PSIHOLOGIJA, 2014, Vol. 47(1), 5–32 UDC: 159.9.07:811.163.41'373 ; 004.9:811.163.41'322

DOI: 10.2298/PSI1401005B

Psychometric evaluation of the Serbian dictionary for automatic text analysis – LIWCser

Jovana Bjekić[1,2], Ljiljana B. Lazarević[3], Marko Živanović,[1] and Goran Knežević[1]

[1]*Department of Psychology, University of Belgrade, Serbia*
[2]*Institute for Medical Research; University of Belgrade, Serbia*
[3]*Institute of Psychology, University of Belgrade, Serbia*

LIWC (Linguistic Inquiry and Word Count) is widely used word-level content analysis software. It was used in large number of studies in the fields of clinical, social and personality psychology, and it is adapted for text analysis in 11 world languages. The aim of this research was to validate empirically newly constructed adaptation of LIWC software for Serbian language (LIWCser). The sample of the texts consisted of 384 texts in Serbian and 141 texts in English. It included scientific paper abstracts, newspaper articles, movie subtitles, short stories and essays. Comparative analysis of Serbian and English version of the software demonstrated acceptable level of equivalence (ICCM=.70). Average coverage of the texts with LIWCser dictionary was 69.93%, and variability of this measure in different types of texts is in line with expected. Adaptation of LIWC software for Serbian opens entirely new possibilities of assessment of spontaneous verbal behaviour that is highly relevant for different fields of psychology.

Key words: *automatic text analysis, Linguistic Inquiry and Word Count – LIWC; LIWCser; psychometric evaluation of LIWCser*

There is a consensus among the authors that words we use map our mental, social and physical states (Frojd, 1969; Tausczik & Pennebaker, 2010). During the history of psychology a number of prominent researchers, pointed out the importance of studying the ways people naturally talk in the real world (e.g., Bradac, 1986; Gottschalk & Gleser, 1969; Gottschalk, Gleser, Daniels, & Block, 1958). Although this idea exists in psychology for more than a century, researchers recently started to systematically investigate relationship between psychological constructs, on one side, and content and style of verbal behaviour, on the other (Hirsh & Peterson, 2009; Mehl, Gosling, & Pennebaker, 2006).

Corresponding author: bjekicjovana@gmail.com

* Research was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (179018; 175012)

Quantitative approach in the analysis of verbal behaviour, seeks objectivity (i.e., measurement equivalence across studies), through explicit criteria on classification of the words and quantification, and is based on extraction of the "psychometrically good data" (Mehl, 2006; Mehl & Gill, 2010). It also offers low-cost and comprehensive research (Pennebaker, Mehl, & Niederhoffer, 2003; Ramirez-Esparza, Pennebaker, Garcia, & Suria, 2007).

Several distinct approaches in the quantitative analysis have been developed, i.e., thematic text analysis, and automatic text analysis – ATA (Pennebaker et al., 2003). ATA has emerged from the development of artificial intelligence and focuses on the frequency (i.e., intensity) of thematic and/or stylistic characteristics of the text (Shapiro & Markoff, 1997; Pennebaker et al., 2003). Methodologically, it has several advantages. First, since computer software analyses data, it provides results that are more objective and replicable, compared to manual coding. Second, measurement error (that usually results from individual differences between raters) is minimal and it allows methodological equivalence of different studies. Finally, these data do not share method variance with the data obtained with other assessment methods that researchers frequently use in psychology (Mehl & Gill, 2010).

It is possible to differentiate two relatively distinct methodological approaches within the ATA. First approach, *Word pattern analysis,* based on complex algorithms, detects how meaning conveyers (words and word phrases) group in large text samples (Wolfe, Schreiner, Rehder, Laham, Foltz, Kintsch, & Landauer, 1998). For example, Latent Semantic Analysis (LSA) enables researchers to determine similarity of the texts based on latent structure of the meaning in the analyzed verbal product – i.e., it is concerned with the use of words in a specific context (Landauer, Foltz, & Laham, 1998). Second approach, *Word count strategies*, focuses on a single word analyses in order to extract both content and style properties of the text. Basic assumption is that individual differences in the frequency of use of specific words or word groups reflect individual differences in feelings, attitudes, and cognition (Pennebaker et al., 2003). Therefore, software designed to perform single word analysis focuses on word counting, according to predefined (grammatical or semantic) word categories.

Software for the Automatic Text Analysis

In the beginning, researchers used ATA dominantly in the field of clinical psychology but the focus has broadened to other fields, e.g., social, occupational, and psychology of individual differences (Pennebaker et al., 2003). With respect to that, several software for the ATA have been developed during the years, e.g., *The General Inquirer* (Stone, Dunphy, Smith, & Ogilvie, 1966), *TAS/C* (Mergenthaler, 1996), and *DICTION* (Hart, 1984; 2001) (for the overview see Bjekić, Lazarević, Erić, Stojimirović, & Đokić, 2012; Pennebaker et al., 2003; Lowe, 2003).

The authors of the most recent software, *Linguistic Inquiry and Word Count* (LIWC) constructed it to overcome issues related to judges' ratings in emotional writing assessment (Tausczik & Pennebaker, 2010). Word-count approach is a basis of LIWC and therefore, this software performs successive text analysis with a single word as unit of analysis. It compares grapheme patterns of each unit in the input text with the patterns in the dictionary incorporated into the software (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007). LIWC dictionary consists of a large number of grapheme patterns (words or word stems[1]) classified into categories, where single pattern can belong to one or several categories. Based on the number of patterns detected, software provides information about the share of each predefined category in the analyzed text. The content of the dictionary and software properties evolved over time – since the first attempt of construction in early '90 to the today's version in 2007 (for details about the process see Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007).

English LIWC2007 dictionary consists of about 4500 word stems, classified into 63 categories, which are relevant to various aspects of human cognitive, emotional, social, and physical functioning (Pennebaker et al., 2007). Authors organized these categories into four groups (Pennebaker et al., 2007)[2]. First group includes various *Linguistic processes,* e.g., verbs, auxiliary verbs, pronouns, adverbs, prepositions, etc., and other categories consisting of words manifesting the way something is said (e.g., negations, quantifiers, informal words, etc.). In the second group, authors included 32 hierarchically organised *Psychological categories,* created specifically for psychological researches (Pennebaker et al., 2007). These include several superordinate categories, i.e., *Social*, *Affective*, *Cognitive*, *Biological processes,* and *Relativity*. Each of these has several lower-level categories. For example, category *Social processes,* includes three lower-level categories: *Family*, *Friends*, and *People*. The third group consists of seven *Current concerns*, representing some of the most frequent themes in various kinds of texts: *Work, Achievement, Leisure, Home, Money, Death*, and *Religion*. Fourth group includes *Spoken categories* that are especially useful for the analysis of oral production (*Fillers, Assents*, and *Nonfluencies*). These were included in order to broaden the analyses beyond pure syntax and content characteristics of the text.

---

1 The term "word stem" has a meaning of the dictionary unit which is not a complete word. For example, some words are coded in all possible forms (*dog – pas* (nominative case, singular)*, psi* (nominative case, plural)*, psu* (dative case, singular), *etc*). On the other hand, some dictionary units, (which are referred to as "word stems") are grapheme patterns with the asterisk at the end, which capture more than one word/word form (e.g. prijatelj* - prijatelj /friend/, prijateljstvo /friendship/, prijateljski /friendly/, etc.). Note here that in this sense "word stem" is not necessary lexical or grammatical entity (e.g. jedrenj* - jedrenje / sailing/, jedrenjak /sailboat/, etc.)

2 For a detailed overview of the structure of the English LIWC2007 dictionary, see Pennebaker, et al., 2007.

In addition, LIWC2007 provides information about *General text descriptors*, e.g., word count, percentage of the text covered with the dictionary, number of the words longer than six letters, and frequency of different punctuation signs (Pennebaker et al., 2007).

Use of LIWC in Psychological Research

Large body of evidence suggested that automatic text analysis is very useful in the wider spectrum of psychology research. For example, in clinical psychology it was used for the evaluation of effectiveness of expressive writing in different clinical populations, e.g., depressive, psychotic, patients suffering from PTSD, etc. (e.g., Bernard, Jackson, & Jones, 2006; Gortner, Rude, & Pennebaker, 2006; Lepore, 1997). In addition, some features of verbal production were found to be related with different symptoms, such as negative affectivity, negative symptoms of schizophrenia, anhedonia, etc. (e.g., Watson & Pennebaker, 1989; Cohen, Alpert, Nienow, Dinzeo, & Docherty, 2008; Cohen, St-Hilaire, Aakres, & Docherty, 2009; etc.). In social psychology, LIWC was used in the research of lying and deception (e.g., Newman, Pennebaker, Berry, & Richards, 2002), interpersonal relationships (e.g., Ireland, Slatcher, Eastwick, Scissors, Finke, & Pennebaker, 2011), attitudes (e.g. Lee, 2009), and political views (e.g., Graham, Haidt, & Nosek, 2009). Researchers, also, demonstrated its usefulness in educational research (e.g., Carroll, 2007) and occupational psychology (e.g., Djikic, Oatley, & Peterson, 2006). In recent years, attention on linguistic markers of basic personality traits is rising. A large body of evidence suggests that personality reflects in linguistic style and that it is possible to assess it with LIWC (e.g., Hirsh & Peterson, 2009; Holtgraves, 2011; Mairesse, Walker, Mehl, & Moore, 2007; Pennebaker & King, 1999; Yarkoni, 2010, etc.)[3].

Even though there is a large body of evidence suggesting that individual differences in word use are related to different important psychological variables, the mechanisms underlying this relationship are yet to be discovered.

