribution of this study since it appears to be representative of a differentiated linguistic attitude. Men formulate longer phrases than women, but without using more verbs, because of using more complex syntactic forms and subordinate clauses. Combining the period length to the number of adjectives it is assumed that men tend to use more adjectives than women and, therefore, more subordinate clauses. The "lexical density" measurements appear to be informative enough and men use more content words than women. Unlike

previous work in gender and lexical density, which proved that lexical density is not statistically significant in differentiating gender (Alami, Sabbah, & Iranmanesh, 2013), in the gender-annotated corpus used in the experiments it appears to be a new finding about the language of men.

The finding regarding the "sentimental language" characteristic confirms the sociolinguistic studies which support evidence that women use more sentimental phrases than men and proves to be important in the measurements. In accordance with the theory seems to be also the politeness, agreement/disagreement phrases used by women, which differs from the male corresponding strategies. The last important characteristic, the "slang types" feature, is also related to the men's language and this measurement confirms linguistic studies which relate the use of coarse language and slang types to male attitudes. The "interrogative forms" and "non-standard types" features are not distinctive and do not confirm the theory, and finally, although the "bad words" characteristic confirms the theoretical studies, it is not significant enough, at least on the data-set we examined.

In Table 5, the measurements for the Class A, female and male, are presented. Based on the *t*-statistic results, there is no difference between the overall measurements (Class E) and the smaller posts of Class A, except for the "interrogative forms" characteristic which is positive over women in this category of the corpus, without being informative. Thus, "syntactic complexity", "vocabulary richness", "adjectives", "sentimental language", "period length" and "tag questions" features are to be considered as the differentiating characteristics between women and men's language in texts of a small size.

In Table 6, the results concerning the female and male posts of the Class B are presented. In this category a difference in the "interrogative forms" feature is observed, which is statistically significant in Class B and a distinctive feature of men's language as discussed in Class E. Moreover, the "syntactic complexity", "tag questions", "vocabulary richness", "period length", "lexical density", "politeness strategies" and "adjectives" features are the most informative characteristics of linguistic differentiation between men and women.

In Table 7 the results of the statistical analysis concerning the Class C for both female and male posts are presented. A first observation is that this is the smallest list of distinctive characteristics among all categories (all different size posts). Besides "tag questions", "lexical density", "syntactic complexity" and "vocabulary richness" features, all other features are not significant enough to be distinctive. This could be explained as follows: as discussed in previous sections, women and men tend to make different linguistic choices even in the same communicative situation. These gender preferential choices can be detected primarily in speech, and most of them are identified at the phonetic level of linguistic analysis. The gender differential characteristics run across all linguistic levels (morphology, lexicon, etc.) and they can be identified in written language too, as long as written remains informal and rather spontaneous. It is

observed consequently that these features in texts carry characteristics of orality, consisting of a sample of spontaneous language, not prepared or processed. In smaller texts the language remains unprocessed with clues of orality, whilst for writing text of a bigger size it is inevitable for most people to reflect on the forms used and correct according to vocabulary, grammar and syntax rules, preferring the standard linguistic structures – regardless of gender (Kakridi-Ferrari, 2005).

In Table 8, the results of the Class D are presented. The most informative features are the "tag questions", the "adjectives", the "period length", the "syntactic complexity", the "slang types", and the "interrogative forms". The characteristics move to the same direction with the other categories' results, and the only additional observation could be the alteration of the "sentimental language" from women preferential characteristic to a men's choice, without being informative enough.

As mentioned above, the Classes C and D highlight some different aspects of the characteristics than the other two categories (Classes A and B), as a result of the men and women writing style related to the remarkable difference of the texts' average size. A useful remark could be that the posts should not be processed and classified without taking into account the document's length. It is observed that texts of different size, even when they belong to the same thematic category, or have the same author, they do not necessarily share all the same significant characteristics. The size turns out to be a clue about the differentiated characteristics under study. After the statistical analysis, we observe that both men and women use longer sentences in texts of a small/medium size. To be more specific, the smaller a text it is, the longer periods it has.

Table 9 lists the most representative characteristics found in more than one categories of our corpus. The results enable us to speak about common differentiated characteristics observed in all categories (Classes A, B, C, D, E), and about common distinctive features found in more than one category of the corpus. They are also classified according to the gender that tent to use them.

