Izvorni znanstveni rad
Rukopis primljen 8. 2. 2017.
Prihvaćen za tisak 18. 10. 2017.
https://doi.org/10.22210/govor.2017.34.02

Marina Olujić, Ana Matić
marina.olujic@erf.hr, ana.matic@erf.hr
Edukacijsko-rehabilitacijski fakultet Sveučilišta u Zagrebu
Hrvatska

Govorni i pisani jezik odraslih: koliko se razlikuju?

Sažetak

Gotovo da nema istraživanja koja uspoređuju jezik spontanoga govorenja i pisanja odraslih govornika hrvatskog jezika. Sukladno tomu, obilježja ovih dvaju oblika jezične proizvodnje nisu uspoređivana. Informacija o zastupljenosti pojedinih leksičkih, morfoloških, sintaktičkih, semantičkih i drugih kategorija u govornom i pisanom jeziku pruža uvid u složenost ovih dvaju modaliteta jezične proizvodnje te opisuje njihova obilježja kod odraslih govornika jezika. Cilj ovog istraživanja jest prikazati i usporediti leksička i sintaktička obilježja spontanoga govornog i pisanog jezika. Ispitana su obilježja i razlike u leksičkoj raznolikosti, leksičkoj gustoći i zastupljenosti pojedinih vrsta riječi te u jednoj od mjera sintaktičke složenosti. Istraživanje je provedeno na nezavisnim skupinama ispitanika, a uzorci spontanoga govora i pisanja preuzeti su iz Hrvatskog korpusa govornog jezika odraslih (HrAL) te Hrvatskog korpusa neprofesionalnog pisanog jezika (HKNPJ). Rezultati analize pokazuju da: 1) je prosječna leksička raznolikost pisanog jezika značajno veća od leksičke raznolikosti govornog jezika; 2) postoje razlike u zastupljenosti određenih vrsta riječi, pri čemu je u govornom jeziku veća zastupljenost glagola, zamjenica, priloga, veznika, čestica i uzvika, dok u pisanom jeziku prevladavaju imenice, pridjevi i prijedlozi; 3) pisani jezik ima veću sintaktičku složenost mjerenu u prosječnoj duljini komunikacijske jedinice.

Ključne riječi: govorni jezik, pisani jezik, korpus, odrasli govornici

1. UVOD

"Pisana riječ liči na sliku. Čini ti se da je slika živa, ali ako je zapitaš, ona dostojanstveno šuti. Drugačija je njezina rođena sestra živa riječ jer ona se sa znanjem zapisuje u dušu onoga koji uči, a umije samu sebe braniti i zna govoriti i šutjeti s kim treba." (Sokrat, Phaedrus: 1111)

Jezik je organizirani sustav znakova i pravila koji se ostvaruje govorom kao zvučnim ili pismom kao vizualnim sredstvom. Govorenje i pisanje dva su modaliteta jezične proizvodnje, tj. ekspresivnog jezika, dok se razumijevanje onoga što se čita ili sluša smatra receptivnim jezikom. Ekspresivni i receptivni jezik u istraživanjima se najčešće promatraju odvojeno jer, primjerice, osoba može prilično dobro razumjeti strani jezik, a da ne progovara niti jednu riječ (Hoff, 2013). U ovom istraživanju promatrat će se dva modaliteta jezične proizvodnje – govorenje i pisanje.

Jezik, tj. njegova obilježja nije jednostavno izmjeriti. Primjerice, u istraživanjima receptivnog jezika isprepliću se različite metode mjerenja, od standardiziranih testova do metoda oslikavanja mozga. Međutim, u mjerenjima ekspresivnog jezika još uvijek dominiraju ispitivanja velikih jezičnih uzoraka, tj. jezičnih korpusa (Kuvač i Palmović, 2007). Stoga, kako bi se govorni i pisani jezik istraživali, potrebno je prikupiti velik broj jezičnih uzoraka ili se pak osloniti na dostupne jezične korpuse. Jezični korpusi opsežan su i reprezentativan izvor znanstvenicima koji proučavaju jezik, tj. njegov sastav i strukturu (npr. lingvistima), ali i onima koji se bave otkrivanjem zakonitosti psiholingvističkog razvoja (npr. logopedima, psiholozima, neuroznanstvenicima). Po strukturi se razlikuju opći korpusi koji se često nazivaju i nacionalnim korpusima, a reprezentativni su za jezik u cjelini, te specijalizirani korpusi koji obuhvaćaju samo jedan jezični varijetet (Klobučar Srbić, 2008). Hrvatski nacionalni korpus (HNK) sadrži odabrane pisane tekstove na hrvatskome jeziku iz svih područja, struka, žanrova i stilova (Tadić, 1998) te ne sadrži govorne uzorke. U specijalizirane jezične korpuse uvrštavaju se oni koji sadrže specijalizirane jezične uzorke, primjerice, jezika spontanoga govora i pisane uzorke neprofesionalnih pisaca. Spontani govor podrazumijeva razgovorni funkcionalni stil (Badurina, 2008) koji se upotrebljava u svakodnevnim neformalnim situacijama (Shriberg, 2005). Ovi uzorci "specijalizirani" su i utoliko što ih proizvode neprofesionalni govornici hrvatskog jezika. Po pitanju pisanih jezičnih uzoraka u Nacionalni korpus uvrštavaju se samo uzorci pisanih djela čiji su autori profesionalni govornici ili pisci pa stoga ne predstavljaju reprezentativan uzorak populacije neprofesionalnih govornika, tj. prosječne populacije.

Trenutačno dostupni korpusi spontanoga govornog hrvatskog jezika jesu Hrvatski korpus dječjeg jezika (Kovačević, 2002), koji sadrži 136 000 pojavnica dječjeg jezika i oko 152 800 pojavnica spontanoga govora odraslih govornika, odnosno pojavnica

djetetovog ulaznog jezika, te Hrvatski korpus govornog jezika odraslih (HrAL; Kuvač Kraljević i Hržica, 2016a) s više od 250 000 pojavnica. Oba su dostupna u bazi CHILDES (talkbank.org).[1]

Spontani jezik odraslih općenito je vrlo malo istražen u hrvatskom jeziku, a jedan od razloga je i taj što u HNK i druge hrvatske jezične korpuse još uvijek nisu uvršteni govorni jezični uzorci[2]. Stoga je većina lingvističkih istraživanja usmjerena na mnogo dostupnije pisane izvore pa su oni i bolje opisani u odnosu na govorne. Većina prethodnih istraživanja koja uspoređuju govorni i pisani jezik potječe iz 80-ih godina prošlog stoljeća, nakon čega se vrlo malo istraživalo o razlikama među ovim dvama jezičnim modalitetima.