Translations and Adaptations of LIWC Dictionary to Different Languages

First LIWC software was using only English dictionary thus; authors used it in psychological research within the English speaking population. Since it proved to be a useful tool in different areas of research, researches started developing dictionaries in different languages. Among the first dictionaries to

3 For detailed overview of research invetigating role of linguistic parameters in different aspects of human functioning, see Bjekic et al., 2012.

be developed were Dutch (Zijlstra, van Meerveld, van Middendorp, Pennebaker, & Geenen, 2004), Italian (Alparone, Caso, Agosti, & Rellini, 2004), Spanish (Ramirez-Esparza et al., 2007), and German (Wolf, Horn, Mehl, Haug, Pennebaker, & Kordy, 2008). It is interesting that first translations differed very slightly from English dictionary due to linguistic similarities between these languages.

However, development of some other dictionaries, like French (Piolat, Booth, Chung, Davids, & Pennebaker, 2011) and Chinese (Huang, Chung, Hui, Lin, Seih, Chen et al., in press) was very time consuming, since it demanded alterations in the software itself in order to make text analysis possible. Namely, Chinese version of the software (C-LIWC) had to be able to make segmentation of the words before processing the text, while French had to allow inclusion of accent markers in the analysis. Beside these, Arabic (Hayeri, Chung, & Pennebaker, 2010), Russian (Kailer & Chung, 2011), Turkish (Murderrisoglu, 2011), and Korean (Lee, Shim, & Yoon, 2005) dictionaries were developed.

All adaptations of the LIWC software, except for the Arabic, Turkish and Russian (to the best of our knowledge) were empirically validated and demonstrated to be useful tool in psychology research beyond English speaking countries. For example, Spanish LIWC demonstrated usefulness in research of depression (Ramírez-Esparza, Chung, Sierra-Otero, & Pennebaker, 2009), bilingualism, and personality (Ramírez-Esparza**,** Gosling, Benet-Martínez, Potter, & Pennebaker, 2006). Korean LIWC was used in the analysis of political speeches (Chung & Park, 2010), research on relations between verbal outputs and age (Lee, Park, & Seo, 2006), and for the investigation of relations between basic personality structure and frequency of different word categories usage (Lee, Kim, Seo, & Chung, 2007).

Serbian LIWC Dictionary–LIWCser

Basis for the development of the LIWCser dictionary was LIWC2007 English dictionary. In addition to, we have used existing adaptations of this software to model Serbian dictionary. Serbian dictionary works with the same software as other LIWC2007 dictionaries. This means that the text analysis is conducted in the same successive manner and that the structure of the output is the same for all LIWC2007 adaptations. LIWCser dictionary corresponds to other dictionaries, with respect to formal and characteristics of the content. LIWCser consists of 12103 words and word stems classified into 65 categories. Table 1 shows LIWCser categories with representative examples of words.

Table 1. *LIWCser categories with representative examples of words.*

1.Linguistic processes	1.1. Word count	
	1.2. Words per sentence	
	1.3. Dictionary words	
	1.4. Words>6 letters	
	1.5. Total function words	
	1.6. Pronouns	
	1.6.1. Personal pronouns	
	1.6.1.1. 1st person singular	*ja*/I/, *moj*/my/
	1.6.1.2. 1st person plural	*mi*/we/, *naš*/our/
	1.6.1.3. 2nd person	*ti*/you/, *vaš*/your/, *tvoj*/your/
	1.6.1.4. 3rd person singular	*on*/he/, *njegov*/his/
	1.6.1.5. 3rd person plural	*oni*/they/, *njihov*/their/
	1.6.2. Impersonal pronouns	*neki*/somebody/,*svako*/everybody/
	1.7. Common verbs	*trčim*/run/, *ići*/go/, *znaju*/know/
	1.8. Auxiliary verbs	*ću*/will/, *smo*/are/
	1.9. Past	*davno*/long ago/, *juče*/yesterday/
	1.10. Present	*sada*/now/, *trenutno*/at the moment/
	1.11. Future	*ubuduće*/in future/, *sutra*/tomorrow/
	1.12. Adverbs	*uvek*/always/, *veoma*/much/
	1.13. Prepositions	*na*/on/, *ka*/to/, *iz*/from/
	1.14. Conjunctions	*dakle*/therefore/, *ali*/but/, *mada*/ although/
	1.15. Negations	*nije*/is not/, *neće*/would not/, *nisam*/ am not/
	1.16. Negative words	*Nesreća*/accident/, *neaktivan* /inactive/
	1.17. Superlatives	*Najbolji*/best/, *najgori* /worst/
	Quantifiers	*Mnogo*/much/, *puno/a lot/*
	Numbers	*Jedan*/one/, *deseti*/tenth/
	Swear /Informal words	*Mrš*/fuck-off/, *muda*/balls/, *omg*/oh my God/
2.Personal concerns	2.1. Work	*Preduzeće*/company/, *plata*/paycheck/
	2.2. Achievement	*Samouveren*/self-confident/, *šampion*/ champion/
	2.3. Leisure	*Hobi* /hobby/, *surfovanje* /surf/, *igra* / play/
	2.4. Home	*Dom* /home/, *kapija* /gate/, *kauč* / couch/
	2.5. Money	*Kupiti* /buy/, *dinar* /dinar/, *plaćati* /pay/
	2.6. Religion	*Pop* /priest/, *pričest* /communion/, *krštenje* /baptism/
	2.7. Death	*Masakr* /massacre/, *mrtav* /dead/, *pokojni* /deceased/

3. Psychological processes	3.1. Social processes	
	3.1.1. Family	*Mama/*mum/, *ujak/*uncle/, *porod*
	3.1.2. Friends	*Cimer/*rommey/, *drug/*friend/, *ortak/*buddy/
	3.1.3. Humans	*Sugrađani/*fellow citizens/, *sused/*neighbour/
	3.2. Affective processes	
	3.2.1. Positive emotion	*Sviđa/*like/, *lepo/*nice/, *sreća/*happiness/
	3.2.2. Negative emotion	*Grozno/*awful/, *prevara/*scam/
	3.2.3. Fear and Anxiety	*Zabrinut/*worried/, *briga/*concern/
	3.2.4. Anger	*Drzak/*rude/, *dovraga/*to hell//
	3.2.5. Sadness	*Plač/*cry/, *jad/*grief/, *lišen/*deprived/
	3.3. Cognitive processes	
	3.3.1. Insight	*Objasni/*explain/, *shvatam/*understand/
	3.3.2. Causation	*Stoga/*therefore/, *izaziva/*cause/
	3.3.3. Discrepancy	*Teže/*harder/, *treba/*should/, *umesto/*instead/
	3.3.4. Tentative	*Otprilike/*roughly/, *eventualno/*possibly/
	3.3.5. Certainty	*Kategorično/*categorically/, *moraš/*must/
	3.3.6. Inhibition	*Barijera/*barrier/, *osujeti/*frustrate/
	3.3.7. Inclusive	*Preuzet/*overtaken/, *prihvaćen/*accepted/
	3.3.8. Exclusive	*Sem/*but/, *stran/*foreign/, *van/*outside/
	3.4. Perceptual processes	
	3.4.1. See	*Belo/*white/, *svetlucav/*sparkling/
	3.4.2. Hear	*Kuc/*knock/, *doziva/*call/, *glas/*voice/
	3.4.3. Feel	*Opipam/*touch/, *kiselo/*sour/
	3.5. Biological processes	
	3.5.1. Body	*Noga/*leg/, *lice/*face/, *malje/*hair/
	3.5.2. Health	*Kašlje/*cough/, *lekar/*medicaldoctor
	3.5.3. Sex and Love	*Orgazam/*orgasm/, *nag/*naked/, *ljubi/*kiss/
	3.5.4. Ingestion	*Pečenje/*roast/, *piće/*drink/, *gutam/*swallow/
	3.6. Relativity	
	3.6.1. Motion	*Prolazi/*go through/, *putujem/*travel/, *ide/*goes/
	3.6.2. Space	*Ring/*ring/, *sever /*north/, *hodnik/*hallway/
	3.6.3. Time	*Ikad/*ever/, *januar/*January/, *kasno/*late/
4. Paralinguistic categories	4.1. Assent	*Svakako/*sure/,*vau/*wow/, *aha/*yeah/
	4.2. Nonfluencies	hmm, mm, uf
	4.3. Fillers	*Bla/*blah/, *brate/*bro'/, *mislimm/*Imean/

The construction of LIWCser has gone through several phases. First, we have translated all the words from English dictionary, and added synonyms, antonyms and jargon words. Content of Linguistic categories was defined upon word-lists for grammatical categories given in Serbian grammar book (Klajn, 2005), so that these categories would be representative for Serbian language. Than we have applied appropriate inflections to all the words from the initial pool. The following step included classification of the words into categories defined by LIWC2007 dictionary. In this step, five raters classified each word into one or more categories

by joined consensus of all five. In the final phase, two independent judges reviewed content of all categories and added some culturally specific words.

In the construction of LIWCser, we have paid a significant attention to linguistic and cultural context of future use of the program. Specific characteristics of Serbian language and culture were included in the dictionary, which resulted in certain deviations from English. For example, due to grammar differences category *Articles* was excluded, while categories *Superlative* and *Negative words* were added to LIWCser, because of their single word representations in Serbian. Furthermore, in English dictionary categories *Present*, *Past,* and *Future* include verbs in deferent tenses, while in Serbian version they were replaced with adverbials since most of the tenses in Serbian do not have single word representation. Finally, adding culturally specific words enriched the content of some categories. For example, words that represent important aspects of Orthodox Christian religion were added to the category *Religion*, words that mark different family relationships were added to category *Family*, most common informal and swear words were added to the category *Swear*, etc. (for details of the LIWCser construction see Bjekić et al., 2012).

With 12103 words and word stems, Serbian dictionary is larger than English (4500), Dutch (6568), and Spanish (7515), but smaller than French (39230). The basic reasons for this are differences between languages. For example, Spanish adjectives are gender specific and it led to a larger number of word stems in the dictionary (Ramirez-Esparza et al., 2007), while French dictionary has almost nine times more word stems compared to English, due to large number of synonyms, and different word forms (Piolat et al., 2011). Large number of words and word stems in Serbian dictionary results from developed inflexional morphology, large number of semantically similar words, slang, and culturally specific words that were included.

