Psihološka istraživanja, Vol. XV (1) 2012.
85–110

UDK 159.9.07 ; 81'23
TIP ČLANKA: Originalni naučni rad

Razvoj srpske verzije rečnika za automatsku analizu teksta (LIWCser)[1]

Jovana Bjekić[2]
Institut za medicinska istraživanja, Univerzitet u Beogradu

Ljiljana Lazarević
Institut za psihologiju, Filozofski fakultet, Univerzitet u Beogradu

Milica Erić, Elena Stojimirović, Teodora Đokić
Beograd

Automatska analiza teksta je metodološki pristup analizi individualnih razlika u verbalnoj produkciji (ponašanju) koji omogućava ekstrakciju statistički manipulabilnih informacija o intenzitetu i/ili frekvenci tematskih i stilističkih karakteristika verbalnih produkata. LIWC (Linguistic Inquiry and Word Count) jedan je od najzastupljenijih programa za automatsku analizu teksta, koji analizu obavlja upoređivanjem odrednica u rečniku sa odrednicama u tekstu i beleženjem relativne zastupljenosti svake od kategorija u datom uzorku teksta. Istraživanja ukazuju na to da mere dobijene obradom verbalnog ponašanja imaju potencijal da objasne odnose između mera dobijenih primenom implicitnih i tradicionalnih eksplicitnih mera, nezavisno od predmeta merenja (stavovi, psihopatološki potencijal, procena bazičnih dimenzija ličnosti itd.). Cilj ovog rada jeste konstrukcija rečnika za srpski jezik LIWCser. Konstrukcija rečnika odvijala se u četiri faze: prevod sadržaja engleskog rečnika, kreiranje odrednica, klasifikacija odrednica pomoću apsolutnog konsenzusa između četiri nezavisna procenjivača (jednu odrednicu moguće je klasifikovati u više kategorija u zavisnosti od konteksta u kojem se upotrebljava) i revizija sadržaja kategorija i formiranje konačnog skupa odrednica. Konačna verzija rečnika LIWCser sadrži 12103 odrednice, kalsifikovane u 65 kategorija (lingvističkih, psiholoških i tematskih). Od ukupnog broja odrednica, samo četiri (0,03%) je klasifikovano u osam kategorija, 22 (0,2%) u sedam, 147 (1,2%) u šest, a 568 (4,7%) u pet kategorija. U četiri kategorije klasifikovano je 1531 (12,6%) odrednica, u tri 2913 (24,1%) odrednica, u dve 4800 (39,7%) odrednica, dok je 2118 (17,5%) odrednica klasifikovano samo u jednu kategoriju. Razvojem rečnika

1 Članak je rezultat rada na projektu „Identifikacija, merenje i razvoj kognitivnih i emocionalnih kompetencija važnih društvu orijentisanom na evropske integracije" (179018), čiju realizaciju finansira Ministarstvo prosvete i nauke Republike Srbije.

2 bjekicjovana@gmail.com

LIWCser otvara se mogućnost prikupljanja mera verbalnog ponašanja i dalja istraživanja odnosa implicitnih i eksplicitnih mera u različitim oblastima psihologije.

Ključne reči: automatska analiza teksta, rečnik srpskog jezika za automatsku analizu teksta LIWCser, verbalno ponašanje, implicitne i eksplicitne mere

Uvod

Ideja da sadržaj verbalne produkcije odslikava mentalna, socijalna i fizička stanja osobe prisutna je od početka razvoja psihologije, pa su tako, na primer, prvi psiholozi, a posebno psihoanalitičari, tvrdili da je ključ za razumevanje ličnosti i nesvesnog moguće naći u jeziku (Frojd, 1969; Tausczik & Pennebaker, 2010). Međutim, istraživači (uglavnom iz oblasti kliničke, socijalne i psihologije individualnih razlika) tek su poslednjih godina počeli da sistematski istražuju način na koji je upotreba reči u vezi s važnim psihološkim konstruktima (Fast & Funder, 2008; Hirsh & Peterson, 2009; Mehl, Gosling & Pennebaker, 2006).

Analizi teksta moguće je pristupiti iz dva široka teorijsko-metodološka okvira, kvalitativnog i kvantitativnog (Pennebaker, Mehl & Niederhoffer, 2003). Kvalitativna analiza ima korene u psihoanalitičkoj tradiciji i bazira se na ideji da je jezik pre svega kontekstualan, te da je izučavanje jezika neophodno vršiti unutar konteksta u kojem se jezička produkcija odvija (Pennebaker & King, 1999). S druge strane, kvantitativni pristup bazira se na ekstrakciji statistički manipulabilnih podataka, putem eksplicitnih kriterijuma klasifikacije i kvantifikacije.

U okviru kvantitativnog pristupa autori su se najpre bavili *tematskom analizom teksta* u kojoj procenjivači, na osnovu unapred definisanog sistema kodovanja, procenjuju postojanje određene teme unutar teksta ili skupa tekstova (Smith, 1992). Ipak, pokazalo se da tematska analiza teksta nosi određene metodološke teškoće (Mehl & Gill, 2010). Naime, u tematskoj analizi teksta teško je postići apsolutno slaganje procenjivača uprkos postojanju eksplicitnih kriterijuma za analizu. Osim toga, sam kontekst nameće problem pristrasnosti procenjivača u proceni sitnijih jedinica teksta (npr. sintagme, reči itd.). Dodatno, ovakav pristup u analizi teksta često iziskuje znatne vremenske i materijalne resurse (Mehl & Gill, 2010).

Međutim, tehnološki razvoj doveo je do razvoja *automatske analize teksta,* specifične po upotrebi informacionih tehnologija u ekstrakciji statistički manipulabilnih informacija o prisustvu, intenzitetu i/ili frekvenci tematskih i/ili stilističkih karakteristika teksta (Shapiro & Markoff, 1997). Posmatrano iz metodološkog ugla, automatska analiza teksta ima nekoliko veoma značajnih prednosti (Mehl & Gill, 2010). Pre svega, pošto analizu vrši računar na osnovu unapred definisanog algoritma, podaci dobijeni na ovaj način objektivni su, proverljivi i replikabilni. Osim toga, primenom ove metode minimizuje se greška merenja, koja je posledica individualnih razlika između procenjivača i omogućava se maksimalna metodološka ekvivalentnost različitih studija

koje koriste isti program za analizu teksta. Takođe, podaci prikupljeni na ovaj način ne dele metodsku varijansu sa eksplicitnim metodama koje se često koriste u psihologiji (npr. mere samoizveštaja, procene od strane eksperta, procene od strane bliskih drugih, upitnici itd.; Mehl & Gill, 2010).

U okviru automatske analize teksta razvijen je veći broj programa koji analizu obavljaju na nivou pojedinačnih reči. Postavka koja leži u osnovi ovakvog pristupa jeste da individualne razlike u učestalosti upotrebe pojedinačnih reči odslikavaju individualne razlike u mišljenju, osećanjima i stavovima (Pennebaker et al., 2003). Naime, kognitivni kapaciteti osobe usmereni su na konceptualizaciju ideje, a način na koji će ideja biti iskazana zavisi pre svega od unutrašnjih stanja i karakteristika osobe. U prilog ovoj pretpostavci govore rezultati istraživanja koji su pokazali da se u svakodnevnoj komunikaciji proces izbora pojedinačnih reči obavlja u velikoj meri automatski i da stoga analiza teksta na nivou pojedinačnih reči omogućava sticanje uvida u psihološki relevantne procese i karakteristike osobe (Hart, 2001; Lazarević, 2012; Pennebaker et al., 2003).

Programi za automatsku analizu teksta

Jedan od prvih programa za automatsku analizu teksta koji je baziran na strategiji prebrojavanja reči bio je *The General Inquirer* (Stone, Dunphy, Smith, & Ogilvie, 1966). Ovaj program radi na platformi tri tematska rečnika, pri čemu je jedan od njih bio konstruisan za potrebe skorovanja TAT protokola (Stone et al., 1966). Prednost ovog programa jeste u tome što je jedinstven po svojoj fleksibilnosti (koja se ogleda u tome da je moguće uključiti nov rečnik za potrebe istraživanja) i u tome što je omogućavao elegantan način rešavanja problema homonimije. Međutim, sadržinski je veoma specifičan, te je njegova primena bila ograničena na tekstove prilagođene tumačenju unutar psihoanalitičke paradigme, jer je konstrukcija novih rečnika izuzetno zahtevan posao (Pennebaker et al., 2003).

Drugi, takođe sadržinski specifičan, program za automatsku analizu teksta novijeg datuma jeste *TAS/C,* konstruisan sa ciljem istraživanja ključnih momenata psihoterapijskog procesa (Margenthaler, 1996). Taj program se u analizi teksta fokusirao na dve nepreklapajuće dimenzije, emocionalni ton i stepen apstrakcije. Međutim, uprkos tome što su navedene dimenzije izuzetno značajne za analizu psihoterapijskih seansi, pokazalo se da ovaj program „pokriva" mali procenat celokupnog teksta koji se analizira (manje od 10%), što ga je činilo nedovoljno obuhvatnim za primenu izvan konteksta psihoterapije (Pennebaker et al., 2003).

