Pertanika J. Soc. Sci. & Hum. 26 (1): 103 - 112 (2018)

PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

Journal homepage: http://www.pertanika.upm.edu.my/

Identification of Gender of the Author of a Written Text using Topic-Independent Features

Tatiana Litvinova[1,2]*, Pavel Seredin[1], Olga Litvinova[1] and Olga Zagorovskaya[2]

[1]*Gosudarstvennyi nauchnyi tsentr 'Kurchatovskii institut', Moscow, Russia*
[2]*Voronezskij gosudarstvennyj pedagogiceskij universitet, Voronezh, Russia*

ABSTRACT

Authorship profiling, which is the process of extraction of information about a text's author through linguistics analysis, is now gaining momentum as an interdisciplinary subject. Scholars who employ this technique (i.e. data analysis specialists, linguists, psychologists) study the identification of demographics, personality traits, education and the native language of authors of texts, among others. Gender, in this context, is the most popular variable. Some studies report accuracy as high as 80% or even higher in identifying the gender of a text's author. However, there are still many issues that must be addressed. Firstly, most of the previous research concerns English texts. Secondly, most of the papers focus on content-based features, which are obviously easily to imitate. Thirdly, many recent papers in the field make use of machine-learning algorithms with emphasis on accuracy, not on the differences between male and female writing. The objective of this paper is to reveal differences in male and female Russian written texts and to design a mathematical model to identify the gender of authors of texts using only high-frequency topic-independent text parameters. Special emphasis is made on comparing the obtained data on the differences in male and female written texts with those previously obtained for Russian and other languages. An original mathematical solution for identification of author's gender is set forth.

Keywords: Authorship profiling, corpus, corpus linguistics, gender attribution, gender identification, Russian language, stylometry

ARTICLE INFO

Article history:
Received: 06 August 2016
Accepted: 05 December 2017

E-mail addresses:
centr_rus_yaz@mail.ru (Tatiana Litvinova),
paul@phys.vsu.ru (Pavel Seredin),
olga_litvinova_teacher@mail.ru (Olga Litvinova),
olzagor@yandex.ru (Olga Zagorovskaya)
*Corresponding author

ISSN: 0128-7702

INTRODUCTION

For decades, scientists have studied the differences in writing done by males and females. These studies indicated a number of differences in the style of writing used by males and females and highlighted the possibility of identifying gender using written texts. However, these studies also argued that all of the differences were not inventory but rather probabilistic, as they manifested themselves in certain features of language use, both qualitatively and quantitatively. In order to identify the gender of an author using his/her text, special methods of analysis are necessary. Mulac and Lundell (1994) revealed that gender may be identified with 50% accuracy i.e. at the level of a random value. Studies concerning the development of methods to identify the gender of a text's author do not only have a practical importance, for instance, in marketing and forensics; indeed, they also have a theoretical significance as they allow one to identify the cognitive activity of males and females as manifested in their language use. Indeed, this gives a wider insight into human cognitive ability. The analysis of context-independent text parameters that are easy to extract by means of methods of natural language processing is vital in developing practically applicable methods of identifying the gender of a text's author.

Of course, sociolinguists have acquired a lot of information about the differences in male and female speech, but as Nini (2014) pointed out, "Little work has been done on relative frequencies of linguistic features. These forms have not been studied traditionally and other disciplines like computational linguistics and corpus linguistics are only now exploring their correlations with social dimensions" (p. 26). In some studies (analysing mostly English texts) it was found that females presented a higher frequency of the use of pronouns and negations, whereas males presented a higher frequency of determiners and prepositions. This was consistent with the proposal of Biber et al. (1998), that males are more informational, whereas females are more involved. Words longer than six letters and articles were found to be among other favourite male features (see Nini (2014) for a thorough review).

Authorship profiling, which is the process of extraction of information about text authors through linguistics analysis, is now gaining momentum as an interdisciplinary subject. Scholars who employ this technique, data-mining specialists and computer linguists, for instance, are dealing with the identification of demographics, personality traits, education and the native language of authors of texts, with gender being the most popular variable to identify (Koppel, Argamon, & Shimoni, 2002; Newman, Groom, Handelman, & Pennebaker, 2008; Argamon, Koppel, Pennebaker, & Schler, 2009). However, there are still many issues that must be addressed (Soler & Wanner, 2014). Most of the previous research studied texts written in English, although recently, some studies have looked at texts written in other languages (Rangel et al., 2015; Litvinova

et al., 2016). Scientists are still divided on what mathematical methods should be used for this purpose. The main issue is selecting the text parameters to analyse. Content-based features are considered the most effective, although it is obvious that they are consciously controlled and therefore, can be easily imitated. Studies employing style-based parameters such as lexical, syntactic and character use, for instance, do not normally provide explanation of the correlations between the parameters of the texts and the gender of their authors.