The largest number of word stems in LIWCser was classified in categories *Affective* and *Cognitive processes*, similarly to other LIWC adaptations (e.g., Alparone et al., 2004; Pennebaker et al., 2007; Ramirez-Esparza et al., 2007; Wolf et al., 2008; Zijlstra et al., 2004), due to psychological relevance of these categories (for the overview see Chung & Pennebaker, 2007; Tausczik & Pennebaker, 2010). In order to avoid misclassification in the text analysis, during the classification of the words into categories, authors decided to exclude from the dictionary all words that would fit into different categories when used with different meanings in different contexts (Bjekić et al. 2012).

Aim of the Research

Variety of information that automatic text analysis, and LIWC specifically provides, influenced expansion of use of this software. Development of the dictionary in several languages, enabled research in non-English speaking countries and cross-language evaluation of the findings obtained in English-speaking regions (Kroner-Herwig, Linkemann, & Morris, 2004; Lee et al., 2007; Yogo & Fujihara, 2008). Further, it enabled cross-cultural comparisons,

bilingualism research, research of second language acquisition, follow-up of the vocabulary development in different communities, and gaining insight into psychologically relevant linguistic aspects of different languages (Kim, 2008; Ramirez-Esparza, Gosling, Benet-Martinez, Potter, & Pennebaker, 2006). Finally, development of the dictionary for automatic text analysis in different languages provides an opportunity to larger number of researchers to investigate relations between psychological phenomena and language.

The aim of this paper is to present data about psychometrical properties of the Serbian dictionary for the LIWC software – LIWCser (Bjekic et al., 2012). In order to assess quality of LIWCser, since it has certain specificities resulting from inter-language differences (e.g., authors had to make specific decisions about certain categories in the process of construction), several aspects of the dictionary were tested. First, equivalence of results obtained with LIWCser and LIWC2007 was analysed. Second, we assessed efficacy of the dictionary when processing different forms of texts, i.e., comprehensiveness of the LIWCser dictionary. In addition, average representation of each of the category in different types of the texts was calculated, in order to gain information about the influence of specific context, which depends on the type of the text that is analysed. Finally, we tested the impact of the homonymous words exclusion on the comprehensiveness of the analyses.

Equivalence of LIWCser and LIWC2007

In order to assess generalizability of the results obtained with Serbian dictionary to the results obtained with English LIWC dictionary, we tested the *equivalence* of dictionaries on the parallel Serbian-English sample of texts.

Method

Sample. For equivalence testing a sample of 141 texts was used, out of which 46 (32.6%) were abstracts of scientific papers, 54 (38.3%) were newspapers articles, and 41 (29.1%) were movie subtitles. Each text was in both Serbian and in English; specifically, abstracts and newspapers articles were originally in Serbian but then translated to English, while movie subtitles were originally in English, and then translated to Serbian by a professional[4].When discussing sample size on the level of words, it is satisfying since it covers more than 35000 words (Wolf, et al. 2008).

Scientific journal abstracts were selected from different issues of journal *Psihologija* published between 2000 and 2008. Criteria for abstract selection were to have texts representing majority of fields in psychology, and to have abstracts with highest quality of translation from Serbian to English.

Newspapers articles were selected from electronic version of *JAT revija* magazine[5], which was chosen for several reasons. First, magazine is bilingual where professionals translated each text in full length to English. Second, magazine covers different topics (e.g. culture, leisure, sport, and politics) and formats (e.g. reports, interviews). These topics are relatively equally represented, which adds to diversity of content and writing styles. Finally,

4 Although, as pointed by the reviewer, it would be ideal to have equal number of both S-E and E-S translations, it was not possible because there are no scientific journals with translation from English to Serbian.

5 http://www.jat.com/active/sr-latin/home/main_menu/travel_info/jat_review.html

all articles have satisfying length, which adds to the reliability of analysis. All articles from the period between October 2010 and May 2012 were analysed.

Movie subtitles and their translations that were included in the analyses were downloaded from the internet[6]. Eleven subtitles of the movies nominated for the American Academy Award from the period between 2007 and 2011 were selected. We have divided each film into 3 to 5 parts equal in length. Subtitles were included in the analysis, because of the similarities between everyday language and the one used in the movies. Therefore, it was possible to observe differences in representativeness of LIWC categories in oral and written language.

Text analysis. No text corrections were made before processing, i.e., we did not correct possible printing errors nor did we exclude words that could be irrelevant for the analysis (e.g., personal names). English and Serbian texts were analysed with LIWC2007 English and with LIWCser, respectively.

Data analysis. Equivalence between dictionaries was conducted in a similar way as in German adaptation of LIWC (Wolf et al., 2008). The overall number of various texts belonging to the three aforementioned types was 282 (141 in Serbian and 141 in English). LIWC categories were calculated both for Serbian and English language and stored in the database (texts in the database emulated subjects, i.e., texts were stored in rows, whilst LIWC categories for both languages were presented in columns). Descriptive statistics indicating representation (% of each category in given text) and variability of different LIWC categories were calculated for all texts separately for English and Serbian versions. For the assessment of equivalence between English and Serbian LIWC dictionaries, rang-correlations were calculated (instead of Pearson correlations), thus avoiding potential problems resulting from extreme values and usually non-normal distributions of the LIWC categories (Wolf et al., 2008). As the primary measure of equivalence of two dictionaries, we used coefficient of intraclass correlation (ICC) Two-way mixed effect model, Consistency type. This measure directly reflects the proportion of between-texts variance (similar LIWC category values for both languages within a particular text) in the overall variance (between-texts + within-texts variance). Both measures of equivalence were calculated for each of the LIWC category across all texts.

Results

Serbian texts have on average 300 words less than parallel texts in English, i.e., on average in English texts there are two words per sentence more than in Serbian. In addition, English texts have higher percentage of function words (about 50% in English in comparison to 30% in Serbian). Highest difference is in the frequency of first person singular pronouns, which in English is 7% of all words in the text, while in Serbian these are about 3%.

Average correlation of pairs of Serbian and English LIWC results was .65, and average intraclass coefficient (ICC) was .70, where 76% of categories had correlations higher than .60. Table 2 presents descriptive statistics for each category in English and Serbian LIWC dictionary and data on dictionary equivalence. Average ICC for *Linguistic* categories was .74, for *Psychological* was slightly lower (ICC=.72), and for *Personal concerns* it was highest (ICC=.75). On the level of specific categories highest equivalence was observed for *Religion* (ICC=.96), *Family* (ICC=.96), *Negations* (ICC=.95), *Sex and love* (ICC=.93), *Sadness* (ICC=.92), *Achievement* (ICC=.91), and *Leisure* (ICC=.90). Categories with the lowest equivalence were *Present* (ICC=.30), *Anger* (ICC=.29), and *Feeling* (ICC=.24), while for categories *Past* and *Inclusion* ICCs were close to zero.

6 www.titlovi.com

Table 2. *Equivalence of LIWC2007 and LIWCser*

	LIWC2007					LIWCser						
	M	SD	Mdn	Min	Max	M	SD	Mdn	Min	Max	r	ICC
Word count	1390.20	1699.45	1049.00	87	13269	1056.74	1064.24	790.00	95	8452	.97**	.86**
Words per sentence	19.82	10.45	20.56	4.10	60.25	17.74	9.81	17.92	3.39	56	.87**	.91**
Dictionary words	80.32	8.28	81.97	58.21	93.67	64.28	5.34	65.22	53.60	73.30	.74**	.81**
Words>6 letters	24.66	10.48	23.38	7.76	48.09	33.14	13.08	32.61	11.76	59.00	.95**	.95**
Total function words	50.94	7.744	52.36	32.26	63.41	31.60	4.50	32.40	20.27	39.50	.82**	.83**
Total pronouns	11.58	7.67	10.45	0.00	24.83	8.69	3.96	8.57	1.49	16.38	.92**	.86**
Personal pronouns	6.90	6.11	5.48	0.00	17.89	2.92	2.63	2.00	0.00	8.51	.93**	.83**
1st pers singular	2.48	2.62	1.51	0.00	7.63	1.27	1.36	0.63	0.00	4.19	.94**	.87**
1rd pers plural	0.94	0.89	0.92	0.00	4.74	0.63	0.61	0.52	0.00	2.34	.68**	.72**
2nd person	1.92	2.40	0.73	0.00	7.49	0.89	1.15	0.28	0.00	3.91	.93**	.86**
3rd pers singular	0.88	0.97	0.57	0.00	4.03	4.11	1.58	4.08	0.87	8.87	.78**	.81**
3rd pers plural	0.67	0.63	0.56	0.00	3.73	0.53	0.50	0.45	0.00	3.51	.62**	.77**
Impersonal pronouns	4.68	2.01	4.67	0.00	9.93	3.83	1.45	3.92	0.00	8.01	.66**	.76**
Common verbs	11.66	6.00	9.70	2.26	24.48	6.18	2.53	5.95	0.89	12.80	.80**	.75**
Auxiliary verbs	7.69	3.50	7.14	1.69	16.06	5.75	1.62	5.96	2.11	9.63	.63**	.64**
Past tense	3.21	1.52	3.00	0.54	7.05	0.15	0.20	0.09	0.00	1.41	.39**	n.s.
Present tense	6.96	4.92	5.95	0.00	17.62	0.72	0.60	0.68	0.00	2.76	.72**	.30*
Future tense	0.73	0.62	0.62	0.00	2.04	0.38	0.40	0.31	0.00	1.48	.73**	.77**
Adverbs	3.23	1.69	3.31	0.00	6.97	2.30	1.28	2.15	0.00	5.12	.81**	.87**
Prepositions	13.76	2.94	13.57	7.33	21.83	9.22	2.84	9.09	3.69	16.17	.74**	.86**
Conjunctions	5.13	1.60	5.26	1.54	8.66	8.56	1.82	8.75	2.40	13.17	.25**	.49**
Negations	1.36	1.22	0.90	0.00	5.17	1.86	1.69	1.29	0.00	6.65	.91**	.95**
Quantifiers	2.72	1.26	2.47	0.00	6.59	2.20	0.99	2.29	0.00	5.26	.34**	.46**
Numbers	2.07	1.64	1.63	0.00	11.19	1.17	0.91	0.96	0.00	6.82	.49**	.67**
Informal/Swear words	0.22	0.53	0.00	0.00	3.23	0.15	0.35	0.00	0.00	2.29	.86**	.86**

