Behavior Research Methods (2020) 52:1817–1835
https://doi.org/10.3758/s13428-020-01361-z

Ambulatory assessment of language use: Evidence on the temporal stability of Electronically Activated Recorder and stream of consciousness data

Ljiljana B. Lazarević[1] · Jovana Bjekić[2] · Marko Živanović[1] · Goran Knežević[1]

Published online: 3 February 2020

Abstract

The ambulatory assessment offers a wide range of methods enabling researchers to investigate psychological, behavioral, emotional, and biological processes. These methods enable us to gather data on individual differences in language use for psychological research. Two studies were conducted with an aim to evaluate and compare the temporal stability of language measures extracted by LIWC software form data obtained by two frequently used methods for assessment of language use, i.e., Electronically Activated Recorder (EAR) and stream of consciousness (SOC) task. Additionally, we examined the amount of variance in language use (assessed by both methods) that can be attributed to intra-individual variability and stable individual differences. Study 1 was focused on investigating language use obtained from 74 respondents using the EAR for 3 consecutive days. Study 2 was conducted on 250 respondents participating in a SOC task where verbal production was collected at ten time points over a 2-month period. Results show that measures obtained using the SOC task have higher temporal stability and consistency, and to a certain extent enable better detection of individual differences. Taking into account certain situational variations improves the reliability of EAR measures.

Keywords Linguistic Inquiry and Word Count (LIWC) · Electronically Activated Recorder · Stream of consciousness task · Temporal stability · Multilevel random coefficient modeling (MRCM)

Introduction

The idea that the words people use can be tapped to assess their mental, social, and physical states has been present since the beginning of psychological science. Today, scholars agree that individual differences in language use reflect important psychological characteristics of the speaker (Hirsh & Peterson, 2009; Mairesse, Walker, Mehl, & Moore, 2007; Pennebaker, Mehl, & Niederhoffer, 2003). Continuous technological advancements enabled the development of novel data collection methods that made everyday language use more accessible to researchers. In this paper, we will address some of the key properties of language data collected using two different procedures: Electronically Activated Recorder (EAR) and stream of consciousness (SOC) task. Specifically, we will explore consistency, i.e., inter- and intra-individual variability in language measures extracted by Linguistic Inquiry and Word Count (LIWC) across multiple time points, and discuss the implications of psychological research of language use.

The first and the second author contributed equally to the manuscript and share first authorship.

✉ Ljiljana B. Lazarević
ljiljana.lazarevic@f.bg.ac.rs

Jovana Bjekić
jovana.bjekic@imi.bg.ac.rs

Marko Živanović
marko.zivanovic@f.bg.ac.rs

Goran Knežević
gknezevi@f.bg.ac.rs

1 Institute of Psychology and Laboratory for the Research of Individual Differences, Faculty of Philosophy, University of Belgrade, Čika Ljubina 18-20, Belgrade 11000, Serbia

2 Institute for Medical Research, University of Belgrade, Dr. Subotića 4, Belgrade 11000, Serbia

Ambulatory assessment of language use

Ambulatory assessment (AA) is an umbrella term encompassing a range of methods used to study people in their natural environment (Trull & Ebner-Priemer, 2013). The AA approach has several characteristics that distinguish it from the other widely used assessment methods: (1) it presents a modern idiographic approach that allows for the examination of multiple individual processes (emotional, psychophysiological, and behavioral); (2) it has high ecological validity since it enables data collection in a real-world environment; (3) it focuses on a respondents' current (rather than past) states, feelings, and behaviors; (4) it enables multiple assessments at multiple time points during the assessment period; (5) it offers continuous, event-based, interactive, and time-prompted or randomly prompted assessment (Le, Hat, & Beal, 2006; Trull & Ebner-Priemer, 2013). There are three wide categories of AA methods: momentary self-report, observational, and physiological/biological/behavioral (Trull & Ebner-Priemer, 2013; Trull & Ebner-Priemer, 2014). We will focus on the first two as they enable the collection of data on language use.

The momentary self-report AA is the most frequently used (Trull & Ebner-Priemer, 2013). Using the experience sampling method (Csikszentmihalyi & Larson 1987, 2014), or ecological momentary assessment (Stone & Shiffman, 1994), respondents provide responses to queries that can be either prompted by the researcher or self-initiated, e.g., after experiencing a craving for a substance. One specific form of self-report AA that is used for the collection of linguistic data is the SOC task (James, 1890; natural stream of thought task, Holleran & Mehl, 2008). The SOC task is frequently used as a protocol when assessing language use in relation to different psychological characteristics, such as personality or emotions (e.g., Holleran & Mehl, 2008; Mehl, Robbins, & Holleran, 2012). In this task, the respondents are instructed to write their thoughts as they come to their mind for 20 min (Pennebaker & King, 1999). The basic assumption behind this task is that if people follow their spontaneous stream of thoughts, they provide us with direct insight into their inner world. The respondents have the opportunity to decide what and in which way they will write (Holleran & Mehl, 2008). The task can be adjusted for both written and oral language, and it is applicable in highly controlled lab conditions and for studies conducted outside the lab. Since the face validity and the flexibility of the SOC task are very high, it has become one the most widely used techniques for collecting data on written verbal production (see Holleran & Mehl, 2008; Lee, Kim, Seo, & Chung, 2007; Mehl et al., 2012; Pennebaker & King, 1999).

The observational AA does not rely on self-reports and offers an assessment of ambient sounds, speech, activity, location, and context (Trull & Ebner-Priemer, 2013). One of the most frequently used data collection techniques in this type of assessment is the EAR (Mehl, Pennebaker, Crow, Dabbs, & Price, 2001; Mehl & Robbins, 2012). Using a wearable device with implemented EAR software (Mehl, 2017), periodic brief snippets of ambient sounds are recorded. The EAR has high ecological validity as it records participants in their natural environment and allows for longitudinal assessments. The EAR differs from self-report AA in the perspective it captures (Mehl & Robins, 2012). Namely, in the SOC task, respondents are prompted to provide self-reports on their momentary feelings, thoughts, and behaviors, and therefore these data are vulnerable to typical self-report biases such as impression management or socially desirable responding. The EAR method captures a bystander's perspective on behavior (Manson & Robbins, 2017). Studies have indicated that data collected via EAR possess substantial inter-rater reliability and within-participant temporal stability in variables related to respondent location, activity, and social interaction (Pennebaker, Francis, & Booth, 2001).

Automatic text analysis with LIWC software

One of the most widely used tools for automatic language analysis in psychological research is LIWC software (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007). The software performs closed-vocabulary word-by-word analyses and provides information on the relative frequency of different types of words. The basic principle is that each word, or word stem, defines one or more word categories which are often hierarchically arranged, e.g., the word sad is coded as Sadness > Negative emotion > Affect (Pennebaker et al., 2007).

The LIWC output provides information on 80 categories – 4 general descriptor categories (e.g., total number of words), 22 Linguistic categories (e.g., Pronouns, Auxiliary verbs, Prepositions), 32 categories tapping Psychological processes (e.g., Affect, Cognition, Biological processes), seven Personal concern categories (e.g., Work, Home, Leisure activities), three Paralinguistic/Spoken categories (Assents, Fillers, Non-fluencies), and 12 punctuation categories. Since LIWC software is primarily intended for psychological research, the majority of words in the dictionary fall into one or more psychological or thematic categories. For a detailed overview of the LIWC categories, see Pennebaker et al. (2007).

Several qualities of LIWC contribute to its wide use in psychology (Mehl & Gill, 2010; Tausczik & Pennebaker, 2010). First, it analyzes both basic grammatical features of texts and psychologically relevant word categories. Second, various studies have contributed to the construct validity of its categories. Third, the software is psychometrically tested in several world languages. Finally, the software is user-friendly and enables quick, cost-effective, and reliable language analysis.

For these reasons, LIWC software has been widely used over the past 20 years, and during this period three versions of software (LIWC2001 – Pennebaker et al., 2001; LIWC2007 – Pennebaker et al., 2007; and LIWC2015 – Pennebaker, Boyd, Jordan, & Blackburn, 2015) and dictionaries in 15 languages, including Serbian (Bjekić, Lazarević, Erić, Stojimirović, &

Djokić, 2012; Bjekić, Lazarević, Živanović, & Knežević, 2014), have been developed. All dictionaries for different languages have the same structure and were developed using the same methodology (e.g., see Bjekić et al., 2014; Ramirez-Esparza, Pennebaker, Garcia, & Suria, 2007; Wolf, Horn, Mehl, Haug, Pennebaker, & Kordy, 2008), thus providing highly comparable outputs of language analysis. Still, dictionaries in different languages take into account language-specific features in both a grammatical and pragmatic sense. Since the majority of studies to date have been conducted using LIWC2007 with an English dictionary, one may justifiably question the cross-linguistic generalizability of those results.

Current study

We focus on two AA techniques used for collecting data on language use: observational AA using EAR and self-report AA using the SOC task. The reason for comparing these two assessment methods lies in the fact that in spite of their numerous advantages and potential to produce high-quality data, both methods are difficult to implement and have some drawbacks. Namely, in order to obtain representative samples of one's habitual language, both EAR and SOC need to be administered during a prolonged period of time or on multiple occasions. To collect data with EAR, a specific device that supports EAR software is needed, data collection lasts for several days, and the collected material has to be transcribed, which is very costly. Additionally, the EAR is somewhat intrusive for the participants, and some countries have introduced all-party consent laws making it very difficult to implement (Manson & Robbins, 2017). On the other hand, running a study in which data are collected using the SOC task requires a considerable amount of time, as data collection for a single participant can take up to 2 months. In addition, in the SOC task, respondents have higher autonomy and control over the content that they will deliver (e.g., the content can be subject to auto-censorship), and the representativeness of the content for one's own inner psychological space depends largely on the respondent's willingness to share it. In other words, when using SOC, it is of utmost importance to have highly motivated respondents that are open to sharing their intimate thoughts, feelings, and emotions. Finally, SOC is a situationally restrictive technique (it includes only one situation – that of SOC production), unlike the EAR, which is a situationally inclusive method (it comprises a variety of situations during a day). In other words, the stability indices of the SOC technique reflect mostly temporal stability, while the stability indices of EAR reflect both temporal and cross-situational stability.

Even though language use is situationally dependent (Mehl et al., 2012) and psychologically important events can disrupt typical language patterns (De Choudhury, Counts, & Horvitz, 2013), the key assumption behind the "search for linguistic fingerprint" is that there must be a certain level of temporal and cross-situational stability in language use (Boyd & Pennebaker, 2015). In other words, the variance of each aspect of verbal production could be decomposed into within- and between-person variability. The first reflects intra-individual differences in language that could be influenced by different situational factors, emotional states, and/or their interaction with personal characteristics. The second reflects differences between people that exist across time and situations, which most likely stem from some relatively stable personal characteristics.