We argue that it is of particular importance to investigate differences at the level of frequently used context-independent text parameters and then to employ the parameters correlating with gender to design prognostic models. It is obvious that a list of such parameters should be expanded and more languages should be employed in identifying universal and language-specific differences in male and female speech.

The current study was performed using material from a specially designed corpus of texts written in Russian. Russian sociolinguists have carried out a lot of research addressing differences in male and female speech as well as gender imitation (see Oschepkova (2003) for detailed review). It was found that for respondents of different social groups (prisoners and university students), the following was typical even for gender imitation: males tended to make more mistakes; females made more use of negations; lexical diversity was higher in male texts, and; men used fewer clichés. However, the authors of these papers made no attempt to identify the gender of text authors.

The Russian language has long been neglected in authorship profiling studies, but lately there have been relevant studies including those dealing with gender identification of text authors (Litvinova, 2014; Litvinova, Seredin, & Litvinova, 2015; Litvinova et al., 2016; Sboev et al., 2016). Note that the main focus has been on the accuracy of the resulting models rather than on differences between male and female writing. In this paper, we made it our objective to identify significant differences in qualitative parameters of Russian written texts by males and females to further design a prognostic model.

MATERIALS AND METHOD

This study utilised a specially designed and constantly growing corpus of Russian written texts, *RusPersonality* (Litvinova et al., 2016), which contained, aside from the texts themselves, rich metadata i.e. information about authors (gender, age, education, psychological testing data etc.). All the texts of the corpus were written by respondents according to the researchers' instructions. For this study, we selected two subcorpora from the corpus: (1) A total of 150 texts by 75 respondents (each respondent was instructed to write two texts, "Describe a Picture" and "What would I Spend a Million Dollars On?"); (2) A total of 1,354 texts by 677 respondents (description of a picture and a letter to a friend). All of the texts contained an average of 130-160 words.

In order to exclude a maximum of other characteristics that might affect the text parameters, we selected a fairly homogeneous group of respondents i.e. of students of large Russian universities aged 19 to 22. Since each respondent was instructed to write two texts on different topics, we used two analysis scenarios: In the first, we viewed each text individually and in the second, both texts by the same author were merged into one.

All the texts were marked using Python script based on a morphological analyser, *pymorphy2*, and processed using an online service, *istio.com*. The text parameters were only those that were not consciously controlled; finite forms of verbs and other clear indicators of an author's gender were not considered. The parameters were indicators of lexical diversity of a text and proportions of parts of speech and their correlations (a total of 78 parameters).

In order to determine the characteristics and type of connection between the text parameters and gender of the author, a correlation analysis was performed using the Pearson correlation coefficient ($p<0.05$). Calculations were done using the IBM SPSS statistics software. We established a number of correlations between the text parameters and the author's gender (0 – woman, 1 – man). A large number of the parameters of the texts and the gender of their authors correlated with r=0.25-0.39. Further, we selected only the parameters that correlated with the text author's gender in both subcorpora and in both scenarios ('merged' and 'individual').

Indeed, this allowed us to design a regression model considering the most significant correlations based on multiparameter linear approximation. However, testing of the quality of the models showed that this type of approximation yields a low level of accuracy as the parameters of texts by male and female are usually in overlapping ranges. Therefore, it was decided to use not a multiparameter regression model as we did in previous studies (Litvinova, 2014; Litvinova, Seredin, & Litvinova, 2015), but to design a few regression models instead.

RESULTS

Let us show the suggested approach using an example of five texts with the parameters correlated with the gender of an author with the highest r:

1. TTR (type-token ratio). This is the most commonly used index of lexical diversity of a text. Given a text t, let Nt be the number of tokens in t and Vt be the number of types in t, then the simplest measure for the TTR of the text t is:

$$TTRt = Vt / Nt. \quad [1]$$

Note that the measure in Eq. (1) is a number defined in [0, 1], since for any text results $1 \leq Vt \leq Nt$. Some interesting attempts to improve the TTR index have been proposed in the literature, although only a few of these variants possess key properties that are essential if they are to be used in our text comparison, and

these properties are harder to calculate (see Caruso et al. (2014) for details).

Since the texts in the corpus were of different length, we calculated TTR in the first 100 words of each text. Indeed, TTR-value is known to depend on the length of the analysed text and therefore, the comparison of values makes sense for the same number of tokens (Caruso et al., 2014, p. 139).

