PSIHOLOGIJA, 2017, Vol. 50(1), 67–84

UDC: 159.923.072
81'27
DOI: 10.2298/PSI161012006J

The relationship between thematic, lexical, and syntactic features of written texts and personality traits

Ivana Jakovljev[1] & Petar Milin[2]

[1]*Laboratory for Experimental Psychology, University of Novi Sad, Serbia*
[2]*Department of Journalism Studies, University of Sheffield, UK*

The relationship between linguistic features of written texts and personality traits was investigated. Linguistic features used in this study were thematic (co-occurrence of the most frequent content words across participants), lexical (the maximum of new words) and syntactic (average sentence length). Personality traits were measured by VP+2 questionnaire standardized for Serbian population. Research was conducted on text materials collected from 114 Serbian participants (age 15–65), in their native tongue. Results showed that participants who gained low scores on Conscientiousness and high scores on Neuroticism and Negative Valence wrote about repeated daily activities and everyday life, but not about job-related matters or life perspective. Higher scores on Aggressiveness and Negative Valence coincided with writing about job-related matters and with the lower lexical richness. By showing that thematic content of text materials is affected by personality traits, these results support and expand previous findings regarding the relationship between personality and linguistic behaviour.

Keywords: linguistic behaviour, personality traits, principal component analysis, canonical correlation analysis

Previous studies have shown that each person uses a unique combination of linguistic features (i.e. grammar, syntax, spelling, vocabulary and phraseology) in his or her spoken and written communication (Juola, 2006; McMenamin, 2002; Van Halteren, Baayen, Tweedie, Haverkort, & Nejit, 2005). These findings encouraged researchers to investigate linguistic behaviour in the context of psychological variables, especially in the context of personality. Bearing in mind that personality traits represent relatively stable dispositions that affect human

Corresponding author: jakovljev.ivana@yahoo.com
Acknowledgements. This research was supported by Ministry of Education and Science, Republic of Serbia (grant no. 179033)

Note. Parts of this work were presented at conferences *Empirical studies in Psychology* and *Current trends in Psychology* in 2011.

behaviour (Mischel & Shoda, 1998), it was expected that linguistic behaviour is, to a certain degree, affected by some of the traits.

The existing literature is in dispute regarding specific features that could be considered as stable indicators of linguistic behaviour, and thus related to personality traits (McMenamin, 2002; Stamatatos, 2009). However, the study by Pennebaker and King (1999) demonstrated that people make systematic choices of words in their written texts, which are stable over time and across different topics. Based on this finding, numerous studies investigated whether personality affects the frequency in which people use specific words. The quantitative method that has been dominantly used in the field is the Linguistic Inquiry and Word Count – LIWC software (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007). LIWC is primarily concerned with lexical content (i.e. context-free) occurrences of words in text materials that fall within pre-defined categories (Pennebaker et al., 2007). Those categories consist of linguistic categories (e.g., articles, prepositions, pronouns), psychological processes (positive and negative emotions, cognitive processes, etc.), and the current concern (i.e. thematic content) dimensions (sex, death, home, occupation, etc; for details, consult the most recent LIWC edition in Pennebaker, Boyd, Jordan, & Blackburn, 2015, and for the Serbian version of LIWC consult Bjekić, Lazarević, Erić, Stojimirović, & Đokić, 2012).

Previous studies found various relationships between personality traits and word categories in LIWC. For example, it was demonstrated that higher Extraversion coincides with the frequent use of pronouns (Gill, Nowson, & Oberlander, 2009; Mairesse, Walker, Mehl, & Moore, 2007) and with the less frequent use of articles and negations (Mairesse et al., 2007; Pennebaker & King, 1999). People who gain high scores on Neuroticism were shown to frequently use first person pronouns (Gill et al., 2009; Oberlander & Gill, 2006; Pennebaker & King, 1999) and to rarely use third person pronouns (Oberlander & Gill, 2006). Also, it was demonstrated that people who score high on Openness to new experience use more articles (Mairesse et al., 2007; Mehl, Gossling, & Pennebaker, 2006) and more second person pronouns (Mehl et al., 2006; Qiu, Lin, Ramsey, & Yang, 2012).

The most consistent finding concerns the relationships between Extraversion and Neuroticism, and words that reflect emotions. Namely, it has been shown that higher Extraversion corresponds with the frequent use of positive emotion words (Mairesse et al., 2007; Pennebaker & King, 1999), while higher Neuroticism is associated with the frequent use of negative emotion words (Hirsh & Peterson, 2009; Mairesse et al., 2007; Pennebaker & King, 1999). These associations were also demonstrated in the online linguistic behaviour, specifically in the e-mail communication (Oberlander & Gill, 2006), Facebook messages (Schwartz et al., 2013), Twitter posts (Qiu et al., 2012) and online blogs (Gill et al., 2009; Li & Chignell, 2010; Yarkoni, 2010).