Although the empirical evidence linking some aspects of verbal production with different psychological variables is growing, the compelling evidence on the degree of the temporal stability of language variables is still lacking. Pennebaker and King (1999) analyzed a sample of written material (essays) on various topics over several years and reported an average consistency for all categories of .59. The authors further stated that more than half of the LIWC categories showed consistency coefficients higher than .60. However, the authors did not report on the range of the consistency coefficients for specific categories and which categories had low reliability. When discussing data collected using EAR, several studies claim that the spontaneous word use remains stable over time – with an average test–retest correlation for standard linguistic variables of .41, and for psychological processes, .24 – and consistent across social contexts (Mehl et al., 2001; Mehl & Pennebaker, 2003). Nonetheless, when calculating reliability coefficients, Mehl and Pennebaker (2003) relied only on a smaller subset of LIWC categories, and obtained these average coefficients only after excluding all LIWC categories that were not sufficiently reliable. Another study that reported the stability of the written text messages showed that absolute stability of the LIWC categories was practically absent, while the rank stability for psychological categories was .35, and for linguistic categories, only .14 (Yee, Harris, Jabon, & Bailenson, 2010).

It is important to note that none of the aforementioned studies was designed to systematically assess temporal stability, i.e., the reliability of LIWC variables. Thus, it does not come as a surprise that stability coefficients differ immensely, and that some important parameters for making judgments about temporal stability were not reported. Furthermore, none of these studies included situational dependency when reporting on temporal stability. To address these issues, the current study aims to evaluate the data obtained using the EAR and SOC task. Namely, we assessed the reliability/temporal stability of language variables as well as their intra- and inter-individual variability. Therefore, two separate studies were conducted. In Study 1, we assessed temporal stability, as well as person-by-situation interaction in everyday language use, captured by EAR. In Study 2, we assessed the reliability of language material collected in a multipoint SOC task.

All aforementioned studies on reliability/temporal consistency were conducted in English and therefore used the original LIWC dictionary. In this paper, studies use the Serbian adaptation of LIWC – LIWCser (Bjekić et al., 2012, 2014). Overall, LIWCser has shown good psychometric properties and dictionary coverage across different types of verbal material (70% on average)[1]. LIWCser and English LIWC2007 have a satisfactory level of equivalence (ICC = .70), one of the highest among non-English dictionaries (Bjekić et al., 2014). Therefore, this paper primarily addresses the question of the temporal stability of LIWCser, but due to its high equivalence with English, it enables tentative conclusions about LIWC as an approach for studying individual differences in language use.[2]

Study 1

In Study 1, we assessed the temporal stability of everyday speech collected by EAR during three consecutive days and processed by LIWCser.

Method

Sample

The final sample in Study 1 consisted of 74 university students. All participants were Serbian native speakers, the average age was 20.05 years (*SD* = 1.25), 77% female. Initially, 100 respondents were recruited, but 26 had to be excluded due to technical malfunctions (i.e., the recorded sounds were completely unclear)[3]. The study was approved by the Ethics Committee of the Serbian Psychological Association at the Faculty of Philosophy, University of Belgrade. All participants provided written informed consent, and all procedures adhered to the principles of the Declaration of Helsinki. Participants could withdraw their consent at any time and could ask for their audio clips to be deleted.

Procedure and EAR protocol

For a collection of ambient sounds, we used iPod touch devices with iEAR software (Mehl, 2017). Respondents wore devices for three consecutive days, in a protective case on their arm (over the clothes). In line with recommendations of Mehl et al. (2012), ambient sounds were recorded each day between 9 AM and 12 PM, with an overnight blackout period from 12 AM to 9 AM, and recording was activated by the device for 30 s every 6 min. Participants were informed about the sampling pattern beforehand. There were several safeguards to protect the privacy of the respondents: recorded snippets were very short, preventing people who transcribed collected materials from grasping the context in which the conversation was happening, and the aggregated recording time comprised a maximum of 8% of each participant's day. In this study, the verbal production of the person carrying the device was transcribed and analyzed using LIWC, with the Serbian dictionary implemented – LIWCser (Bjekić et al., 2012, 2014).

Serbian LIWC dictionary–LIWCser

LIWCser is based on the LIWC2007 English dictionary, and it works within the same software as other LIWC2007 dictionaries. The LIWCser dictionary corresponds to other LIWC dictionaries with respect to the formal characteristics of the content. It contains 12,103 words and word stems that are classified into 65 categories. A complete overview of the development and the LIWCser dictionary is available in Bjekić et al. (2014).

Situational coding

In addition to linguistic analysis, participants' social environments from all EAR snippets were coded. We employed the Social Environment Coding of Sound Inventory (SECSI) (e.g., Mehl, Gosling, Pennebaker, 2006; Mehl & Pennebaker, 2003), that comprises the person's *interaction* (e.g., alone, talking to others), current *activity* (e.g., listening to music, eating, on the computer), and current *location* (e.g., in an apartment, outdoors, in transit). Two experts were trained and conducted contextual coding by listening to audio snippets (i.e., acoustic cues rather than text were used to increase the accuracy of coding). These cues included the noise of a running engine or a voice from the machine announcing the next bus stop (in transit), the sounds of wind blowing (outdoors), the voice of the professor (lecture), voices of local TV reporters or speakers (*TV on*), and sounds of chewing or jingling of cutlery (eating). Raters also used context information from previous and consecutive intervals to increase the accuracy of coding. For example, if after being in transit a person enters the apartment, it can be inferred that the few consecutive snippets would be in the same apartment, and

[1] It is important to note that differences in words covered by LIWC dictionaries in different languages do not directly reflect the quality of the dictionary, but rather changes in proportion of unique words – the linguistic property that differs immensely across languages. The average coverage of non-English dictionaries is usually lower than the English one – French 54% (Piolat, Booth, Chung, Davids, & Pennebaker, 2011), German 63% (Wolf et al., 2008), Spanish 66% (Ramírez-Esparza et al., 2007), and Dutch 66% (Zijlstra, Van Meerveld, Van Middendorp, Pennebaker, & Geenen, 2004).

[2] The specificities of Serbian language and limitations of LIWCser are discussed in detail later in the section "Cross-linguistic generalizability of the obtained results".

[3] Unfortunately, several devices used for data collection in this study had problems with the microphones, and all they recorded was noise, which made the snippets unusable.

the certainty further increases by the information from the subsequent snippet when TV sounds are registered.

Analytical strategy

When planning the studies, and to allow for reliable analysis of LIWCser categories, we opted for one of the frequently used units of analysis: aggregated one-day recordings for EAR (e.g., Ireland, Slatcher, Eastwick, Scissors, Finkel, & Pennebaker, 2011; Mehl & Pennebaker, 2003), and single writing for the SOC task (e.g., Mehl et al., 2012). To assess the temporal stability of the LIWCser variables, we calculated the intraclass correlation coefficient (ICC), using a two-way random effect model, absolute agreement type. The ICC can be interpreted as reliability across multiple measurements (McGraw & Wong, 1996). We report the ICC for both the single measure (*ICCsing*), i.e., the reliability of one-day word samples, and the average measure (*ICCavrg*), i.e., reliability based on the mean of 3 days (Koo & Li, 2016). As an additional measure of reliability, we calculate the Cronbach alpha for each LIWC variable.

As we had days of assessment nested within an individual, the multilevel approach was more appropriate here than ordinary least squares (OLS). Usual strategies employing traditional OLS (treating days or snippets as units of the analysis while neglecting the fact that the measurements are nested within individuals, or aggregating days or snippets within individuals while neglecting the fact that the measurements vary within the individuals) are less appropriate than multilevel random coefficient modeling (MRCM) in the case of hierarchical data structures (Nezlek, 2001). The main advantage of MRCM is the ability to model random effects, thus enabling more accurate parameter estimates and tests of significance than OLS (Raudenbush & Bryk, 2002). This analysis is ideally suited here because it enables a straightforward and accurate estimation of the amount of intra-individual variations in LIWCser categories across days that could be ascribed to situational variations ($\% \ s^2_{.sit}$, obtained by comparing intra-individual variance when situations are not introduced – unconditional model, and when situations are introduced – random coefficient model). The unique advantage of MRCM is the possibility to investigate whether the relationship between a situation and LIWCser category varies across the participants. It also enables decomposing the amount of variance stemming from the stable (inter-individual differences, $\% \ s^2_{.ind}$) and the variance originating from unstable factors (intra-individual differences, $\% \ s^2_{intra} = 100 - \% \ s^2_{.ind}$). HLM 6.06 software was used (Raudenbush, Bryk, & Congdon, 2000) for this analysis.

In Study 1, we calculated the following parameters. First, the consistency of two coders in assessing situations was calculated using the intraclass correlation coefficient (ICC), using a two-way random effect model, absolute agreement type. Since interrater consistency was high (see Table 2), these codes were averaged and used in the subsequent analyses. After that, the average presence of each situation per respondent, that is, the proportion of the snippets in which the specific situation was coded as present during each day, was calculated.

Second, to understand the influence of situations on LIWCser categories, we compared the amount of variance of LIWCser categories, as well as the precision of measurement (reliability) of LIWCser categories, before and after situations were taken into account. For this, we employed MRCM analyses, and first, we ran the unconditional model for each LIWCser category. It enabled the decomposition of the variance of each LIWCser category into the variance stemming from inter-individual variability ($\% \ s^2_{.ind}$) and intra-individual differences ($100 - \% \ s^2_{.ind}$). After that, we introduced situations as predictors at level 1 (i.e., analyzed how daily variations in situations influence daily variations in LIWCser categories) and calculated the percentage of the explained variance ($\% \ s^2_{.sit}$), as well as the improvement in the precision of measurement of LIWCser categories after introducing situations as predictors. For each of the LIWCser categories, regressions were done for all 18 situations, and the one with the largest predictive value was selected to be presented in Table 1. The regression equations at both levels are given below:

Level-1 Model[4]

$$Y = B0 + B1*(\text{Situation}) + R$$

Level-2 Model

$$B0 = G00 + U0$$
$$B1 = G10 + U1$$

The influences of situations on LIWCser categories were allowed to vary across the participants, thus reflecting person-by-situation interaction. This is achieved by modeling U1 as a random coefficient.

[4] Y – LIWCser category score; B0 – intercept (mean score of a participant on a LIWCser category across measurement occasions); B1 – slope (regression of a LIWCser category on the situation predicting it the best); R – error variance at level 1 (intra-individual variance of measurement occasions when situation is taken into account); G00 – mean of a LIWCser category across all participants (LIWCser category grand mean); U0 – error variance for the intercept at level 2; G10 – average regression slope across participants; U1 – error variance of slope at level 2 (reflecting Person × Situation interaction, i.e., whether the regression of a LIWC category on a situation varies across the participants).