S obzirom na to da su govorni i pisani jezik dva modaliteta jezične proizvodnje, odnosno dijelovi iste jezične djelatnosti koji služe za komunikaciju, čini se da bi oni trebali biti vrlo slični. Međutim, prethodna istraživanja provedena u drugim jezicima pronalaze značajne sličnosti, ali i razlike. Pisana je komunikacija jednomedijska jer upotrebljava samo pismo, dok je govorna multimedijska jer sadrži više medija koji pomažu u prijenosu poruke, primjerice glas, geste, mimiku te auditivne vrednote govornog jezika. Gramatičko-stilistički izraz pisanog jezika je zahtjevniji i složeniji, dok govor dopušta gramatičko-stilistički slobodniji izraz, a da se pri tome ne naruši razumijevanje poruke (Pavličević-Franić, 2005). Najprirodnija razlika između govornog i pisanog jezika leži u spontanosti jezične proizvodnje (Yabuuchi, 1998) te u činjenici da se u govornom jeziku nalazi mnogo više ponavljanja informacija i mnogo više materijala koji pruža dodatne informacije negoli je to slučaj kod pisanog teksta (Curnow, 1979) kojeg pak karakterizira mnogo precizniji izbor riječi (Biber, 1988), bolja organiziranost i jezgrovitost te brže uvođenje novih i potpunijih informacija (Smith, 1987; Chafe, 1992; Bartsch, 1997). Također, pisani tekst najčešće ima bogatiji rječnik ili leksičku raznolikost te koristi više apstraktnih termina (Drieman, 1962; Halliday, 1985; Biber, 1988). Kada se raspravlja o razlikama između govornog i pisanog jezika, svakako se treba naglasiti da one proizlaze iz činjenice da je primarno riječ o

[1] *Talkbank* je trenutačno najveća baza spontanih uzoraka govornog jezika. U sklopu projekata *Jezična obrada u odraslih govornika* (HRZZ; UIP-11-2013-2421) te *Računalni asistent za pomoć pri unosu teksta osobama s jezičnim poremećajima* (RAPUT; EU Strukturni fondovi RC.2.2.08-0050), oba započeta pod voditeljstvom izv. prof. dr. sc. Jelene Kuvač Kraljević, tim stručnjaka Laboratorija za psiholingvistička istraživanja (Edukacijsko-rehabilitacijski fakultet Sveučilišta u Zagrebu) krajem 2014. godine započeo je izgradnju dvaju korpusa proizvodnje spontanog jezika: već spomenutog *Hrvatskog korpusa govornog jezika odraslih* (HrAL; Kuvač Kraljević i Hržica, 2016a) te *Hrvatskog korpusa neprofesionalnog pisanog jezika* (HKNPJ; Kuvač Kraljević i Hržica, 2016b). U oba su korpusa, osim uzoraka osoba urednog jezičnog statusa, zastupljeni i jezični uzorci osoba s jezičnim poremećajima.

[2] Govorni jezični uzorak odnosi se na transkript audio ili videozapisa govora ili razgovora nekoga govornika.

dvama različitim medijima (govorno-auditivnom i grafičko-vizualnom) koji funkcioniraju prema vlastitim zakonitostima, odnosno imaju svoje prednosti i ograničenja, a navedene razlike u spontanosti, ponavljanju informacija i organiziranosti proizlaze upravo iz ovih osnovnih obilježja medija. Svakako je važno da se pri usporedbi govornog i pisanog jezika uspoređuju uzorci istog žanra. Što su govorni i pisani materijali sličniji, moguća je preciznija usporedba (Bartsch, 1997).

Pri analizi jezičnog produkta koriste se objektivne lingvističke mjere za analizu diskursa. U ovom će se radu govorni i pisani uzorci promatrati i uspoređivati na leksičkoj i sintaktičkoj razini korištenjem mjera leksičke raznolikosti, leksičke gustoće te zastupljenosti pojedinih vrsta riječi, kao i jedne od mjera sintaktičke složenosti.

Leksička raznolikost (engl. *lexical diversity*) nekog jezičnog produkta, u ovom slučaju uzoraka govornog i pisanog jezika, izražava se kao indeks omjera natuknica (osnovni oblik riječi) i pojavnica (ukupan broj riječi u jezičnom produktu). Ovaj indeks daje podatak o tome koliko je jezik složen ovisno o broju ponavljanih riječi, tj. iskazuje udio različitih riječi koje se pojavljuju u jezičnom produktu. Što je omjer natuknica (N) i pojavnica (P) veći (bliži vrijednosti 1) to je veća leksička raznolikost jer je u jezičnom produktu više novih riječi, a što je taj omjer manji (bliži vrijednosti 0) to je leksička raznolikost manja jer je veći broj riječi koje se ponavljaju (Kuvač i Palmović, 2007). Yu (2010) je u svom istraživanju uspoređivao govorne i pisane uzorke istih govornika filipinskog, kineskog, ruskog i perzijskog jezika koristeći parametar D[3] kao mjeru leksičke raznolikosti, a rezultati su pokazali približno jednaku razinu leksičke raznolikosti govornih i pisanih jezičnih uzoraka. Međutim, autor je istaknuo kako leksička raznolikost uvelike ovisi o poznavanju teme o kojoj se govori/piše, mjeri leksičke raznolikosti koja se koristi te veličini uzorka. U nekim drugim istraživanjima, primjerice u onom Bibera (1988), koji je kao mjeru leksičke raznolikosti koristio omjer natuknica i pojavnica (N/P), nađena je veća leksička raznolikost pisanog jezika u odnosu na govorni.