Istraživači iz oblasti socijalne psihologije su 80-ih godina XX veka razvili program za analizu političkih govora, specifičan za procenu verbalnog tona političkih govora (DICTION, Hart, 1984; 2001). Taj program vršio je analizu u okviru pet statistički nezavisnih nadređenih varijabli (aktivitet, optimizam, sigurnost, realizam i uobičajnost), od kojih je svaka bila sastavljena iz niza

lingvističkih subkarakteristika (npr. optimizam: pohvala, zadovoljstvo, inspiracija, krivica, negiranje itd.). Pošto nadređene kategorije nisu bile specifične samo za političke govore, i pošto je program mogao da se poboljšava upotpunjavanjem baze reči na osnovu podataka koji su se dobijali procesiranjem novih tekstova, njegova upotreba znatno se proširila i na marketing, analizu medija, javne debate i slično (Hart, 2001).

Program za automatsku analizu teksta LIWC

Trenutno najzastupljeniji program za automatsku analizu teksta koji se bazira na analizi pojedinačnih reči jeste LIWC (eng. *Linguistic Inquiery and Word count*). Do danas su razvijene dve verzije ovog programa, LIWC2001 (Pennebaker, Francis, & Booth, 2001) i revidirana verzija LIWC2007 (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007), koje su pokazale veoma dobre metrijske karakteristike.

Taj program inicijalno je razvijen za potrebe istraživanja uticaja ekspresivnog pisanja na duševno zdravlje (Pennebaker et al., 1993), ali je tokom godina prilagođavan istraživačkim potrebama, te je uključen veći broj semantičkih kategorija i proširen njihov sadržaj (Pennebaker et al., 2007). Na taj način, autori programa omogućili su istraživačima iz praktično svih oblasti psihologije primenu automatske analize teksta (Chung & Pennebaker, 2007; Pennebaker et al., 2003; Pennebaker et al., 2007; Tausczik & Pennebaker, 2010).

LIWC program obavlja rednu analizu teksta, pri čemu je jedinica analize pojedinačna reč. Dizajniran je tako da upoređuje grafemski složaj svake jedinice u tekstu sa grafemskim složajima u rečniku inkorporiranom u program. Rečnik se sastoji iz mnoštva grafemskih složajeva – odrednica, koje su klasifikovane u određene kategorije, i to tako da jedna odrednica može pripadati jednoj ili većem broju kategorija kako bi se u što većoj meri omogućilo prevazilaženje problema konteksta u kojem se pojedina reč javlja. Kada program pronađe odrednicu koja je korespondentna grafemskom složaju iz rečnika, on registruje da se javila kategorija kojoj pripada data odrednica. Nakon procesiranja celokupnog teksta program daje informaciju o proporciji javljanja svake od definisanih kategorija u analiziranom tekstu. Drugim rečima, konačan ishod analize jeste kvantitativna mera zastupljenosti svake od kategorija u datom uzorku teksta.

LIWC2007 Rečnik. Engleski rečnik programa LIWC2007 sadrži oko 4500 odrednica, klasifikovanih u četiri skupine, koje sačinjavaju 63 kategorije[3] (Pennebaker et al., 2007). Prvu grupu čini 21 *lingvistička kategorija* u koje spadaju standardne gramatičke kategorije kao što su glagoli, pomoćni glagoli, zamenice, prilozi, predlozi, itd. Drugu grupu čine 32 hijerarhijski organizovane *psihološke kategorije*. U nadređene psihološke kategorije svrstani su *socijalni*, *afektivni*, *kognitivni*, *biološki* procesi i kategorije koje se odnose na relativitet. U okviru svake od navedenih nadređenih kategorija postoji veći broj kate-

3 U prilogu 1 se nalazi prikaz strukture engleskog rečnika za program LIWC2007.

gorija nižeg reda. Na primer, u okviru kategorije *socijalni procesi* postoje tri pod-kategorije: *porodica*, *prijatelji* i *ljudi*. Treću grupu čini 7 *tematskih kategorija*: *posao*, *postignuće*, slobodno vreme, kuća, novac, *smrt*, i *religija*. Četvrtu grupu čine tri *paralingvističke kategorije*, specifične za analizu usmenog govora (*poštapalice*, *potvrđivanje* i *nefluentnosti*). Međutim, osim prethodno navedenih, ispis analize daje i informacije o *opštim deskriptorima teksta*, kao što su ukupan broj reči, prosečan broj reči u rečenici, procenat teksta obuhvaćen rečnikom, broj reči dužih od šest slova, kao i učestalost upotrebe različitih intrepunkcijskih znakova (Pennebaker et al., 2007).

Pozicija i značaj jezičkih pokazatelja LIWC-a u psihološkim istraživanjima

Uključivanjem velikog broja kategorija LIWC je omogućio analizu gramatičkih i sintaksičkih karakteristika testa, čime je bar delimično bio prevaziđen problem sadržaja i konteksta. Dosadašnja istraživanja su pokazala da to omogućava, pre svega, analiza *lingvističkih kategorija* i *nepunoznačnih reči*, kao što su zamenice, članovi, konjukcije, pomoćni glagoli, pa čak i tzv. *poštapalice*, pri čemu se ima u vidu da upravo ove kategorije reči imaju ulogu da povezuju sadržaj reči (Chung & Pennebaker, 2007; Pennebaker et al., 2003).

O značaju *lingvističkih* kategorija za psihološka istraživanja, govore nalazi da se reči koje pripadaju ovim kategorijama najdoslednije upotrebljavaju bez obzira na kontekst i sadržaj verbalne produkcije (Pennebaker et al., 2003). Rezultati istraživanja pokazuju veću stabilnost i konzistentnost u upotrebi reči koje pripadaju lingvističkim kategorijama tokom vremena (sa prosečnim test-retest korelacijama za standardne lingvističke varijable 0.41), u odnosu na reči koje mapiraju psihološke procese (gde test-retest korelacije iznose 0.24), kao i da su lingvističke kategorije konzistente u pogledu socijalnog konteksta (Mehl, Pennebaker, Crow, Dabbs, & Price, 2001; Mehl & Pennebaker, 2003; Pennebaker & King, 1999).

Takođe, kao što je poznato, najveći broj jezika sadrži ograničeni broj ličnih zamenica, koje se razlikuju u pogledu lica i broja. Da učestalost upotrebe različitih zamenica može biti psihološki relevantna, pokazuju nalazi da je učestalost upotrebe zamenica u prvom licu jednine povezana sa uzrastom, polom i depresijom (Pennebaker et al., 2003), dok je upotreba zamenica u prvom licu množine povezana sa grupnim identitetom i emocionalnim distanciranjem (Pennebaker & Lay, 2002). Dodatno, pokazalo se da učestalost upotrebe zamenica u drugom i trećem licu predstavlja marker socijalne svesnosti i angažovanosti govornika (Pennebaker et al., 2003), učestalija upotreba zamenica u prvom licu sugeriše da je osoba u verbalnoj produkciji centrirana na sebe, dok učestalija upotreba zamenica u drugom licu sugeriše centriranost na druge osobe (Pennebaker et al., 2003).

Istraživanja ukazuju i na to da upotreba nepunoznačnih reči govori o psihološkim stanjima nezavisno o konteksta, odnosno o kognitivnim procesi-

ma, emocijama i motivima (Chung & Pennebaker, 2007; Newman, Groom, Handelman, & Pennebaker, 2008). Na osnovu ovih rezultata, čini se da su lingvističke kategorije verovatno najpouzdaniji pokazatelji lingvističkog stila, a samim tim i stabilnih karakteristika osobe, kao što su bazične dimenzije ličnosti. Tako je dimenzija Ekstraverzija povezana sa ukupnim brojem reči koje osoba produkuje, upotrebom članova, negacija, brojeva i ličnih zamenica (Heylighen & Dewaele, 2002; Hirsh & Peterson, 2009; Mairesse, Walker, Mehl & Moore, 2007; Mehl et al, 2006; Pennebaker & King, 1999). Takođe, utvrđeno je da je Ekstraverzija u vezi sa nižim odnosom globalnih i lokalnih verovatnoća (*type/token,* koji ukazuje na raznovrsnost jezika koji osoba koristi i koji se računa kao količnik broja različitih reči i ukupnog broja reči) i manje formalnim jezikom, dok je Introverzija u vezi sa širim opsegom jezika koji osoba koristi (Mairesse et al., 2007). Kada je reč o odnosu stilističkih kategorija i dimenzije Neuroticizam, utvrđeno je da je ova dimenzija u vezi sa upotrebom zamenica u prvom licu jednine (Pennebaker & King, 1999). Mel i saradnici su utvrdili da je dimenzija Otvorenost u negativnoj korelaciji sa upotrebom zamenica trećeg lica i sa glagolima u prošlom vremenu (Mehl et al., 2006). Dimenzija Saradljivosti je u negativnoj korelaciji sa zamenicama jednine u prvom licu (Pennebaker & King, 1999), dok je dimenzija Savesnosti u pozitivnoj korelaciji sa upotrebom zamenica u drugom licu (Mehl et al., 2006). Kada je reč o dimenziji Savesnost istraživanje Mela i saradnika pokazalo je i polne razlike u upotrebi *poštapalica* (tj. filera), gde su muškarci sa višim skorovima na ovoj dimenziji produkovali, za razliku od žena, i više reči iz ove kategorije (Mehl et al., 2006).Isti autori navode i da je ova dimenzija u negativnoj korelaciji s količinom upotrebe psovki i u pozitivnoj korelaciji sa upotrebom zamenica drugog lica (Mehl et al., 2006).