Table 2. *Equivalence of LIWC2007 and LIWCser(continued)*

	LIWC2007					LIWCser						
	M	SD	Mdn	Min	Max	M	SD	Mdn	Min	Max	r	ICC
Social processes	9.20	4.94	8.18	1.00	21.06	4.49	1.89	4.46	0.00	12.20	.56**	.50**
Family	0.36	0.50	0.11	0.00	2.43	0.39	0.53	0.14	0.00	2.82	.92**	.96**
Friends	0.15	0.22	0.06	0.00	1.49	0.11	0.16	0.00	0.00	0.88	.49**	.64**
Humans	1.17	1.07	1.00	0.00	6.71	0.76	0.82	0.61	0.00	5.51	.57**	.71**
Affective processes	4.96	2.08	5.13	0.00	11.76	5.43	2.04	5.48	0.00	11.97	.76**	.85**
Positive emotions	3.48	1.73	3.65	0.00	9.66	3.37	1.44	3.39	0.00	8.66	.69**	.81**
Negative emotions	1.46	1.26	1.14	0.00	6.62	1.41	1.43	1.16	0.00	9.15	.72**	.88**
Fear and anxiety	0.24	0.54	0.11	0.00	4.12	0.23	0.58	0.10	0.00	5.97	.53**	.84**
Anger and resentment	0.57	0.84	0.34	0.00	6.62	0.14	0.24	0.00	0.00	1.50	.59**	.29*
Sadness	0.30	0.63	0.15	0.00	5.39	0.16	0.70	0.00	0.00	6.34	.42**	.92**
Cognitive processes	15.08	3.26	14.91	5.25	27.12	14.63	4.23	14.93	5.18	27.09	.53**	.73**
Insight	2.83	1.75	2.38	0.28	8.21	2.22	1.75	1.67	0.00	9.09	.71**	.82**
Causation	2.03	1.44	1.63	0.00	10.46	1.80	1.02	1.72	0.00	6.90	.52**	.79**
Discrepancy	1.06	0.79	1.03	0.00	3.35	1.55	0.86	1.52	0.00	4.29	.52**	.62**
Tentative	1.95	1.05	1.87	0.00	6.78	1.85	1.22	1.70	0.00	9.85	.55**	.57**
Certainty	1.16	0.73	1.18	0.00	4.15	1.87	0.93	1.84	0.00	6.32	.40**	.47**
Inhibition	0.52	0.56	0.41	0.00	4.37	2.00	1.27	1.67	0.00	6.02	.28**	.42**
Inclusive	4.79	1.40	4.61	1.55	9.60	0.83	0.54	0.82	0.00	2.99	n.s.	n.s.
Exclusive	1.48	0.93	1.39	0.00	3.95	1.47	0.70	1.59	0.00	3.31	.68**	.80**
Perceptual processes	1.63	1.15	1.59	0.00	5.66	2.03	1.73	1.64	0.00	11.05	.58**	.73**
See	0.61	0.74	0.51	0.00	4.52	0.88	1.40	0.53	0.00	10.50	.46**	.74**
Hear	0.48	0.66	0.37	0.00	5.33	0.64	0.86	0.48	0.00	5.73	.75**	.86**
Feel	0.31	0.32	0.26	0.00	1.64	0.13	0.18	0.07	0.00	1.05	.45**	.24*
Biological processes	1.48	1.30	1.22	0.00	7.19	1.40	1.29	1.22	0.00	7.58	.70**	.78**
Body	0.43	0.47	0.34	0.00	2.26	0.49	0.63	0.29	0.00	3.87	.61**	.75**
Health	0.54	0.77	0.36	0.00	4.92	0.24	0.71	0.00	0.00	5.22	.41**	.62**
Sex and love	0.28	0.68	0.06	0.00	6.27	0.24	0.63	0.00	0.00	6.64	.83**	.93**

Table 2. *Equivalence of LIWC2007 and LIWCser (continued)*

	LIWC2007					LIWCser						
	M	SD	Mdn	Min	Max	M	SD	Mdn	Min	Max	r	ICC
Ingestion	0.28	0.72	0.06	0.00	6.77	0.19	0.36	0.00	0.00	2.78	.80**	.80**
Relativity	12.46	3.36	12.39	3.45	24.73	11.68	3.10	11.25	5.91	20.58	.40**	.58**
Motion	1.64	1.13	1.44	0.00	6.07	1.03	0.73	0.94	0.00	3.50	.49**	.67**
Space	6.49	2.17	6.10	1.83	17.58	3.89	1.78	3.44	0.00	10.17	.52**	.71**
Time	4.10	1.83	4.14	0.00	9.77	2.43	1.35	2.40	0.00	6.49	.83**	.88**
Work	3.03	2.95	2.30	0.00	21.16	1.81	2.21	1.20	0.00	16.07	.71**	.89**
Achievement	2.63	2.61	2.07	0.00	24.87	1.92	2.16	1.40	0.00	20.54	.70**	.91**
Leisure	1.83	1.83	1.10	0.00	7.68	1.11	1.35	0.55	0.00	6.83	.80**	.90**
Home	0.30	0.43	0.18	0.00	2.99	0.13	0.19	0.05	0.00	0.76	.68**	.44**
Money	0.62	0.65	0.45	0.00	3.75	0.27	0.35	0.15	0.00	1.97	.43**	.41**
Religion	0.44	1.05	0.17	0.00	6.93	0.39	1.02	0.14	0.00	7.45	.79**	.97**
Death	0.15	0.25	0.00	0.00	1.83	0.14	0.26	0.00	0.00	1.79	.85**	.71**
Assent	0.38	0.63	0.00	0.00	2.40	0.09	0.12	0.00	0.00	0.56	.46**	n.s.
Nonfluencies	0.20	0.26	0.11	0.00	1.61	0.00	0.02	0.00	0.00	0.02	.31**	n.s.
Fillers	0.15	0.19	0.11	0.00	1.01	0.03	0.09	0.00	0.00	0.09	.39**	.58**

Note: ** p<.01, * p<.05, n.s. p>.05; Mean (M), Standard deviation (SD), Median (Mdn), Minimum (Min), Maximum (Max), Spearman's coefficient of correlation (r), Interclass coefficient of correlation (ICC)

Analysis of LIWC2007-LIWCser equivalence across various text types revealed that the high level of equivalence exists across all three types of texts, and the types of text influenced LIWC2007-LIWCser equivalence to some extent (Appendix 1). Thus, for scientific articles equivalence is .69 on average, for movie subtitles .71, and for newspaper articles .75.

Discussion

When we compare formal characteristics of the texts in English and in Serbian, differences in total word count and average number of words per sentence are noticeable. This is a result of grammar differences in the languages. For example, English has articles that do not exist in Serbian. In addition, there is a difference between proportions of function words in the text between Serbian and English. This is the consequence of two factor. First, some function words in Serbian are homonyms (e.g., "da" is a conjunction ("to") and assertive word ("yes")), and those words were not included in the dictionary.[7] Second, having in mind that Serbian is highly inflective language considerable differences in syntax structure exist between Serbian and English. For example, verbs in Serbian have suffices marking person in all verb forms. Consequently, in sentence construction it is not necessary to use pronouns, while in English, use of pronouns is obligatory. It leads to the smaller number of function words in Serbian[8].

Average equivalence between the LIWCser and LIWC2007 is satisfying compared to same measures between English and some other LIWC dictionaries. For example, German version on the standardized sample of the texts demonstrated almost the same level of equivalence with English dictionary as Serbian dictionary (average ICC=.70, and average correlation .68) (Wolf et al., 2008). Demonstrated level of equivalence between LIWCser and LIWC2007 can be considered very good having in mind the differences in the dictionaries itself (i.e., languages are different and there are differences in the classification of the words into different categories), and differences in the quality of the translation of various forms of texts. The results of the equivalence analyses of different types of texts testify about the difference in the quality of the translation. Namely, highest level of equivalence was in newspapers articles translated by professional translators and the lowest was for abstracts of scientific papers where authors were more preoccupied with presenting basic data about the research than with the stylistic and formal characteristics of the translation.

Linguistic categories in LIWCser were classified according to grammar rules. Therefore, the differences in linguistic categories between Serbian and English LIWC versions will reflect the difference in grammar rules of the languages. For example, compared to LIWC2007, LIWCser contains relatively

7 Impact of homonymous words exclusion on the comprehensiveness of the LIWCser is discussed in further section.

8 For example, in Serbian both sentences ***Ja idem kući*** *kolima* (***I go home*** *by car)* and ***Idem kući*** *kolima* (***Go home*** *by car)* are gramatically correct, where construction with the pronoun is less often used, since pronoun *I* is gramatically redundant in this example.

small number of word stems representing auxiliary verbs (144 compared to 28, respectively). In addition, Serbian dictionary contains lower number of prepositions (60 compared to 49, respectively), but larger number of adverbs (69 compared to 154, respectively). Number of word stems in other linguistic categories is relatively equal in LIWCser and LIWC2007.

On the level of specific categories, *Present* and *Past* have lower level of equivalence. This is probably due to differences in content of these categories in LIWCser and LIWC2007 (Bjekić et al., 2012). In addition, results for the category *Inclusion* do not indicate equivalence of the two dictionaries. Possible reason for this is that authors of LIWC2007 did not provide an explicit criterion for classification of words into categories. Therefore, it is possible that in LIWCser construction we have used different criteria than LIWC2007 constructors when selecting words for this category. Similar issue was noticeable in some other LIWC adaptations (e.g., Ramirez-Esparza et.al, 2007).

When it comes to paralinguistic categories, lower equivalence is a result of small sample of words belonging to this category in the text (which is expected since we did not analyse spontaneous speech) and of differences in transcribing. On the other hand, categories filled-in with culturally specific words belonging to categories *Religion*, *Family,* and *Leisure*, demonstrated high equivalence, which speaks in favour of the decision to add those words during the process of dictionary development.