Nadalje, razlika između govornog i pisanog jezika često se tumači i učestalošću pojave pojedinih jezičnih konstrukcija koje će se u pisanom tekstu javiti prije nego u govornom jeziku, i obrnuto. Proporcija pojedinih vrsta riječi u odnosu na ukupan broj pojavnica predstavlja mjeru leksičke gustoće (engl. *lexical density*) te označava zastupljenost punoznačnih riječi u cjelokupnom jezičnom uzorku (Johansson, 2008). Korištenjem ove mjere pokazalo se da govorni jezik odraslih ima manju leksičku gustoću u odnosu na pisani (Ure, 1971; Johansson, 2008), pri čemu je Ure (1971) dodatno upozorio da leksička gustoća može značajno varirati među jezicima s obzirom na njihove

[3] D je parametar matematičke funkcije koji modelira pad krivulje omjera natuknica i pojavnica s obzirom na njezinu ovisnost o veličini uzorka (Malvern i sur., 2004).

tipološke specifičnosti (vidi raspravu u Johansson, 2008). Allwood (1998) je otišao korak dalje i detaljnije promotrio zastupljenost pojedinih vrsta riječi u dvama švedskim korpusima: transkribiranom govornom jezičnom korpusu i pisanom jezičnom korpusu. Govorni korpus sačinjavali su prijepisi iz 14 aktivnosti, a pisani tekstovi iz pripovijetki i novina. Pronađeno je nekoliko sličnosti i razlika između govornog i pisanog jezika: (1) zamjenice i glagoli najučestaliji su u oba modaliteta; (2) u oba modaliteta najučestalije su funkcionalne riječi te je (3) broj vrsta riječi manji u govornom nego u pisanom jeziku. U ovome je radu analizirana standardna mjera leksičke gustoće kao i zastupljenost pojedinih vrsta riječi u govornom i pisanom jeziku.

Kada pak govorimo o sintaktičkoj složenosti, prvi korak u njezinoj analizi kod govornog jezika jest određivanje osnovnih jedinica za analizu sintaktičke složenosti. Naime, pisani je jezik vrlo jasno podijeljen na rečenice, dok se u govornom jeziku rečenice ne izdvajaju tako jasno. Ovisno o cilju analize, razlikuju se strukturalni i funkcionalni sustavi (Chaudron, 1988). Crookes (1990) opisuje T-jedinice (engl. *T-units*; Hunt, 1965) kao osnovne jedinice analize govornog jezika te njihove varijacije poput C-jedinica (engl. *C-units*; Loban, 1976) koje predstavljaju komunikacijske jedinice. Podjela teksta na T- ili C-jedinice zasniva se na sintaktičkim obilježjima. T-jedinica određuje se kao nezavisna rečenica i njezini modifikatori. To je iskaz koji se ne može dalje dijeliti, a da se ne izgubi njegovo osnovno značenje. Glavna rečenica može stajati zasebno i može biti određena kao jedna T-jedinica. Za razliku od glavne, zavisne rečenice ne mogu stajati zasebno jer time gube osnovno značenje pa ne mogu tvoriti samostalne T-jedinice (Miller i sur., 2015). C-jedinica je usko povezana s T-jedinicom, ali uključuje i iskaze koji se sastoje od izoliranih fraza koje ne sadrže glagole (kao što su, primjerice, odgovori na pitanja). Stoga korištenje C-jedinica kao osnovnih jedinica analize omogućuje uključivanje ključnih komunikacijskih elemenata, istovremeno se oslanjajući na sintaksu kao glavni kriterij analize. Prosječna duljina komunikacijske jedinice – PDKJ (MLCU, engl. *mean length of communication unit*) odražava sintaktičku složenost preciznije nego prosječna duljina drugih mogućih jedinica jer u obzir uzima ne samo broj riječi već i hijerarhiju organizacije teksta na sintaktičkoj razini. Iako je ovo samo jedna od mjera sintaktičke složenosti, potvrdila se njezina umjerena i pozitivna povezanost s drugim mjerama složenosti (primjerice, mjerom sintaktičke sofisticiranosti; više u istraživanju Matić i sur., u tisku), zbog čega se smatra valjanom i pouzdanom. Nadalje, ova je mjera dobar pokazatelj produktivnosti (Foster i sur., 2000).

O'Donnell (1973) je analizirao sintaktička obilježja govornog i pisanog jezika kako bi utvrdio je li pisani jezik sintaktički složeniji. Govorni uzorci bili su prikupljeni iz televizijskog programa u kojem je osoba odgovarala na pitanja novinara, a pisani uzorci bili su objavljeni novinski članci istoga govornika. Rezultati su pokazali kako se u pisanom jeziku koriste složenije sintaktičke strukture.

Cilj je ovog istraživanja prikazati i usporediti leksička obilježja (leksičku raznolikost, leksičku gustoću i zastupljenost pojedinih vrsta riječi) i sintaktička obilježja (prosječnu duljinu komunikacijske jedinice kao jedne od mjera sintaktičke složenosti) spontanoga govorenja i pisanja u hrvatskome jeziku. Pretpostavlja se da je pisani jezik složeniji u odnosu na govorni, što će biti vidljivo putem veće leksičke raznolikosti, leksičke gustoće i sintaktičke složenosti. Pretpostavka je i da će se modaliteti razlikovati u zastupljenosti vrsta riječi.

2. METODE

2.1. Ispitanici

U istraživanju je korišten uzorak od ukupno 105 ispitanika, od kojih je 44 (17 muških i 27 ženskih) preuzeto iz Hrvatskog korpusa govornog jezika odraslih (HrAL; Kuvač Kraljević i Hržica, 2016a), a 61 (23 muška i 38 ženskih) iz Hrvatskog korpusa neprofesionalnog pisanog jezika (HKNPJ; Kuvač Kraljević i Hržica, 2016b). Prosječna dob ispitanika u govornom uzorku je M = 35,48 (SD = 14,31), a u pisanom M = 36,57 (SD = 13,63) godina. U govornom uzorku zastupljena su 22 ispitanika nižeg obrazovanja (OŠ, SŠ) i 22 višeg obrazovanja (VŠS, VSS), dok su u pisanom uzorku 33 ispitanika nižeg obrazovanja i 28 višeg obrazovanja. Dakle, ove dvije skupine ispitanika, tj. dva jezična uzorka, pretežito su ujednačena po spolu, dobi i obrazovanju. Osim toga, jezični uzorci, govorni i pisani, ujednačeni su i po broju pojavnica: u govornom ih je ukupno 19 352, a u pisanom 19 484.

2.2. Postupak odabira jezičnih uzoraka iz korpusa

Uzorci govornog i pisanog jezika preuzeti su iz HrAL-a (Kuvač Kraljević i Hržica, 2016a) i HKNPJ-a (Kuvač Kraljević i Hržica, 2016b). Analizirana su ukupno 44 govorna i 122 pisana uzorka, od kojih po dva pisana teksta svakog ispitanika.