The index was calculated using *istio.com*. The correlation coefficient r=0.39. The resulting regression equation took the following form:

$$GENDER_1 = -0.669 + (2.622\ TTR). \quad [2]$$

2. Formality of a text that was calculated using the following formula (Nini, 2014):

$$F = (noun + adjective + preposition - pronoun - \\ - verbs - participles - adverbs - \\ - conjunction - interjections) + 100) / 2. \quad [3]$$

The correlation coefficient r=0.315.

The regression equation was as follows:

$$GENDER_2 = -0.637 + (0.971\ Formality). \quad [4]$$

3. Proportion of prepositions and pronoun-like adjectives in a text (r=0.243):

$$GENDER_3 = -0.188 + \\ + (0.0432\ preposition + pronoun - like\ adjective) \quad [5]$$

4. Proportion of the 100 most frequently used Russian words in a text (Lyashevskaya & Sharov, 2009), r=-0.322.

The regression equation was as follows:

$$GENDER_4 = 1.500 - (0.0303 Frequent\ ones). \quad [6]$$

5. The index of the functional density based on the ratio of function words to content words (r=-0.295).

$$GENDER_5 = 1.392 - (0.0229 Function). \quad [7]$$

In order to properly estimate the obtained result, let us determine the average arithmetic values from the solutions obtained in the five equations:

$$GENDER = \frac{\sum_{i=1}^{5} GENDER_i}{5}. \quad [8]$$

Let us assume that a design value in the range [0; 0.499] indicates that the author of a text is female and in the range [0.500; 1]

shows that the author is male. In order to estimate the suggested approach, we used a corpus of texts with contributions from 553 individuals (368 women and 185 men, while two texts from each respondent were considered as one text). Their topic and length were identical to those used to design the regression models.

Let us determine the accuracy of the approach. Accuracy, in this context, was the ratio of the number of test documents that were correctly predicted to the total number of test documents. The calculations suggested that gender was correctly identified in 65% of the texts written by females and 63% of the texts written by males. Thus, the accuracy of the approach was 64%.

DISCUSSION

The analysis showed that in texts written in Russian by men compared to those written by women, the index of lexical diversity and the proportion of prepositions and pronoun-like adjectives were higher; in addition, the proportion of 100 most frequently used Russian words as well as the index of functional density was lower. Texts written by males were found to be more formal than texts written by females.

Overall, the data were in good agreement with the results obtained for texts written in English. Hence, as noted above, many scientists have argued that texts by men have on average more nouns and adjectives as well as prepositions and demonstrative and relative pronouns; in contrast, those by women have more verbs and personal pronouns (see a detailed review in Nini, 2014). According to the literature, this is indicative of profound cognitive differences in the linguistic profiles of men and women: reporting is more important for men, while rapport is more significant for women. Therefore, texts by men seem more 'formal', while those by women seem more 'contextual' (see Heylighen & Dewaele, 2002 for more details). It is interesting to compare this with the paper by Saily, Siirtola and Nevalainen (2011), which shows that the prevalence of nouns in texts by men as opposed to pronouns in those by women was common in personal letters written in English from 1415 to 1681. Indeed, this shows that the above gender differences are universal.

Nini (2014) has shown that "the more personal a text becomes, the less likely it is to show a gender pattern of the rapport/report type. In other words, in a register in which individuals are already pressed to be involved and person-centred, there is no room for variation between rapport and report discourse, thus blocking the gender pattern from emerging" (p. 132). However, our analysis has shown that this effect is retained in personal texts such as letters to a friend.

We argue that a higher index of lexical diversity in texts by men is due to the above differences: In texts by males, there are fewer most frequently used words, the majority of which are function words; in addition, there are fewer repetitions and more unique vocabulary units occur in a text at one time. Mikros (2013), who analysed

Greek texts, found that texts written by men presented less lexical repetition and avoidance of standardised lexical patterns and a higher percentage of *hapax legomena*. Mikros also stated that woman used more function words than men.

These data are in good agreement with the results obtained for texts written in Russian (Oschepkova (2003), see above). It is interesting that the level of lexical diversity and the number of clichés were one of the few distinguishing parameters that were preserved in texts by females and males of different social groups and even in gender imitation.

CONCLUSION AND FUTURE WORK

The present study identified the differences between texts written in Russian by males and females using a range of context-independent parameters by means of a text corpus that was controlled simultaneously for the author's gender, age, education, text topic, genre and medium. The obtained results were in good agreement with those from previous studies on Russian and other languages. The use of only five linguistic parameters as part of the suggested approach showed that it is possible to identify the gender of text authors with accuracy above the random value.