Penebeker i saradnici posvetili su posebnu pažnju, osim lingvističkim kategorijama, razvijanju kategorija za koje veruju da nose esencijalno psihološko značenje, odnosno analizi sadržinskih karakteristika teksta (Pennebaker et al., 2001). One su od posebne važnosti jer pružaju uvid u doživljaje, osećanja i način mišljenja osobe. Do sada se najveći broj istraživanja, posebno u kliničkoj psihologiji, fokusirao na kategoriju *Afektivni procesi.* Tako je ovaj program korišćen za istraživanja afektivnih procesa kod depresivnih (Gortner, Rude, & Pennebaker, 2006), post-traumatskog stresnog poremećaja (Orsillo, Batten, Plumb, Luterek & Roessner, 2004) i psihotičnih poremećaja (Cohen, Alpert, Nienow, Dinzeo & Docherty, 2008; Cohen, St-Hilaire, Aakres & Docherty, 2009), u kojima je utvrđeno da postoje veze između određenih psihopatoloških stanja i verbalnog ponašanja.

Kada je reč o istraživanju nekliničkih fenomena, jedan broj istraživanja pokazao je da postoje relacije između frekventnosti upotrebe reči koje pripadaju psihološkim kategorijama i bazičnih dimenzija ličnosti (Mairesse et al., 2007). Pokazano je, na primer, da je dimenzija Neuroticizam u pozitivnoj korelaciji sa upotrebom reči koje odražavaju negativne emocije i u negativnoj korelaciji s rečima koje odražavaju pozitivne emocije (Pennebaker & King, 1999). Takođe,

istraživanja pokazuju da je dimenzija Ekstraverzija u pozitivnoj korelaciji s rečima koje mapiraju pozitivne emocije, kao i s rečima koje se odnose na socijalne procese (Pennebaker & King, 1999). Utvrđeno je takođe da je dimenzija Saradljivost u pozitivnoj korelaciji s rečima koje odražavaju pozitivne emocije (Mehl et al., 2006; Pennebaker & King, 1999), i u negativnoj korelaciji s rečima koje odražavaju negativne emocije (Pennebaker & King, 1999).

Neki od dosadašnjih nalaza pokazuju da *tematske* kategorije pružaju informaciju o zastupljenosti određene teme u uzorku teksta, a pretpostavlja se da je zastupljenost teme dobar indikator značaja koji osoba pripisuje toj temi (Stirman & Pennebaker, 2001). U prilog tome govore nalazi dobijeni analizom poezije suicidalnih i nesuicidalnih pesnika. Naime, istraživanje je pokazalo značajno veću zastupljenost reči koje pripadaju tematskoj kategoriji *smrt* u grupi suicidalnih autora (Stirman & Pennebaker, 2001). Međutim, odnos jezičkih pokazatelja iz grupe tematske kategorije i psiholoških konstrukata i dalje predstavlja polje s velikim istraživačkim potencijalima.

Brojni pomenuti nalazi ukazuju na veliki potencijal koji imaju istraživanja odnosa jezika i psiholoških fenomena pomoću LIWC programa. Do danas je razvijen veći broj LIWC rečnika za analizu tekstova na različitim jezicima: holandski (Zijlstra, van Meerveld, van Middendorp, Pennebaker, & Geenen, 2004), italijanski (Alparone, Caso, Agosti, & Rellini, 2004), korejski (Lee, Shim, and Yoon, 2005), španski (Ramirez-Esparza, Pennebaker, Garcia, & Suria, 2007), nemački (Wolf, Horn, Mehl, Haug, Pennebaker, & Kordy, 2008), arapski (Hayeri, Chung, & Pennebaker, 2010), francuski (Piolat, Booth, Chung, Davids, & Pennebaker, 2011), ruski (Kailer, & Chung, 2011), turski (Murderrisoglu, 2011), i kineski (Huang, Chung, Hui, Lin, Seih, Chen et al, in press).

Cilj rada

Velika raznovrsnost informacija koju pruža LIWC uticala je na ekspanziju upotrebe ovog programa u cilju razumevanja načina na koji se različiti psihološki fenomeni manifestuju u verbalnoj produkciji. Razvoj rečnika na različitim jezicima omogućio je sprovođenje istraživanja na neengleskom govornom području i, pre svega, kros-jezičku evaluaciju nalaza dobijenih analizom tekstova na engleskom jeziku (Kroner-Herwig, Linkemann & Morris, 2004; Lee, Kim, Seo & Chung, 2007; Yogo & Fujihara, 2008). Dalje, na taj način bili su stvoreni uslovi za kros-kulturalna poređenja, istraživanja bilingvalizma i usvajanja drugog jezika, praćenje procesa razvoja rečnika kod pripadnika različitih jezičkih zajednica i sticanje uvida u psihološki relevantne lingvističke aspekte različitih jezika (Kim, 2008; Ramirez-Esparza, Gosling, Benet-Martinez, Potter, & Pennebaker, 2006). Konačno, razvoj rečnika za analizu verbalnog materijala na različitim jezicima omogućio je većem broju istraživača upotrebu ovog pristupa u istraživanju psiholoških pojava.

U početku je ovaj program uspešno korišćen u oblastima kliničke i socijalne psihologije, kao i psihologije ličnosti, da bi se poslednjih godina opseg

njegove primene proširio i na pedagošku psihologiju (npr. u istraživanju kritičkog mišljenja kod učenika, Carroll, 2007) i organizacionu psihologiju (npr. u istraživanjima odabira zanimanja, Djikic, Oatley & Peterson, 2006).

Poslednjih godina, naročito u oblasti individualnih razlika, javljaju se studije koje ukazuju na značaj mera verbalnog ponašanja koje su dobijene automatskom analizom teksta i sugerišu da bi ove mere mogle da imaju značaj u razjašnjenju odnosa između implicitnih (npr. Test Implicitnih Asocijacija) i tradicionalnih, eksplicitnih mera (npr. mere samoizveštaja, procene od strane bliskih drugih, procene od strane eksperata) (Back, Schmukle & Egloff, 2010; Bosson, Swann, & Pennebaker, 2000; Cohen, Beck, Brown, & Najolia, 2010; Lazarević, 2012). Naime, istraživanja pokazaju da je priroda odnosa automatskih i kontrolisanih procesa, bar kada je reč o proceni dimenzija ličnosti, još uvek nedovoljno jasna i da postoji slabo preklapanje između istih konstrukata merenih pomoću ovih dveju grupa metoda (Lazarević, 2012; Nosek & Smyth, 2007). Međutim, neka od skorašnjih istraživanja ukazuju na to da bi mere verbalnog ponašanja (kao vrsta spontanog ponašanja) mogle da budu veza u daljem razjašnjenju automatskih i kontrolisanih procesa (Cohen et al., 2010; Lazarević, 2012).

Cilj ovog rada jeste razvoj srpske verzije rečnika za automatsku analizu teksta (LIWCser), čime bi se omogućilo prikupljanje i analiza podataka koji ne dele metodsku varijansu s tradicionalno korišćenim metodama u psihologiji. U okviru ovog rada predstavljen je proces konstrukcije rečnika za srpski jezik i opis konačne verzije rečnika LIWCser. Osim toga, biće diskutovane i sadržinske karakteristike srpskog rečnika. Takođe, osvrnućemo se i na rezultate dosadašnjih istraživanja koji ukazuju na veze važnih psiholoških konstrukata i elemenata verbalne produkcije.

Metod

S obzirom na to da je cilj ovog rada adaptacija programa LIWC za upotrebu na srpskom jeziku, u metodološkom delu predstavljen je proces konstrukcije srpskog LIWC rečnika. Konstrukcija srpskog rečnika odvijala se u četiri faze: formiranje početnog skupa reči, kreiranje odrednica, klasifikacija odrednica i revizija sadržaja kategorija.

I faza: Formiranje početnog skupa reči

U prvoj fazi istraživanja najpre su sve odrednice koje su klasifikovane u psihološke i tematske kategorije iz engleskog rečnika prevedene na srpski, pri čemu je korišćeno nekoliko dvojezičnih rečnika[4]. Osim toga, u ovoj fazi zabeleženi su svi prevodi reči koji se nalaze u rečnicima, nezavisno od učestalosti njihove upotrebe i stepena semantičke bliskosti sa engleskom rečju, kako bi

4 Osim on-line rečnika, od dvojezinih rečnika, korišćeni su: Džaković, D. & Knežević, M. (2010). *Rečnik englesko-srpski, srpsko-engleski*. Jasen; Bjelica, N (2009). Englesko-srpski srpsko-engleski rečnik sa gramatikom. Book

početni skup reči bio što obuhvatniji. Kad je reč o gramatičkim kategorijama, važno je naglastiti da one nisu prevođene sa engleskog jezika, već je njihov sadržaj preuzet iz Gramatike srpskog jezika (Klajn, 2005). Ovaj korak učinjen je kako bi sadržaj svake od gramatičkih kategorija na najbolji način reprezentovao srpski jezik. Takođe, na ovaj način izbegnute su potencijalne greške koje bi bile posledica razlika u gramatičkim klasifikacijama dva jezika. Na primer, srpski rečnik ne uključuje kategoriju članova, jer oni ne postoje u srpskom; zamenice u trećem licu množine u engleskom nisu rodno specifične, dok u srpskom jesu, te ova kategorija u srpskom rečniku sadrži znatno veći broj odrednica itd. Na kraju ove faze rečnik je dopunjen semantički srodnim rečima, uglavnom sinonimima, antonimima i žargonizmima, na osnovu Rečnika sinonima i tezaurusa srpskog jezika (Ćosić, 2008).