Uzorci govornog jezika u HrAL-u prikupljeni su snimanjem zvučnog zapisa spontanog razgovora tijekom neke neformalne situacije (npr. obiteljski ručak, druženje s prijateljima). Prikupljeni audio zapisi transkribirani su u programu *CHAT Transcription Format* (CHAT; MacWhinney, 2000). U HKNPJ su uvršteni pisani tekstovi različitih razina strukturiranosti čime se varira spontanost proizvodnje što utječe na krajnju izvedbu; od najmanje strukturiranih tekstova kao što su eseji na zadanu temu koji zahtijevaju komponiranje teksta, preko nešto više strukturiranih tekstova poput pripovijedanja i pisanja pisama, pa sve do visoko strukturiranog teksta kao što je diktat

koji uopće ne zahtijeva komponiranje te je kao takav kognitivno najmanje zahtjevan s obzirom da podrazumijeva reprodukciju unaprijed lingvistički uređenog teksta. Kako bi se pisani tekstovi što više približili diskursu spontanoga govora iz HrAL-a, iz HKNPJ-a su odabrani isključivo tekstovi najniže razine strukturiranosti, tj. eseji pisani na zadanu temu (*Moja kuća/stan*, *Moj susjed/susjeda*), dok primjerice diktati nisu odabrani zbog visoke razine strukturiranosti, odnosno minimalne mogućnosti spontane proizvodnje. Jezik najniže strukturiranog teksta opravdano je uspoređivati s jezikom spontanoga govora i s obzirom na to da su ovi žanrovi funkcionalno najsličniji. Prikupljeni pisani uzorci također su transkribirani u programu CHAT (MacWhinney, 2000).

3. REZULTATI I RASPRAVA

Cilj je ovog istraživanja prikazati i usporediti leksička i sintaktička obilježja spontanoga govora i pisanja u hrvatskom jeziku.

Iz prethodnih istraživanja proizlazi kako je pisani jezik unaprijed planiran, dobro organiziran i strukturiran, dok je govorni jezik mahom neplaniran, manje strukturiran i interaktivan. Govorni jezik jednostavniji je i kraći te koristi nestandardne gramatičke oblike. Čini se da je stil govornog jezika stil prvog koncepta, tj. prve verzije produkta, dok pisani jezik stilski podsjeća na završni koncept, odnosno završni produkt koji je stoga bogatiji sadržajem, složeniji i bolje strukturiran (Ghasemi i Jahromi, 2014).

3.1. Obrada podataka

Govorni i pisani jezični uzorci analizirani su u programu *Computerized Language Analysis programme* (CLAN; MacWhinney, 2000) te SPSS 22 statističkim paketom (*Statistical Package for the Social Sciences*). Podatci su obrađeni deskriptivnom i inferencijalnom analizom. Kod govornog i pisanog jezika uspoređeni su leksička raznolikost, leksička gustoća i zastupljenost pojedinih vrsta riječi te sintaktička složenost izražena u prosječnoj duljini komunikacijskih jedinica, sve na razini pojavnica.

3.2. Leksička raznolikost

Kako je ranije spomenuto, leksička raznolikost izražava se kao indeks omjera natuknica (osnovni oblik riječi) i pojavnica (ukupan broj riječi u jezičnom produktu) – N/P.

Kako bi se opisala veličina prikupljenih jezičnih uzoraka, u Tablici 1. prikazan je ukupan broj pojavnica i natuknica za uzorke govornog i pisanog jezika. Vidljivo je kako je broj pojavnica u govornom i pisanom jezičnom modalitetu približno ujednačen, što omogućuje njihovu valjanu usporedbu. Također, prosječan broj prikupljenih pojavnica po ispitaniku je do 400 ili blago više od 400, što donekle osigurava pouzdanost omjera N/P kao mjere leksičke raznolikosti, budući da je ovaj omjer često manji što je promatrani govorni ili pisani uzorak dulji, čime mjera gubi na pouzdanosti (vidi Johansson, 2008). U istoj je tablici za oba jezična modaliteta prikazana prosječna leksička raznolikost.

Prije inferencijalne analize Kolmogorov-Smirnovim testom provjeren je preduvjet normalnosti distribucije rezultata prosječne leksičke raznolikosti za govorni (K-S Z = 0,69; p > 0,05) i pisani (K-S Z = 0,71; p > 0,05) uzorak; u oba se rezultati raspoređuju prema normalnoj distribuciji. T-testom za nezavisne uzorke provjerena je značajnost razlike u prosječnoj leksičkoj raznolikosti govornog i pisanog jezika te je dobivena statistički značajna razlika u korist pisanog jezika (Tablica 1.), što se vidi i iz grafičkog prikaza (Slika 1.).

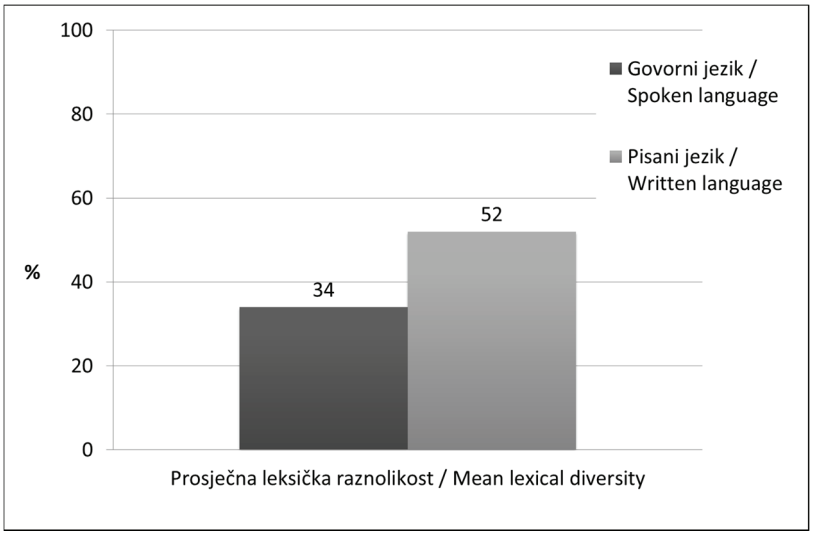
Tablica 1. Ukupan broj pojavnica i natuknica te leksička raznolikost govornog i pisanog jezika

Table 1. Total number of types and tokens and lexical diversity of spoken and written language

	Govorni jezik / Spoken language		Pisani jezik / Written language	
	Pojavnice / Types	Natuknice / Tokens	Pojavnice / Types	Natuknice / Tokens
Ukupan broj / Total number	19 484	6 685	19 352	10 061
M	439,71	151,93	319,41	170,52
SD	353,86	89,86	133,16	57,25
Prosječna leksička raznolikost (N/P) / Average lexical diversity (N/P)	34 %		52 %	
M	0,43	/	0,55	/
SD	0,13	/	0,05	/
t-test	t = -6,71** (df = 101; Cohenov *d* = -1,34)			