Findings showed that LIWCser has satisfying equivalence with LIWC2007 dictionary, with the exception of few categories.

Comprehensiveness and Representation of the LIWCser Categories

If we want to have reliable results in the automatic text analysis, it is necessary to include in the dictionary words that are representative for specific category. However, representativeness of the categories is not possible to assess directly (Pennebaker et al. 2007). Usually, measure of comprehensiveness of the dictionary, i.e., percentage of the text covered with a dictionary[9], serves as an indicator of software's "goodness of fit". If the percentage of the words not covered by the dictionary is relatively small, the analysis is more comprehensive and therefore results are considered as more reliable.

So far, research demonstrated that comprehensiveness of the English LIWC dictionary is about 82% (Pennebaker et al, 2007), while the same measure ranges between 50% and 70% for other LIWC dictionaries (Alparone et al., 2004; Hayeri et al., 2010; Huang et al., in press; Kailer & Chung, 2011; Lee et al., 2005; Murderrisoglu, 2011; Piolat et al., 2011; Ramirez-Esparza et al., 2007; Wolf et al., 2008; Zijlstra et al., 2004). As part of evaluation of the LIWCser, we have assessed comprehensiveness of Serbian LIWC dictionary on different types of texts, i.e., the stability of these parameters across texts.

9 This is automatically generated measure which can be read form *Dictionary words* variable in the LIWC output.

When applying automatic text analysis in psychology, researchers often have problems with the interpretation of the results obtained for different categories. Namely, relative representation of each category partly results from the type of the text that is analysed and from its style. In order to have insight into expected values of different categories, we have investigated differences in representation of different categories depending on the type of the analysed text.

Method

A sample of 386 texts was used, out of which 141 was used for the assessment of the equivalence of LIWCser and LIWC2007 (i.e., scientific abstracts, newspapers articles and movie subtitles). Of the remaining 245 texts, 140 were short stories[10] written by psychology students as part of research conducted by Lazarević (2012) and 105 were short essays where respondents were reporting about their attitude towards homosexuals (Bjekić, Živanović, & Žeželj, 2012). To sum up, five different types of texts were analysed: abstracts of scientific papers, newspaper articles, movie subtitles, short stories, and essays. Each text from the corpus of short stories and short essays was processed with LIWCser.

Results

LIWCser dictionary covers on average 69.93% of words in the texts. As seen from the Table 3, representation of the categories differs depending on the type of the text that is analysed. Comprehensiveness of LIWCser dictionary is highest for essays and short stories, while it is lowest for abstracts of the scientific papers.

Table 3.*Comprehensiveness of LIWCser dictionary for different types of texts*

	Short stories	Essays	Newspapers articles	Scientific papers abstracts	Movie subtitles
M	73.09	74.36	63.81	61.38	68.16
SD	5.33	4.10	4.49	5.44	3.81
Mdn	73.58	74.52	64.87	61.10	68.82
Min	54.46	61.84	54.09	53.60	56.57
Max	85.56	82.98	70.95	73.30	73.04

Note: Mean (M), Standard deviation (SD), Median (Mdn), Minimum (Min), Maximum (Max)

Depending on the type of the text, differences in the average representation of different LIWCser categories occur. Largest differences occur in linguistic categories, which are the best indicator of the writing style, and in psychological categories. For example, frequency of first person pronouns is higher in essays about specific topic than in other types of texts (i.e., in abstracts of scientific papers words from these categories are almost absent).When it comes to psychological categories, slightly higher values are obtained for essays, short stories and movie subtitles, compared to abstracts of scientific papers and newspapers articles. Table 4 presents the representation of LIWCser categories for different types of texts.

10 Students were supposed to write short story that would include five specific words.

Table 4. *Descriptives on LIWCser categories for different kinds of texts and results of Kruskal-Wallis test*

	Scientific papers abstracts		Newspapers articles		Movie subtitles		Short stories		Essays		χ^2
	M	SD	M	SD	M	SD	M	SD	M	SD	
Word count	207.04	76.72	903.17	323.37	2212.34	1260.17	174.16	40.83	190.28	51.23	184.89**
Words per sentence	26.98	7.15	19.15	4.22	5.52	1.29	16.28	5.44	20.20	7.24	154.04**
Words>6 letters	48.06	5.94	32.30	5.62	17.49	3.06	22.33	6.19	24.33	6.24	174.67**
Total function words	26.77	3.34	33.15	3.07	34.97	1.98	38.66	4.92	39.53	4.34	133.95**
Total pronouns	4.83	2.11	8.46	2.44	13.32	1.64	11.70	2.71	13.31	3.05	144.20**
Personal pronouns	0.46	0.49	2.23	1.13	6.58	0.93	3.29	2.05	3.51	1.84	145.43**
1st pers singular	0.08	0.41	0.92	0.83	3.07	0.51	0.97	1.51	2.36	1.78	112.34**
1rd pers plural	0.06	0.15	0.59	0.45	1.31	0.41	0.68	0.90	0.81	0.84	92.72**
2nd person	0.04	0.12	0.34	0.31	2.58	0.56	0.15	0.34	0.24	0.43	153.75**
3rd pers singular	2.59	1.15	4.39	1.02	5.43	1.13	7.89	4.54	4.16	1.53	92.12**
3rd pers plural	0.43	0.65	0.50	0.33	0.68	0.44	0.99	1.02	1.48	1.29	19.73**
Impersonal pronouns	2.72	1.42	4.26	1.26	4.52	0.91	5.56	2.01	7.00	2.16	74.09**
Common verbs	3.97	1.51	5.73	1.34	9.27	1.37	8.87	3.49	6.58	2.03	111.13**
Auxiliary verbs	4.53	1.42	5.79	1.42	7.05	0.91	9.11	3.20	5.84	1.93	102.33**
Past tense	0.12	0.27	0.17	0.17	0.17	0.15	0.27	0.38	0.06	0.17	11.88**
Present tense	0.22	0.49	0.61	031	1.42	0.30	1.18	1.00	1.16	1.05	86.39**
Future tense	0.17	0.37	0.28	0.23	0.74	0.37	0.86	1.13	0.35	0.46	51.44**
Adverbs	1.15	0.84	2.23	0.82	3.69	0.72	3.77	1.83	3.52	1.80	113.18**
Prepositions	11.37	2.46	9.85	1.71	5.99	1.04	7.62	2.40	7.28	2.08	111.06**
Conjunctions	7.42	2.06	9.07	1.54	9.18	1.18	10.84	2.80	12.57	2.47	63.63**
Negations	0.54	0.66	1.25	0.79	4.14	0.93	2.77	1.79	4.97	2.32	27.86**

Table 4. *Descriptives on LIWCser categories for different kinds of texts and results of Kruskal-Wallis test (continued)*

	Scientific papers abstracts		Newspapers articles		Movie subtitles		Short stories		Essays		χ^2
	M	SD	M	SD	M	SD	M	SD	M	SD	
Negative words	0.24	0.38	0.15	0.16	0.08	0.07	0.22	0.50	0.26	0.75	15.31**
Superlative	0.16	0.35	0.27	0.19	0.08	0.06	0.16	0.39	0.13	0.30	58.82**
Quantifiers	1.61	1.18	2.61	0.80	2.33	0.61	3.79	1.66	3.88	1.71	72.62**
Numbers	1.34	1.38	1.32	0.60	0.80	0.30	0.76	0.79	0.73	0.64	31.24**
Informal/Swear words	0.00	0.00	0.01	0.04	0.49	0.51	0.12	0.34	0.13	0.38	124.93**
Social processes	4.27	2.79	4.09	1.17	5.27	1.02	5.80	2.70	6.49	2.25	29.37**
Family	0.24	0.60	0.22	0.29	0.76	0.50	0.35	0.77	0.28	0.50	62.32**
Friends	0.09	0.21	0.08	0.01	0.18	0.11	0.29	0.65	0.19	0.35	32.91**
Humans	0.86	1.27	0.53	0.41	0.98	0.40	1.33	1.40	2.28	1.33	23.53**
Affective processes	4.54	2.85	5.51	1.35	6.31	1.14	6.11	2.17	5.60	2.56	22.76**
Positive emotion	2.77	1.89	3.94	1.05	3.31	0.95	3.20	1.88	3.12	1.69	21.56**
Negative emotion	1.49	2.28	0.96	0.64	1.90	0.53	2.29	1.57	2.13	1.61	52.00**
Fear and anxiety	0.32	0.99	0.14	0.16	0.25	0.18	0.44	0.58	0.32	0.53	22.77**
Rage and anger	0.07	0.29	0.10	0.15	0.28	0.20	0.30	0.51	0.23	0.42	48.54**
Sadness	0.31	1.20	0.09	0.15	0.09	0.09	0.21	0.41	0.09	0.23	19.18**
Cognitiveprocesses	17.79	3.85	11.13	3.14	15.70	1.90	16.29	4.38	21.69	4.21	70.08**
Insight	4.06	1.93	1.23	0.57	1.45	0.49	1.75	1.30	2.66	1.46	74.41**
Causation	2.47	1.36	1.36	0.63	1.65	0.45	1.78	1.19	2.19	1.31	23.79**
Discrepancy	1.22	1.06	1.36	0.63	2.17	0.49	2.17	1.29	2.60	1.52	44.87**
Tentative	2.26	1.88	1.41	0.72	1.96	0.37	2.62	1.47	3.39	1.86	35.04**
Certainty	1.76	1.18	1.47	0.65	2.51	0.47	2.29	1.58	3.06	1.53	31.68**
Inhibition	1.81	1.42	1.19	0.45	3.27	0.75	1.63	1.03	3.24	1.46	75.54**

Table 4. *Descriptives on LIWCser categories for different kinds of texts and results of Kruskal-Wallis test (continued)*