Particularly interesting are relationships demonstrated between personality traits and words that reflect the thematic content of texts. Studies showed that people with higher Extraversion frequently use words related to humans, social processes and family (Hirsh & Peterson, 2013; Pennebaker & King, 1999; Schwartz et al., 2013). Higher scores on Openness were shown to correspond

with the frequent use of words related to perceptual processes (Hirsh & Peterson, 2009, but also consult Yarkoni, 2010) and with the less frequent use of words reflecting work, family and home topics (Mairesse et al., 2007). People with high Conscientiousness were shown to use more job-related and time-related words, especially words related to the future (Gill et al., 2009; Mairesse et al., 2007) and to use less death and body-related words (Hirsh & Peterson, 2009).

Based on the abovementioned findings, several authors suggested that personality affects the topics that people are motivated to write about. For example, the frequent usage of first person pronouns in people with higher Neuroticism was discussed as their tendency to write about themselves (Argamon, Dhawle, Koppel, & Pennebaker, 2005; Gill et al., 2009). The frequent usage of pronouns and social process words in Extraverts was believed to reflect their motivation to write both about themselves and about other people (Gill et al., 2009; Yarkoni, 2010). Also, it was suggested that people with higher Openness show tendency to write about activities involving perceptual processes (art, television, culture), while people with higher Conscientiousness are motivated to write about work and job-related matters (Gill et al., 2009).

These interpretations, however, should be taken with caution for several reasons. First, described studies mainly used single words as units of analysis, therefore neglected the thematic context in which words were used. This might have resulted in the loss of valuable information about the thematic content of texts. Also, previous studies often predetermined topics that participants wrote about (Hirsh & Peterson, 2009; Pennebaker & King, 1999; Li & Chignell, 2010), which in return placed a focus on linguistic style (how do we say something) rather than the content (what we are saying; Yarkoni, 2010). It should also be noted that most of the recent studies have investigated specific online linguistic behaviour, such as Facebook messages or Twitter posts (Schwartz et al., 2013; Qiu et al., 2012). As suggested by several authors, this type of online linguistic behaviour is less rich and less emotional, unedited and informal in tone (Oberlander & Gill, 2006; see also Baron, 2003; Crystal, 2001), which makes it rather specific form of written language. Additionally, it has remained unclear to what extent results obtained on specific samples, such as bloggers (Gill et al., 2009; Yarkoni, 2010) could be generalized to the general population.

In this study, we aimed to further explore whether personality traits affect the thematic content of written language by investigating the topics people choose to write about. In contrast with previous studies that investigated occurrences of single words predefined in LIWC, we applied the principal component analysis to extract thematic components consisted of words that co-vary across texts. This approach is similar to the so-called *open-vocabulary approach*, which extract language features from the texts that are being analyzed (for more information about this approach consult Park et al., 2014). Compared to LIWC, the open-vocabulary approach showed more reliable prediction of individual differences in personality based on the analysis of Facebook statuses (Schwartz et al., 2013). The approach itself is clearly inspired by approaches that are gaining popularity in quantitative linguistics, psycholinguistics, and related fields, and which are commonly named distributed semantic models (also vector semantic models;

see, for example, Griffiths, Steyvers, & Tenenbaum, 2007; Marelli & Baroni, 2015; Shaoul & Westbury, 2010).

We applied this analysis to 5000-words texts materials collected from a rather wide range of participants. That way, we aimed to overcome limitations of previous studies, which collected data on the constrained sample of participants such as students (Hirsh & Peterson, 2009) or bloggers (Gill et al., 2009; Yarkoni, 2010). Also, text materials analyzed in this study did not include any form of short online communication (such as Facebook posts, Facebook messages, Twitter posts etc).

In addition, we explored some syntactic and lexical features of written texts, which were rarely investigated in the context of personality. Still, some previous studies succeeded to show lower lexical diversity in Extraverts (Dewaele & Furnham, 2000; Mairesse et al., 2007). Also, a recent study demonstrated that Aggressiveness and Depressiveness were positively correlated with the average sentence length, and negatively correlated with the lexical diversity (Litvinova, Zagorovskaya, Litvinova, & Seredin, 2016). Keeping that in mind, we additionally explored the relationship between personality traits on the one side, and sentence complexity and lexical richness, on the other.

Method

Sample

The sample consisted of 114 participants (61 females), divided into five age groups spanning the range from 15 to 65 years (Table 1). Given that each participant was asked to prepare an original authorial text of minimum 5000 words, we decided to apply the convenience sampling, i.e. to invite volunteers.

In this study we did not test participants themselves, but only analyzed texts and questionnaires they have submitted anonymously. Consequently, we did not use additional consent form but assumed that our participants, who volunteered rather long pieces of writing made exactly according to our specifications, have de facto agreed to participate, after they have been fully informed about the aims of the study.