6. Discussion

The aim of this study was to identify differential linguistic markers between women and men. The present study relied on the sciences of sociolinguistics, statistics and text mining. It was an interdisciplinary research effort in order to extract the different linguistic choices that people of different gender make. A bibliographic research was made in the field of sociolinguistics, and more specifically in studies dealing with the language and gender relationship. Then, all the characteristics that linguists have identified after their theoretical and empirical studies were collected and a long enough list was created, with all linguistic markers that are supposed to distinguish women's from men's language.

These characteristics were descriptive enough to allow their measuring in a large corpus. For this reason, direct and indirect ways were used to convert

Table 9. The most important differentiated quantitative characteristics in terms of gender.

	Female features	Male features
Universal (all categories)	Syntactic complexity Tag questions	
In 4/5 categories		Period length Adjectives Vocabulary richness
In 3/5 categories		Lexical density
In 2/5 categories	Politeness strategies Sentimental language	Slang types Interrogative forms

them into quantitative values, which can be measured by text mining tools. The corpus used was an already gender-annotated corpus, and it was separated into male and female corpora in order to allow the examination of two different text samples. A statistical analysis of the feature values was performed and the *t*-statistic test highlighted the most distinctive linguistic features between women and men. Nine over twelve of the measured characteristics proved to be statistically significant in more than two corpus categories, and two of them, the "tag questions" and the "syntactic complexity" are distinctive and female-preferential features in the totality of the corpus and its four subsets, the A, B, C and D.

These results in most cases confirm the theoretical differential markers of "syntactic complexity", "tag questions", "sentimental language" and "politeness strategies". These characteristics are measured and, as expected after the sociolinguistic studies, their female-preferential nature has been confirmed in this study. On the other hand, female markers as the use of "interrogative forms" proved to be a more male than a female linguistic choice. The "adjectives" marker is not confirmed as a female characteristic, and after the measurement, the increased use of adjectives proved to be a male choice. The theoretical and empirical characteristics related to men's language are all confirmed in our study, and the "vocabulary richness", the "slang types" features, the – previously considered as neutral – characteristic of "lexical density", turn to be important male features. Finally, it is observed that the increased length of the period is also a choice that men prefer to make, since they formulate longer phrases than women.

An important parameter concerning the results of this study which should be taken into account is the text genre of the corpus used. As mentioned previously, the theoretical and empirical studies concern spoken language and in most cases the research database is recorded speech. This language type carries all the characteristics of oral linguistic choices of the speaker and the differential characteristics observed are influenced by the orality. The corpus used in this study is not formed according to the norms of formal language use, but it consists of samples of informal-like written language. Although there has been an effort to detect the sociolinguistic features in fear of having not confirmed

the theory, the results were quite encouraging. Another important result, to be further investigated, concerns a tendency of reduction of the genderized characteristics in the Classes B and C (containing blog posts of a medium and longer size), while in the case of the Classes A and D (containing smaller and very long posts), the differences remain rather accentuated. More specifically, in Classes A and D, 6 out of 12 differential features appear, while in B and C appear 8 and 4, accordingly, out of 12 differential characteristics. The results of our study could be seen as a contribution to the field of gender identification, and in a further step these sociolinguistic features could be used to perform gender classification experiments.

As a general conclusion, it is observed that even written language differs between women and men. Women tend to use more complicated syntactic forms, but men are more analytical and they use longer phrases and more adjectives. Men also use more content words than women and their vocabulary is "richer" than the women's vocabulary. In women's text, on the other hand, tag questions, sentimental and polite phrases tend to be dominant, without enriching their vocabulary though. In a future study, these characteristics could be investigated in a greater depth and the markers, which demonstrate the different choices that men and women make in discourse, could be further enriched.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Alami, M., Sabbah, M., & Iranmanesh, M. (2013). Male-female discourse difference in terms of lexical density. *Research Journal of Applied Sciences, Engineering and Technology*, *5*(23), 5365–5369.

Archakis, A., & Kondyli, M. (2004). *Introduction to Sociolinguistic Issues*. Athens: Nisos. (in Greek).

Bagavandas, M., & Manimannan, G. (2008). Style consistency and authorship attribution: A statistical investigation. *Journal of Quantitative Linguistics*, *15*(1), 100–110.

Bucholtz, M. (1998). Geek the girl: Language, femininity, and female nerds. In: N. Warner, J. Ahler, L. Bilmes, M. Oliver, S. Wertheim & M. Chen (Eds), *Gender and Belief Systems: Proceedings of The Fourth Berkeley Women and Language Conference* (pp. 119–131). Berkeley: Berkeley Women and Language Group.

Bucholtz, M. (1999). You da man: Narrating the racial other in the production of white masculinity. *Journal of Sociolinguistics*, *3*(4), 443–460.