	Scientific papers abstracts		Newspapers articles		Movie subtitles		Short stories		Essays		χ^2
	M	SD	M	SD	M	SD	M	SD	M	SD	
Inclusive	0.63	0.65	1.00	0.56	0.81	0.23	1.41	1.03	1.35	1.08	30.46**
Exclusive	1.08	0.78	1.52	0.68	1.85	0.34	1.96	1.53	3.02	1.27	22.11**
Perceptual processes	1.94	2.48	2.26	1.53	1.82	0.59	2.31	1.81	1.27	0.91	7.95*
See	1.07	2.25	1.02	0.80	0.49	0.28	0.98	1.10	0.46	0.56	15.11**
Hear	0.34	0.96	0.77	0.99	0.83	0.33	0.73	1.12	0.33	0.43	43.02**
Feel	0.07	0.21	0.16	0.17	0.17	0.13	0.13	0.30	0.09	0.24	52.50**
Biological processes	1.41	1.96	1.11	0.76	1.76	0.73	2.19	1.33	4.39	1.72	41.68**
Body	0.61	0.95	0.25	0.24	0.66	0.43	0.20	0.42	0.25	0.50	57.25**
Health	0.38	1.18	0.08	0.16	0.29	0.29	0.11	0.31	0.34	0.50	71.97**
Sex and love	0.16	0.98	0.16	0.24	0.44	0.41	1.70	1.25	3.76	1.61	177.80**
Ingestion	0.03	0.14	0.23	0.51	0.31	0.21	0.10	0.31	0.01	0.09	91.70**
Relativity	12.14	3.39	13.22	2.52	9.13	1.46	11.89	3.28	7.08	2.69	48.95**
Motion	0.67	0.77	0.94	0.45	1.57	0.66	1.26	1.04	0.74	0.71	35.74**
Space	4.10	2.06	4.47	1.83	2.89	0.60	3.20	1.61	2.30	1.31	29.53**
Time	1.53	1.28	3.32	1.20	2.27	0.74	3.75	1.46	1.49	1.10	78.87**
Work	3.08	3.39	1.38	0.81	0.97	0.61	0.42	0.64	0.31	0.46	84.87**
Achievement	3.02	3.18	1.92	1.13	0.68	0.47	0.82	0.84	0.45	0.52	84.65**
Leisure	0.14	0.48	2.36	1.38	0.57	0.41	0.87	1.18	0.09	0.25	96.31**
Home	0.00	0.00	0.14	0.18	0.27	0.19	0.22	0.64	0.06	0.23	75.95**
Money	0.15	0.37	0.27	0.33	0.40	0.30	0.08	0.27	0.06	0.21	94.32**
Religion	0.01	0.04	0.72	1.57	0.38	0.26	0.08	0.31	0.16	0.38	130.19**
Death	0.04	0.26	0.11	0.21	0.30	0.25	0.03	0.16	0.01	0.08	126.38**

Note: ** p<.01, * p<.05, n.s. p>.05; Mean (M), Standard deviation (SD),Kruskal Wallis test (χ^2)

Discussion

Results on comprehensiveness of LIWCser dictionary demonstrate that it is possible to extract reliable information about the texts that are analysed. When percent of words covered by LIWCser dictionary is compared to other LIWC dictionaries, we observe that Serbian dictionary covers on average larger percentage of the text than French (54%) (Piolat et al., 2011), German (63%) (Wolf et al., 2008), and Spanish (66%) (Ramirez-Esparza et al., 2007), and the same as Dutch (70%) (Zijlstra et al., 2004). Therefore, we can conclude that Serbian LIWC dictionary is quite successful when it comes to dictionary comprehensiveness, i.e., reliability of the information obtained.

The differences in percent of words covered by dictionary for the different types of texts are in line with the expectations. Namely, the lowest coverage is for scientific abstracts, while the highest is for the short stories and essays. It is quite understandable since scientific abstracts mostly consist of specific terms, and LIWC does not contain professional terminology because its primary purpose is the analysis of everyday verbal output (Pennebaker et al., 2007). Style of short stories and student essays is relatively informal and closest to the everyday speech.

The coverage of different types of texts in Serbian is similar to the results obtained in English. The results of the validation of LIWC2007 demonstrate lowest percentage of coverage for scientific abstracts (53%), and highest for oral speech (91%), and emotional writing (93%) (Pennebaker et al., 2007). These results suggest that LIWC software is largely adapted for the analysis of everyday oral and written language, both in English and in Serbian.

Displayed results about representation of each LIWCser category in different types of texts provide insight into how values for different categories vary across different text types. These values are descriptive. One should bear in mind that they are not obtained on representative sample of specific type of the text, and that they serve more as a general tendency than as a norm. In other words, it is advisable to use this information as a general guideline about the basic characteristics of different types of verbal outputs when interpreting results. For example, scientific abstracts usually have longer sentences, lower proportion of function words, relatively rare use of pronouns and negations and lack of informal words. These tendencies are in accordance with linguistic characteristics of scientific style, e.g., monolog character, use of normed speech, and higher saturation of the text with a meaning (Simić, 2001). Characteristics of newspaper articles are middle long sentences; use of less affective words and words marking cognitive processes in comparison to other kinds of texts, which indicate objectivity and restraint in expression, which are standard characteristics of these kinds of writings (Katanić-Bakaršić, 1999). Movie subtitles were included because they are highly representative of everyday speech. Therefore, they usually have short sentences, frequent use of personal pronouns, content refers to present tense, have informal words, etc. Short stories and essays represent written form of everyday speech. Characteristics of these kinds of texts

are higher frequency of function words, and more frequent use of pronouns and verbs (i.e., sentences with basic structure)[11].

Differences in percentages of LIWC categories depending on the type of the text stress the importance of both the context in which verbal communication takes place, and of validity of content of specific categories (Pennebaker et al., 2007). In other words, it is expected that texts written with different aims and in different contexts diverge in style and content. If a software for ATA assesses those differences and if they are interpretable (i.e., if the results provide information in line with general characteristics of specific type of the text), we can consider specific software as a valid instrument.

Impact of Homonymous Words Exclusion on the Comprehensiveness of the LIWCser

Unlike authors of LIWC, during dictionary construction we have decided to exclude all the words that could be classified into different categories depending on the context (i.e., homographs and homophones). Although this decision resulted in lower number of words in the dictionary and led to lower percentage of the text coverage in the analyses, we have avoided misclassification of the words as much as possible and consequently lowered the measurement error. In order to have an idea about the percentage of the words that were left out from the analyses due to exclusion of the homonyms, the percentage of the excluded homonymous words across texts has been calculated.

Method

Additional dictionary for homonyms was constructed and it included 323 word stems that were initially excluded from LIWCser due to homonymy. In this dictionary, 8.7% were function words. The same sample of 386 texts was processed again.

Results

Analyses demonstrated that average 5.27% (SD=2.30) of all words in texts were homonyms. The highest percentage of homonymous words was found in the essays (M=6.67, SD=2.08), movie titles (M=5.97, SD=1.53) and short stories (M=5.63, SD=2.10), while lower percentage was found in newspapers articles (M=3.30, SD=1.35) and scientific paper abstracts (M=2.94, SD=1.42).

Discussion

Results show relatively low number of homonymous words in analysed texts. If the homonyms were included in LIWCser dictionary, its comprehensiveness would be on average 75%, instead of 70% as demonstrated in previous analyses using LIWCser dictionary. In other words, exclusion of this

11 Sentences with basic structure consist of a minimum number of words that can convey certain meaning. Basic structure of the sentence usually consists of three constituents in canonical word order.

type of words did not significantly reduce the quality of LIWCser dictionary in terms of its comprehensiveness (from 75% with homonyms included to 70% without homonyms). In addition, it empirically supports primary decision to exclude homonymous words in order to avoid the possibility of misclassification of such words during the text analysis. However, since authors of other LIWC adaptations did not report results on homonyms analyses, question remains whether these results can be cross-linguistically generalised.

General Discussion and Conclusion

Use of automatic text analysis, and specifically LIWC software recently became more frequent in psychological research in English and in non-English speaking countries (see Pennebaker et al., 2003; Tausczik & Pennebaker, 2010). This kind of text analysis has several advantages. ATA enables researchers to have objective quantitative data on large number of different content and stylistic characteristics of the text, and application of various statistical analyses. In addition, analysis is simple, reliable, low-cost, and sample is relatively easy to assemble (i.e., we can use internet, e-mails, literature, speeches, etc.).

All analyses demonstrated a satisfying level of equivalence between Serbian and English version of the dictionary, which enables cross-language evaluation. Empirical evidence from this study validates LIWCser as a method strong enough to analyse texts in Serbian with the same quality as LIWC2007 processes English verbal products. Although some categories did not have high level of equivalence, results revealed that overall LIWCser shows similar level of equivalence as other translations of the dictionary.

High percentage of coverage of the text, and stability in the percentage of coverage depending on the type of the text, provides more evidence on validity of LIWCser as assessment method in psychology research. Overall, LIWCser performs similar to LIWC2007. Specifically, results demonstrated that LIWCser performs better when processing texts with more informal style, compared to more formal texts. This adds to the validity of the LIWCser as an instrument designed to analyse texts saturated with psychologically relevant content.

Final study related to homonymous analysis demonstrated that the decision to exclude relatively small percentage of words so possible wrong classification could be avoided, proved to be good. On average, only 4–5% of the words that were not initially classified with LIWCSser belong to the group of homonyms. This result supports the decision to add on reliability of the classification by excluding potentially misclassified words.

To conclude, several arguments go in favour of LIWCser as a good instrument for the analysis of the texts in Serbian. First, since the basis for the development of LIWCser was English dictionary, researchers have clear theoretical and methodological framework. Second, all analyses indicate good psychometric properties of the instrument. In addition, LIWCser is very user-friendly and it offers possibility to create new categories depending on the need of the researcher. Finally, during development of LIWCser, significant attention

was paid to cultural and linguistic specificities of Serbian language. It would be a useful tool for all professionals interested in studying various aspects of linguistic behaviour, especially spontaneously produced verbal material.

