Jelena Kuvač Kraljević, Gordana Hržica, Marina Olujić,
Lana Kologranić Belić, Marijan Palmović i Ana Matić
Edukacijsko-rehabilitacijski fakultet u Zagrebu

Uzorkovanje specijaliziranih korpusa govornog i pisanog jezika odraslih: izazovi i nedoumice

Važnost jezičnih korpusa u suvremenoj je lingvistici nedvojbena i dobro dokumentirana. Iako je većina općih korpusa sastavljena od pisanih tekstova različitih žanrova, tendencija je svjetskih nacionalnih korpusa uključiti do 10 % uzoraka govornog jezika prikupljenih na temelju neformalne konverzacije. Da bi se to uspjelo u hrvatskoj korpusnoj lingvistici, potrebno je izraditi specijalizirane govorne korpuse. Također, s povećanjem broja istraživanja posebnih skupina govornika javlja se i sve veća potreba za drugim specijaliziranim korpusima. Takvi korpusi oblikuju se ograničavanjem uzorkovanja na određenu skupinu (primjerice djeca, dvojezični govornici, osobe s afazijom...) ili vrstu jezičnog uzorka (primjerice spontani govor, pisanje tekstova različite razine strukturiranosti), a mogu se podijeliti na govorne i pisane korpuse. Pri prikupljanju specijaliziranih korpusa nameću se brojna metodološka pitanja kao što su odabir vrste tekstova, način unosa teksta, način označavanja, specifičnosti komunikacijske situacije i slično. U sklopu dvaju projekata *Jezična obrada u odraslih govornika* (Hrvatska zaklada za znanost UIP-11-2013-2421) i *Računalni asistent za pomoć pri unosu teksta osobama s jezičnim poremećajima* (EU Strukturni fondovi RC.2.2.08-0050), oba dodijeljena *Laboratoriju za psiholingvistička istraživanja,* u izradi su dva specijalizirana korpusa. Jedan je *Hrvatski korpus govornog jezika odraslih*, a drugi *Hrvatski korpus neprofesionalnog pisanog jezika*, pri čemu su u oba korpusa uključene osobe bez teškoća u jezičnoj obradi, to jest uredni govornici, ali i osobe s različitim, razvojnim i stečenim, jezičnim poremećajima. U radu se iznose metodološke nedoumice u procesu uzorkovanja, ponajprije u odabiru načina uzorkovanja, kontroli izvanjezičnih čimbenika te oblikovanju materijala za uzorkovanje.

Ključne riječi: specijalizirani korpusi, Hrvatski korpus govornog jezika odraslih, Hrvatski korpus neprofesionalnog pisanog jezika, klinički korpusi

1. Uvod

Unatoč brojnim prigovorima koje je Noam Chomsky šezdesetih godina prošlog stoljeća imao na korpuse, oni su se neprekidno do danas potvrđivali kao vrijedan izvor jezičnih podataka. Naime, Chomskijev mentalistički pristup lingvističkim temama temeljio se na nužnosti objašnjavanja mentalne pozadine jezičnog ponašanja, ali ne i samog jezičnog ponašanja (Chomsky 1965) koje je upravo

internalizirano putem korpusa. Danas ne samo da korpusi predstavljaju vrijedan izvor jezičnih podataka, a što ih čini i pogodnim alatom za brojna jezična istraživanja, već su i široko primjenjivani u različitim znanstvenim područjima. Pri tome treba naglasiti da se ne primjenjuju samo u različitim lingvističkim disciplinama nego, štoviše, svoju primjenjivost svakodnevno potvrđuju i u drugim područjima, posebice interdisciplinarnim, kao što su kognitivna neuroznanost i logopedija.

Korpus je zbirka tekstova ili zbirka manjih jezičnih uzoraka koja je organizirana prema jasnim lingvističkim kriterijima kako bi, u cjelini, činila veći i reprezentativan jezični uzorak (Tadić 2003). Zbog tog je obilježja korpus pogodan za računalna pretraživanja i objektivne lingvističke analize (Francis 1992), ali može služiti i kao baza za izradu različitih jezičnih tehnologija. Mogućnost pretraživanja osnovnih informacija, poput evidencije (sadrži li korpus određenu jezičnu jedinicu ili ne), frekvencije (učestalost pojavljivanja jezične jedinice u korpusu) i relacije (odnos jedne jezične jedinice s drugim jedinicama u korpusu), smatra se temeljnom svrhom korpusa (Tadić 2003).