Table 1 Intraclass coefficients for single measurements, average intraclass coefficients for three measurements on data collected using EAR of LIWCser categories, and precision of measurement using MRCM before and after introducing situations, the percentage of intra-individual variance and percentage of intra-individual variance explained by situations

LIWC category	ICC_{sing}	[95% CI], p	ICC_{avrg}	[95% CI], p	α	MRCM reliability	MRCM reliability$_{.sit}$	% $s^2_{.ind}$	% s^2_{intra}	% $s^2_{.sit}$
Word count	.339	[.193, .488] < .001	.606	[.418, .741] < .001	.604	.605	.889	34.1	65.9	71.9 (Talking)**
WPS	-	-	-	-	-	-	-	-	-	-
Function words	.130	[−.007, .286] = .032	.310	[−.021, .546] = .032	.308	.314	.479	13.3	86.7	24.2 (Eating)
Pronouns	.100	[−.031, .253] = .070	.251	[−.099, .504] = .070	.253	.225	.447	8.9	91.1	28.8 (Outdoor)**
Personal pronouns	.110	[−.021, .262] = .052	.271	[−.066, .516] = .052	.276	.240	.388	9.5	90.5	17.3 (To others)*
I	.194	[.055, .348] = .003	.419	[.149, .615] = .003	.425	.418	.460	19.4	80.6	6.3 (Outdoor)*
We	.137	[−.001, .292] = .026	.322	[−.002, .553] = .026	.321	.322	.478	13.5	86.5	23.1 (Outdoor)
You	.053	[−.075, .204] = .217	.143	[−.264, .434] = .217	.143	.085	.268	3.1	96.9	19.9 (Alone)
S/he	.160	[.021, .316] = .011	.364	[.061, .581] = .011	.363	.371	.516	16.5	83.5	23.3 (Alone)
They	.153	[.015, .308] = .015	.351	[.042, .572] = .015	.351	.336	.504	11.2	88.8	25.5% (Alone)*
Impersonal pronouns	.108	[−.026, .263] = .059	.267	[−.082, .516] = .059	.267	.257	.576	10.4	89.6	43.3 (Alone)
Common verbs	.111	[−.023, .266] = .054	.273	[−.073, .520] = .054	.273	.278	.499	11.4	88.6	28.2 (Other public places)*
Auxiliary verb	.229	[.086, .384] = .001	.472	[.221, .652] = .001	.472	.462	.679	22.5	77.5	40.5 (Alone)
Past	.069	[−.062, .222] = .157	.182	[−.211, .462] = .157	.181	.184	.288	5.2	94.8	9.5 (Reading)*
Present	.152	[.013, .308] = .016	.349	[.037, .571] = .016	.347	.350	.405	14.6	85.4	8.5 (Alone)
Future	.120	[−.013, .273] = .040	.290	[−.040, .529] = .040	.294	.278	.425	8.4	91.6	16.9 (Music)
Adverbs	−.048	[−.156, .088] = .766	−.160	[−.682, .225] = .766	−.167	.002	.303	0.1	99.9	30.0 (Laughing)
Prepositions	.239	[.095, .394] < .001	.485	[.239, .661] < .001	.483	.494	.588	24.5	75.5	17.1 (Other public places)*
Conjunctions	.135	[−.003, .292] = .027	.320	[−.008, .553] = .027	.317	.369	.439	16.4	83.6	10.6 (Lecture)
Quantifiers	.051	[−.072, .198] = .217	.139	[−.253, .426] = .217	.143	.110	.667	3.6	96.4	62.7 (Alone)
Numbers	−.013	[−.132, .134] = .564	−.039	[−.541, .318] = .564	−.038	.018	.487	0.6	99.4	44.9 (Music)
Informal/Swear words	.301	[.155, .452] < .001	.564	[.356, .713] < .001	.562	.545	.599	27.3	72.7	12.1 (TV on)*
Social processes	.034	[−.091, .185] = .301	.096	[−.334, .404] = .301	.096	.128	.390	4.5	95.5	29.7 (Eating)
Family	−.040	[−.155, .103] = .717	−.132	[−.675, .255] = .717	−.131	.002	.037	0.1	99.9	5.7 (Eating)
Friends	.343	[.197, .491] < .001	.610	[.423, .743] < .001	.607	.622	.717	35.0	65.0	25.2 (To others)
Humans	.023	[−.098, .169] = .360	.065	[−.364, .378] = .360	.066	.079	.320	2.4	97.6	21.6 (Talking)
Affective processes	.051	[−.078, .203] = .226	.139	[−.275, .433] = .226	.138	.135	.278	5.0	95.0	16.3 (Laughing)*
Positive emotions	.152	[.015, .307] = .015	.350	[.044, .571] = .015	.351	.345	.421	14.7	85.3	11.7 (Lecture)
Negative emotions	.059	[−.069, .211] = .189	.159	[−.239, .445] = .189	.160	.157	.768	5.5	94.5	72.8 (Alone)
Fear and anxiety	−.076	[−.185, .061] = .189	−.270	[−.880, .164] = .189	−.269	.001	.060	0.1	99.9	2.7 (Amusement)*
Anger and resentment	.084	[−.046, .236] = .107	.216	[−.150, .481] = .107	.218	.206	.279	6.7	93.3	5.8 (Phone)*
Sadness	.303	[.157, .454] < .001	.566	[.358, .714] < .001	.563	.566	.754	31.6	68.4	40.3 (Outdoor)*
Cognitive processes	.192	[.053, .347] = .003	.417	[.143, .615] = .003	.419	.436	.665	20.7	79.3	39.7 (Transit)
Insight	.248	[.105, .402] < .001	.497	[.260, .668] < .001	.498	.512	.554	26.2	73.8	6.8 (Lecture)*
Causation	.339	[.194, .487] < .001	.606	[.419, .740] < .001	.605	.629	.678	35.9	64.1	13.2 (Tv on)

Table 1 (continued)

LIWC category	ICC_{sing}	[95% CI], p	ICC_{avrg}	[95% CI], p	α	MRCM reliability	MRCM reliability$_{.sit}$	% $s^2_{.ind}$	% s^2_{intra}	% $s^2_{.sit}$
Discrepancy	.073	[−.057, .225] = .142	.190	[−.192, .465] = .142	.191	.188	.788	6.9	93.1	74.2 (Alone)
Tentative	.233	[.089, .387] = .001	.476	[.227, .655] = .001	.475	.396	.455	18.0	82.0	9.6 (Alone)*
Certainty	.128	[−.009, .284] = .034	.306	[−.028, .543] = .034	.304	.326	.429	13.4	86.6	15.5 (Alone)*
Inhibition	.158	[.019, .313] = .013	.359	[.054, .578] = .013	.358	.358	.705	15.7	84.3	54.1 (Alone)
Inclusive	.117	[−.018, .271] = .046	.284	[−.055, .528] = .046	.285	.263	.382	10.0	90.0	16.2 (Laughing)
Exclusive	.122	[−.014, .278] = .041	.294	[−.045, .536] = .041	.292	.297	.550	12.5	87.5	36 (Outdoor)
Perceptual processes	.066	[−.065, .219] = .169	.174	[−.223, .457] = .169	.173	.174	.340	6.0	94.0	17.2 (Laughing)
See	.114	[−.022, .270] = .052	.279	[−.069, .526] = .052	.276	.269	.527	10.7	89.3	30.6 (Talking)
Hear	−.056	[−.170, .085] = .791	−.191	[−.770, .219] = .791	−.188	.002	.281	0.1	99.9	29.7 (Alone)
Feel	.067	[−.064, .220] = .165	.76	[−.219, .458] = .165	.175	.240	.351	9.1	90.9	12.1 (Laughing)
Biological processes	.220	[.077, .376] = .001	.459	[.199, .643] = .001	.456	.450	.520	21.1	78.9	13.3 (TV on)*
Body	.205	[.065, .359] = .002	.437	[.173, .627] = .002	.440	.428	.536	15.4	84.6	16.5 (Talking)
Health	.007	[−.115, .154] = .452	.020	[−.447, .354] = .452	.020	.013	.348	0.4	99.6	33.3 (Laughing)*
Sex and love	.075	[−.056, .228] = .138	.195	[−.190, .470] = .138	.194	.189	.271	7.0	93.0	10.2 (To others)*
Ingestion	.041	[−.085, .191] = .270	.113	[−.309, .415] = .270	.113	.115	.396	3.8	96.2	33.3 (TV on)*
Negations	.259	[.115, .412] < .001	.512	[.280, .678] < .001	.512	.528	.656	27.4	72.6	26.7 (Talking)*
Relativity	.083	[−.048, .237] = .113	.214	[−.161, .483] = .113	.214	.216	.494	8.5	91.5	32.6 (Alone)
Motion	.051	[−.077, .202] = .206	.138	[−.273, .432] = .206	.138	.129	.208	4.7	95.3	8.3 (Computer)*
Space	.044	[−.084, .196] = .257	.121	[−.304, .423] = .257	.120	.142	.202	5.2	94.8	5.4 (Phone)*
Time	.008	[−.114, .156] = .445	.023	[−.441, .356] = .445	.023	.024	.229	0.7	99.3	22.0 (Alone)
Work	.079	[−.051, .231] = .124	.204	[−.172, .474] = .124	.205	.121	.209	4.5	95.5	8.6 (Phone)*
Achievement	.017	[−.107, .167] = .392	.050	[−.408, .376] = .392	.050	.063	.428	1.9	98.1	39.1 (To others)*
Leisure	.062	[−.069, .215] = .184	.165	[−.239, .451] = .184	.163	.161	.258	6.1	93.9	7.1 (TV on)*
Home	.150	[.012, .305] = .016	.346	[.035, .569] = .016	.345	.346	.508	15.7	84.3	20.6 (Reading)*
Money	.118	[−.014, .271] = .042	.287	[−.045, .527] = .042	.291	.270	.411	11.6	88.4	19.4 (Alone)
Religion	.180	[.039, .337] = .006	.398	[.109, .604] = .006	.395	.383	.481	16.7	83.3	15.9 (Eating)*
Death	.009	[−.115, .158] = .441	.026	[−.446, .360] = .441	.025	.018	.373	0.7	99.3	25.6 (Transit)*
Assent	.333	[.188, .482] < .001	.600	[.409, .737] < .001	.598	.593	.644	32.7	67.3	12.1 (Laughing)*
Non-fluencies	.249	[.104, .403] < .001	.498	[.259, .669] < .001	.497	.511	.773	26.3	73.7	0
Fillers	.288	[.143, .439] < .001	.548	[.334, .702] < .001	.549	.534	.560	27.8	72.2	5.5 (Reading)*

Note: ICC_{sing} – intraclass coefficients for single measurements; ICC_{avrg} – average intraclass coefficients for three measurements; α – Cronbach alpha; % $s^2_{.ind}$ – the percentage of the variance of inter-individual differences (MRCM); % s^2_{intra} – the percentage of the intra-individual variance (MRCM) (100-% $s^2_{.ind}$); MRCM reliability – precision of measurement using MRCM; MRCM reliability$_{.sit}$ – precision of measurement after introducing the situation with the highest predictive value for each LIWCser category; % $s^2_{.sit}$ – the percentage of the intra-individual variance in LIWCser category explained by the situation having the highest predictive value. **$p < 0.01$; *$p < 0.05$ (these significance values are related to regression coefficients between LIWCser categories and situations). One should bear in mind that although some percentages of variance in LIWCser categories explained by situations look high, they are not significant because of high standard errors of these regression coefficients.