Table 1
Sample structure by age and gender

Age group	Number of male participants	Number of female participants	Total number of participants
15–24	10	12	22
25–34	11	11	22
35–44	9	18	27
45–54	12	13	25
55–64	11	7	18

Instruments

VP+2 questionnaire (Smederevac, Mitrović, & Čolović, 2010). VP+2 was used for measuring personality traits. This questionnaire is designed for the assessment of seven major personality dimensions and contains 184 items with responses in the format of five-level

Likert-type scale. Items are grouped in seven main scales corresponding to seven personality dimensions: Extraversion, Neuroticism, Conscientiousness, Aggressiveness[1], Openness to experience, Positive valence and Negative valence (Smederevac et al., 2010). VP+2 questionnaire is derived from the LEXI questionnaire (Smederevac, Mitrović, & Čolović, 2007), which was based on the psycho-lexical study and the methodology of Tellegen and Waller and their Seven Factor Model of Personality Description (Waller, 1999). Compared to the five-factor model of personality (for example, see McCrae & John, 1992), the seven-factor model includes two additional evaluative categories named Positive valence and Negative valence (Waller, 1999). The VP+2 was standardized for Serbian population, which made it obvious choice for the present study.

Linguistic features used in the present research

Thematic features (co-occurrence of the most frequent content words across participants). To extract *thematic features* from a raw text material, we applied a text mining technique, which allows words analysis, words grouping and classification, and investigation of relations between text information and other variables (Han & Kamber, 2006; Hearst, 1999; StatSoft, 2010).

In the present study, text mining was used in form of analyzing the coincidence of participants in the "space" defined by the most frequent content words (i.e., nouns, verbs, adjectives, and adverbs). By using the Principal component analysis, we extracted several principal components, defined by converging words. Finally, we related extracted principal components with participants' personality traits.

Technically, the procedure that we have applied can be divided into a few simple steps. First, the texts written by a sample of participants were concatenated into a single text unit, with additional code markings added for each participant. This collection was segmented into words and sentence-end markers (dot, exclamation mark, and question mark). Then, words were lemmatized, using procedures by Ilić and Kostić (2002), and Milin (2004). The aim of lemmatization was to bring Serbian inflected variants under the same lemma (or lexeme; canonical form). For example, nominal inflected forms "kuće" and "kućom" were subsumed under the lemma "kuća" (*house*). Further, a stop-list was formed, to remove all function word, thus, leaving only nouns, verbs (excluding auxiliary verbs), adjectives and adverbs. Remaining lemmata were counted for frequency and ranked in descending order. In the penultimate step, 100 most frequent words, used by 10 or more participants, were retrieved for the later analyses. Finally, coincidence matrix was formed, with the most frequent words as columns, and participants' codes as rows. Cells of this matrix contained frequencies of occurrence of a lemma (column) in the text of a given participant (row). Since this type of matrices is typically sparse (contains many zero-cells), we applied a simple transformation procedure to minimize risks: cell values were increased by one and transformed into log-ratio of observed and expected frequencies (c.f., Siegel & Castellan, 1988):

$$X_{ij} = \log\left(\frac{F_{ij} \times F_{TOTAL}}{F_i \times F_j}\right)$$

where F_{ij} is the cell frequency, F_i and F_j are corresponding marginal values (row-frequency and column frequency, respectively), and F_{TOTAL} is the summed frequency of all words. The ratio of the observed and expected values formalizes the degree of informativeness:

1 Dimension Aggresiveness in VP+2 questionnaire corresponds with the negative pole of the Agreeableness dimension (as defined in the five-factor and seven-factor personality models).

in cases where observed and expected frequencies are close, the ratio will show this lack of surprise and vice versa (Siegel & Castellan, 1988). At the same time, log-transformation conveniently improves symmetry of the distribution and centers ratio values at zero.

Syntactic feature (average sentence length). As a *syntactic feature* we used the average sentence length for each participant. This feature indicates sentence complexity because longer sentences tend to have more complex structure. Given the elongated right-tail of the sentence length distribution, median was used as a measure of central tendency.

Lexical feature (the maximum of new words). Lexical richness represents a measure of the estimated size of author's vocabulary (Juola, 2006). In this study, we used characteristic constant of text, which was proposed by Milin and Ilić (2003). Between the two measures that authors proposed, we singled *the maximum of new words* that represents the positive maximum of the ratio between new words (appearing for the first time) and the old words (already appeared in a given text). Milin and Ilić (2003) showed that the maximum of new words represents the influx of old, repeating words in the text: the slower the influx, the higher the positive maximum value, therefore, the greater the lexical richness.

Procedure

Initially, we contacted a larger pool of potential volunteers via e-mail. After their response, the final list of participants, balanced by age and gender, has been sampled. We decided not to control the education level of participants because it would additionally aggravate sampling process, which would produce a smaller sample for the study. After the sample of participants was formed, we sent back instructions for the text writing, and answering the VP+2 questionnaire.

While collecting participants' text material the main requirement was the length of a text sample since previous studies showed that the reliability of analysis grows with the length of the text (c.f., Stamatatos, Fakotakis, & Kokkinakis, 2001). Given that we used a sample of volunteers, we made a decision that text material consisting of 5000 to 7000 words from each participant would be sufficient to provide reliable results, but also not to induce participants' drop-off. We instructed participants to compile a combination of their own writings, both existing and newly written. The content of the text material was not predetermined since it would limit the thematic variation, which was in the focus of this study. However, instruction provided detailed specification aiming at the widest possible range of functional styles (e.g. diary entries, written correspondences, essays, newspaper articles, creative writing samples etc.) and it also proposed a list of provisional topics to write about (e.g. description of the last couple of days/weeks, description of some previous events in life, comments and observations on any matter etc.). Participants were also instructed that text materials should not contain any form of the short online content (such as Facebook messages/posts, Twitter posts, etc.).