**p < 0,01

Leksička raznolikost smatra se jednim od najvažnijih parametara za ovladavanje govorenjem (Zechner i sur., 2007) i pisanjem (Chodorow i Burstein, 2004). Veća leksička raznolikost pisanog teksta dobivena u ovom istraživanju u skladu je s nekim prethodnim istraživanjima (npr. Biber, 1988). Pisani jezik ima mnogo manji opseg mogućih medija kojima se prenose informacije, tj. one se prenose samo putem grafema i interpunkcijskih znakova (Pavličević-Franić, 2005) te se stoga potrebno više osloniti na jezične resurse kako bi poruka bila uspješno prenesena čitatelju. Koliko je govorni jezik različit od pisanoga, gledamo li samo njegovu jezičnu osnovu, može se lako uvidjeti iz prijepisa govornih uzoraka, gdje se vrlo lako uočava obilje uporabe nestandardnih jezičnih oblika, nedovršenost rečenica/iskaza i slično. Stoga, kako bi pisani tekst nadomjestio sve one resurse za prijenos informacija koje govorni modalitet ima, za taj mu je prijenos potreban primarno velik broj različitih riječi, a čime se izravno povećava broj natuknica, kao i leksička raznolikost.



Slika 1. Leksička raznolikost govornog i pisanog jezika
Figure 1. Lexical diversity of spoken and written language

3.3. Leksička gustoća i zastupljenost pojedinih vrsta riječi

Vrste riječi vrlo su složen sustav u teoriji jezika i postoji nekoliko pravaca koji zastupaju donekle različite kriterije za određivanje vrsta riječi (Brlobaš, 2001). U ovom radu vrste riječi određivane su prema morfološkim kriterijima pa su glagolski

pridjev trpni i poimeničeni pridjevi svrstani u kategoriju pridjeva. Također, riječi se u ovom istraživanju promatraju i s obzirom na kategoriju promjenjivosti. Promjenjivost je morfološko obilježje u kojem se promjenjive riječi mijenjaju u odnosu na svoju ulogu u rečenici koja se iskazuje gramatičkim značenjskim kategorijama, dok nepromjenjive riječi ne mijenjaju oblik u rečenicama. Prema tome, promjenjive riječi su imenice, pridjevi, brojevi, zamjenice, prilozi i glagoli, a nepromjenjive su prijedlozi, veznici, čestice i uzvici. S obzirom na funkciju značenja, riječi se dijele na punoznačne (leksičke) – riječi koje izriču nekakav sadržaj vanjskog ili unutrašnjeg svijeta, te odnošajne (pomoćne) – riječi koje izriču odnose između onoga što znače riječi iz skupa punoznačnih riječi. Punoznačne riječi ujedno su i promjenjive ili djelomično promjenjive, a odnošajne su nepromjenjive (Barić i sur., 2005).

Uspoređena je proporcija promjenjivih (punoznačnih) riječi, što je istovjetno mjeri leksičke gustoće, u govornom i pisanom jeziku te je u tu svrhu izračunat t-test za nezavisne uzorke parova. Rezultati pokazuju kako ne postoji razlika u leksičkoj gustoći između govornog (M = 0,78; SD = 0,06) i pisanog (M = 0,80; SD = 0,05) jezika (t = – 1,91; df = 89; p = 0,06; Cohenov *d* = 0,40). Premda statistički značajna razlika nije utvrđena na strožoj razini značajnosti od 5 %, na 6 % ona jest značajna te postoji tendencija k većoj leksičkoj gustoći u pisanom jeziku, a što je u skladu s rezultatima dobivenim u istraživanju Ure (1971) i Johansson (2008). Ovu tendenciju k razlici trebalo bi pomnije istražiti na većem uzorku te uzimajući u obzir različite mogućnosti definiranja punoznačnih riječi, tj. vrsta riječi koje se u ovu skupinu ubrajaju (vidi raspravu o leksičkoj gustoći u Johansson, 2008).

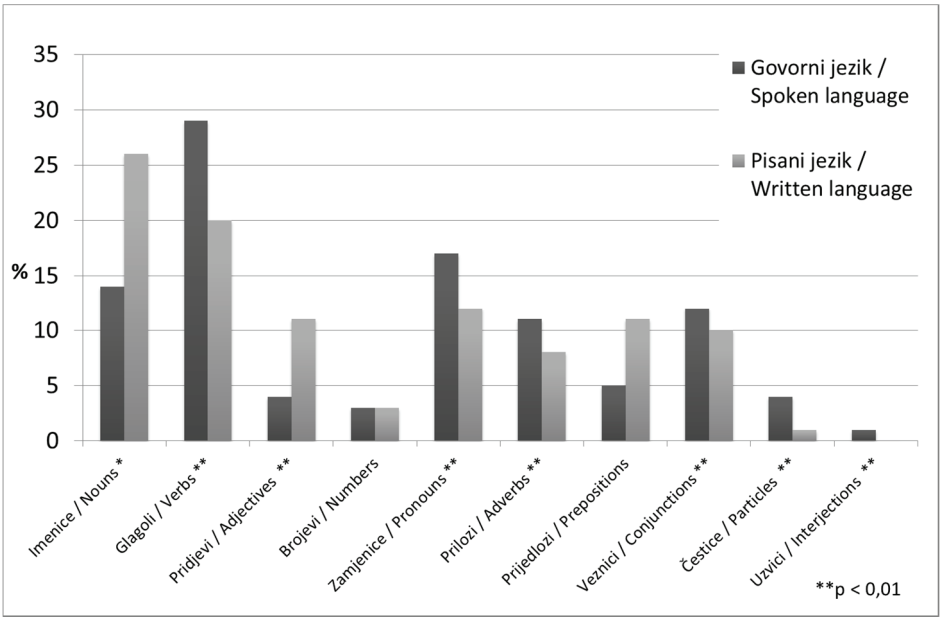
Kako bi se dodatno opisala zastupljenost vrsta riječi u govornom i pisanom jeziku, t-testom za nezavisne uzorke uspoređena je proporcija zastupljenosti svake pojedine vrste riječi. K-S testom provjeren je i preduvjet normalnosti distribucije pojedinih vrsta riječi za govorni i pisani jezični modalitet te sve distribucije pokazuju normalnu ili približno normalnu raspodjelu rezultata. U Tablici 2. prikazana je proporcija zastupljenosti pojedinih vrsta riječi unutar pojavnica u govornom i pisanom jeziku, kao i značajnost dobivenih razlika. Rezultati su prikazani i grafički (Slika 2.).