Ishod ove faze bio je početni skup reči koji se sastojao od 9127 reči u osnovnim oblicima (u infinitivu za glagole, nominativu jednine za imenice i nominativu jednine muškog roda za prideve).

II faza: Kreiranje odrednica

S obzirom na to da program u koji se rečnik implementira upoređuje grafemske složaje u ulaznom tekstu sa grafemskim složajima u rečniku, bilo je potrebno specifikovati sve oblike pojedinačnih reči koji se mogu javiti u tekstu. Kako srpski jezik, za razliku od engleskog, ima razvijenu inflektivnu morfologiju, morali su biti specifikovani svi padežni oblici za imenice, zamenice i prideve, rod za prideve i vremena za glagole. Važna prednost ovog programa jeste u tome što omogućava postavljanje „zvezdice" (*) na kraj reči, čime je omogućeno da svaka reč koja sadrži zadati koren bez prefiksa bude klasifikovana u kategoriju kojoj odrednica sa zvezdicom pripada u rečniku. Ova karakteristika programa omogućila je ekonomičnost u kreiranju odrednica. Naime, odrednice su konstruisane tako što je svaka reč svedena na minimalni broj grafema, oduzimanjem neophodnog broja grafema sa kraja reči, pri čemu je nužan uslov bio taj da je tako konstruisanu odrednicu bilo moguće nedvosmisleno kategorisati, a da, pri tom, pokriva što je moguće veći broj oblika reči koje pripadaju istoj kategoriji. Tako je, na primer, reč *nezasitost* reprezentovana odrednicom *nezasit**, čime su pokriveni svi oblici date reči (*nezasitosti, nezasitostima*, itd.), ali i pridev *nezasit* i svi oblici tog prideva (*nezasita, nezasiti, nezasitim*, itd.). S druge strane, reč *drug* nije reprezentovana odrednicom *drug** zato što u tom slučaju ne bi postojala razlika između reči *drugi* i *drugarstvo*, s obzirom da obe reči sadrže isti koren. U ovom slučaju odrednice su kreirane dodavanjem derivacionih[5] i inflektivnih sufiksa[6], čime je omogućeno kako razdvajanje reči koje imaju isti koren, a različito značenje u izvedenim oblicima, tako i obogaćivanje skupa odrednica izvedenim leksemama. Skup reči koji je formiran nakon ove faze sastojao se iz 16012 odrednica.

5 Nastavci za građenje reči, koji mogućavaju izvođenje semantički srodnih reči.

6 Nastavci kojima se markiraju različiti gramatički oblici iste reči.

III faza: Klasifikacija odrednica

Klasifikacija odrednica vršena je metodom apsolutnog konsenzusa između pet procenjivača. Iako se usaglašenost nezavisnih procenjivača često koristi kao metod procene, u ovom slučaju opredelili smo se za metod apsolutnog konsenzusa iz nekoliko razloga. Prvo, ako se kao metod klasifikacije koristi slaganje nezavisnih procenjivača, postoji mogućnost idiosinkratične interpretacije kriterijuma za klasifikaciju, kao i samih kategorija. Drugo, osnovna prednost metode nezavisnih procenjivača nad konsenzusom (tj. mogućnost kvantifikacije stepena slaganja), nije bila od značaja za ovakav vid rada. Konačno, primenom metode konsenzusa izbegnut je problem isključivanja značajnog broja reči usled nepostojanja potpune usaglašenosti nezavisnih procenjivača.

Svaka odrednica mogla je biti klasifikovana u jednu ili više kategorija, što je u skladu sa prirodom jezika. Na primer, odrednica *uspeh* istovremeno ukazuje na postignuće, predstavlja imenicu i vezana je za pozitivne emocije. Stoga je pri klasifikaciji svake odrednice razmatrana pripadnost svakoj od kategorija, a uslov za klasifikaciju bio je postizanje konsenzusa procenjivača o pripadnosti odrednice datoj kategoriji.

Procenjivači su se u klasifikaciji vodili principom minimalne greške, odnosno uzimanjem u obzir samo onih reči koje u svim značenjima mogu biti nedvosmisleno klasifikovane. Dakle, kad je reč o homografima, odrednice su isključivane iz rečnika u svim slučajevima osim ukoliko je jedno od značenja izrazito frekventnije od drugog. Iako je ovaj postupak doveo do smanjenja broja reči koje rečnik sadrži, do nešto nižeg procenta pokrivenosti teksta i do procentualno manjeg udela svake od kategorija u ukupnom tekstu, osigurano je da svaka odrednica uvek bude ispravno klasifikovana, i isključena je mogućnost javljanja greške merenja koja bi bila posledica homografije.Takođe, pri klasifikaciji odrednica vodilo se računa o upotrebi reči u pisanom i u usmenom govoru, i prenesenom i metaforičkom značenju.

Nazivi i semantički okvir kategorija preuzeti su iz verzije engleskog rečnika LIWC2007 uz nekoliko izmena. Prvo, dodata je kategorija *negativne reči*, koja je uključila sve odrednice izvedene dodavanjem prefiksa *ne* na reči koje korenski označavaju pozitivne emocije (npr. *ne+srećan, ne+uspeh, itd.*). Cilj je bio da se ovakve reči razlikuju od onih koje imaju korenski negativno značenje (npr. tužan). Drugo, dodata je kategorija *superlativi*, koja uključuje prideve i priloge u superlativnom obliku, sa ciljem preciznog markiranja intenziteta emocija ili stavova. Treće, kategorija koja u engleskoj verziji nosi naziv *sex*, preimenovana je u *ljubav i seks* kako bi naziv kategorije više odgovarao njenom sadržaju. Naime, u srpskoj kao i u engleskoj verziji rečnik sadrži odrednice kao što su *ljubav, poljubac, zagrljaj*, itd. Četvrto, kategorije *budućnost*, *sadašnjost* i *prošlost*, koje u engleskom rečniku sadrže glagole, u srpskom sadrže priloge i prideve koji imaju vremensko značenje. Osnovni razlog za uvođenje ove izmene jeste

taj što u srpskom jeziku nije moguće odrediti vreme na nivou pojedinačne reči, već se za to koriste i složeni glagolski oblici. Na primer, glagolski oblik perfekta *je otišao* sastoji se iz pomoćnog glagola u prezentu (*je*) i radnog glagolskog prideva (*otišao*). Prepreka koju ovakva konstrukcija stavlja pred analizu na nivou pojedinačnih reči jeste taj što se ovaj pomoćni glagol koristi i za određivanje sadašnjeg vremena u imenskim predikatima (*je lep*), a radni glagolski pridev i za građenje futura II (*budem otišao*). Poslednje, kategorija koja u engleskom rečniku nosi naziv *swear words* u srpskom je preimenovana u *informalizmi*, jer su u nju uključene odrednice koje označavaju kako psovke, tako i razne neformalne izraze. Ova izmena je učinjena zbog toga što nema jasne granice između te dve vrste reči i što je omogućena jasnija interpretacija te kategorije u terminima formalnog-neformalnog izražavanja.

Nakon ove faze formiran je rečnik LIWCser, koji se sastojao od 11230 odrednica klasifikovanih u 65 kategorija, pri čemu je svaka odrednica klasifikovana u minimalno jednu, a maksimalno u osam kategorija. Dakle, u ovoj fazi konstrukcije rečnika bilo je obavljeno preko 30 hiljada klasifikacija.

Faza IV: Revizija sadržaja kategorija

U ovoj fazi je sadržaj svake od kategorija revidiran, sa ciljem dopune kategorija kulturno specifičnim rečima, kako bi rečnik što bolje reprezentovao jezičku kulturu srpskog jezika. Najveće izmene unete su u kategorije: a) *porodica,* u koju su dodati nazivi različitih rodbinskih odnosa koji nisu specifikovani u engleskom jeziku (npr. ujna, tetka, šurak itd.); b) *religija*, u kojoj su u engleskom rečniku bile zastupljene odrednice koje markiraju samo najosnovnije pojmove svetskih religija, s posebnim akcentom na različite varijetete hrišćanskih, posebno protestantskih religija. Ova kategorija u srpskoj verziji dopunjena je odrednicama specifičnim za pravoslavnu crkvu i običaje na osnovu teološkog rečnika (Dobrić, 2008); c) *zabava*, koja je dopunjena odrednicama koje se odnose na različite sportove, muziku i načine provođenja slobodnog vremena karakterističnim za našu kulturu; d) *informalizmi*, koja je dopunjena neformalnim izrazima specifičnim za srpski jezik. Za sve reči dodate u ovoj fazi ponovljen je postupak formiranja odrednica i klasifikacije. Rezultat ove faze bila je konačna verzija rečnika LIWCser.