The collection of text materials lasted for two months. The participation in the research was anonymous and participants were free to choose whether to submit the material in paper or electronic form. In the end, we retained only those participants who fulfilled all the criteria specified beforehand and stated in the instruction.

Results

Preliminary analyses

Principal component analysis. We applied principal component analysis to reduce the dimensionality of coincidence matrix and to make it viable for succeeding analyses. According to the Scree-test criterion for determining the

optimal number of dimensions (Cattell, 1966), we set aside six major components. Dimensions were Varimax rotated and interpreted based on the factor loadings matrix (Appendix A). We obtained the most important thematic components from participants' text materials, in decreasing order of explained variance (i.e., importance). They were described, in order from the first to sixth, as:

– Repeating daily activities (actions) and the avoidance of contents that indicate duration and perspective
– The avoidance of the domestic policy matters
– Everyday life and the avoidance of professional, occupational themes
– The most painful domestic socio-economic issues with general and diffused xenophobia
– A matter of the existence and life perspective
– The avoidance of topics related to culture, literature, and language

Age and gender as predictors of linguistic features. Preliminary analyses also included testing the effects of age and gender onto main linguistic variables: thematic (scores on six principal components), syntactic (average sentence length) and lexical (the maximum of new words). We found marginally significant positive effect of age to the maximum of new words ($\beta = .18$; $t = 1.94$; $p = .055$): the maximum of new words was increasing with age. This result is in line with some previous studies suggesting the increase of language complexity with age (Pennebaker, Mehl, & Niederhoffer, 2003; Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014). At the same time, older participants wrote less about *repeating daily activities* ($\beta = –.25$; $t = –2.67$; $p = .009$), with less *avoidance of the domestic policy matters* ($\beta = –.21$; $t = –2.23$; $p = .028$). From these results, it seemed that the older participants had been dealing with more specific topics, political in particular, rather than daily activities.

Relations between linguistic features used in this study. Before the main analyses, aimed at relationships between linguistic features and personality traits, we examined relations *within* the set of linguistic features using partial correlation coefficients. Analysis showed mostly low correlations between measures of language richness and thematic components (Table 2, Appendix B). This was not to our surprise, considering that those are relatively different and independent linguistic variables. Furthermore, isolated thematic components are relatively complex and indirect indicators, while the maximum of new words and sentence length are simple measures, obtained directly from the raw texts. Some of the observed partial correlations reached statistical significance: *the maximum of new words* and the thematic component related to *the most painful domestic socio-economic issues with general and diffused xenophobia* ($r = .55$; $p < .001$); and *average sentence length* and the thematic component reflecting *repeating daily activities (actions) and the avoidance of contents that indicate duration and perspective* ($r = –.61$; $p < .001$). These significant relations will be discussed in the following sections.

Given the overall low correlations between thematic components on the one side, and lexical and syntactic features, on the other, we carried out two separate analyses, independently relating these two sets of linguistic features with personality traits.

Main analyses

Canonical correlation between thematic components and personality traits. In the analysis relating six thematic components with seven personality traits, the first pair of canonical roots reached statistical significance: $R = .51$; chi-square = 63,03; $p = .02$. Among thematic components, high loadings on extracted canonical root showed *repeating daily activities (actions) and avoidance of contents that indicate duration and perspective* (β = .77), and *everyday life and the avoidance of professional, occupational themes* (β = .51). Amongst personality traits, results showed high loadings on Conscientiousness (β = –.93), Neuroticism (β = .63) and Negative valence (β = .511). Detailed results of the Canonical Correlation Analysis with canonical factor loadings are given in Appendix C.

The results suggested that texts of participants who achieved significantly lower scores on Conscientiousness, and higher scores on Neuroticism and Negative valence contained daily activities and plans, and did not include professional or life's perspective themes. This seemed plausible: people who achieve high scores on Neuroticism are often worried and have difficulties in decision making and in copying with new situations (Smederevac et al., 2010). Hence, it was not to our surprise that they avoided writing about changes and future events in general. Negative self-perception expressed through the Negative valence was probably another reason for these participants to write mostly about everyday topics, without focusing on professional issues and matters of perspective. Conversely, Conscientiousness concerns attitude toward responsibilities (Smederevac et al., 2010). Thus, profession topics would not be expected if the scores on the dimension were low.

Canonical correlation between lexical and syntactic features and personality traits. Results of the canonical correlation analysis showed that there were no significant correlations between lexical and syntactic features, on the one hand, and personality traits, on the other. These results will be elaborated in the Discussion section in details.