It is important to highlight that this approach focuses on how intra-individual consistency in language use changes when language data are collected across different situations. Therefore, it differs from the approach employed by Baddeley, Pennebaker, & Beevers (2013) and Mehl & Pennebaker (2003) who studied situational dependency of the LIWC variables regardless of intra-individual differences in language use.

Results

Transcribed one-day snippets on average contained 682 words (*SD* = 438.10), with the shortest text being only seven words long, while the longest text contained 2360 words. Aggregated texts (i.e., total oral production of a single person recorded during the 3-day period) had on average 2028 words (*SD* = 958.81), where the smallest sample had 374 words, and the longest 5154 words. On average, the dictionary coverage of a single text was 70.1% (*SD* = 3.86, *Min* = 47.9%, *Max* = 83.9%).

Results on temporal stability, reliability coefficients, and the variance of individual differences for all LIWCser categories are presented in Table 1. The average intraclass correlation coefficients for three measurements taken together (*ICCavrg*) range from *ICC* = −.27 to *ICC* = .75, with the average ICC across all categories of *ICCavrg* = .28 and the average Cronbach alpha, .27.

The intraclass correlation coefficients for single measurements (*ICCsing*) are generally low and range between −.08 and .34, with the average *ICCsing* = .13. The results of the MRCM analysis show that on average, 12.9% of the variance could be attributed to inter-individual differences, while the rest represents within-person variability. Overall, it seems that 3-day EAR recordings do not produce very reliable language measures in terms of LIWCser categories.

As results show, LIWCser categories differ greatly in terms of their temporal stability. Namely, the most consistent aspects of everyday speech are the total number of produced words and paralinguistic properties of the speech, such as Assents, Fillers, and Non-fluencies. Out of Linguistic categories, among the most consistent are First-person singular pronouns, Auxiliary verbs, Preposition, Swear (informal) words, and Negations. Words reflecting Cognitive processes (as well as the subordinated categories Insight, Causation, and Tentative words) tend to be the most stable in time among psychological categories. Categories Friends, Sadness, and Biological process/Body also showed some consistency over the course of 3 days. On the other hand, if higher-order dictionary categories are compared, results show that all three groups have highly similar yet relatively low consistency parameters, i.e., Linguistic processes (mean *ICCavrg* = .29; mean $S^2_{.ind}$ = 13.4%), Psychological processes (mean *ICCavrg* = .25; mean $S^2_{.ind}$ = 11.5%), and Personal concerns (mean *ICCavrg* = .21; $S^2_{.ind}$ = 8.3%).

Verbal behavior and participants' social environments

Finally, one might ask to what extent situational differences reduce the consistency of verbal output, i.e., how much does the temporal stability differ if one takes the context into account. To answer this question, participants' social environments from all EAR snippets were coded and then averaged between coders for each day. The ratings of two experts for all situations had high inter-rater reliability. The lowest intraclass correlation coefficient was *ICC* = .944 for codes Music and Amusement, while the highest was *ICC* = 1.000 for code Working. The average *ICC* for situational coding was *ICC* = .988 (detailed results of this analysis are available in Table 2). Additionally, we have tested the stability of SECSI categories across 3 days. The average *ICC* for all situations across days was *ICC* = .564, but as can be seen from Table 2, these coefficients vary, and range between *ICC* = .123 (for Amusement) and *ICC* = .856 (for Reading). These findings are in line with previous reports (Mehl et al., 2006; Mehl & Pennebaker, 2003). We also report the average presence of each situation across 3 days in Table 2, while the detailed representation of each situation per respondent (i.e., the proportion of the snippets in which the specific situation was coded as present for each of the 3 days per respondent, and the average proportion across 3 days per respondent) is provided in Table S1 in the Online Supplementary Material (https://osf.io/hm7ky/).

To assess how the precision of measurement[5] of different LIWCser categories was affected when situations were introduced, we performed MRCM analyses and compared reliability coefficients before and after the situations were taken into account. The percentages of intra-individual variance explained by the situations were calculated for those situations that were significantly predicting a particular LIWCser category (all MRCM reliabilities for all LIWCser categories after introducing situations and percentages of the explained intra-individual variance in LIWCser categories by situation are available in Table S2 in the Online Supplementary Material https://osf.io/6wd3c/). Table 1 shows side-by-side reliabilities for each of the LIWCser categories before and after introducing situations that significantly predict each language category. In addition to that, the last column in Table 1 shows the percentage of intra-individual differences in each LIWCser category that can be explained by the situation with the highest predictive value.

[5] MRCM reliability is calculated as: $\lambda = \tau_{00}/(\tau_{00} + \sigma/n_k)$; τ_{00} – variance of inter-individual differences; σ – variance of intra-individual differences; n_k – number of measurement points at level 1. This coefficient gives similar information as the percentage of the variance explained by the inter-individual differences.

Table 2 Intraclass correlation coefficients for SECSI situations and average representation of situations in snippets across three days

SECSI situations	ICC_{coders}	ICC_{days}	Average % across 3 days
Alone	.993	.548	54.3
Talking	.995	.458	25.3
To others	.996	.518	21.9
On the phone	.999	.534	2.5
Laughing	.991	.579	3.7
Music on	.944	.649	7.5
TV on	.997	.710	6.7
Computer	.998	.816	8.2
Reading	.997	.856	7.8
Working	1.000	.762	1.5
Eating	.975	.596	1.7
Lecture	.997	.521	6.6
Amusement	.944	.123	1.0
Apartment	.998	.700	57.0
Outdoor	.994	.201	5.6
Transit	.998	.581	4.2
Restaurant	.996	.308	1.5
Other public places	.997	.748	11.5

Note: ICC_{coders} – average Intraclass coefficients between coders for each situational category; ICC_{days} – stability across days; average % across 3 days – the representation of each SECSI situation in EAR snippets collected during the 3-day period

A careful look at the results reveals that reliability increases for LIWCser categories that were found to be more temporally stable even prior to introducing situations such as Word Count, Pronouns, Prepositions, and Negations. The reliability of the category Word Count increased when the situation *Talking* was taken into account: being in a situation coded as *Talking* led to a higher number of words produced, and thus there was less "unexplained variance" left in the model. The same situation increased the reliability of the category Negations.

In people who tend to *talk to others* more, the reliability of the category Personal pronouns is higher, i.e., the reliability significantly increases when the frequency of *talking to others* is entered into the model. Similarly, *talking to other people* increases the temporal stability of the category Sex and Love. Our sample consisted of university students; thus, it might be that a significant amount of conversation was conducted with romantic partners and friends. Moreover, when engaged in personal communication *with others*, the reliability of the words belonging to the category Achievement increases – this is expected as students substantially discuss with their classmates curriculum topics and scholastic achievement.

Furthermore, the high frequency of talking over the *phone* increased the reliability of several LIWCser categories: Anger and Resentment, Space, and Work. These results suggest that when engaged in the *phone* conversation, people are more prone to exhibit negative affect, frustration, and consistent use of word mapping anger. Additionally, it is possible that a substantial amount of communication via *phone* deals with making arrangements regarding where to meet, leading to a higher frequency of words mapping space, and therefore increasing temporal stability of the Space category. Furthermore, communication over the *phone* increased temporal consistency of the usage of words related to Work – as already noted, our respondents are students, and a majority of their daily activities and preoccupations are related to the university and school activities.

Interestingly, when more situations were coded as *outdoors,* the reliability of several LIWCser categories, i.e., Pronouns, I, and Sadness, increased significantly. Our findings also suggest that the more people are in situations coded as *Other public places,* the greater the reliabilities of Prepositions and Common verbs. The majority of the communication happening outdoors involves hanging out with friends, sitting in a bar or a restaurant, or spending time at the university, so people in this context usually communicate with friends, romantic partners, or classmates. Therefore, it is highly likely that when people are outdoors, communication tends to be more personal and psychologically involved. Previous studies identified markers of verbal immediacy and verbal emotional expression (e.g., Gill, Oberlander, & Austin, 2006; Mehl et al., 2012; Pennebaker & King, 1999). Verbal immediacy is defined as personal, involved, and experiential language, or a degree to which participants' language use reflects psychological engagement, and is mapped by function words, a high percentage of first-person pronouns and prepositions, and a higher frequency of verbs. Verbal emotional expression is mapped by words tapping affective processes. Our results are in line with the viewpoint of Mehl et al. (2012) and indicate that when involved in close communication, people tend to consistently use words describing personal involvement (Function words: Pronouns, I), words expressing spatial and temporal relations (e.g., near, beside, under) or marking various semantic roles (e.g., for, of), words describing inner processes (Affective processes), and other peoples' behaviors (Common verbs). These findings support the standpoint of Mehl et al. (2012) and highlight the relevance of communication context when studying the psychological implications of natural language use.

Context dependency of natural language is also visible in the fact that the reliability of the category Insight increased significantly when respondents were attending *Lecture* at the university. The informal situation, such as *TV on*, emerged as a significant predictor of several LIWCser categories: Leisure, Informal words and swears, Biological processes, and Ingestion. Lastly, when situations were coded as *Laughing*, the reliability of categories Affective processes, Health, and Assents increased.

Finally, some of our results require further investigation to be fully understood. For example, when respondents were in a situation that usually does not contain verbal communication, like being *alone*, *reading*, or using a *computer*, the reliability of several LIWCser categories significantly increased. Specifically, the temporal stability of the categories They, Tentative, and Certainty increased when respondents were *alone*. *Reading* increased the reliability of categories Past, Home, and Fillers. Reliability of the category Motion was significantly increased when people used *computers,* while the situation *eating* increased reliability of the LIWCser category Religion.

Study 2

Study 2 was conducted with the same goal as Study 1, with the exception that data were obtained using the SOC task, i.e., the target of analysis was written verbal production collected at ten separate time points over the course of 2 months. Study 2 received the same ethics committee approval and adhered to the same principles as Study 1, i.e., participants could withdraw their consent at any time, and could ask their data to be removed from the data set.

Method

Sample

The final sample in Study 2 consisted of 250 university students, all native Serbian speakers, average age 22 years (*SD* = 2.17), 77% female. Initially, 310 respondents were recruited, but 18 dropped out, 37 were excluded due to irregular completion of the SOC task (did not complete the task within the expected time frame, i.e., missed deadlines for completion of the SOC task), and five had invalid verbal products in terms of task requirements (e.g., wrote song lyrics instead of their own stream of thought)[6].