Rezultati

Formalne odlike rečnika LIWCser

Konačna verzija rečnika LIWCser sastoji se iz 12103 odrednice, klasifikovane u 65 kategorija. U Tabeli 1 dat je prikaz kategorija LIWC rečnika za srpski jezik (LIWCser) i tipičnih primera za svaku od njih.

Tabela 1: Prikaz svih kategorija LIWCser rečnika sa tipičnim primerima

1.Lingvističke kategorije		**3.Psihološke kategorije**	
1.1. Ukupan broj reči		3.1. Socijalni procesi	
1.2. Prosečan broj reči u rečenici		3.1.1. Porodica	*mama, ujak, porod*
1.3. Broj rečničkih reči		3.1.2. Prijatelji	*cimer, drug, ortak*
1.4. Reči duže od 6 slova		3.1.3. Ljudi	*sugrađani, sused*
1.5. Ukupan broj funkcijskih reči		3.2. Afektivni procesi	
1.6. Zamenice		3.2.1. Pozitivne emocije	*sviđa, lepo, sreća*
1.6.1. Lične zamenice		3.2.2. Negativne emocije	grozno, prevara
1.6.1.1. Prvo lice jednine	ja, moj	3.2.3. Strah i ankisoznost	zabrinut, briga
1.6.1.2. Prvo lice množine	mi, naš	3.2.4. Bes i ljutnja	drzak, dovraga
1.6.1.3. Drugo lice	ti, vaš, tvoj	3.2.5. Tuga	plač, jad, lišen
1.6.1.4. Treće lice jednine	*on, njegov*	3.3. Kognitivni procesi	
1.6.1.5. Treće lice množine	oni, njihov	3.3.1. Uvid	objasni, shvatam
1.6.2.Nelične zamenice	neki, niko, svaki	3.3.2. Kauzacija	stoga, izaziva
1.7. Glagoli	trčim, ići, znaju	3.3.3. Diskrepanca	teže, treba, umesto
1.8. Pomoćni glagoli	ću, bih, smo	3.3.4. Nesigurnost	otprilike, eventualno
1.9. Prošlost	davno, juče	3.3.5. Sigurnost	kategorično, moraš
1.10. Sadašnjost	sada, trenutno	3.3.6. Inhibicija	barijera, osujeti
1.11.Budućnost	ubuduće, sutra	3.3.7. Inkluzija	preuzet, prihvaćen
1.12. Prilozi	uvek, veoma	3.3.8. Ekskluzija	sem, stran, van
1.13.Predlozi	*na, ka, iz*	3.4. Perceptivni procesi	
1.14. Veznici	dakle, ali, mada	3.4.1. Vid	belo, svetlucav
1.15. Negacije	nije, neće, nisam	3.4.2. Sluh	kuc, doziva, glas
1.16. Negativne reči	nesreća,neaktivan	3.4.3. Osećaj/ oset	opipam, kiselo
1.17. Superlativi	*najbolji, najgori*	3.5. Biološki procesi	
Kvantifikatori	mnogo, puno	3.5.1. Telo	noga, lice, malje
Brojevi	jedan, deseti	3.5.2. Zdravlje	kašlje, lekar, brufen
Psovke/informalizmi	mrš, muda, omg	3.5.3. Seks i ljubav	*orgazam, nag, ljubi*
2.Tematske kategorije		3.5.4. Ingestiv	*pečenje, piće, gutam*
2.1. Posao	*preduzeće, plata*	3.6. Relativitet	
2.2. Postignuće	samouveren, šampion	3.6.1.Kretanje	prolazi, putujem, ide
2.3. Razonoda	hobi, surfovanje, igra	3.6.2. Prostor	ring, sever, hodnik
2.4. Kuća	dom, kapija, kauč	3.6.3. Vreme	ikad, januar, kasno
2.5. Novac	*kupiti, dinar, plaćati*	**4.Kategorije u usmenom govoru**	
2.6. Religija	pop, pričest, krštenje	4.1. Potvrđivanje	svakako,vau, aha
2.7. Smrt	masakr, mrtav, pokojni	4.2. Nefluentnost	hmm, mm, uf
		4.3. Fileri/ poštapalice	bla, brate,mislimm

Od ukupnog broja odrednica, samo četiri od njih (0,03%) klasifikovano je u osam kategorija, 22 (0,2%) u sedam, 147 (1,2%) u šest i 568 (4,7%) u pet kategorija. Najveći broj odrednica klasifikovan je u četiri ili manje od četiri kategorija. U četiri kategorije klasifikovano je 1531 (12,6%) odrednica, u tri kategorije 2913 (24,1%) odrednica, u dve kategorije 4800 (39,7%) odrednica, dok se 2118 (17,5%) odrednica nalazi samo u jednoj kategoriji. Dakle, iako rečnik sadrži nesto više od 12 hiljada odrednica, ukupno je obavljeno 30489 klasifikacija.

Sadržinske karakteristike rečnika LIWCser

Kao i u engleskoj verziji LIWC rečnika, sve kategorije srpske verzije grupisane su u četiri domena: lingvističke kategorije, tematske kategorije, psihološke kategorije i kategorije u usmenom govoru.

Lingvističke kategorije

Lingvističke kategorije odnose se na sintaksičke i gramatičke karakteristike teksta i u njih je svrstano 13.61% klasifikovanih odrednica. U Tabeli 2 dat je prikaz kategorija koje pripadaju segmentu *Lingvističke kategorije,* sa brojem odrednica[7].

Tabela 2: Prikaz segmenta Lingvističke kategorije sa brojem klasifikovanih odrednica

Lingvističke kategorije	Broj klasifikovanih odrednica	% klasifikacija
1.5.Ukupan broj funkcijskih reči	463	1.52
1.6. Zamenice	136	0.45
1.6.1. Lične zamenice	49	0.16
1.6.1.1. Prvo lice jednine	12	0.04
1.6.1.2. Prvo lice množine	11	0.04
1.6.1.3. Drugo lice	26	0.08
1.6.1.4. Treće lice jednine	17	0.06
1.6.1.5. Treće lice množine	9	0.03
1.6.2. Nelične zamenice	81	0.27
1.7. Glagoli	1 610	5.28
1.8. Pomoćni glagoli	28	0.09
1.9. Prošlost	25	0.08
1.10. Sadašnjost	26	0.08
1.11.Budućnost	17	0.06
1.12. Prilozi	154	0.51
1.13. Predlozi	49	0.16
1.14. Veznici	26	0.08
1.15. Negacije	43	0.14
1.16. Negativne reči	234	0.77
1.17. Superlativi	382	1.25
1.18. Kvantifikatori	165	0.54
1.19. Brojevi	107	0.35
1.20. Psovke/informalizmi	479	1.57
Ukupno klasifikacija	4149	13.61

Ovaj segment rečnika uključio je standardne gramatičke kategorije, kao što su zamenice, glagoli, brojevi, prilozi, predlozi i veznici. Osim toga, uključene su i odrednice koje pripadaju kategorijama *negacije, Negativne reči, Su-*

7 Napomena: Suma klasifikovanih odrednica nadređene kategorije nije jednaka zbiru klasifikacija odrednica pojedinačnih kategorija zbog toga što su odrednice klasifikovane u više kategorija unutar nadređene kategorije. Takođe, kada je reč o homonimima, kako bi se izbegla konfundacija, homonimni oblici koji pripadaju različitim kategorijama klasifikovani su samo u nadređenu kategoriju.

perlativi, Kvantifikatori, i *Informalizmi.* Za kategorije *zamenice, Predlozi* i *Veznici* specifično je to što su uzeti svi oblici koji postoje u srpskom jeziku.

Tematske kategorije

Tematske kategorije obuhvataju 2182 odrednice, što čini 7.15% svih klasifikovanih odrednica u rečniku. Ove kategorije odnose se na oblasti ličnog funkcionisanja osobe, kao što su *posao, razonoda, kuća, religija,* itd. U Tabeli 3 dat je prikaz *tematskih* kategorija sa brojem klasifikovanih odrednica.

Tabela 3: Prikaz Tematskih kategorija sa brojem klasifikovanih odrednica

Tematske kategorije	Broj klasifikovanih odrednica	% klasifikacija
2.1. Posao	414	1.36
2.2. Postignuće	401	1.31
2.3. Razonoda	449	1.47
2.4. Kuća	146	0.48
2.5. Novac	331	1.08
2.6. Religija	317	1.04
2.7. Smrt	124	0.41
Ukupno klasifikacija	2182	7.15

Kategorija *posao* sadrži opšte odrednice kao što su *karijera, zaposlenje, posao, otkaz;* odrednice koje se odnose na obrazovanje (*diploma, biografija, škola*); pozicije u radnoj organizaciji (*direktor, kolega, menadžer*); prostor i dokumentaciju (*kancelarija, ugovor)*; uspešnost u obavljanju posla (*produktivan, profesionalan*), i slično.

Kategorija *postignuće* sadrži odrednice koje se odnose na motivisanost, želju za uspehom, takmičenje, visoki položaj (*ambiciozan, elita, ugled, izazov*), ali i neuspeh (*bezuspešan, nazadovanje*).

Kategorija *razonoda* sadrži odrednice koje se odnose na različite sportove i rekreaciju (*fudbal, reket, aerobik*), umetnost (*film, džez*), lokacije na kojima se može provoditi slobodno vreme (*pozorište, vikendica, kafić*), predmete koje ljudi koriste u slobodno vreme (*„playstation", domine*).