Joint Principal Component Analysis. As our last step in the analysis we applied the Principal Component Analysis combining all three sets of variables from our study: (a) personality traits (seven dimensions); (b) thematic components (six isolated dimensions); and (c) measures of lexical and syntactic richness (the maximum of new words and average sentence length). Our motivation was two-folded. First, this appears to be the simplest yet appropriate approach to explore the common structure of more than two sets of variables (we are, however, well-aware of the existence of other, more advanced techniques such

as, for example, OVERALS: van der Burg, de Leeuw, & Dijksterhuis, 1994). Second, Knežević and Momirović (1996) convincingly discussed problems that Canonical Correlation Analysis can bring, and for that reason this joint analysis served to verify (i.e., replicate) relationship structures observed in two separate canonical analyses.[2]

We used the Kaiser-Guttman criterion to decide upon number of components (Guttman, 1954; Kaiser, 1961), since Scree-test curve deviated from the typical, negative decelerating function with a point of infraction, thus, not allowing unambiguous conclusion about the number of dimensions that should be kept. Six components had satisfied this criterion (having eigenvalues greater than 1). Overall, 67% of the variance of the set of 15 variables had been explained. Isolated dimensions were Varimax rotated to obtain the structure convenient for interpretation. For details, see Appendix D.

Two out of the six isolated dimensions were not considered in the final interpretation. The first dimension had high loadings on seven personality dimensions only, not reflecting any relations with linguistic variables. The fifth dimension had only one significant loading and was treated as the so-called *single factor*, which is usually excluded from final Principal Component solution (see, Harris, 1975). Four remaining dimensions not only confirmed the findings from Canonical Analysis but also revealed some important additional relations across sets of variables.

The second component had high loadings with the maximum of new words ($\beta = .67$), and the thematic component related to *the most painful domestic socio-economic issues with general and diffused xenophobia* ($\beta = .77$). This was in line with the preliminary analysis, which showed significant partial correlation, in the same direction, between the abovementioned variables. From this, we concluded that more frequent writing about socio-economic matters coincided with higher lexical richness and vice versa.

The third isolated component was characterized with the high positive loading of the average sentence length ($\beta = .86$), and high negative loading of the thematic component reflecting *repeating daily activities (actions) and the avoidance of contents that indicate duration and perspective* ($\beta = –.88$). These results showed that participants, who wrote about daily activities without dealing with more complex topics in their texts, also wrote in shorter, simpler sentences. Again, this finding was in line with our preliminary analysis, showing significant negative partial correlation for the same variables.

The fourth isolated component revealed important additional relations across sets of investigated variables. This component had high loadings with thematic component about *everyday life and the avoidance of professional, occupational themes* ($\beta = –.64$) and Aggressiveness ($\beta = .76$), and somewhat lower loadings with the Negative valence ($\beta = .38$) and the maximum of new words ($\beta = –.38$). These results showed that higher scores on Aggressiveness

2 Note, however, that Knežević and Momirović (1996) suggested Quasi-Canonical Analysis as the best alternative, but the method suffers from the same limitation as the original Canonical Analysis, since it cannot be applied to more than two sets of variables.

and Negative Valence coincided with more frequent writing about professional topics, and with a more modest vocabulary.

Finally, on the sixth isolated dimension, two thematic components showed high loadings: *a matter of the existence and life perspective* (β = .86) and *the avoidance of topics related to culture, literature and language* (β =.41). This result indicated that the existential themes coincided with the absence of topics related to culture.

Discussion

Previous studies have repeatedly demonstrated relationships between personality traits and words people tend to use most frequently in their written language (Hirsh & Peterson, 2009; Pennebaker & King, 1999; Yarkoni, 2010). In this study, we demonstrated that personality also (co)affects the choice of thematic context in which words appear in texts, i.e. the topics people choose to write about.

The key thematic components in participants' texts that showed significant relationship with personality traits were (a) topics related to everyday life and everyday actions and activities, (b) issues related to the job and professional matters and problems, and (c) issues related to the development and the life perspective. Participants, who gained low scores on Conscientiousness and high scores on Neuroticism and Negative valence, wrote about everyday topics but did not write about professional and life perspective topics. People with low scores on Conscientiousness show lack of organization, they are inert and passive about their duties and their work (Smederevac et al., 2010), and hence it is not surprising that these people would not write about professional issues. This result is in line with previous findings suggesting that people with higher scores on this dimension tend to use more work related words in their written language (Gill et al., 2009; Hirsh & Peterson, 2009). Additionally, our finding that low scores on Conscientiousness coincided with the avoidance of writing about life perspective is in line with the finding that people who score high on this dimension frequently use words related to the future (Mairesse et al., 2007). Anxiety, worry and a negative self-perception that coincide with higher scores on Neuroticism and Negative valence prod an avoidant behaviour in general, which in this study emerged as writing about daily activities and as avoiding of writing about life perspective and professional issues. These results provide a valuable contribution to the understanding of relations between Neuroticism and linguistic behaviour. Namely, not only that high Neuroticism coincides with the frequent use of words related to negative emotions (Pennebaker & King, 1999; Schwartz et al., 2013; Yarkoni, 2010), but it seems that it also coincides with the avoidance of complex topics and writing about perspective. To our surprise, even though previous studies have systematically demonstrated the relationship between Extraversion and the usage of specific words in texts (Gill et al., 2009; Pennebaker & King, 1999), this personality dimension did not show a significant relationship with the choice of topics. This result might be the consequence of the cross-cultural differences in the content of this personality dimension (for

example, see Čolović, Mitrović, & Smederevac, 2005; Smederevac et al., 2007) but further studies are needed in order to draw any valid conclusion.