References

Alparone, F., Caso, S., Agosti, A., &Rellini, A. (2004).*The Italian LIWC2001 Dictionary*. Austin, TX: LIWC.net

Bradac, J. J. (1986). Threats to generalization in the use of elicited, purloined, and contrived messages in human communication research.*Communication quarterly, 34*, 55–65.

Bernard, M., Jackson, C., & Jones, C. (2006).Written emotional disclosure following first-episode psychosis: Effects on symptoms of post-traumatic stress disorder. *British journal of clinical psychology, 45*, 403–415.

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.

Bjekić, J., Živanović, M., i Žeželj, I. (2012). Lingvistički korelati implicitnih stavova prema homoseksualnosti. 60.naučno stručni skup "Sabor psihologa Srbije*"*, Beograd,30.maj–02. jun 2012, str.. 184–185, Filozofski fakultet u Beogradu.

Carroll, D. W. (2007). Patterns of student writing in a critical thinking course: A quantitative analysis. *Assessing writing, 12*, 213–227. doi:10.1016/j.asw.2008.02.001

Chung, C. J., & Park, H. W. (2010). Textual analysis of a political message: The inaugural addresses of two Korean presidents. *Social science information, 49,* 215–239. doi: 10.1177/0539018409359370

Chung, C. K., & Pennebaker, J. W. (2007). The psychological function of function words. In K. Fiedler (Eds), *Social communication: Frontiers of social psychology* (pp. 343–359). New York: Psychology Press.

Cohen, A., Alpert, M., Nienow, T., Dinzeo, T., & Docherty, N. (2008). Computerized measurement of negative symptoms in schizophrenia. *Journal of psychiatric research, 42*(10), 827–836.

Cohen, A. S., St-Hilaire, A., Aakre, J. M., & Docherty, N. M. (2009). Understanding anhedonia in schizophrenia through lexical analysis of natural speech. *Cognition and emotion, 23*, 569–586. doi:10.1080/02699930802044651

Djikic, M., Oatley, K., & Peterson, J. B. (2006). The bitter-sweet labour of emoting: The linguistic comparison of writers and physicists. *Creativity research journal, 18*, 191–197. doi:10.1207/s15326934crj1802_5

Frojd, S. (1969). *Uvod u psihoanalizu*. Matica srpska: Novi Sad.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology, 96*, 1029–1046. doi:10.1037/a0015141

Gortner, E. M., Rude, S. S., & Pennebaker, J. W. (2006). Benefits of expressive writing in lowering rumination and depressive symptoms. *Behavior therapy, 37*, 292–303. doi:10.1016/j.beth.2006.01.004

Gottschalk, L. A., & Gleser, G. C. (1969). *The measurement of psychological states through the content analysis of verbal behavior*. Berkeley: University of California Press.

Gottschalk, L. A., Gleser, G. C., Daniels, R., & Block, S. (1958). The speech patterns of schizophrenic patients: a method of assessing relative degree of personal disorganization and social alienation. *Journal of nervous and mental disease*, *127*, 153–166.

Hart, R. P. (1984). *Verbal Style and the Presidency: A Computer-Based Analysis.* NewYork: Academic.

Hart, R. P. (2001). Redeveloping DICTION: Theoretical Considerations. In West, M. D. (Ed.), *Theory, method and practice in computer content analysis*. (pp. 43–60). New York: Ablex.

Hayeri, N., Chung, C. K., & Pennebaker, J. W. (2010). *The development of Linguistic Inquiry and Word Count (LIWC) for Arabic texts*. Austin, TX: LIWC.net

Hirsh, J., & Peterson, J. (2009). Personality and language use in self-narratives. *Journal of research in personality, 43*, 524–527. doi:10.1016/j.jrp.2009.01.006

Holtgraves, T. (2011). Text messaging, personality and social context. *Journal of research in personality, 45,* 92–99.

Huang, C. L., Chung, C. K., Hui, N., Lin, Y. C., Seih, Y. T., Chen, W. C., Lam, B., Bond. M., & Pennebaker, J. W. (in press). The development of the Chinese Linguistic Inquiry and Word Count dictionary. *Chinese journal of psychology.*

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological science, 22*, 39–44. doi:10.1177/0956797610392928

Kailer, A., & Chung, C.K. (2011). *The Russian LIWC2007 dictionary*. Austin, TX: LIWC.net

Katanić-Bakaršić, M. (1999). *Lingvistička stilistika*. Budapest: Open Society Institute

Kim, Y. (2008). Effects of expressive writing among bilinguals: Exploring psychological well-being and social behaviour. *British journal of health psychology, 13*(1), 43–47.

Klajn, I. (2005). *Gramatika srpskog jezika*. Zavod za udžbenike i nastavna sredstva, Beograd.

Kroner-Herwig, B., Linkemann, A., & Morris, L. (2004). Selbst öffnung beim Schreiben über belastende Lebensereignisse: Ein Weg in die Gesundheit? *Zeitschrift für klinische psychologie und psychotherapie, 33*, 183–190. doi:10.1026/1616–3443.33.3.183

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse processes*, *25*, 259–284.

Lazarević, Lj. (2012). *Relations between implicit and explicit measures of personality – Prospects of Implicit Association Test (IAT) in assessment of basic personality traits.* (Unpublished doctoral dissertation). Faculty of philosophy, Belgrade.

Lee, C. H, Kim, K., Seo, Y. S., & Chung, C. (2007). The relations between personality and language use. *The journal of general psychology, 134*(4), 405–413.

Lee, C. H., Park, J., & Seo, Y. S. (2006). An analysis of linguistic styles by inferred age in TV dramas. *Psychological reports, 99*, 351–356. doi:10.2466/pr0.99.2.351–356

Lee, C. H., Shim, J., & Yoon, A. (2005). The review about the development of Korean linguistic inquiry and word count. *Korean journal of cognitive science, 16*(4), 93–121.

Lee, Y. (2009). Measures of student attitude on aging. *Educational Gerontology, 35*, 121–134.

Lepore, S. J. (1997). Expressive writing moderates the relation between intrusive thoughts and depressive symptoms. *Journal of personality and social psychology, 73*, 1030–1037.

Lowe, W. (2003).Software for content analysis – A Review. Retrieved from 17 September 2011 from http://kb.ucla.edu/system/datas/5/original/content_analysis.pdf

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*(1), 457–500.

Mehl, M. R. (2006). Quantitative text analysis. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp.141–156). Washington, DC: American Psychological Association.

Mergenthaler, E. (1996). Emotion-abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of consultation and clinical psychology, 64*, 1306–15. doi:10.1037/0022–006X.64.6.1306.

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestation and implicit folk theories of personality in daily life. *Journal of personality and social psychology, 90,* 862–877. doi: 10.1037/0022–3514.90.5.862

Mehl, M. R., & Gill, A. J. (2010). *Automatic text analysis*. In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods in conducting online behavioural research* (pp. 109–127). Washington, DC: American Psychological Association.

Murderrisoglu, S. (2011).*The Turkish LIWC2007 dictionary.* Austin, TX: LIWC.net

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin, 29*, 665–675. doi:10.1177/0146167203029005010

Pennebaker, J. W. & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology, 77*(6)*,* 1296–1312.

Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., & Booth, R.J. (2007). *The development and psychometric properties of LIWC2007*. Austin, TX: LIWC.net

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*, 547–577. doi:10.1146/annurev.psych.54.101601.145041

Piolat, A., Booth, R. J., Chung, C. K., Davids, M., & Pennebaker, J. W. (2011). La version française du LIWC: modalités de construction etexemplesd' application*. Psychologie française, 56*, 145–159. doi:10.1016/j.psfr.2011.07.002

Ramírez-Esparza, N., Chung, K. C., Sierra-Otero, G., & Pennebaker J. W. (2009). El lenguaje de la depresión: Categorías linguísticas y temasusados en foros de discusión en Internet en inglés y en español. *Revista de la asociación de psicoterapia de Argentina.* Retrieved from*http://www.revistadeapra.org.ar/ant_julio09.htm*

Ramírez-Esparza, N., Gosling, S. D., Benet-Martínez, V., Potter, J. P., & Pennebaker, J. W. (2006). Do bilinguals have two personalities? A special case of cultural frame switching. *Journal of research in personality, 40*, 99–120. doi:10.1016/j.jrp.2004.09.001

Ramirez-Esparza, N., Pennebaker, J.W., Garcia, F.A., & Suria, R. (2007). La psicología del uso de las palabras: Un programa de computador a que analizatextos en Español. *Revista mexicana de psicología, 24*(1)*,* 85–99.

Shapiro, G., & Markoff, J. (1997).A matter of definition. In C. W. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. Mahwah, NJ: Lawrence Erlbaum Associates.

Simić, P. (2001). *Opšta stilistika*. Jasen

Simmons, R. A., Gordon, P. C., & Chambless, D. L. (2005). Pronouns in marital interaction. *Psychological science, 16*, 932–936.

Stone P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A computer approach to content analysis.* Cambridge, MA: MIT Press.

Tausczik, Y. R., & Pennebaker J. W. (2010).The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology, 29*, 25–54. doi:10.1177/0261927X09351676.

Watson, D., & Pennebaker, J. W. (1989). Health complaints, stress, and distress: exploring the central role of negative affectivity*. Psychological review, 96*, 234–254.

Wolf, M., Horn, A., Mehl, M., Haug, S., Pennebaker, J. W., & Kordy, H. (2008). Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count*. Diagnostica, 2*, 85–98. doi:10.1026/ 0012–1924.54.2.85

Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch,W., & Landauer, T. K. (1998). Learning from text: Matching readers and text by Latent Semantic Analysis. *Discourse processes, 25*, 309–336. doi:10.1080/01638539809545030

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*, 363–373. doi:10.1016/j.jrp.2010.04.001Yogo, M., & Fujihara, S. (2008). Working memory capacity can be improved by expressive writing: A randomized experiment in a Japanese sample. *British journal of health psychology, 13*(1), 77–80.