Korpusi se razlikuju sadržajem, veličinom i strukturom. I dok neki, posebice opći korpusi, broje i više stotina milijuna pojavnica (na primjer *The Bank of English*, *COBUILD*), specijalizirani korpusi obično su mnogo manji pa sadrže nekoliko tisuća pojavnica (na primjer korpus pripovjednih tekstova *Narrative English Gopnik Corpus*, *CHILDES*).

Po svojoj strukturi razlikuju se opći korpusi, reprezentativni za jezik u cjelini, i specijalizirani korpusi koji obuhvaćaju samo jedan jezični varijetet (Klobučar Srbić 2008). Opći korpusi téžē obuhvatiti što više jezične raznolikosti određenog jezika. Sastoje se od velika broja pojavnica koje odražavaju trenutačno ili povijesno stanje određenog jezika. Tradicionalno se takvi korpusi temelje na pisanim tekstovima i strukturirani su prema unaprijed određenim žanrovima iz različitih vremenskih razdoblja njihova nastanka. Međutim, današnji su, primjerice engleski opći korpusi, još širi jer uključuju tekstove šireg opusa. Na primjer *Corpus of Contemporary American English* (*COCA*) uključuje u svoje uzorke i akademski potkorpus, filmske scenarije te uzorke govornog jezika preuzete iz televizijskih emisija.

Najveći je hrvatski korpus *Hrvatski nacionalni korpus* (*HNK*) koji trenutačno obaseže 216,8 milijuna pojavnica (Tadić 2009). Osim *HNK-a* postoji i *Hrvatski jezični korpus Instituta za hrvatski jezik i jezikoslovlje* koji broji 84,9 milijuna pojavnica. Kada se učini pregled tih dvaju korpusa, vidljivo je da je riječ o uzorkovanju već postojećih tekstova, novina, knjiga, časopisa i(li) drugih pisanih izvora. Danas je dostupan i hrvatski internetski korpus naziva *hrWAC* veličine 1,2 milijarde pojavnica u verziji 1,0 (Ljubešić, Erjavec 2011) te 1,9 milijardi pojavnica u verziji 2.0 (Ljubešić, Klubička 2014). Iako su svi do sada dostupni hrvatski korpusi oblikovani isključivo na temelju pisanih tekstova različitih žanrova, tendencija je drugih nacionalnih korpusa poput *British National Corpusa* (*BNC*) da u svoju strukturu

uključe do 10 % uzoraka govornog jezika prikupljenih na temelju neformalne konverzacije. Da bi se ta struktura prenijela i u spomenute hrvatske korpuse, potrebno je uspostaviti specijalizirane korpuse spontanog jezika.

2. Specijalizirani korpusi govornog i pisanog jezika odraslih govornika

Interes za izradu specijaliziranih korpusa češće proizlazi iz korpusnoj lingvistici srodnih lingvističkih disciplina kao što su primjerice leksikologija i psiholingvistika. Razlog tomu podudarnost je sadržajne usmjerenosti korpusa sa središnjim temama i istraživačkim pitanjima tih disciplina. U slučaju hrvatskih specijaliziranih korpusa, koji će biti opisani u nastavku, svi do sada oblikovani takvi korpusi razvijeni su u *Laboratoriju za psiholingvistička istraživanja Odsjeka za logopediju*, odnosno razvijeni su u području psiholingvistike potaknuti potrebama logopedske struke.