Bucholtz, M. (2002). From 'sex differences' to gender variation in sociolinguistics. *University of Pennsylvania Working Papers in Linguistics*, 8(3), 33–45.

Bucholtz, M. (2003). Theories of discourse as theories of gender: Discourse analysis in language and gender studies. In: J. Holmes & M. Meyerhoff (Eds), *The Handbook of Language and Gender* (pp. 43–68). Oxford: Blackwell.

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, *7*(4–5), 585–614.

Bucholtz, M., Liang, A. C., & Sutton, L. A. (1999). *Reinventing Identities: The Gendered Self in Discourse* (p. 273). Nova Iorque: Oxford University Press.
Cameron, D. (1998). Gender, language, and discourse: A review essay. *Journal of Women, Culture and Society*, *23*(4), 945–960.
Cameron, D. (2005). Language, gender, and sexuality: Current issues and new directions. *Applied Linguistics*, *26*(4), 482–502.
Chen, M., Jin, X., & Shen, D. (2011). Short text classification improved by learning multi-granularity topics. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)* (pp. 1776–1781). Citeseer.
Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, *41*, 87–100.
Eckert, P., & McConnell-Ginet, S. (1999). New generalizations and explanations in language and gender research. *Language in society*, *28*(02), 185–201.
Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (Vol. 6, pp. 417–422).
Fishman, P. M. (1983). *Interaction: The work women do*. In B. Thorne, K. Cheris & N. Henley (Eds), *Language, Gender, and Society* (pp. 89–102). Rowley, MA: Newbury House.
Gordon, E. (1997). Sex, speech, and stereotypes: Why women use prestige speech forms more than men. *Language in Society*, *26*(01), 47–63.
Hadley, E. (1995). *Melodramatic Tactics: Theatricalized Dissent in the English Marketplace, 1800–1885*. Stanford University Press.
Irvine, J. (2001). Style as distinctiveness: The culture and ideology of linguistic differentiation. In P. Eckert & J. Rickford (Eds), *Stylistic Variation in Language* (pp. 21–43). Cambridge: Cambridge University Press.
Jespersen, O. (1922). *Language. It's Nature, Development and Origin*. London: Allen & Unwin.
Kakridi-Ferrari, M. (2005). Language and social environment: Sociolinguistic issues (Part A). *Parousia*, Annex No 64, Athens (in Greek).
Kobayashi, D., Matsumura, N., & Ishizuka, M. (2007). Automatic estimation of bloggers' gender. In *Proceedings of International Conference on Weblogs and Social Media*. Boulder: Omnipress.
Kubát, M., & Milička, J. (2013). Vocabulary richness measure in genres. *Journal of Quantitative Linguistics*, *20*(4), 339–349.
Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia, PA: University of Pennsylvania Press.
Lakoff, R. (1973). Language and woman's place. *Language in Society*, *2*(01), 45–79.
Lakoff, R. (1975). *Language and Women's Place*. New York, NY: Harper and Row.
Lakoff, R. (1990). *Talking Power: The Politics of Language in Our Lives*. New York, NY: Basic Books.
Makri-Tsilipakou, M. (2010). 'Women's language' and the language of women. In V. Kantsas, V. Moutafis & E. Papataxiarchis (Eds), *Gender and Social Sciences in Modern Greece* (pp. 119–146). Athens: Alexandria Editions. (in Greek).
Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of Linguistics*, *21*(02), 339–384.
Moore, E., & Podesva, R. J. (2009). Style, indexicality and the social meaning of tag questions. *Language in Society*, *38*, 447–485.
Mukherjee, A., & Liu, B. (2010). Improving Gender Classification of Blog Authors. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP'10).

Przybyła, P., & Teisseyre, P. (2014). Analysing utterances in polish parliament to predict speaker's background. *Journal of Quantitative Linguistics*, *21*(4), 350–376.
Savoy, J. (2012). Authorship attribution: A comparative study of three text corpora and three languages. *Journal of Quantitative Linguistics*, *19*(2), 132–161.
Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, *26*(4), 471–495.
Sun, A. (2012). Short text classification using very few words. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 1145–1146). ACM.
Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society*, *1*(2), 179–195.
Trudgill, P. (2000). *Sociolinguistics: An Introduction to Language and Society*. London: Penguin.
Vo, D. T., & Ock, C. Y. (2015). Learning to classify short text from scientific documents using topic models with various types of knowledge. *Expert Systems with Applications*, *42*(3), 1684–1698.
Welch, B. L. (1947). The generalization of student's' problem when several different population variances are involved. *Biometrika*, *34*, 28–35.