Our results also showed that the presence of professional issues in texts related to higher scores on Aggressiveness and Negative valence. It is interesting to notice that our findings indicate a general conative negativity regarding professional matters. The absence of professional themes in texts was accompanied with high Neuroticism, while the presence of these themes was accompanied with high scores on Aggressiveness. Such findings suggested that the professional domain was troublesome for most of the participants. Despite the fact that the sample of participants was not random, participants with different educational levels and occupational statuses were included, enabling us to conclude that the *sensitivity* about the participants' professional life was equally present across age, education and specific job situation.

Our results showed that topics related to daily activities explained most frequent words variation. It seems that such topics were the most common and typical for our participants, especially to younger ones. Additionally, it appears that topics related to the perspective, as well as political and socio-economic topics concerned our participants the most. This assumption is also indirectly supported by the fact that the same component explained how the *existential* topics were related to the absence of language and culture topics. Based on these results, it seems that the choice of topics might be influenced by socio-economic and political circumstances, and moderated by age. Nevertheless, our results suggest that personality traits affect the relationship between mentioned situational factors and the choice of the message content. This hypothesis is supported by a group of personality theories, which emphasize personality and situation interaction – the stable characteristic of personality are reflected in the way a person behavior varies as a function of specific features of situations (Read & Miller, 1998). Considering our results in general, we can assume that the influence of personality traits on the topics people choose to write about might be happening in the same way.

Even though our results demonstrated complex relationships within investigated linguistic features, lexical and syntactic features showed poor relations with personality traits. However, in the joint analysis, we demonstrated the relationship between Aggressiveness and Negative valence on the one side and a lower lexical richness, on the other, which is in line with previous findings suggesting more modest vocabulary in people with higher Aggressiveness and higher Depressiveness (Litvinova et al., 2016).

A possible reason for the lack of correlation between measures of language richness and personality traits might be the fact that the functional styles of participants' texts were not controlled. It is possible that this benefited the thematic variations, but reduced lexical and syntactic variations, which led to the significant relations between the thematic components and personality traits only. Also, when choosing the adequate linguistic features, characteristics of language itself should be considered. This is important, since most of the previous studies were conducted in English, which has relatively low morphological richness, hence the usage of syntactic features would provide higher discriminability between authors (compare

with Manning & Schuetze, 2000, regarding problems of natural language processing, in general). Conversely, in inflected languages, like Serbian, morphological analysis is the central problem in language studies (c.f., Baayen & Sproat, 1996). Hence, we believe that future studies in Serbian language should explore additional markers of lexical richness in the context of personality traits.

Bearing in mind that we used a novel approach to text analysis and collected our data outside of the English speaking area, we believe that our results represent a valuable contribution to the understanding of the linguistic behaviour in the context of personality. However, couple of limitations of this study should be considered. The most important one concerns the size of the collected written material. Having in mind that we aimed to include participants of different age and backgrounds and to investigate the *offline* linguistic behaviour, we were not able to provide samples of hundreds of thousands words for each participant, as required by some authors in the field (for example Juola, 2006; McMenamin, 2002). Larger language sample would, for certain, provide much more detailed insight into the relationship between thematic variations and personality traits. Also, as we have already noted, it is possible that we have chosen quite coarse measures of lexical and syntactic richness, which could have led to non-significant correlations with personality traits.

In conclusion, by showing that personality affects the choice of topics in written language, our results support previous studies conducted in this field of research. Future studies should continue to search for answers regarding the connection between the particular way we write and speak and the personality domain. But also, future studies should be encouraged to explore linguistic features in the context of other psychological structures. Bearing in mind that the language is one of the most complex forms of human behavior, we can expect that the answer will not be simple and that we cannot search for it in a single psychological domain.

References

Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. W. (2005). Lexical predictors of personality type. *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*. Retreived from: https://goo.gl/cnVSw3

Baayen, H., & Sproat, R. (1996). Estimating lexical priors for low-frequency morphologically ambiguous forms. *Computational Linguistics*, *22*(2), 155–166.

Baron, N. S. (2003). Language of the Internet. In A. Farghali (Ed.), *The Stanford handbook for language engineers* (pp. 59–127). Stanford, California: CSLI.

Bjekić, J., Lazarević, Lj., Erić, M., Stojimirović, E., & Đokić, T. (2012). Razvoj srpske verzije rečnika za automatsku analizu teksta (LIWCser).Psihološka istraživanja, *15*(1), 85–110.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Crystal, D. (2001). *Language and the Internet*. Cambridge University Press.