Procedure and SOC task

For data collection, we employed a slightly modified version of the SOC task. As mentioned before, in the original version of the task, respondents have 20 min to write about whatever comes to their mind (Holleran & Mehl, 2008; Lee et al., 2007; Pennebaker & King, 1999). In this study, the standard SOC protocol was modified in terms of its length and data collection setting. Time for writing was shortened to 10 min so that participants would make a smaller number of pauses during writing in order to ensure relatively continuous verbal production. The length of SOC in this study was based on the results of a small pilot ($N = 8$) in which we administered the SOC task without time limits and recorded the frequency of short breaks during writing. We learned that participants started taking breaks longer than 5 s between 9 and 12 min into the writing.

Previous studies using the SOC task were conducted in a laboratory setting (Holleran & Mehl, 2008; Mehl et al., 2012). In light of all the problems in studies investigating habitual verbal production in artificial conditions, in this study the respondents completed the task in their natural environment. We designed the online version of the task which enabled participants to complete it from different personal devices such as laptops or smartphones. The instructions for the SOC task were exactly the same as in Holleran & Mehl: "During the next 10 min, you can write whatever springs to your mind. You can write freely about your thoughts, feelings, ideas, etc., in other words, everything that comes to your mind. Write your thoughts as they come to you, do not go back and change the text. Write all 10 min and do not stop!" (2008, p. 750). Previous findings indicated that this instruction can provoke respondents to share their intimate thoughts and feelings (e.g., Holleran & Mehl, 2008).

The respondents completed the SOC task ten times over the course of 2 months (5 days apart). For each writing, participants received an email with the link directing them to the task. The timeline of the emails was proportionally distributed to cover the working days and weekends as well as different parts of the day (morning, midday, afternoon, evening). In line with the recommendations of Mehl et al. (2006), when the respondents received an email, they could choose the time and the environment in which they would do the task. However, they had to do it within 24 h from the time they received the email. All texts were anonymized before being read by the researchers. After correction for typos, each of the 2500 texts (10 per participant) was analyzed using LIWCser (Bjekić et al., 2012).

Analytical strategy

The same analyses used in Study 1 are employed here. The difference is that MRCM is used here to estimate intra- and inter-individual variance estimations only. Since all ten occasions of measurement assumed the same SOC situation, the design of the study did not allow us to study the influence of situations on the temporal variations in LIWCser categories.

[6] In line with the suggestion of the anonymous reviewer, we have explored potential reasons for dropout. As SOC data were collected as part of a larger study (Bjekić, 2016), we were able to compare scores of the respondents who completed the SOC task and those who dropped out on the Big Five personality traits (assessed via Serbian NEO-PI R; Đurić-Jočić, Džamonja-Ignjatović, & Knežević, 2004; Knežević et al. 2004). No significant differences in any of the Big Five personality traits were detected.

Results

Single texts on average contained 244 words (SD = 105.37), where the shortest text contained only 24 words, while the longest had 786 words. Aggregate texts, from the ten time points, on average had 2254 words (SD = 920.21), where the smallest sample had 546 words and the longest 5282 words. On average, the coverage of single texts by the dictionary was 72.4% (SD = 4.62, Min = 48.3%, Max = 88.5%).

The intraclass coefficients for all measurements taken together (*ICCavrg*) range from ICC = .33 to ICC = .97, depending on the LIWCser category, with the overall average ICC = .60. These results indicate that measures of verbal production based on ten writing samples taken together could be considered relatively consistent. However, the intraclass coefficients for single measurements (*ICCsing*) are mostly low and range between ICC = .05 and ICC = .30 (with the exception of Word Count, Words per sentence, and the usage of punctuation in the text, which tends to be much higher), with the average *ICCsing* = .16. These results imply that verbal production collected at one time point contains more variance that is specific for that particular measurement than the variance reflecting the linguistic style of a person. In line with that, the results of MRCM show that on average, 15.5% of the variance can be attributed to differences between persons, i.e., individual differences, while the rest represent within-person differences between measurements. The intraclass coefficients together with Cronbach alphas and the variance of individual differences for each LIWC category are presented in Table 3.

The results show that almost all LIWCser categories derived from ten SOC tasks possess temporal stability above .40. The most consistent categories are formal characteristics of the text, such as Word Count, the average number of Words per sentence, and the usage of punctuation signs (*ICCavrg* > .90). Similar to the results obtained in Study 1, the larger differences in terms of time consistency are obtained between individual categories within the specific higher-order category of the dictionary, than between the types of the categories. Namely, in terms of higher-order dictionary categories, all three major groups of variables show similar consistency, i.e., Linguistic categories (mean *ICCavrg* = .63; $S^2_{.ind}$ = 15.4%), Psychological processes (mean *ICCavrg* = .57; $S^2_{.ind}$ = 13.2% variance of individual differences), and Personal concerns (mean *ICCavrg* = .53; $S^2_{.ind}$ = 11.6%). If specific categories are explored, the most stable in time (*ICCavrg* > .70) are Function words, Pronouns, Personal pronouns, First-person pronouns, Conjunctions, Adverbs, Prepositions, Informal words, and higher-order psychological categories like Positive emotions, Affect, Cognitive, and Relative processes.

Categories with relatively low representation in the text, such as Third-person personal pronouns, Past, Fear, Anger, Sadness, Health, etc., tend to be less stable in time (*ICCavrg* < .50). However, it would not be correct to infer that temporal stability results from the representation of the specific category in the text since some of the less frequent categories demonstrate a high level of consistency, such as Informal words, Second-person pronouns, categories Humans, Body, Work, etc. (see Table 4 with average representation of each category in the Appendix). Finally, to allow for a more straightforward comparison of results from Study 1 and Study 2, we have calculated *ICCavrg* for three SOC samples, i.e., first, second, and third writing. As seen from Table 3, all average ICC dropped as anticipated and produced the mean *ICCavrg3* = .37.

General discussion

The study aimed to evaluate the temporal stability of language production variables extracted by LIWC software, i.e., the Serbian version of the dictionary, LIWCser, using two frequently employed contemporary methods for collecting samples of habitual language use: EAR and SOC. The growing interest in the ambulatory assessment of language use can be primarily attributed to its ecological validity (Trull & Ebner-Priemer, 2013, 2014). The application of these methods together with the development of cost-effective automatic text analysis tools provided researchers with a real possibility to study habitual language use. Even though language properties operationalized as relative frequency of one or more LIWC categories have been shown to correlate with different aspects of human functioning, such as personality traits (Holtgraves, 2011; Kern et al., 2014; Schwartz et al., 2013; Sumner, Byers, Boochever, & Park, 2012), social interactions (Park et al., 2015; Pfeil, Arjan, & Zaphiris, 2009), personal values (Chen, Hsieh, Mahmud, & Nichols, 2014; Vaisey & Miles, 2014), and psychological disorders (McMain et al., 2013; Molendijk, Bamelis, van Emmerik, Arntz, Haringsma, & Spinhoven, 2010; Rosenbach & Renneberg, 2015), there is still limited knowledge about some of the basic properties of these language variables, especially across different languages.

The consistency of language outputs from EAR and SOC

Overall, our results demonstrated that 3-day EAR produces less consistent language data than ten SOC writings over the course of 2 months. That is, the mean consistency coefficient across all LIWCser variables when data are collected by EAR is .28 and .60 for SOC. These results are in line with consistency coefficients reported by Pennebaker and King (1999), Mehl et al. (2001), as well as Mehl and Pennebaker (2003).

Table 3 Intraclass coefficients for single measurements, average intraclass coefficients for ten measurements, and average intraclass coefficients for three measurements on data collected using the stream of consciousness (SOC) task

	ICC_{sing}	[95% CI], p	ICC_{avrg}	[95% CI], p	α	% $s^2_{.ind}$	p_y	ICC_{avrg3}
Word count	.746	[.709, .782] < .001	.967	[.961, .973] < .001	.967	73.98	<.001	.888
WPS	.560	[.511, .610] < .001	.927	[.913, .940] < .001	.929	23.81	<.001	.849
Function words	.197	[.159, .241] < .001	.710	[.653, .761] < .001	.710	19.34	<.001	.357
Pronouns	.191	[.153, .235] < .001	.702	[.644, .754] < .001	.702	18.67	<.001	.399
Personal pronouns	.208	[.169, .253] < .001	.724	[.670, .772] < .001	.725	17.51	<.001	.486
I	.239	[.198, .287] < .001	.759	[.712, .801] < .001	.759	21.57	<.001	.518
We	.128	[.096, .166] < .001	.595	[.516, .666] < .001	.595	12.21	<.001	.258
You	.131	[.099, .169] < .001	.601	[.522, .671] < .001	.600	14.30	<.001	.296
S/he	.157	[.123, .198] < .001	.651	[.583, .712] < .001	.651	13.89	<.001	.462
They	.048	[.025, .076] < .001	.336	[.206, .452] < .001	.335	6.04	<.001	.185
Impersonal pronouns	.145	[.112, .185] < .001	.629	[.557, .694] < .001	.629	14.82	<.001	.282
Common verbs	.137	[.104, .176] < .001	.614	[.538, .681] < .001	.614	14.07	<.001	.337
Auxiliary verb	.128	[.096, .166] < .001	.595	[.516, .666] < .001	.595	11.59	<.001	.366
Past	.051	[.028, .079] < .001	.349	[.222, .463] < .001	.349	2.79	= .005	.003
Present	.096	[.068, .131] < .001	.516	[.421, .600] < .001	.516	9.19	<.001	.224
Future	.110	[.080, .147] < .001	.554	[.467, .632] < .001	.553	11.10	<.001	.327
Adverbs	.214	[.174, .260] < .001	.731	[.679, .778] < .001	.731	22.96	<.001	.529
Prepositions	.200	[.161, .244] < .001	.714	[.658, .764] < .001	.715	19.15	<.001	.568
Conjunctions	.250	[.208, .299] < .001	.769	[.724, .810] < .001	.769	27.19	<.001	.564
Quantifiers	.115	[.084, .151] < .001	.565	[.479, .641] < .001	.564	11.54	<.001	.135
Numbers	.136	[.103, .175] < .001	.611	[.535, .679] < .001	.611	15.14	<.001	.441
Informal/Swear words	.296	[.251, .347] < .001	.808	[.770, .842] < .001	.808	25.75	<.001	.645
Social processes	.172	[.136, .215] < .001	.676	[.612, .732] < .001	.676	17.87	<.001	.459
Family	.082	[.056, .115] < .001	.473	[.371, .565] < .001	.474	11.20	<.001	.318
Friends	.112	[.082, .148] < .001	.558	[.471, .635] < .001	.557	12.12	<.001	.366
Humans	.182	[.145, .225] < .001	.690	[.629, .744] < .001	.690	18.20	<.001	.554
Affective processes	.189	[.152, .233] < .001	.700	[.641, .753] < .001	.700	18.67	<.001	.431
Positive emotions	.200	[.161, .244] < .001	.714	[.658, .764] < .001	.715	19.84	<.001	.343
Negative emotions	.140	[.107, .179] < .001	.619	[.545, .686] < .001	.620	14.07	<.001	.311
Fear and anxiety	.074	[.048, .105] < .001	.442	[.333, .540] < .001	.442	6.97	<.001	−.025
Anger and resentment	.047	[.024, .075] < .001	.329	[.198, .446] < .001	.330	5.67	<.001	.194
Sadness	.064	[.040, .095] < .001	.408	[.292, .511] < .001	.409	4.02	<.001	.258
Cognitive processes	.242	[.200, .290] < .001	.761	[.715, .803] < .001	.761	24.67	<.001	.504
Insight	.170	[.134, .212] < .001	.672	[.607, .729] < .001	.672	18.17	<.001	.442
Causation	.095	[.066, .129] < .001	.511	[.416, .597] < .001	.512	10.15	<.001	.345
Discrepancy	.126	[.094, .164] < .001	.591	[.511, .662] < .001	.590	13.57	<.001	.393
Tentative	.148	[.114, .188] < .001	.635	[.564, .699] < .001	.635	13.22	<.001	.352
Certainty	.134	[.101, .172] < .001	.607	[.530, .676] < .001	.608	13.57	<.001	.429
Inhibition	.100	[.071, .134] < .001	.525	[.432, .608] < .001	.526	11.77	<.001	.341
Inclusive	.116	[.086, .153] < .001	.568	[.484, .644] < .001	.568	12.75	<.001	.278
Exclusive	.141	[.108, .180] < .001	.621	[.547, .688] < .001	.621	12.00	<.001	.276
Perceptual processes	.157	[.122, .198] < .001	.650	[.582, .712] < .001	.651	17.97	<.001	.477
See	.085	[.058, .118] < .001	.482	[.380, .572] < .001	.482	9.01	<.001	.171
Hear	.082	[.055, .114] < .001	.471	[.368, .564] < .001	.471	10.62	<.001	.396
Feel	.060	[.035, .089] < .001	.388	[.268, .495] < .001	.387	7.22	<.001	.385
Biological processes	.167	[.131, .209] < .001	.667	[.601, .725] < .001	.666	17.67	<.001	.527
Body	.135	[.103, .174] < .001	.610	[.534, .678] < .001	.610	14.69	<.001	.449
Health	.055	[.031, .084] < .001	.367	[.243, .478] < .001	.367	4.98	<.001	.398