There are plans to use the material of our newly designed *Russian Gender Imitation Corpus* to check whether the differences we have identified would be retained in a gender-imitation scenario as well as to carry on searching for more differences in texts written by male and female authors that would remain even in a gender-imitation scenario.

In addition, rich metadata of the corpus would allow us to investigate the effect of biological and social gender as independent variables on text parameters (Chambers, 1992) as well as to evaluate the joint impact of these factors and a range of personality traits, functional cerebral asymmetry profile etc. on linguistic parameters. As correctly pointed out by Nini, it can be assumed that "the real differences in the linguistic patterns adopted by people depend on their personality and/or hormone levels and that genders are different to the extent that on average different genders are prone to different personality orientations and/or hormone levels" (2014, p. 34).

We also seek to employ language-independent text parameters for gender identification of text authors using the material of our corpus and freely available text corpora in other languages to identify universal differences in texts written by males and females.

This analysis to be conducted during further research would allow one to develop a more current and deeper insight into the way gender is manifested in written texts and to develop more accurate methods of identifying the gender of individuals based on the quantitative parameters of their texts.

ACKNOWLEDGEMENT

This research was financially supported by the Russian Science Foundation, project

No 16-18-10050, "Identifying the Gender and Age of Online Chatters Using Formal Parameters of their Texts".

REFERENCES

Argamon, S., Koppel, M., Pennebaker, J., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM, 52*(2), 119–123.

Caruso, A., Folino, A., Parisi, F., & Trunfio, R. (2014). *A statistical method for minimum corpus size determination*. Paper presented at JADT 2014, Paris, France.

Chambers, J. K. (1992). Linguistic correlates of gender and sex. *English World-Wide, 13*(2), 173–218.

Heylighen, F., & Dewaele, J. (2002). Variation in the contextuality of language: An empirical measure. *Foundations of Science, 7*(3), 293–340.

Koppel, M., Argamon, S., & Shimoni, A. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing, 17*(4), 401–412.

Litvinova, T. A. (2014). Profiling the author of a written text in Russian. *Journal of Language and Literature, 5*(4), 210–216.

Litvinova, T. A., Seredin, P. V., & Litvinova, O. A. (2015). Using part-of-speech sequences frequencies in a text to predict author personality: A corpus study. *Indian Journal of Science and Technology, 8*(9), 93–97.

Litvinova, T., Litvinlova, O., Zagorovskaya, O., Seredin, P., Sboev, A., & Romanchenko, O. (2016). "Ruspersonality": A Russian corpus for authorship profiling and deception detection. In *Proceedings of International FRUCT Conference on Intelligence, Social Media and Web* (ISMW FRUCT) (pp. 1-7). St. Petersburg.

Lyashevskaya, O., & Sharov, S. (2009). *Frequency dictionary of modern Russian language (on materials of the Russian national corpus)*. Moscow, Russia: Azbukovnik.

Mikros, G. K. (2013). Systematic stylometric differences in men and women authors: A corpus-based study. In R. Köhler & G. Altmann (Eds.), *Issues in Quantitative Linguistics* (pp. 206–223). Lüdenscheid: RAM-Verlag.

Mulac, A., & Lundell, T. L. (1994). Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects. *Language and Communication, 14*(3), 299–309.

Newman, L. M., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14.000 text samples. *Discourse Processes, 45*(3), 211–236.

Nini, A. (2014). *Authorship profiling in a forensic context*. (PhD dissertation). Aston University. Retrieved from http://eprints.aston.ac.uk/25337/1/Nini_Andrea_2015.pdf

Oschepkova, E. S. (2003). *Written text author identification: Lexicogrammatical aspect*. (PhD thesis). Moscow State Linguistic Uni. (in Russian).

Rangel, F., Celli, F., Rosso, P., Pottast, M., Stein, B., & Daelemans, W. (2015, September 8-11). *Overview of the 3rd author profiling task at PAN 2015*. Paper presented at CLEF 2015, Toulouse, France.

Saily, T., Siirtola, H., & Nevalainen, T. (2011). Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing, 26*(2), 167–188.

Sboev, A., Litvinova, T., Gudovskikh, D., Rybka, R., & Moloshnikov, I. (2016). Machine learning models of text categorization by author gender using topic-independent features. *Procedia Computer Science, 101*, 135–142. https://doi.org/10.1016/j.procs.2016.11.017

Soler, J., & Wanner, L. (2014). How to use less features and reach better performance in author gender identification. In *Proceedings of the 9th International Conference on Language Resources and Evaluation* (pp. 1315–1319). Reykjavik, Iceland. Reykjavik: European Language Resources Association.