Zijlstra, H., van Meerveld, T., van Middendorp, H., Pennebaker, J.W., & Geenen R. (2004). De Nederlandse versie van de Linguistic Inquiry and Word Count (LIWC), eenge computeriseerd tekst analyse programma.[Dutch version of the Linguistic Inquiry and Word Count (LIWC), a computerized text analysis program].*Gedrag & gezondheid, 32*, 273–283.

Appendix 1 Equivalence of LIWC2007 and LIWCser for different text formats

	Scientific papers abstracts				Newspapers articles				Movie subtitles			
	LIWC2007	LIWCser	r	ICC	LIWC2007	LIWCser	r	ICC	LIWC2007	LIWCser	r	ICC
Word count	193.09	207.04	.44**	.40**	1119.02	903.17	.98**	.98**	3090.46	2212.34	.83**	.71**
Words per sentence	27.78	26.98	.40**	.55**	22.59	19.15	.86**	.92**	7.24	5.52	n.s.	n.s.
Dictionary coverage	72.61	61.38	.57**	.71**	80.85	63.81	.88**	.92**	88.26	68.16	.45**	n.s.
Words>6 letters	37.00	48.06	n.s.	.46*	23.29	32.30	.87**	.89**	12.62	17.49	.62**	.79**
Total function words	42.52	26.77	.36*	.46*	52.88	33.15	.81**	.81**	57.83	34.97	.34*	n.s.
Total pronouns	4.31	4.83	.54**	.75**	10.00	8.46	.85**	.84**	21.82	13.32	.44**	.60**
Personal pronouns	1.07	0.46	.42**	.49*	5.52	2.23	.84**	.70**	15.25	6.58	.55**	.69**
1st pers singular	0.05	0.08	n.s.	n.s.	2.04	0.92	.93**	.76**	5.78	3.07	.50**	.63**
1rd pers plural	0.44	0.06	n.s.	n.s.	1.03	0.59	.73**	.67**	1.38	1.31	.43**	.59**
2nd person	0.00	0.04	-	-	0.88	0.34	.72**	.66**	5.46	2.58	.69**	.78**
3rd pers singular	0.04	2.59	n.s.	n.s.	0.86	4.39	.33*	.64**	1.86	5.43	.79**	.84**
3rd pers plural	0.55	0.43	.50**	.69**	0.71	0.50	.56**	.72**	0.77	0.68	.89**	.96**
Impersonal pronouns	3.24	2.72	.44**	.60**	4.48	4.26	.74**	.86**	6.57	4.52	.36*	.44*
Common verbs	6.35	3.97	n.s.	n.s.	9.87	5.73	.55**	.57**	19.97	9.27	.60**	.72**
Auxiliary verbs	4.60	4.53	n.s.	n.s.	6.85	5.79	.34*	.43*	12.24	7.05	.51**	.65**
Past tense	3.01	0.12	.30*	n.s.	3.13	0.17	.53**	n.s.	3.55	0.17	n.s.	n.s.
Present tense	2.68	0.22	n.s.	n.s.	5.54	0.61	n.s.	n.s.	13.63	1.42	n.s.	n.s.
Future tense	0.22	0.17	.54**	.58**	0.56	0.28	.49**	.62**	1.51	0.74	.44**	.60**
Common verbs	1.57	1.15	n.s.	n.s.	3.18	2.23	.62**	.75**	5.15	3.69	.51**	.64**
Prepositions	16.08	11.37	n.s.	.49*	14.52	9.85	.30*	.44*	10.17	5.99	.55**	.68**
Conjunctions	4.85	7.42	.28*	.53**	6.37	9.07	.36	.62	3.80	9.18	.58**	.73

Appendix 1 Equivalence of LIWC2007 and LIWCser for different text formats

	Scientific papers abstracts				Newspapers articles				Movie subtitles			
	LIWC2007	LIWCser	r	ICC	LIWC2007	LIWCser	r	ICC	LIWC2007	LIWCser	r	ICC
Negations	0.51	0.54	.57**	.75**	0.83	1.25	.89**	.90**	3.02	4.14	.72**	.84**
Quantifiers	2.90	1.61	.32*	n.s.	3.07	2.61	.49**	.70**	2.06	2.33	.54**	.75**
Numbers	2.96	1.34	.36*	.66**	2.05	1.32	.57**	.63**	1.08	0.80	.61**	.71**
Informal/Swear words	0.01	0.00	-	-	0.00	0.01	-	-	0.73	0.49	.85**	.80**
Social processes	5.65	4.27	.62**	.79**	7.35	4.09	.52**	.55**	15.62	5.27	.37*	.54**
Family	0.23	0.24	.93**	.96**	0.20	0.22	.86**	.96**	0.70	0.76	.89**	.95**
Friends	0.10	0.09	.41**	.73**	0.13	0.08	.38**	.43*	0.22	0.18	n.s.	n.s.
Humans	1.42	0.86	.51**	.69**	0.75	0.53	.61**	.77**	1.43	0.98	.37*	.48*
Affective processes	3.93	4.54	.69**	.83**	5.17	5.51	.69**	.81**	5.84	6.31	.63**	.76**
Positive emotions	2.40	2.77	.56**	.81**	4.24	3.94	.61**	.72**	3.70	3.31	.57**	.78**
Negative emotions	1.40	1.49	.42**	.88**	0.94	0.96	.62**	.86**	2.22	1.90	.74**	.76**
Fear and anxiety	0.34	0.32	.55**	.85**	0.16	0.14	.31*	.67**	0.24	0.25	n.s.	.70**
Anger and resentment	0.37	0.07	n.s.	n.s.	0.28	0.10	.35**	.53**	1.18	0.28	n.s.	n.s.
Sadness	0.43	0.31	.34*	.93**	0.20	0.09	.44**	.67**	0.27	0.09	.34*	.56**
Cognitive processes	16.37	17.79	.38*	.62**	14.20	11.13	.87**	.92**	14.78	15.70	.42**	.52**
Insight	4.49	4.06	.39**	.55**	1.89	1.23	.69**	.76**	2.20	1.45	.49**	.69**
Causation	3.16	2.47	.56**	.77**	1.65	1.36	.53**	.72**	1.25	1.65	n.s.	.44*
Discrepancy	0.65	1.22	n.s.	n.s.	0.84	1.36	.64**	.78**	1.83	2.17	n.s.	.45*
Tentative	1.89	2.26	.49**	.49*	1.82	1.41	.64**	.81**	2.19	1.96	.53**	.69**
Certainty	0.75	1.76	n.s.	n.s.	1.24	1.47	.61**	.78**	1.51	2.51	n.s.	.47*
Inhibition	0.53	1.81	n.s.	.49*	0.40	1.19	n.s.	n.s.	0.66	3.27	n.s.	n.s.
Inclusive	5.01	0.63	n.s.	n.s.	5.40	1.00	n.s.	n.s.	3.72	0.81	n.s.	n.s.
Exclusive	0.76	1.08	.46**	.68**	1.39	1.52	.81**	.88**	2.41	1.85	.36*	.46*

Appendix 1 Equivalence of LIWC2007 and LIWCser for different text formats

	Scientific papers abstracts				Newspapers articles				Movie subtitles			
	LIWC2007	LIWCser	r	ICC	LIWC2007	LIWCser	r	ICC	LIWC2007	LIWCser	r	ICC
Perceptual processes	0.89	1.94	.47**	.71**	1.71	2.26	.72**	.87**	2.35	1.82	.58**	.84**
See	0.34	1.07	.65**	.82**	0.60	1.02	.48**	.71**	0.94	0.49	.37*	.64**
Hear	0.09	0.34	.46**	.59**	0.59	0.77	.77**	.96**	0.78	0.83	.40**	.59**
Feel	0.10	0.07	n.s.	n.s.	0.35	0.16	.35**	n.s.	0.49	0.17	n.s.	n.s.
Biological processes	1.04	1.41	.56**	.78**	1.35	1.11	.73**	.79**	2.15	1.76	.58**	.77**
Body	0.23	0.61	.49**	.74**	0.27	0.25	.51**	.73**	0.87	0.66	.66**	.85**
Health	0.64	0.38	n.s.	.61**	0.49	0.08	.49**	.49**	0.49	0.29	.68**	.87**
Ingestion	0.18	0.16	.40**	.98**	0.14	0.16	.72**	.93**	0.57	0.44	.74**	.64**
Relativity	0.00	0.03	-	-	0.47	0.23	.71**	.81**	0.35	0.31	.71**	.74**
Motion	10.23	12.14	.30*	.55**	14.12	13.22	.69**	.82**	12.77	9.13	.68**	.74**
Space	0.97	0.67	n.s.	n.s.	1.47	0.94	.29*	.62**	2.61	1.57	.80**	.85**
Time	6.30	4.10	.33*	.52**	7.22	4.47	.60**	.88**	5.75	2.89	.58**	.67**
Work	2.43	1.53	.68**	.83**	5.29	3.32	.75**	.84**	4.40	2.27	.76**	.85**
Achievement	4.96	3.08	.70**	.90**	2.80	1.38	.58**	.66**	1.17	0.97	.80**	.89**
Leisure	3.31	3.02	.47**	.92**	3.30	1.92	.69**	.85**	1.00	0.68	.71**	.87**
Home	0.71	0.14	n.s.	n.s.	3.45	2.36	.81**	.85**	0.95	0.57	.69**	.90**
Money	0.15	0.00	-	-	0.33	0.14	.43**	.50**	0.43	0.27	.62**	.77**
Religion	0.84	0.15	n.s.	n.s.	0.44	0.27	.60**	.82**	0.60	0.40	.73**	.83**
Death	0.04	0.01	n.s.	n.s.	0.84	0.72	.75**	.97**	0.36	0.38	.72**	.88**
Assent	0.04	0.04	n.s.	n.s.	0.10	0.11	.73**	.85**	0.36	0.30	.90**	.90**

Note: **– p<.01, *-p<.05, n.s.– p>.05, "-"– parameters were not calculated because of the absence of the category in the text; average frequency of the category in the text in English (LIWC2007), average frequency of the category in Serbian (LIWCser), Spearman's correlation coefficient (r), Interclass coefficient of correlation (ICC)