Kategorija *kuća* uključuje odrednice koje se odnose na delove kuće ili stana (*kupatilo, tavan, trpezarija*), nameštaj i kućne aparate (*sofa, cipelarnik, šporet*), osobe koje su vezi sa domaćinstvom (*spremačica, komšija, domaćin*) i druge reči koje referiraju na kuću (*kirija, renovirati, odseliti*).

Kategorija *novac* uključuje odrednice različitih valuta (*dinar, dolar*) i opšte odrednice vezane za novac (*kupiti, jeftin, honorar*). Takođe, ova kategorija uključuje odrednice koje se odnose na aktivnosti u vezi s novcem (*pozajmica*, *klađenje, investirati*), i razne profesije u vezi sa ekonomijom (*broker, revizor, nadničar*).

Kategorija *religija* uključuje opšte pojmove dominantnih svetskih religija (*jidiš, džihad, psalm*), odrednice koje se odnose na hrišćansku pravoslavnu religiju (*slava, Nikoljdan*) i opšte religijske odrednice (*đavo, bog, hram, molitva*).

Poslednja kategorija u ovom delu rečnika jeste *smrt,* koja uključuje odrednice kao što su *testament, linč, ucmekati* itd.

Psihološke kategorije

Psihološke kategorije obuhvataju šest kategorija višeg reda: socijalni procesi, afektivni procesi, kognitivni procesi, perceptivni procesi, biološki procesi i relativitet. U ove kategorije uključen je najveći broj odrednica (77.92%) (Tabela 4). *Socijalni procesi* uključuju sve odrednice iz subkategorija *porodica, prijatelji i ljudi,* kao i odrednice koje ukazuju na socijalnu interakciju (*dobrodošlica, dogovor, intervju*), društvene pojave i institucije (*kriminal, ministarstvo*), socijalne stavove (*konzervativac, demokratski*) i emocije koje su usko povezane sa drugim ljudima (*ljubomora, empatija*). Specifično, subkategorija *porodica* uključuje odrednice kao što su *mama, stric, brat*; subkategorija *prijatelji* uključuje odrednice kao što su *ortak, drug, pajtos;* subkategorija *ljudi* uključuje opšte odrednice koje referiraju na osobe, kao što su *klinac, dama, muško.*

Tabela 4: Prikaz Psiholoških kategorija sa brojem klasifikovanih odrednica

Psihološke kategorije	Broj klasifikovanih odrednica	% klasifikacija
3.1. Socijalni procesi	578	1.89
3.1.1. Porodica	259	0.85
3.1.2. Prijatelji	32	0.10
3.1.3. Ljudi	180	0.59
3.2. Afektivni procesi	4 328	14.19
3.2.1. Pozitivne emocije	1 489	4.88
3.2.2. Negativne emocije	2 370	8.95
3.2.3. Strah i ankisoznost	263	0.86
3.2.4. Bes i ljutnja	367	1.20
3.2.5. Tuga	251	0.82
3.3. Kognitivni procesi	2 444	8.02
3.3.1. Uvid	426	1.40
3.3.2. Kauzacija	256	0.84
3.3.3. Diskrepanca	350	1.15
3.3.4. Dvoumljenje	269	0.88
3.3.5. Sigurnost	363	1.19
3.3.6. Inhibicija	259	0.85
3.3.7. Inkluzija	92	0.30
3.3.8. Ekskluzija	132	0.43
3.4. Perceptivni procesi	1 202	3.94
3.4.1. Vid	300	0.98
3.4.2. Sluh	222	0.73
3.4.3. Osećaj	406	1.33
3.5. Biološki procesi	1 831	6.00
3.5.1. Telo	519	1.70
3.5.2. Zdravlje	503	1.65
3.5.3. Seks i ljubav	308	1.01
3.5.4. Ingestiv	404	1.32
3.6. Relativitet	1 663	5.45
3.6.1.Kretanje	564	1.85
3.6.2. Prostor	441	1.45
3.6.3. Vreme	343	1.12
Ukupno klasifikacija	23414	77,92

Nadređena kategorija *afektivni procesi* sadrži najviše odrednica u rečniku i uključuje subkategorije: *pozitivne emocije, negativne emocije, strah i anksioznost, bes i ljutnja, tuga,* kao i opšte deskriptore emotivnih stanja (*osećanje, afekat*). *Pozitivne emocije* se odnose na odrednice koje opisuju (*srećan, radostan*), pobuđuju pozitivne emocije (*grljenje, ljubiti*), stanja i aktivnosti koje su posledica pozitivnih osećanja (*osmeh, hahaha, ozaren*) i odrednice sa pozitivnim konotativnim značenjem (*uživancija, raj, slatkiš*). *Negativne emocije* uključuju odrednice koje se odnose na negativna osećanja i stanja (*apatija, bolan, jad, mržnja*), akcije koje su u vezi sa negativnim emocijama (*linčovati, nepoštovanje*, *zloupotreba*), kao i reči sa negativnom konotacijom (*cmizdravac, đubre, gnjida*) i nepristojne izraze. *Strah i anksioznost* obuhvataju odrednice koje su markirane i u okviru kategorije *negativne emocije* (*napet, strašan*), ali i one koje se odnose na stanja koja izazivaju anksioznost i neizvesnost koje ne moraju nužno biti negativno obojene (*iščekivati, oprez*). *Bes i ljutnja* obuhvataju odrednice kao što su *gnev, jarost, m'rš.* U kategoriju *tuga* svrstane su odrednice tipa *beda, jauk, očaj.* Važno je istaći da su gotovo sve odrednice koje pripadaju kategorijama *tuga* i *bes i ljutnja* markirane i u okviru kategorije *negativne emocije.*

U nadređenu kategoriju *kognitivni procesi* svrstane su odrednice koje se odnose na mišljenje i pamćenje (*dokučiti, sećanje, pažnja*), kao i odrednice klasifikovane unutar subkategorija: *uvid* (*logičan, otkriće*), *kauzacija* (*dakle, uzrok*), *diskrepanca* (*kompenzovati, međutim, suprotan*), *dvoumljenje* (*izgleda, najverovatnije*), *sigurnost* (*apsolutno, garantovano*), *inhibicija* (*obuzdan, kontrolisan*), *inkluzija* (*obuhvaćen, uračunati*) i *ekskluzija* (*osim, izuzetak*).

Nadređena kategorija *perceptivni procesi* uključuje subkategorije: *vid*, *sluh*, i *osećaj*. *Vid* obuhvata odrednice koje se odnose na vizuelnu percepciju, boje, svetlost itd. (*vidim, crveno, taman*), a *sluh* obuhvata odrednice koje se odnose na auditivnu percepciju (*čujem, kuckanje, zvučan*). Kategorija *osećaj* obuhvata odrednice koje se odnose na taktilnu (*maženje,stisak*), olfaktornu (*smrdi, nozdrva*) i gustativnu (*kiseo, gorak*) percepciju.

Nadređena kategorija *biološki procesi* obuhvata subkategorije: *telo* (*kuk, telesni, puls*), *zdravlje* (*lekovi, hroničan, bubuljica*), *seks i ljubav* (*ljubavnik, libido, strastven*). Takođe, u ovu nadređenu kategoriju spada i *ingestiv,* odnosno subkategorija koja se odnosi na varenje i obradu različitih materija u telu (*nahraniti, žvakati, povraćati, drogirati*).

Poslednja nadređena kategorija u ovom segmentu rečnika jeste *relativitet.* Ona obuhvata subkategorije: *kretanje* (*pratiti, uzimati, zgrčiti*), *prostor* (*spolja, vakuum, dublji*) i *vreme* (*ranije, januar, brzo*).

Paralingvističke kategorije

Paralingvističke karakteristike govora reprezentovane su u rečniku najmanjim brojem odrednica, tačnije manje od 1% klasifikacija obavljeno svrstano je u Paralingvističke kategorije: *Potvrđivanje, Nefluentnost* i *Poštapalice* (Tabela 5).

Tabela 5: Prikaz Paralingvističkih kategorija sa brojem klasifikovanih odrednica

Kategorije u usmenom govoru	Broj klasifikovanih odrednica	% klasifikacija
4.1. Potvrđivanje	20	0.06
4.2. Nefluentnost	9	0.03
4.3. Fileri/ poštapalice	14	0.04
Ukupno klasifikacija	23	0.13%

Za razliku od ostalih kategorija, odrednice koje se nalaze u ovom delu rečnika izabrane su tako da odražavaju najčešće načine na koje ljudi u komunikaciji izražavaju slaganje sa sagovornikom (*da, ok, aha*), različite emotivne i kognitivne procese koji nisu izraženi putem reči (*grrr*, *mmm*, *hm*), kao i različite poštapalice koje se najčešće sreću u govoru (*mislimmm, brate*).