Table 3 (continued)

	ICC_{sing}	[95% CI], *p*	ICC_{avrg}	[95% CI], *p*	α	% $s^2_{.ind}$	p_y	ICC_{avrg3}
Sex and love	.093	[.065, .128] < .001	.508	[.411, .594] < .001	.507	11.22	<.001	.153
Ingestion	.138	[.105, .177] < .001	.615	[.539, .682] < .001	.614	14.11	<.001	.393
Negations	.145	[.111, .184] < .001	.628	[.556, .693] < .001	.629	14.82	<.001	.372
Relativity	.197	[.158, .241] < .001	.710	[.653, .761] < .001	.710	18.50	<.001	.474
Motion	.119	[.088, .156] < .001	.575	[.492, .649] < .001	.574	11.14	<.001	.270
Space	.093	[.065, .127] < .001	.507	[.410, .593] < .001	.507	6.87	<.001	.248
Time	.181	[.145, .225] < .001	.689	[.628, .743] < .001	.690	18.73	<.001	.371
Work	.159	[.124, .200] < .001	.654	[.587, .715] < .001	.654	14.57	<.001	.355
Achievement	.117	[.087, .154] < .001	.571	[.487, .646] < .001	.571	12.57	<.001	.298
Leisure	.104	[.075, .139] < .001	.537	[.447, .618] < .001	.537	12.88	<.001	.348
Home	.153	[.119, .193] < .001	.643	[.574, .706] < .001	.643	15.94	<.001	.363
Money	.068	[.043, .099] < .001	.422	[.309, .524] < .001	.422	7.40	<.001	.104
Religion	.095	[.067, .129] < .001	.512	[.417, .598] < .001	.512	7.79	<.001	.226
Death	.060	[.035, .089] < .001	.389	[.269, .496] < .001	.388	6.94	<.001	.236
Assent	-	-	-	-	-	-	-	
Non-fluencies	-	-	-	-	-	-	-	
Fillers	-	-	-	-	-	-	-	
All punctuation	.536	[.488, .587] < .001	.920	[.905, .934] < .001	.921	53.09	<.001	.754

Note: ICC_{sing} - intraclass coefficients for single measurements; ICC_{avrg} - average intraclass coefficients for 10 measurements; α - Cronbach alpha; % $s^2_{.ind}$ - the percentage of the variance of inter-individual differences from MRCM; ICC_{avrg3} - average intraclass coefficients for three measurements

In order to adequately interpret the results obtained, we need to look at the differences between these two AA methods of language data collection. Firstly, EAR collects language data continuously over the course of several days, while the SOC task is completed at separate time points. Second, EAR provides data on oral production, while SOC is a written language production task. Third, the SOC task is consistent in its form and instruction every time a person completes it, while the EAR collects language produced in different contexts and situations throughout the day. As already mentioned above, EAR data include much more cross-situational variability than SOC data. Furthermore, language captured by EAR is almost always conversational (i.e., as part of the interaction with another person), while SOC language takes the form of a monologue. Finally, it is well known that data gathered at more time points tend to produce higher consistency coefficients. It is important to note that there is no difference in the overall amount of language data collected using these two methods: an average participant tends to produce just over 2000 words during a 3-day EAR (when participants are recorded for 30 s every 6 min during 15 h of the day), which is just slightly less than the number of words obtained by ten SOC tasks (2200 words).

In light of these differences, it seems that written language in the form of a monologue obtained in a typical SOC production situation tends to provide researchers with more stable language products than oral production recorded during regular daily activities of a person comprising a variety of situations. As is well known from the long-standing situation-personality debate, indices of temporal stability are much higher and more robust than the indices of cross-situational behavioral consistency (Epstein, 1979; Mischel & Peake, 1982). To support this inference, one could look at the consistency of only three SOC writing samples, which are still higher than those obtained by EAR, despite being several times smaller in terms of the total number of produced words.

Natural language use and its situational dependency. One can reasonably expect that lower consistency of EAR data stems from high contextual variability. Our results firmly support the standpoint of Mehl et al. (2012) highlighting the relevance of communication context when investigating natural language use and its psychological implications. The contextual variability is easily observable in our results, indicating the high intra-individual variance in LIWCser measures. Our findings suggest that introducing situations increased indices of stability in almost half of the LIWCser categories, but mostly those that were initially more temporally stable, i.e., with higher ICC_{avrg} coefficients. For example, the total Word Count, which shows the highest temporal stability, i.e., some people are simply more talkative than others, becomes even more consistent when we take into account the relative frequency of situations in which the person is talking to someone throughout the day. Situations like *talking, talking to others, talking over the phone*, were amongst the best predictors.

Our findings support the results of Ireland & Pennebaker (2010) showing that across social contexts, people tend to consistently and actively respond to others and synchronize with other participants in communication in the number of words, usage of function words, and words mapping psychologically relevant content like affective processes or personal concern. This synchronization in verbal interaction is called language style matching, or LSM (Ireland & Pennebaker, 2010; Ireland et al., 2011). Although we did not register the verbal output of the communicating partners, our findings are in line with the standpoint that natural communication with others increases the consistency of LSM categories across situations and time.

The results also revealed that another cluster of situations, i.e., being *outdoors* and being in *other public places* like sports halls, shopping malls, grocery stores, arcades, etc., significantly increases the reliability of LIWCser categories representing verbal immediacy and verbal emotional expression (see Mehl et al., 2012). Given that our sample consists of university students who spend time outdoors with close-others (romantic partners, friends, and classmates), the results indicating more personal and psychologically involved communication in these situations do not come as a surprise.

Another straightforward confirmation of the contextual dependency of natural language is the finding that the reliability of the category Insight increases significantly when respondents are attending *Lecture* at the university, that is, when in a formal setting like the university, people tend to use more words mapping cognitive processes like think, know, or consider.

The situation that significantly predicted the temporal stability of several LIWCser categories was a *TV on*. The consistency of usage of words related to Leisure increased when people were engaged in informal activities like watching TV. Additionally, this informal situation prompted respondents to more consistently use Informal words and swears. Finally, *TV on* increased the consistency of the category Biological processes, and in particular the category Ingestion. It could be that respondents in their spare time, when watching TV, also prepare meals, eat, and discuss food, therefore leading towards the use of words related to digestion, dieting, and drinking.

When respondents are expressing emotions, i.e., situations coded as *Laughing*, the temporal consistency of the categories Affective processes and Health increases. Our results support the evidence showing that laughter is a prototypical indicator of (joyful) affect in all human cultures (Panksepp, 2000). Also, when people are laughing, they tend to more consistently use words semantically related to the concept of health. Lastly, another verification that laughing is tightly related to verbal emotional expression is the result that when laughter is present, communication is more saturated with Assents. This indicates that when in entertaining and funny situations, people actively respond to the other participants in the communication.

Here we have to acknowledge that after introducing situations as predictors, the reliabilities of some categories remained the same – to be precise, the reliability of a substantial number of LIWCser categories did not change after introducing any of the SECSI situations in the model. This means, for example, that none of the coded situations affected the temporal consistency of the utterance of words representing social process or perceptual processes. Additionally, our results suggest that being *alone, reading*, or using a *computer* increased the temporal consistency of some LIWCser categories. This unexpected result calls for future investigation of the situational dependency of natural language. Consequently, it seems that the interplay between individuals, situations, and language use is quite complex, as it is not uniform across all aspects of language production. Still, we argue that introducing situations (or at least the limited number of situations as we did in this study) indeed has an effect on the temporal stability of the majority of LIWCser measures.

One should also bear in mind that in this study, we performed situational coding for a selected number of categories of aggregated data (both linguistic and situational) in order to ensure reliable estimates of the effects (our smallest temporal unit of analysis was a day, not a snippet). We can expect that situations would play a more prominent role if verbal behaviors were collected over a longer period of time (the reader should bear in mind that we had EAR data in only three time points). Furthermore, it is worth noting that our analysis was performed on a rather situationally homogeneous sample (i.e., more than half of the EAR snippets were recorded while the participant was in an apartment), which further limited the predictive power of situations.

Additional proof of a situational dependency of language use can be derived from a comparison between the results of Study 1 and Study 2. Namely, when assessing language in the same situation (i.e., SOC task), temporal stability of verbal behavior tends to be higher than when language data are collected across different contexts throughout the day (i.e., EAR). Based on the findings from both studies, we can conclude that taking into account situations when studying language use is very important, as the situational context imposes restrictions and introduces regulations and order into our verbal behavior.