Čolović, P., Mitrović, D., & Smederevac, S. (2005). Evaluacija modela Pet velikih u našoj kulturi primenom upitnika FIBI. Psihologija, *38*(1), 55–76.

Dewaele, J. M., & Furnham, A. (2000). Personality and speech production: a pilot study of second language learners. *Personality and Individual differences*, *28*(2), 355–365.

Gill, A. J., Nowson, S., & Oberlander, J. (2009). What are they blogging about? Personality, topic and motivation in blogs. In E. Adar, M. Hurst, T. Finin, N. Glance, N. Nicolov, &

B. Tseng (Eds.), *Proceedings of the 3rd international AAAI conference on weblogs and social media* (ICWSM09) (pp. 18–25). Menlo Park, CA: The AAAI Press.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika, 19*(2), 149–161.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. San Francisco: Elsevier.

Harris, R. J. (1975). *A Primer of Multivariate Statistics*. New York: Academic Press.

Hearst, M. (1999). Untangling Data Mining. In *Proceesings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 3–10).

Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of research in personality*, *43*(3), 524–527.

Ilić, N., & Kostić, A. (2002). Problem homografije pri automatskoj lematizaciji [Problem of homography in automatic lemmatization]. Paper presented at the *The 8th Empirical Studies in Psychology*, Faculty of Philosophy, Belgrade, Serbia.

Juola, P. (2006). Authorship Attribution, *Foundation and Trends in Information Retrieval, 1*(3), 233–334.

Kaiser, H. F. (1961). A note on Guttman's lower bound for the number of common factors. *British Journal of Psychology, 14*(1), 1–2.

Knežević, G. D., & Momirović, K. (1996). Algoritam i program za analizu relacija kanoničke korelacijske analize i kanoničke analize kovarijansi kovarijansi [Algorithm and program for the analysis of the relations between the canonical correlation analysis and canonical analysis of covariance]. U P. Kostić (Ur.), *Merenje u psihologiji* 2 (str. 57–73). Beograd: Institut za kriminološka i sociološka istraživanja/IKSI.

Li, J., & Chignell, M. (2010). Birds of a feather: How personality influences blog writing and reading. *International Journal of Human-Computer Studies*, *68*(9), 589–602.

Litvinova, T., Zagorovskaya, O., Litvinova, O., & Seredin, P. (2016). Profiling a Set of Personality Traits of a Text's Author: A Corpus-Based Approach. In A. Ronzhin, R. Potapova, & G. Nemeth (Eds.), *Speech and Computer: Proceedings of the 18th International Conference, SPECOM 2016* (pp. 555–562). Springer International Publishing.

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research, 30*, 457–500.

Manning, C. D., & Schuetze, H. (2000). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Marelli, M., & Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological review*, *122*(3), 485.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality*, *60*(2), 175–215.

McMenamin, G. R. (2002). *Forensic linguistics: Advances in forensic stylistics*. CRC press.

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, *90*(5), 862.

Milin, P. (2004). *Probabilistic approach to morphological disambiguation and cognitive strategies in language processing* (Doctoral thesis). University of Belgrade, Serbia.

Milin P., & Ilić, N. (2003). Text as Binary Sequence: A Case of Characteristics Constant of Text. In *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora*; 10th Conference of the European Chapter of the Association for Computational Linguistics (pp. 47–53).

Mischel, W., & Shoda, Y. (1998). Reconciling processing dynamics and personality dispositions. *Annual Review of Psychology, 49,* 229–258.

Oberlander, J., & Gill, A. J. (2006). Language with character: A corpus-based study of individual differences in e-mail communication. *Discourse Processes, 42*(3), 239–270.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, *108*(6), 934.

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC 2007 [Software manual]. Austin, TX.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, *77*(6), 1296.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54*(1), 547–577.

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality*, *46*(6), 710–718.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in cognitive science*, *6*(1), 5–42.

Read, S. J., & Miller, L. C. (1998). *Connectionist models of social reasoning and social behavior.* Mahwah, NJ: Erlbaum.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8*(9), e73791.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*(2), 393–413.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Smederevac, S., Mitrović, D., & Čolović, P. (2007). The structure of the lexical personality descriptors in the Serbian language. *Psihologija, 40*(4)*,* 485–508.

Smederevac, S., Mitrović, D., & Čolović, P. (2010). *Velikih pet plus dva: Primena i interpretacija* [Big five plus two: Manual for administration and interpretation]. Beograd, Srbija: Centar za primenjenu psihologiju.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, *60*(3), 538–556.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, *35*(2), 193–214.

StatSoft (2010). *Electronic Statistics Textbook*. Tulsa: StatSoft.

Van der Burg, E., de Leeuw, J., & Dijksterhuis, G. (1994). OVERALS: Nonlinear canonical correlation with k sets of variables. *Computational Statistics & Data Analysis*, *18*(1), 141–163.

Van Halteren, H., Baayen, H., Tweedie, F., Haverkort, M., & Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, *12*(1), 65–77.