Evaluacija instrumenta LIWCser

Imajući u vidu obuhvatnost rezultata koji se odnose na samu konstrukciju rečnika i na psihometrijsku evaluaciju, rezultati empirijske provere rečnika LIWCser biće prikazani u narednom radu. U okviru psihometrijske evaluacije proveravaju se karakteristike rečnika koje su relevantne za njegovu upotrebu u psihološkim istraživanjima, na uzroku većem od 1000 različitih tekstovnih materijala. Pre svega, utvrđuje se obuhvatnost rečnika, odnosno procenat teksta koji je korespondentan odrednicama u rečniku. Preliminarne analize pokazale su da je obuhvatnost rečnika LIWCser između 62% i 89% (AS=74%). Takođe, analize obuhvataju zastupljenost kategorija u različitim vrstama tekstova, sa ciljem formiranja neke vrste normi za srpski rečnik kad je reč o tekstovima koji su pisani različitim funkcionalnim stilovima (naučni, umetnički, novinarski, itd.) i kako bi se rezultati uporedili sa onima koji su dobijeni na engleskom rečniku. Analizira se i zastupljenost kategorija u usmenom i pisanom jeziku kako bi se stekao uvid u specifičnosti koje proizlaze iz modaliteta verbalne produkcije, sa ciljem da se uzmu u obzir te specifičnosti pri obradi podataka koji su dobijeni iz različitih modaliteta verbalne produkcije. Evaluacija instrumenta obuhvata i proveru validnosti, interne konzistentnosti i test-retest pouzdanosti.

Diskusija

Jedna od ključnih zamerki istraživača koji dolaze iz kvalitativne paradigme jeste da analiza verbalnog materijala koja se zasniva na analizi pojedinačnih reči zanemaruje kontekst (Pennebaker et al., 2007). Međutim, pošto je LIWC jedan od retkih programa koji istovremeno vrši analizu i stilističkih (gramatičkih i sintaksičkih) i sadržinskih karakteristika teksta, omogućeno je bar delimično prevazilaženje ove zamerke. Iako na prvi pogled reči koje

pripadaju lingvističkim kategorijama ne deluju kao reči koje nose psihološki relevantne informacije, istraživanja pokazuju da one čine više od polovine celokupne verbalne produkcije (Pennebaker, et al., 2003). Imajući to u vidu, autori smatraju da je za iscrpnu analizu verbalnog materijala, osim analize sadržinskih aspekata, neophodno obratiti pažnju i na gramatičke i sintaksičke kategorije jer učestalost javljanja različitih lingvističkih kategorija pruža informacije o načinu na koji osoba nešto saopštava, nezavisno od sadržaja samog teksta, što je i pokazano u većem broju empirijskih studija (za pregled videti Chung & Pennebaker, 2008). Iz tih razloga je prilikom konstrukcije srpskog rečnika posebna pažnja bila posvećena stvaranju korpusa odrednica koji bi na najadekvatniji način omogućili zahvatanje (svih) specifičnosti sintakse srpskog jezika. Takođe, imajući u vidu nalaze koji ukazuju na veliki značaj psihološki relevantnih kategorija (kao što su *afektivni procesi* ili *kognitivni procesi*) u različitim oblastima psihologije, prilikom konstrukcije rečnika formiran je veoma veliki broj odrednica koje pripadaju ovim kategorijama (one čine gotovo polovinu svih formiranih odrednica).

Izuzetan značaj automatske analize teksta čini i to što on pruža uvide i u kognitivne procese (odnosno, procese mišljenja) koji, u tom obliku, nisu dostupni drugim metodama. Naime, eksperimentalna istraživanja pružaju uvid u kognitivne procese angažovane u rešavanju određenog zadatka, ali ne pružaju informaciju o tome koliko često osoba dolazi do različitih uvida, izvodi zaključke i sa sigurnošću iznosi svoje mišljenje itd. Uzevši u obzir usku vezu između jezika i mišljenja (Carroll, 1999; Ivić, 1987), jasno je da analiza verbalne produkcije može da pruži značajne informacije o učestalosti angažovanja različitih procesa mišljenja i zaključivanja u određenim situacijama. Na primer, pokazano je da se razvoj kritičkog mišljenja manifestuje u obrascima upotrebe reči koje pripadaju kategorijama kognitivnih procesa tako što se povećava učestalost reči koje označavaju uviđanje, nedoslednost i kauzaciju, a smanjuje učestalost reči koje označavaju nesigurnost i dvoumljenje (Carroll, 2007). Takođe, u jednom istraživanju u našoj sredini gde je primenjen LIWCser, dobijeno je da su pozitivniji stavovi prema homoseksualnosti praćeni učestalijom upotrebom reči koje pripadaju kategoriji *uvid* iređom upotrebom reči koje pripadaju kategoriji *dvoumljenje* (Živanović, Bjekić, & Žeželj, submitted). S obzirom na rezultate ovih i drugih istraživanja, može se reći da automatska analiza teksta ima poseban potencijal u istraživanju kognitivnih procesa koji stoje u osnovi različitih psiholoških fenomena, kao što su stavovi, donošenje odluka, atribucije, itd.

Kad je reč o nadređenim kategorijama *biološki procesi, perceptivni procesi i relativitet,* čini se da dosadašnja istraživanja nisu dala konačne zaključke. Naime, za sada još uvek nema dovoljno empirijskih dokaza za nedvosmisleno tumačenje odnosa učestalosti upotrebe reči koje pripadaju tim kategorijama i psiholoških procesa koji stoje u njihovoj osnovi. Čini se da bi istraživanja na specifičnim uzorcima (npr. hronični bolesnici, umetnici, sportisti itd.) mogla da pruže nove uvide.

Iako se opravdano može prigovoriti da domen *tematskih* kategorija, kako je trenutno koncipiran, nije u potpunosti iscrpan i da obuhvaćene kategorije nisu jedine koje bi mogle da budu relevantne za psihološka istraživanja, istraživačima je omogućeno da pored postojećih kategorija konstruišu dodatne kategorije koje će odgovarati cilju istraživanja. Jedan od primera upotrebe dodatno konstruisanih kategorija jeste istraživanje načina na koji osobe s konzervativnim i liberalnim stavovima govore o moralnim pitanjima (Graham, Haidt, & Nosek, 2009). Za potrebe tog istraživanja autori su definisali pet kategorija koje nisu definisane osnovnom verzijom ovog programa (*nasilnost*, *pravednost*, *autoritet*, *čistota*, *grupni identitet*). Dakle, ukoliko postoji potreba, moguće je konstrusati dodatne kategorije koje će odgovarati specifičnim potrebama istraživanja[8].

Zaključak

Primena automatske analize teksta u psihološkim istraživanjima poslednjih godina pružila je značajne uvide u različite psihološke fenomene. Razvoj LIWCser rečnika čini srpski jezik jednim od 12 jezika na kojima je moguća primena ove metodologije. Primena ovog metoda, osim teorijskog, ima nekoliko značajnih praktičnih prednosti. Pre svega, automatska analiza teksta omogućava dobijanje objektivnih kvantitativnih podataka o velikom broju različitih sadržinskih i stilističkih karakteristika teksta, te omogućava kako primenu statističkih analiza, tako i kvalitativnu interpretaciju podataka. Osim toga, srpski LIWC rečnik sadrži veliki broj odrednica (znatno veći nego engleski ili rečnici drugih jezika), a u konstrukciji poseban akcenat stavljen je na uključivanje kulturno specifičnih sadržaja, čime je ovaj program dodatno prilagođen za upotrebu na srpskom jeziku. Takođe, analiza se odvija na jednostavan, pouzdan i relativno ekonomičan način, a uzorci teksta se mogu obezbediti iz veoma raznovrsnih izvora, kao što su transkripti govornog jezika, internet, elektronska pošta, itd.

Imajući u vidu da rezultati velikog broja dosadašnjih istraživanja govore da je verbalno ponašanje moguće dovesti u vezu s ključnim psihološkim konstruktima (kao što su dimenzije ličnosti, psihopatološki potencijali i kognitivno funkcionisanje), konstrukcijom rečnika za obradu verbalnog materijala na srpskom jeziku omogućenaa je jednostavna primena u različitim oblastima psihologije (npr. klinička, socijalna, ljudski resursi, pedagoška itd.) i u našoj sredini.

Kritičari ovakvog pristupa mogli bi da navedu bar dve zamerke (Mehl & Gill, 2010; Pennebaker et al., 2003). Prvo, analiza verbalnih produkata zasnovana na analizi pojedinačnih reči susreće se sa nekim izazovima, kao što su teškoće u „hvatanju" ironije, sarkazma, razumevanje konteksta ili višestrukog značenja reči. Drugo, ovako koncipirana analiza verbalnog ponašanja visoko

8 Uputstvo za konstrukciju novih kategorija se može naći u Pennebaker et al 2007.

je zavisna od kvaliteta i obuhvatnosti predefinisanog rečnika. Međutim, ove zamerke ne diskredituju automatsku analizu teksta u potpunosti, jer ukoliko se pokaže da ekstrahovana lingvistička informacija daje nedvosmislen odgovor na istraživačko pitanje, validnost metoda je potvrđena. Na primer, može se postaviti pitanje da li će izloženost stresnim situacijama dovesti do porasta u upotrebi zamenice prvog lica jednine. Za donošenje zaključaka u vezi sa ovako formulisanim istraživačkim pitanjem, neće biti od presudnog značaja to u kojim se specifičnim kontekstima upotrebljava zamenica prvog lice jednine (Chung & Pennebaker, 2007; Mehl & Gill, 2010). Osim toga, izuzetno je značajno to da je upravo upotreba *nepunoznačnih* reči (a ne samo reči koje nedvosmisleno mapiraju emotivna i kognitivna stanja) u vezi s različitim mentalnim stanjima i psihološkim procesima, a da je LIWC (i specifično LIWCser) jedan od retkih programa koji upravo omogućava analizu ove vrste reči.