Another important insight that the results provide is that not all LIWCser categories have a similar level of consistency. This does not mean that less consistent LIWCser categories are not worth studying. Our standpoint is that researchers should be aware of this when formulating expectations and interpreting results in studies using single-time measured LIWCser categories as indicators of habitual language. Furthermore, results show that the consistencies for the same category could differ greatly depending on the data collection method (e.g., words reflecting Sadness tend to be more consistent if collected by EAR than in the SOC task). Still, some aspects of language use, such as relative frequency of words reflecting cognitive processes, tend to be one of the most stable regardless of the data collection method. Finally, it should be noted that some language properties such as word count or usage of punctuation show high temporal consistency.

Choosing the method for collecting data on language use

Even though our results favor the SOC task over the EAR as a method for data collection, when opting for one of these techniques, researchers should be aware that both have advantages and drawbacks (other than temporal consistency), which may be important depending on the aim and the type of the research. Here we will discuss the main strong and weak points as well as some practical issues in order to enable researchers interested in studying the relationship between language use and different psychological phenomena to make an informed decision.

The EAR has many advantages. One of them is the potential to provide information that participants do not share easily or are unable to report due to limitations in self-awareness, social desirability biases, etc. (Wilson & Dunn, 2004). Furthermore, it is not as vulnerable to reactivity effects as SOC is, and it makes impression management much more difficult. Moreover, the EAR is the most direct observational method for collecting data on habitual language use in one's natural environment. In addition, the EAR is irreplaceable if one wants to study a coincidence between different environments and contexts and language use or everyday conversational language in relation to different social interactions. In spite of these advantages EAR, as observational AA, has some important limitations. It requires a certain level of cooperation since the device has to be worn as instructed. In addition, technical characteristics such as the duration of the battery or the storage size of the devices, may impose practical difficulties. Furthermore, collecting data using EAR is expensive in terms of time and financial resources, since each participant has to wear a device for several days, and it requires time-consuming and expensive preparation of the data (i.e., transcribing verbal output of participants). Finally, the EAR is to some extent intrusive, as it requires participants' compliance and is subject to self-selection bias (Manson & Robbins, 2017).

Compared to other techniques used for the collection of verbal material, the SOC task has several advantages. First, it is not structured in a way that limits the content or the style of verbal production and therefore enables the expression of individual differences in language use. Second, a product of this task could be considered a transient form between oral and written language (Celli & Polonio, 2015), as the context of writing is very informal, and therefore stylistic restrictions implicitly imposed on the respondent are minimal. Furthermore, the way a person follows his/her stream of thoughts represents the inner states and experiences in the closest way. To corroborate this claim, results show that independent raters assess the personality of the person based on the SOC more precisely than when transcription of the everyday speech is used (Holleran & Mehl, 2008). Finally, completing the SOC task, even ten times, is rated as a pleasant experience by the vast majority of participants (on average, pleasantness was rated 4.82 on the scale ranging from 1 to 5). That is because many find it beneficial to occasionally have a task that fosters them to focus on their inner thoughts and find it similar to keeping an intimate diary. On the flip side, collecting high-quality data depends on participants being open and motivated to share their intimate thoughts. As such, SOC is highly susceptible to impression management, which jeopardizes the validity of the task. Therefore, when using SOC, it is of utmost importance to motivate participants and ensure full confidentiality of their written content. Previous studies have shown that diary-like writing is beneficial to one's mental health (e.g., Niederhoffer & Pennebaker, 2009; Pennebaker & Seagal, 1999). In addition, providing participants with the opportunity to read all of their texts after the experiment serves as a good incentive.

LIWC as a tool for language analysis

Even though in our studies we used only one dictionary available as part of LIWC software (i.e., Serbian LIWC), we believe that the obtained results could provide useful insights into the potentials and limitations of LIWC as a general approach to studying individual differences in language use. Therefore, we will first focus on this type of language analysis in general, and then discuss the bandwidth of cross-linguistic generalization of the results.

There are several important properties of this type of language analysis that one should keep in mind when using LIWC or its adaptations to different languages. The basic idea is that the vocabulary people habitually use represents their stylistic behavior, which stands for "one's manner of performing adaptive acts", and is unintentional, spontaneous, and difficult to change (Pennebaker & King, 1999). Thus, each individual has their own unique stylistic behavior, such as the way they walk or smile. In accordance with this view, the assumption is that people have their own characteristic ways of expressing themselves through language. These stylistic differences are considered s individual differences (Pennebaker & Graybeal, 2001; Pennebaker et al., 2003). This idea is deeply embedded in the development of LIWC as a language analysis tool that performs a simple word count of what are psychologically relevant words (e.g., emotions, pronouns, etc.). Therefore, very important parts of the word count approach are decisions as to which words are going to be counted, and how they will be coded.

Overall, a predefined dictionary makes the analysis far less flexible than, for example, the open-vocabulary approach (OVA)[7]. It is important to keep in mind that LIWC's set of categories is just one of many possible ways to structure language products and categorize words. Nonetheless, closed-vocabulary word-category analysis such as LIWC enables studies on smaller samples and straightforward comparison of the results across studies, which is not the case in OVA. Furthermore, some criticize this approach because the words are analyzed in a decontextualized manner (Mehl, 2006). However, one could argue that this could even be an advantage over judge-based thematic content analysis, since the software by definition performs unbiased analysis, while human raters are susceptible to contextual influences. Another obstacle is that word-based analysis cannot grasp someone's irony, or sarcasm, or idioms (Mehl, 2006). In response to this, one could argue that most of the time, people use literal rather than figurative meaning. In addition, the content of the figurative meaning is tightly related to the thinking style of the person (Shen, 2006), and therefore we can assume that the use of a specific metaphor reflects certain characteristics of that person.

On the positive side, many agree that word-based analyses have made it possible to investigate aspects of psychological functioning that were unavailable prior to the development of this methodology (Ireland & Mehl, 2014; Mehl, 2006; Mehl & Gill, 2010; Pennebaker & King, 1999; Pennebaker et al., 2003; Tausczik & Pennebaker, 2010). We can now assess the characteristics of the verbal production that were completely neglected in traditional content analysis (Hart, 2001; Pennebaker & King, 1999; Pennebaker et al., 2003). The data derived from automatic text analysis do not share common method variance with other methods used in psychological research. These measures are objective: they ensure measurement equivalence across studies, and their metrics are not arbitrary (Ireland & Mehl, 2014; Mehl & Gill, 2010).

Cross-linguistic generalizability of the obtained results

Even though the data presented in this paper are based on one of the (to date) largest datasets of habitual language use (more than 700,000 words), it needs to be acknowledged that the majority of the research on individual differences in language use has been conducted in English. Therefore, one could question the cross-linguistic and cross-cultural generalizability of the findings presented. Naturally, some differences in relative frequencies of LIWC categories can be expected across languages due to differences in syntax and pragmatics. For example, when a text is translated from English to Serbian (or any other Slavic language), it will have 20% fewer function words and 50% more unique words and/or word forms. It is our opinion that these cross-linguistic differences would not significantly affect the relative differences in language use between people of the same linguistic background. In other words, we would not expect major differences in terms of overall consistency across different data collection methods. To that point, it should be noted that a comparable level of consistency was obtained in studies conducted in English. In addition, we see no reason to assume that the mechanisms driving language production are different between two Indo-European languages, i.e., English and Serbian. Nonetheless, in order to provide compelling evidence on cross-linguistic generalizability, we strongly encourage multisite replication of both studies, as LIWC is available in more than 15 languages.

Conclusion

Our findings showed that intra- and inter-individual differences can be detected in habitual language use. Thus, pursuing language correlates in personality, mental health, social behaviors, or cognitive functioning seems to be a promising path. Based on our findings, the SOC task can provide reliable data if administered at multiple time points, while EAR data should probably be aggregated across longer periods of time, to account for high within-person variability, and combined with situational analysis to increase the reliability of some language categories. Overall, our data suggest that relatively consistent and reliable estimates of language use can be extracted from SOC and EAR using LIWCser software.

Acknowledgements This study was supported by the Ministry of Education, Science and Technological Development, Serbia, project no. 179018.

[7] The open-vocabulary approach is a set of language analysis techniques that perform comprehensive exploration of word patterns without a priori defined categories (see e.g,. Schwartz et al., 2013).

Appendix

Table 4 Average representation of LIWC categories in EAR and SOC task

LIWC category	EAR	*SOC*
Function words	38.83	39.28
Pronouns	14.39	13.24
Personal pronouns	6.91	4.77
I	3.99	4.03
We	1.29	1.78
You	1.64	0.29
S/he	5.68	3.96
They	0.52	0.49
Impersonal pronouns	5.94	5.44
Common verbs	9.07	7.50
Auxiliary verb	6.31	5.63
Past	0.14	0.18
Present	2.18	1.59
Future	0.62	0.84
Adverbs	4.17	5.19
Prepositions	4.55	6.27
Conjunctions	13.46	13.27
Quantifiers	2.44	3.94
Numbers	2.09	1.00
Informal/Swear words	0.34	0.13
Social processes	4.09	3.17
Family	0.27	0.31
Friends	0.08	0.24
Humans	0.41	0.70
Affective processes	4.54	6.49
Positive emotions	2.74	3.59
Negative emotions	1.27	2.11
Fear and anxiety	0.20	0.32
Anger and resentment	0.11	0.20
Sadness	0.06	0.23
Cognitive processes	16.73	19.22
Insight	1.85	1.69
Causation	3.12	1.99
Discrepancy	1.67	2.42
Tentative	2.56	3.01
Certainty	2.10	2.85
Inhibition	2.58	2.68
Inclusive	0.65	1.09
Exclusive	2.22	2.41
Perceptual processes	1.92	2.14
See	0.66	0.68
Hear	0.89	0.60
Feel	0.17	0.26
Biological processes	1.08	1.44
Body	0.34	0.40
Health	0.17	0.20
Sex and love	0.20	0.42
Ingestion	0.37	0.37
Negations	4.31	3.81
Relativity	7.75	10.58
Motion	1.22	1.41
Space	2.30	2.46
Time	2.31	4.02
Work	0.94	1.37
Achievement	0.38	0.77
Leisure	0.38	0.76
Home	0.19	0.29
Money	0.32	0.20
Religion	0.15	0.12
Death	0.05	0.06
Assent	1.26	
Nonfluencies	0.07	
Fillers	0.79	

Note: All measures represent % of a total number of words in the text

References

Baddeley, J. L., Pennebaker, J. W., & Beevers, C. G. (2013). Everyday social behavior during a major depressive episode. *Social Psychological and Personality Science, 4,* 445–452. doi: https://doi.org/10.1177/1948550612461654

Bjekić, J. (2016). *Stilističke i sadržinske karakteristike verbalne produkcije kao indikatori bazične ličnosti* (*The style and content of verbal production as indicators of basic personality structure*). Doctoral dissertation. Department of Psychology, Faculty of Philosophy, University of Belgrade.