Tablica 2. Zastupljenost pojedinih vrsta riječi u govornom i pisanom jeziku
Table 2. The distribution of particular parts of speech in spoken and written language

	Vrsta riječi / Parts of speech	Uzorak / Sample	M	SD	t-test (df)	Cohenov *d* / Cohen's *d*
Punoznačne / Content words	Imenice / Nouns	Govorni jezik	0,14	0,05	-13,54** (89)	-2,87
		Pisani jezik	0,26	0,03		
	Glagoli / Verbs	Govorni jezik	0,29	0,07	6,73** (89)	1,43
		Pisani jezik	0,20	0,05		
	Pridjevi / Adjectives	Govorni jezik	0,04	0,02	-12,44** (89)	-2,64
		Pisani jezik	0,11	0,03		
	Brojevi / Numbers	Govorni jezik	0,03	0,03	0,08 (89)	/
		Pisani jezik	0,03	0,01		
	Zamjenice / Pronouns	Govorni jezik	0,17	0,05	5,85** (89)	1,24
		Pisani jezik	0,12	0,03		
	Prilozi / Adverbs	Govorni jezik	0,11	0,04	4,88** (89)	1,03
		Pisani jezik	0,08	0,02		
Odnošajne / Grammatical words	Prijedlozi / Prepositions	Govorni jezik	0,05	0,02	-12,90** (89)	-2,73
		Pisani jezik	0,11	0,02		
	Veznici / Conjunctions	Govorni jezik	0,12	0,04	3,47** (89)	0,74
		Pisani jezik	0,10	0,02		
	Čestice / Particles	Govorni jezik	0,04	0,02	10,03** (89)	2,13
		Pisani jezik	0,01	0,01		
	Uzvici / Interjections	Govorni jezik	0,01	0,01	6,95** (89)	1,48
		Pisani jezik	0,00	0,00		

**$p < 0,01$

Iz Tablice 2. i Slike 2. vidljivo je kako postoje značajne razlike između govornog i pisanog jezika u zastupljenosti imenica, glagola, pridjeva, zamjenica, priloga, prijedloga, veznika, čestica i uzvika. U govornom jeziku veća je zastupljenost glagola, zamjenica, priloga, veznika, čestica i uzvika, dok je u pisanom jeziku veća zastupljenost imenica, pridjeva i prijedloga. Značajna razlika nije dobivena samo u zastupljenosti brojeva, što upućuje na podjednaku zastupljenost ove vrste riječi unutar oba modaliteta. Općenito gledajući, u govornom modalitetu najzastupljeniji su glagoli, a u pisanom imenice, dok su u oba modaliteta najmanje zastupljeni brojevi i uzvici.



Slika 2. Usporedba zastupljenosti pojedinih vrsta riječi unutar pojavnica u govornom i pisanom jeziku

Figure 2. Comparison of the distribution of particular parts of speech within tokens in spoken and written language

Imenice, pridjevi i prijedlozi značajno su zastupljeniji u pisanom jeziku. Imenice označavaju bića, stvari i pojave, dok pridjevi opisuju osobine tih imenica. Ove se vrste riječi najviše koriste pri opisivanju i preciznijem iznošenju informacija koje upravo karakteriziraju pisani jezik (Biber, 1988). Funkcija zamjenica jest da zamjenjuju druge riječi, primjerice već spomenute osobe ili mjesta, što je češće u govornom jeziku vrlo vjerojatno jer su u govornom jeziku, zbog njegove multimedijalnosti, informativnije i jasnije negoli u pisanom. Naime, u govornom se jeziku mogu kombinirati s neverbalnim znakovima poput gesti, pogleda, pokreta i sl., dok se u pisanome, u nedostatku jasnoće izraza i informativnosti, izbjegavaju. Nepromjenjive riječi, kao što je već prethodno spomenuto, izriču odnose između značenja punoznačnih riječi. Ovoj skupini pripadaju veznici koji stvaraju odnose među riječima ili rečenicama, potom čestice koje omogućavaju oblikovanje ili preoblikovanje rečenice, uzvici koji naglašavaju osjećaj ili raspoloženje, a češće se pojavljuju u govornom jeziku moguće zbog toga što se njima nerijetko formiraju oklijevajuće i nedovršene sintaktičke strukture koje su karakteristične za govorni jezik (Pietilä, 1999) (primjerice, "parkira se negdje (.) tamo u onome. tu smo znači +/. tak [: tako], ma dobro. +< <aha> [!]".

"jel [: je li]? "; vidi više primjera na TalkBank-u – http://talkbank.org/browser/index.php?url=CABank/Croatian/), a izbjegavaju se u pisanom.

3.4. Sintaktička složenost izražena u PDKJ-u

Već je spomenuto da je prosječna duljina komunikacijske jedinice – PDKJ (MLCU, engl. *mean length of communication unit*) jedna od mjera koja prikazuje sintaktičku složenost.

Kako bi se provjerilo razlikuje li se značajno sintaktička složenost mjerena u duljini komunikacijske jedinice govornog i pisanog jezika, u oba su uzorka označene C-jedinice, zatim je izračunat PDKJ za svakog ispitanika i u konačnici je t-testom za nezavisne uzorke uspoređen PDKJ govornog i pisanog jezika. Prije računanja t-testa, K-S testom provjeren je preduvjet normalnosti distribucije rezultata PDKJ-a za govorni (K-S Z = 0,88; p > 0,05) i pisani (K-S Z = 0,84; p > 0,05) uzorak te se u oba rezultati raspoređuju prema normalnoj distribuciji. Deskriptivna statistika i značajnost testa prikazani su u Tablici 3.

Tablica 3. Sintaktička složenost (prosječna duljina komunikacijske jedinice; PDKJ) u govornom i pisanom jeziku; usporedba sintaktičke složenosti u govornom i pisanom jeziku

Table 3. Syntactic complexity (mean length of communication unit; MLCU) in spoken and written language; comparison of syntactic complexity in spoken and written language

Varijabla / Variable	Uzorak / Sample	Min. / Min	Maks. / Max	M	SD	t (df)	Cohenov *d* / Cohen's *d*
Sintaktička složenost (PDKJ)	Govorni jezik	2,55	6,89	4,56	1,14	-13,95** (103)	-2,75
	Pisani jezik	6,05	12,15	8,50	1,60		

**p < 0,01

T-test za nezavisne uzorke pokazuje značajno veću sintaktičku složenost pisanog u odnosu na govorni jezik. Dakle, pisani jezik ima veću sintaktičku složenost mjerenu u duljini komunikacijske jedinice u odnosu na govorni jezik. Slični rezultati dobiveni su i u istraživanjima navedenima u nastavku.