Waller, N. G. (1999). Evaluating the structure of personality. In R. C. Cloninger (Ed.), Personality and psychopathology (pp.155–197). Washington: American Psychiatric Press.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, *44*(3), 363–373.

RECEIVED 12.10.2016.
REVISION RECEIVED 08.11.2016.
ACCEPTED 12.11.2016.

Appendix A

Eigenvalues, explained variances, component loadings, and component interpretations, from the transformed coincidence matrix of participants and the most frequent content words

Ordinal number of the component, its eigenvalue and explained variance (in brackets)	The name of the thematic component	Words with high positive loadings	Words with high negative loadings
I 3.491 (23.27%)	Repeating daily activities (actions) and the avoidance of contents that indicate duration and perspective	– to ask (*pitati*): 0 .678 – to think (*misliti*): 0.667 – to watch (*gledati*): 0.658 – to talk (*pričati*): 0.648 – to hear (*čuti*): 0.579 – to go (*ići*): 0.579 – to leave (*otići*): 0.582 – to say (*kazati*): 0.565 – to love (*voleti*): 0.563	– development (*razvoj*): – 0.613 – course (*tok*): –0.573 – basic (*osnovni*): –0.547 – number (*broj*): –0.543 – period (*period*): –0.547
II 1.693 (11.29%)	The avoidance of the domestic policy matters		– Serbia (*Srbija*): –0.775 state (*država*): –0.767 – political (*politički*): – 0.709 – Vojvodina (*Vojvodina*): –0.688 – law (*zakon*): –0.610 – legal right (*pravo*): –0.547 – Serbian (*srpski*): –0.458 – land (*zemlja*): –0.458
III 1.465 (9.76%)	Everyday life and the avoidance of professional, occupational themes	– house (*kuća*): 0.633 – to return (*vratiti*): 0.564 – thing (*stvar*): 0.513 – city (*grad*): 0.484 – to live (*živeti*): 0.484 – year (*godina*): 0.469 – child (*dete*): 0.445 – day (*dan*): 0.598	– process (*proces*): – 0.604 – relation (*odnos*): –0.572 – case (*slučaj*): –0.517 – problem (*problem*): –0.497 – system (*sistem*): –0.506
IV 1.280 (8.54%)	The most painful domestic socio-economic issues with general and diffused xenophobia	– to work (*raditi*): 0.628 – job (*posao*): 0.532 – to get (*dobiti*): 0.479	– world (*svet*): –0.443
V 1.090 (7.28%)	A matter of the existence and life perspective	– man (*čovek*): 0.542 – life (*život*): 0.541 – young (*mlad*): 0.424 – a lot (*mnogo*): 0.577	– the next one (*sledeći*): –0.541
VI 1.024 (6.83%)	The avoidance of topics related to culture, literature and language		– book (knjiga): –0.706 – culture (kultura): –0.543 – language (jezik): –0.549

Note. Total variance explained was 66.95%; Loadings bellow 0.4 are not displayed

Appendix B

Correlations between thematic components and measures of language richness

Thematic components	Average sentence length	Maximum of new words
Repeating daily activities (actions) and the avoidance of contents that indicate duration and perspective	-.61**	-.04
The avoidance of the domestic policy matters	-.04	-.22*
Everyday life and the avoidance of professional, occupational themes	.07	.28*
The most painful domestic socio-economic issues with general and diffused xenophobia	.11	.55**
A matter of the existence and life perspective	-.02	-.11
The avoidance of topics related to culture, literature and language	-.03	.13

Note. *p <.05, ** p <.001

Appendix C

Canonical loadings for left (personality traits: variance extracted 0.259; redundance 0.068) and right (thematic components: variance extracted 0.167; redundance 0.043) set of variables

Personality traits	Canonical Loadings	Thematic Components	Canonical Loadings
neuroticism	0.631	Repeating daily activities (actions)	0.768
extraversion		The avoidance of the domestic policy matters	
conscientiousness	-0.929	Everyday life and the avoidance of professional, occupational themes	0.507
aggressiveness		The most painful domestic socio/ economic issues	
openness to experience		A matter of the existence and life perspective	
positive valence	-0.311	The avoidance of topics related to culture, literature and language	
negative valence	0.512		

Note. Loadings bellow 0.3 are not displayed

Appendix D

Component loadings of three sets of variables on 6 isolated principal components

Variables	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
neuroticism	-0.692					
extraversion	0.802					
conscientiousness	0.609					
aggressiveness				0.763		
openness to experience	0.782					
positive valence	0.756					
negative valence	-0.595			0.384		
maximum of new words		0.673		-0.397		
average sentence length			0.859			
Repeating daily activities (actions)			-0.887			
The avoidance of the domestic policy matters					0.829	
Everyday life and the avoidance of professional, occupational themes				-0.638		
The most painful domestic socio/ economic issues		0.767				
A matter of the existence and life perspective						0.860
The avoidance of topics related to culture, literature and language						0.406
Explained variance	0.207	0.0984	0.114	0.102	0.077	0.0710

Note. Loadings bellow 0.3 are not displayed