Posmatrano iz metodološkog ugla, stvaraju se uslovi za prikupljanje izuzetno značajnih podataka koji ne dele metodsku varijansu sa ostalim tradicionalnim metodama u psihologiji. Osim toga, omogućeno je dalje istraživanje odnosa implicitnih i eksplicitnih mera, i to nezavisno od merenog konstrukta (npr. stavovi, bazične dimenzije ličnosti, psihopatološki potencijali).

Literatura:

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

Back, M. D., Schmukle, S. C. & Egloff, B. (2010). Predicting actual behavior from the explicit and implicit self-concept of personality. *Journal of Personality and Social Psychology, 97*, 533–548. doi: 10.1037/a0016229.

Bjelica, N (2009). *Englesko-srpski srpsko-engleski rečnik sa gramatikom*. Book.

Bosson, J. K., Swann, Jr. W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79,* 631–43. doi: 10.1037//0022–3514.79.4.631.

Carroll, D. W. (1999). *Psychology of language (*3rdEd.*)*. New York: Brooks/Cole.

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

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

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

Cohen, A. S., Beck, M. R., Brown, L. A., & Najolia, G. M. (2010). Decoupling implicit measures of pleasant and unpleasant social attitudes. *Journal of Behavior Therapy and Experimental Psychiatry, 41,* 24–30. doi:10.1016/j.jbtep.2009.08.007.

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

Ćosić, P. (2008). *Rečnik sinonima i tezaurus srpskog jezika*. Kornet, Beograd.

Dobrić, A. (2008). *Srpsko-engleski i englesko-srpski teološki rečnik*. Hrišćanski kulturni centar, Beograd.

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

Džaković, D. & Knežević, M. (2010). *Rečnik englesko-srpski, srpsko-engleski*. Jasen.

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: Correlations with self-reports, acquaintance report, and behavior. *Journal of Personality and Social Psychology, 94,* 334–346. doi: 10.1037/0022–3514.94.2.334.

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

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

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and Conservatives Rely on Different Sets of Moral Foundations. *Journal of Personality and Social Psychology, 96*, 1029–1046. doi: 10.1037/a0015141.

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

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

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

Heylighen, F., & Dewaele, J. M. (2002). Variation in the contextuality of language: An empirical measure. *Context in context, Special Issue of foundations of Science, 7,* 293–240.

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

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

Ivić, I. (1987). *Čovek kao animal symbolicum*. Nolit, Beograd.

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

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

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

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

Lazarević, Lj. B. (2012). Relations between implicit and explicit measures of personality – Prospects of Implicit Association Test (IAT) in assessment of basic personality traits. Doctoral dissertation. University of Belgrade, Department of Psychology.

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

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

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.

Matthias, 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*, 857–870. doi: 10.1037/0022–3514.84.4.857.

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

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

Mehl, M. R., Pennebaker, J. W., Crow, M. D., 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.*

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

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

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples, *Discourse Processes***,** 45, 211–236. doi: 10.1080/01638530802073712.

Nosek, B. A. & Smyth, F. L. (2007). A multitrait-multimethod validation of the Implicit Association Test: Implicit and explicit attitudes are related but distinct constructs. Experimental psychology, 54, 14–29. doi: 10.1027/1618–3169.54.1.14.

Orsillo, S. M., Batten, S. V., Plumb, J. C., Luterek, J. A., & Roessner, B. M. (2004). An Experimental Study of Emotional Responding in Women With Posttraumatic Stress Disorder Related to Interpersonal Violence. *Journal of traumatic stress, 17*, 241–248. Doi: 10.1023/B:JOTS.0000029267.61240.94.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A. & Booth, R. J. (2007). *The Development and Psychometric Properties of LIWC2007, Manual.* The University of Texas at Austin and The University of Auckland, New Zealand.

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

Pennebaker, J. W. & King, L. A. (1999). Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology, 77* (6)*,* 1296–1312.

Pennebaker, J. W., &Lay, T. C. (2002). Language use and personality during crisis: analyses of Mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality,36,*271–82.Doi:10.1006/jrpe.2002.2349.

Pennebaker, J. W., Mehl, M. R. & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: our words, our selves. *Annual Review of Psychology, 54,* 547–577. doi: 10.1146/annurev.psych.54.10161.145041.

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

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

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

Shapiro, G., & Markoff, J. (1997). A Matter of Definition. In C. W. Roberts (Ed.). *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum Associates.

Smith, C. P. (1992). Motivation and Personality: Handbook of Thematic Content Analysis. Cambridge, MA: Cambridge University Press.

Stone P. J., Dunphy, D. C., Smith, M. S., &Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis.* Cambridge, MA: MIT Press.

Striman, S. W. & Pennebaker, J. W. (2001). Word Use in Poetry of Suicidal and Nonsuicidal Poets. *Psychosomatic Medicine, 63*, 517–522. doi:0033–3174/01/6304–0517.

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), 25–54. Doi: 10.1177/0261927X09351676.

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

Yogo, M., & Fujihara, S. (2008). Working memory capacity can be improved by expressive writing: A randomized experiment in a Japanese sample. *British Journal of Health Psychology, 13 (1)*, 77–80.

Zijlstra, H., van Meerveld, T., van Middendorp, H., Pennebaker, J.W., & Geenen R. (2004). De Nederlandse versie van de Linguistic Inquiry and Word Count (LIWC), een gecomputeriseerd tekstanalyseprogramma. [Dutch version of the Linguistic Inquiry and Word Count (LIWC), a computerized text analysisprogram]. *Gedrag & Gezondheid, 32*, 273–283.

Živanović, M., Bjekić, J., & Žeželj, I. (submitted). Lingvistički stil u izražavanju stavova prema homoseksualnosti.

DATUM PRIHVATANJA RADA: 25. 05. 2012.

Prilog 1:
Ispis iz LIWC2007 sa inkorporiranim engleskim rečnikom

Linguistic Processes	Psychological Processes
Word count (mean)	Social processes
Words/sentence	Family
Dictionary words	Friends
Words>6 letters	Humans
Total function words	Affective processes
Total pronouns	Positive emotion
Personal pronouns	Negative emotion
1st person singular	Anxiety
1st person plural	Anger
2nd person	Sadness
3rd person singular	Cognitive processes
3rd person plural	Insight
Impersonal pronouns	Causation
Articles	Discrepancy
Common verbs	Tentative
Auxiliary verbs	Certainty
Past tense	Inhibition
Present tense	Inclusive
Future tense	Exclusive
Adverbs	Perceptual processes
Prepositions	See
Conjunctions	Hear
Negations	Feel
Quantifiers	Biological processes
Numbers	Body
Swear words	Health
	Sexual
Current Concerns	Ingestion
	Relativity
Work	Motion
Achievement	Space
Leisure	Time
Home	
Money	**Punctuation**
Religion	
Death	Total Punctuation
	Periods

Spoken categories	Commas
	Colons
Assent	Semicolons
Nonfluencies	Question marks
Fillers	Exclamation marks
	Dashes
	Quotation marks
	Apostrophes
	Parentheses
	Other punctuation

Development of Serbian Dictionary for Automatic Text Analysis (LIWCser)

Jovana Bjekić
Institute for Medical Research, University of Belgrade

Ljiljana Lazarević
Institute of Psychology, Faculty of Philosophy, University of Belgrade

Milica Erić, Elena Stojimirović, Teodora Đokić
Belgrade

Automatic text analysis is a methodological approach in the analysis of individual differences in verbal behaviour. It enables extraction of statistically manipulable information about intensity and/or frequency of thematic and stylistic characteristics of verbal output. LIWC (Linguistic Inquiry and Word Count), one of the widely used software solutions for automatic text analysis, performs analyses by matching word stems from incorporated software dictionary with those from text input. It provides information about the percentage of each of the predefined categories in the analyzed text. Research suggests that data obtained by automatic text analysis have potential in explaining the relationship between implicit and explicit measures, independently of the object of measurement (attitudes, pathological potential, assessment of basic personality traits etc.). The topic of this paper is the development of the Serbian LIWC dictionary. Development of the dictionary was performed in four phases: translation of English LIWC dictionary, forming lemmas, classification of word stems according to absolute consensus among four independent raters (where word stems could be categorized in more than one category, depending on the context), and revision of the content of categories and creation of final set of word stems. The final version of the LIWCser dictionary contains 12103 word stems classified into 65 categories (linguistic, psychological and personal concerns). Only four word stems (0.03%) were classified into eight categories, 22 (0.2%) into seven, 147 (1.2%) into six, and 568 (4.7%) into five. 1531 (12.6%) word stems were classified into four categories, 2913 (24.1%) into three, 4800 (39.7%) into two, while 2188 (17.5%) word stems were classified into only one category. Development of the LIWCser dictionary allows researchers to collect and analyze data on verbal behaviour and to study the relationship between implicit and explicit measures in different fields of psychology.

Key words: automatic text analysis, Serbian dictionary LIWCser for automatic text analysis, verbal behaviour, implicit and explicit measures.