Kada se govori o sintaktičkoj složenosti, nije preporučljivo osloniti se samo na jednu mjeru, tj. jednu njezinu dimenziju, međutim konsenzus oko toga koje mjere

čine sveobuhvatan set još uvijek ne postoji. Wolfe-Quintero i suradnici (1998) i Ortega (2003) iznijeli su preglede velikog broja mjera sintaktičke složenosti korištenih u istraživanjima pa Ortega (2003) u konačnici predlaže šest različitih mjera, dok Wolfe-Quintero i suradnici (1998) predlažu i neke dodatne. Dakle, istraživanja široko variraju u uporabi mjera sintaktičke složenosti, što otežava njihovu međusobnu usporedbu. Najčešće upotrebljavana opća mjera sintaktičke složenosti je prosječna duljina iskaza koji se najčešće definira ili kroz T-jedinice (Scott i Windsor, 2000) ili kroz C-jedinice.

Sintaktička složenost općenito se smatra kvalitativnom dimenzijom jezika te se pokazalo kako vještiji govornici koriste složenije sintaktičke strukture (Housen i sur., 2012), što govori u prilog tome da upravo one sugeriraju i zahtjevniji jezik nekog diskursa. Neka prethodna istraživanja u drugim jezicima pokazuju kako upravo pisani jezik sadrži složenije sintaktičke strukture u odnosu na govorni. Naime, u govoru su vrlo česta oklijevanja i nedovršeni iskazi (Pietilä, 1999), što ga čini sintaktički jednostavnijim. Zhang (2013) također smatra da su rečenice koje se proizvode u pisanom jeziku znatno dulje i složenije od odgovarajućih sintaktičkih jedinica koje se proizvode u govornom jeziku.

3.5. Metodološki nedostatci i praktične implikacije

Premda su iz ovog rada proizašli neki zanimljivi podatci, ne treba zanemariti metodološke nedostatke istraživanja. Naime, jezični uzorci od dvadesetak tisuća pojavnica, premda djeluju veliko, za jezična su istraživanja prilično maleni, a time i manje reprezentativni. Također, nisu kontrolirani prostorni jezični varijeteti te su analizirane samo neke od mogućih mjera (primjerice, PDKJ kao mjera sintaktičke složenosti). U daljnjim istraživanjima ove nedostatke svakako treba uzeti u obzir.

Analize na specijaliziranim jezičnim korpusima kao što su HrAL i HKNPJ posebno su zanimljive logopediji kao znanstvenoj disciplini koja se bavi temeljnim i primijenjenim istraživanjima u području jezika i jezične patologije. Primjerice, vrlo su korisni pri izradi dijagnostičkih instrumenata, za procjenu govornog i pisanog jezika ili za utvrđivanje različitih jezičnih normi. Također, komparativne studije govornog i pisanog jezika vrlo su korisne u proučavanju i poučavanju drugog jezika (J2). Primjerice, za engleski jezik postoje gramatički priručnici za govorni i pisani jezik (npr. *Student Grammar of Spoken and Written English*; Biber i sur., 2002). Kako bi se formirale ovakve jezične norme i opisala jezična obilježja pojedinih jezičnih poremećaja, potrebno je provesti mnogo komparativnih istraživanja na

specijaliziranim korpusima govornog i pisanog jezika urednih govornika i govornika s jezičnim poremećajima. Ovakva su istraživanja u hrvatskom jeziku na samom začetku, a ovaj je rad svojevrstan doprinos tim nastojanjima.

4. ZAKLJUČAK

Govorni i pisani jezik kod govornika hrvatskog jezika gotovo da nisu uspoređivani. Rezultati ovog istraživanja u skladu su s istraživanjima provedenima u drugim jezicima i pokazuju kako je prosječna leksička raznolikost pisanog jezika značajno veća od leksičke raznolikosti govornog jezika. Razlike u leksičkoj gustoći između govornog i pisanog jezika marginalno su značajne u smjeru veće gustoće pisanog jezika. Zastupljenost imenica, pridjeva i prijedloga veća je u pisanom jeziku, dok je u govornom jeziku veća zastupljenost glagola, zamjenica, priloga, veznika, čestica i uzvika. Također, postoji značajna razlika u sintaktičkoj složenosti između govornog i pisanog jezika te pisani jezik ima veću sintaktičku složenost mjerenu u duljini komunikacijske jedinice u odnosu na govorni jezik.

Zahvale: Ovaj rad nastao je u sklopu dvaju projekata: "Jezična obrada u odraslih govornika" (*Adult Language Processing*, ALP; HRZZ; UIP-11-2013-2421) Hrvatske zaklade za znanost te "Računalni asistent za pomoć pri unosu teksta osobama s jezičnim teškoćama" (*Computer assistant for text input for persons with language impairment*; RAPUT, EU, RC.2.2.08-0050), financiran iz Europskih strukturnih i investicijskih fondova.

REFERENCIJE

Allwood, J. (1998). Some frequency based differences between spoken and written Swedish. *Proceedings of the 16th Scandinavian Conference of Linguistics* (ur. T. Haukioja), 18–29.

Badurina, L. (2008). *Između redaka: studije o tekstu i diskursu*. Zagreb: Hrvatska sveučilišna naklada.

Barić, E., Lončarić, M., Malić, D., Pavešić, S., Peti, M., Zečević, V., Znika, M. (2005). *Hrvatska gramatika.* Zagreb: Školska knjiga.

Bartsch, C. (1997). Oral style, written style, and Bible translation. *Notes on Translation* **11**, 3, 41–48.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., Leech, G. (2002). *Longman Student Grammar of Spoken and Written English.* Harlow, Essex: Longman.

Brlobaš, Ž. (2001). Teorijska promišljanja o vrstama riječi. *Suvremena lingvistika* **51,** 1–2, 267–279.

Chafe, W. L. (1992). The flow of ideas in a sample of written language. U W. C. Mann i S. A. Thompson (ur.), *Discourse Description: Diverse Linguistic Analyses of a Fund-raising Text*, 267–294. Amsterdam: J. Benjamins.

Chaudron, C. (1988). *Second Language Classrooms: Research on Teaching and Learning.* New York: Cambridge University Press.

Chodorow, M., Burstein, J. (2004). *Beyond Essay Length: Evaluating e-rater's Performance on TOEFL Essays (Research Report No. 73).* Educational Testing Service.

Crookes, G. (1990). The utterance, and other basic units for second language discourse analysis. *Applied Linguistics* **11**, 2, 183–199.

Curnow, A. (1979). Analysis of written style: An imperative for readable translations. *READ* **14**, 2, 75–83.

Drieman, G. H. (1962). Differences between written and spoken language: An exploratory study. *Acta Psychologica* **20**, 78–100.