Bjekić, J., Lazarević, Lj. B., Erić, M., Stojimirović, E., & Djokić, T. (2012). Razvoj srpske verzije rečnika za automatsku analizu teksta (LIWCser) (Development of Serbian dictionary for automatic text analysis-LIWCser). *Psihološka Istraživanja, 15*(1), 85-110.

Bjekić, J., Lazarević, Lj. B., Živanović, M., & Knežević, G. (2014). Psychometric evaluation of the Serbian dictionary for automatic text analysis—LIWCser. *Psihologija, 47*(1), 5-32. doi:https://doi.org/10.2298/PSI1401005B

Boyd, R. L., & Pennebaker, J. W. (2015). Did Shakespeare write Double Falsehood? Identifying individuals by creating psychological signatures with text analysis. *Psychological Science, 26*(5), 570-582. DOI:https://doi.org/10.1177/0956797614566658

Celli, F., & Polonio, L. (2015). Facebook and the Real World: Correlations between online and offline conversations. In *Natural language processing* (pp. 1–5).

Chen, J., Hsieh, G., Mahmud, J. U., & Nichols, J. (2014). Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM conference on computer supported cooperative work & social computing* (pp. 405–414). ACM: New York.

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *Journal of Nervous and Mental Disease, 175*(9), 526-537. DOI:https://doi.org/10.1097/00005053-198709000-00004

Csikszentmihalyi, M., & Larson, R. (2014). Validity and reliability of the experience-sampling method. In M. Csikszentmihalyi (Ed.), *Flow and the foundations of positive psychology* (pp. 35-54). Berlin: Springer Netherlands. DOI:https://doi.org/10.1007/978-94-017-9088-8_3

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3267-3276). New York, NY: ACM. DOI:https://doi.org/10.1145/2470654.2466447

Đurić-Jočić, D., Džamonja-Ignjatović, T., Knežević, G. (2004). *NEO PI-R. Primena i interpretacija. (NEO PI-R. Administration and interpretation).* Centar za Primenjenu Psihologiju: Beograd.

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37*(7), 1097-1126. DOI:https://doi.org/10.1037/0022-3514.37.7.1097

Gill, A. J., Oberlander, J., & Austin, E. (2006). Rating e-mail personality at zero acquaintance. *Personality and Individual Differences, 40*, 497-507. doi:https://doi.org/10.1016/j.paid.2005.06.027

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

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

Holleran, S. E., & Mehl, M. R. (2008). Let me read your mind: Personality judgments based on a person's natural stream of thought. *Journal of Research in Personality, 42*(3), 747-754. doi:https://doi.org/10.1016/j.jrp.2007.07.011

Holtgraves, T. (2011). Text messaging, personality, and the social context. *Journal of Research in Personality, 45,* 92-99. DOI:https://doi.org/10.1016/j.jrp.2010.11.015

Ireland, M. E., & Mehl, M. R. (2014). Natural language use as a marker of personality. In T. Holtgraves (Ed.), *Oxford handbook of language and social psychology* (pp. 22-35). New York, NY: Oxford University Press.

Ireland, M. E., & Pennebaker, J. W. (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology*, *99*(3), 549-571. doi: https://doi.org/10.1037/a0020386

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*(1), 39-44. doi: https://doi.org/10.1177/0956797610392928

James, W. (1890). *The principles of psychology.* New York: H. Holt and Company.

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Dziurzynski, L., Ungar, L. H., Stillwell, D. J., … & Seligman, M. E. (2014). The online social self: An open vocabulary approach to personality. *Assessment*, *21*(2), 158-169. doi:https://doi.org/10.1177/1073191113514104

Knežević, G., Džamonja-Ignjatović, T., & Đurić-Jočić, D. (2004). *Petofaktorski model ličnosti (The five-factors model of personality).* Beograd: Centar za Primenjenu Psihologiju.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. DOI:https://doi.org/10.1016/j.jcm.2016.02.012

Le, B., Hat, N. C., & Beal, D. J. (2006). Pocket-sized psychology studies: Exploring daily diary software for PalmPilots. *Behavior Research Methods, 38*, 325-332. DOI:https://doi.org/10.3758/BF03192784

Lee, C. H., Kim, K., Seo, Y. S., & Chung, C. K. (2007). The relations between personality and language use. *The Journal of General Psychology, 134*(4), 405-413. doi:https://doi.org/10.3200/GENP.134.4.405-414

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. DOI:https://doi.org/10.1613/jair.2349

Manson, J. H., & Robbins, M. L. (2017). New evaluation of the electronically activated recorder (EAR): Obtrusiveness, compliance, and participant self-selection effects. *Frontiers in Psychology, 8,* 658. doi:https://doi.org/10.3389/fpsyg.2017.00658

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46. doi:https://doi.org/10.1037/1082-989X.1.4.390

McMain, S., Links, P. S., Guimond, T., Wnuk, S., Eynan, R., Bergmans, Y., & Warwar, S. (2013). An exploratory study of the relationship between changes in emotion and cognitive processes and treatment outcome in borderline personality disorder. *Psychotherapy Research*, *23*(6), 658-673. DOI:https://doi.org/10.1080/10503307.2013.838653

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

Mehl, M. R. (2017). The electronically activated recorder (EAR): A Method for the naturalistic observation of daily social behavior. *Current Directions in Psychological Science, 26*, 184-190. doi: https://doi.org/10.1177/0963721416680611

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

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), 862-77. doi:https://doi.org/10.1037/0022-3514.90.5.862

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology, 84*(4), 857-870. doi:https://doi.org/10.1037/0022-3514.84.4.857

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The electronically activated recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, *33*(4), 517-523. doi:https://doi.org/10.3758/BF03195410

Mehl, M. R., & Robbins, M. L. (2012). Naturalistic observation sampling: The electronically activated recorder (EAR). In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 176-192). New York, NY: Guilford.

Mehl, M. R., Robbins, M. L., & Holleran, S. E. (2012). How taking a word for a word can be problematic: Context-dependent linguistic markers of extraversion and neuroticism. *Journal of Methods and Measurement in the Social Sciences, 3*(2), 30-50. doi:https://doi.org/10.2458/azu_jmmss_v3i2_mehi

Mischel, W., & Peake, P. K. (1982). Beyond deja vu in the search for cross-situational consistency. *Psychological Review, 89*(6), 730-755.

Molendijk, M. L., Bamelis, L., van Emmerik, A. A., Arntz, A., Haringsma, R., & Spinhoven, P. (2010). Word use of outpatients with a personality disorder and concurrent or previous major depressive disorder. *Behavior Research and Therapy*, *48*(1), 44-51. doi: https://doi.org/10.1016/j.brat.2009.09.007

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event-and interval-contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin*, *27*(7), 771-785. DOI:https://doi.org/10.1177/0146167201277001

Niederhoffer, K. G., & Pennebaker, J. W. (2009). Sharing one's story: On the benefits of writing or talking about emotional experience. In S. J. Lopez & C. R. Snyder (Eds.), *Oxford library of psychology. Oxford handbook of positive psychology* (pp. 621-632). New York, NY: Oxford University Press.

Panksepp, J. (2000). The riddle of laughter: Neural and psychoevolutionary underpinnings of joy. *Current Directions in Psychological Science*, *9*(6), 183-186.

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), 934-952. https://doi.org/10.1037/pspp0000020

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 LIWC2007, manual.* New Zealand: The University of Texas at Austin and The University of Auckland. LIWC.net

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC 2001*. Mahwah, NJ: Erlbaum.

Pennebaker, J. W., & Graybeal, A. (2001). Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science, 10*(3), 90-93.

Pennebaker, J. W., & King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 12961-1312.

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:https://doi.org/10.1146/annurev.psych.54.10161.145041

Pennebaker, J. W., & Seagal, J. D. (1999). Forming a story: The health benefits of narrative. *Journal of Clinical Psychology*, *55*(10), 1243-1254.

Pfeil, U., Arjan, R., & Zaphiris, P. (2009). Age differences in online social networking—A study of user profiles and the social capital divide among teenagers and older users in MySpace. *Computers in Human Behavior*, *25*(3), 643-654.

Piolat, A., Booth, R. J., Chung, C. K., Davids, M., & Pennebaker, J. W. (2011). La version francaise du dictionnaire pour le LIWC: Modalites de construction et exemplesd'utilisation. *Psychologie Française, 56*(3), 145-159. DOI:https://doi.org/10.1016/j.psfr.2011.07.002

Ramírez-Esparza, N., Pennebaker, J. W., García, F. A., & Suriá, R. (2007). The psychology of word use: A computer program that analyzes texts in Spanish. *Revista Mexicana de Psicologia, 24*(1), 85-99.

Raudenbush, S., Bryk, T., & Congdon, R. (2000). *HLM 6 hierarchical linear and nonlinear modeling.* Scientific Software International, Inc: Lincolnwood.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

Rosenbach, C., & Renneberg, B. (2015). Remembering rejection: Specificity and linguistic styles of autobiographical memories in borderline personality disorder and depression. *Journal of Behavior Therapy and Experimental Psychiatry*, *46*, 85-92. DOI: https://doi.org/10.1016/j.jbtep.2014.09.002

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. https://doi.org/10.1371/journal.pone.0073791

Shen, Y. (2006). Figures of speech. In K. Brown (Ed.), *Encyclopedia of language & linguistics* (pp. 459-464). Oxford: Elsevier. doi:https://doi.org/10.1016/B0-08-044854-2/00510-1

Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment in behavioral medicine. *Annals of Behavioral Medicine, 16*(3), 199-202.

Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012). Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In *2012 11th International Conference on Machine Learning and Applications* (ICMLA), (Vol. 2, pp. 386–393). IEEE.

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*(1), 24-54. doi:https://doi.org/10.1177/0261927X09351676

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, *9*, 151-176. DOI:https://doi.org/10.1146/annurev-clinpsy-050212-185510

Trull, T. J., & Ebner-Priemer, U. (2014). The role of ambulatory assessment in psychological science. *Current Directions in Psychological Science*, *23*(6), 466-470. DOI:https://doi.org/10.1177/0963721414550706

Vaisey, S., & Miles, A. (2014). Tools from moral psychology for measuring personal moral culture. *Theory and Society*, *43*(3-4), 311-332. doi:https://doi.org/10.1007/s11186-014-9221-8

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology*, *55,* 493-518. DOI:https://doi.org/10.1146/annurev.psych.55.090902.141954

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

Yee, N., Harris, H., Jabon, M., & Bailenson, J. N. (2010). The expression of personality in virtual worlds. *Social Psychological and Personality Science, 2*(1), 5-12. doi:https://doi.org/10.1177/1948550610379056

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), een gecomputeriseerd tekstanalyzeprogramma. *Gedrag & Gezondheid, 32*, 273-283.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.