Foster, P., Tonkyn, A., Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics* **21**, 3, 354–375.

Ghasemi, H., Jahromi, M. K. (2014). The differences between spoken and written discourses in English. *International Journal of Language Learning and Applied Linguistics World* **6**, 4, 147–155.

Halliday, M. A. K. (1985). *Spoken and Written Language.* Geelong Vict.: Deakin University.

Hoff, E. (2013). *Language Development*. Cengage Learning.

Housen, A., Kuiken, F., Vedder, I. (ur.) (2012). *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Amsterdam: Benjamins.

Hunt, K. W. (1965). *Grammatical Structures Written at Three Grade Levels*. NCTE Research Report No. 3: Champaign, Illinois, USA.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working Papers* **53**, 61–79.

Klobučar Srbić, I. (2008). Obol korpusne lingvistike suvremenoj leksikografiji. *Studia Lexicographica* **2**, 3, 39–51.

Kovačević, M. (2002). Hrvatski korpus dječjeg jezika. Dostupno na http://childes.psy.cmu.edu/ [posljednji pristup 18. siječnja 2017.].

Kuvač Kraljević, J., Hržica, G. (2016a). Croatian adult spoken language corpus (HrAL). *Fluminensia: Journal for Philological Research* **28**, 2. Dostupno na http://talkbank.org/access/CABank/Croatian.html [posljednji pristup 27. siječnja 2016.].

Kuvač Kraljević, J., Hržica, G. (2016b). *Hrvatski korpus neprofesionalnog pisanog jezika*. Zagreb: Edukacijsko-rehabilitacijski fakultet, Laboratorij za psiholingvistička istraživanja.

Kuvač, J., Palmović, M. (2007). *Metodologija istraživanja dječjega jezika*. Jastrebarsko: Naklada Slap.

Loban, W. (1976). *Language Development: Kindergarten through Grade Twelve*. NCTE Committee on Research Report No. 18.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*, 3. izd. Mahwah, NJ: Lawrence Erlbaum Associates.

Malvern, D., Richards, B., Chipere, N., Durán, P. (2004). *Lexical Diversity and Language Development. Quantification and Assessment.* Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Matić, A., Hržica, G., Kuvač Kraljević, J., Olujić, M. (u tisku) Syntactic complexity of spontaneous spoken language of adult Croatian speakers. *Proceedings of Croatian Applied Linguistics Society.* Frankfurt am Main: Peter Lang.

Miller, J. F., Andriacchi, K., Nockerts, A. (2015). *Assessing Language Production Using Salt Software: A Clinician's Guide to Language Sample Analysis*, 2. izd. Middleton, WI: SALT Software LLC.

O'Donnell, R. C. (1973). Some syntactic characteristics of spoken and written discourse. *Studies in Language Education Report n. 4*, 0–15.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* **24**, 4, 492–518.

Pavličević-Franić, D. (2005). *Komunikacijom do gramatike.* Zagreb: Alfa d.d.

Pietilä, P. (1999). L2 Speech. Oral proficiency of students of English at university level. *Anglicana Turkuensia* **19**, 1236–4754. Turku: University of Turku.

Scott, C. M., Windsor, J. (2000). General language performance measures in spoken and written narrative and expository discourse of school-age children with

language learning disabilities. *Journal of Speech, Language, and Hearing Research* **43**, 2, 324–339.

Shriberg, E. (2005). Spontaneous speech: How people really talk and why engineers should care. *Interspeech*, 1781–1784.

Smith, J. (1987). A search for naturalness in translated material. *SIL-Mexico Branch Workpapers* **9**, 101–106.

Tadić, M. (1998). *Hrvatski nacionalni korpus*. Zagreb: Filozofski fakultet, Zavod za lingvistiku.

Ure, J. (1971). Lexical density and register differentiation. U G. Perren i J. L. M. Trim (ur.), *Applications of Linguistics*, 443–452. London: Cambridge University Press.

Wolfe-Quintero, K., Inagaki, S., Kim, H.-Y. (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity.* Honolulu, HI: University of Hawaii Press.

Yabuuchi, A. (1998). Spoken and written discourse: What's the true difference? *Semiotica* **120**, 1–2, 1–37.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics* **31**, 2, 236–259.

Zechner, K., Bejar, I. I., Hemat, R. (2007). *Toward an Understanding of the Role of Speech Recognition in Non-Native Speech Assessment (ETS Research Report Series).* Educational Testing Service.

Zhang, B. (2013). An analysis of spoken language and written language and how they affect English language learning and teaching. *Journal of Language Teaching and Research* **4**, 4, 834–838.

Marina Olujić, Ana Matić
marina.olujic@erf.hr, ana.matic@erf.hr
Faculty of Education and Rehabilitation Sciences, University of Zagreb
Croatia

Spoken and written language of adult speakers: how much do they differ?

Summay

Spontaneous spoken language, as well as written language, of adult speakers of Croatian language has not been compared often. Therefore, the characteristics of these two forms of language production have also not been explored. The distribution, i.e. the presence of particular lexical, morphological, syntactic, semantic, etc. characteristics in spoken and written language gives us information about the complexity of these two modalities of language production, and also provides insight into characteristics of language production of adult speakers of Croatian language.

The aim of the current paper is to present and compare lexical and syntactic features of spontaneous spoken and written language. For this reason, features and differences in lexical diversity, lexical density and the distribution of parts of speech, as well as one measure of syntactic complexity have been explored. The research has been conducted using the independent samples of participants. The samples of spoken and written language have been taken over from the Croatian Adult Spoken Language Corpus (HrAl) and Croatian Corpus of Non-professional Written Language (HKNPJ).

The results suggest that 1) lexical diversity of written language is significantly higher than lexical diversity of spoken language; 2) there are differences in the distribution of parts of speech; spoken language has greater representation of verbs, pronouns, adverbs, conjunctions and particles, while written language has greater representation of nouns, adjectives and prepositions; and 3) written language has higher syntactic complexity measured in the length of c-units.

The analyses of specialised language corpora such as HrAl and HKNPJ are especially interesting to experts in the field of speech and language pathology, a scientific discipline oriented to fundamental and applied research in the area of language and language pathology. For example, they are useful for the development of diagnostic instruments, for conducting an authentic assessment of spoken and written language or for developing different language

norms. Comparative studies of spoken and written language are also useful for investigating and teaching second language (L2).

In order to create language norms and describe characteristics of certain language impairments, studies of specialised corpora of spoken and written language need to be conducted. This study is a step forward when it comes to corpora analyses in Croatian language.

Key words: spoken language, written language, corpus, adult speakers