Tako je devedesetih godina prošlog stoljeća započela izrada *Hrvatskog korpusa dječjeg jezika* (Kovačević 2002). To je trenutačno jedini dostupan hrvatski korpus govornog jezika koji sadrži 136 000 pojavnica dječjeg jezika i oko 152 800 pojavnica govornog jezika odraslih govornika, odnosno pojavnica djetetova ulaznog jezika. S odmakom od 12 godina od uspostavljanja *Hrvatskog korpusa dječjeg jezika*, to jest 2014. godine, započela je izrada dvaju novih specijaliziranih korpusa: *Hrvatskog korpusa govornog jezika odraslih* te *Hrvatskog korpusa neprofesionalnog pisanog jezika*. Sva tri navedena korpusa proizišla su iz istog istraživačkog pitanja: Koja su obilježja pisanog i govornog jezika djece i odraslih govornika hrvatskog jezika? Izrada dvaju novih korpusa ostvaruje se u sklopu nacionalnog projekta *Jezična obrada u odraslih govornika* (Hrvatska zaklada za znanost UIP-11-2013-2421) te *Računalni asistent za pomoć pri unosu teksta osobama s jezičnim poremećajima* (EU Strukturni fondovi RC.2.2.08-0050), oba dodijeljena *Laboratoriju za psiholingvistička istraživanja Odsjeka za logopediju*.

Korpus spontanoga govornog jezika pruža jedinstvene podatke o jeziku, poput učestalosti i distribucije riječi i struktura, varijacije jezičnih jedinica i kontekstualnih odrednica (Pusch 2006). Takvi su korpusi veličinom manji od prethodno opisanih općih korpusa, ali i rjeđi od njih, kao i od korpusa govornog jezika koji nastaju inkorporacijom gotovih materijala poput govornih sadržaja televizijskih emisija, filmova, predavanja i slično. Trenutačno je najveća baza spontanih uzoraka govornog jezika *TalkBank* (talkbank.org). Sastoji se od različitih korpusa sličnih po tome što su sastavljeni od uzoraka govornog i, barem u određenoj mjeri, spontanog jezika. Dio *TalkBanka* je i *Hrvatski korpus dječjeg jezika* (Kovačević 2002) kao dio baze *CHILDES*[1] koja je javno dostupna i namijenjena za objedinjavanje uzoraka dječjeg jezika.

1 Pretraživanje baze CHILDES moguće je putem poveznice http://childes.psy.cmu.edu/, a odabirom *Browsable Database* dolazi se na izbornik korpusa.

Osim izrade korpusa govornog i pisanog jezika odraslih govornika uredne jezične obrade, u sklopu istih korpusa razvijaju se paralelno i klinički korpusi. Izrada *Kliničkog korpusa govornog jezika* primarno uzorkovanog u konverzaciji osobe s afazijom i logopeda tijekom logopedske terapije započet će u jesen 2015. godine, odnosno u drugoj godini provođenja projekta *Jezična obrada u odraslih govornika*. Takav način uzorkovanja spontanoga govora osoba s afazijom uobičajen je u svijetu, primjerice *A Corpus of Dutch Aphasic Speech - CoDAS* (Westerhout 2006), a najveći broj takvih korpusa oblikovanih u različitim jezicima (primjerice engleski, katalonski) te strukturiranih prema unaprijed određenom protokolu postavljen je u riznicu podataka *AphasiaBank* u sklopu baze *TalkBank*. Krajnji je cilj u izradi hrvatskog *Kliničkog korpusa govornog jezika osoba s afazijom* da i on postane dio te baze.

Dostupni podatci znanstvenih istraživanja pokazuju da osobe s jezičnim teškoćama na drugačiji način obrađuju tekst, bilo pri čitanju ili pri pisanju (Salmelin i sur. 1996, Ramus 2014). Iz tog je razloga započela izgradnja *Kliničkog korpusa neprofesionalnog pisanog jezika*, međutim njegova je namjera višestruka: omogućiti lingvističke analize pisanog jezika osoba različita jezičnog statusa različite dobi, razviti jezične tehnologije, i to one koje će omogućiti predviđanje teksta pri njegovu unosu, kao i uočavanje i ispravljanje pogrešaka te u konačnici na temelju razvijene tehnologije razviti mrežnu aplikaciju. Da bi se postavljeni ciljevi i ostvarili, potrebno je osigurati dva preduvjeta: a) sinergijski pristup dvaju područja, logopedije i računalne lingvistike, što je i osigurano u projektu *Računalni asistent za pomoć pri unosu teksta osobama s jezičnim poremećajima* u sklopu kojeg se taj korpus izrađuje te 2) dovoljan broj pojavnica. Do kraja projekta trebao bi biti osiguran korpus velik pola milijuna pisanih pojavnica, od čega bi polovica trebala biti prikupljena na uzorcima pisanih tekstova govornika uredna jezičnog statusa, a polovica na uzorcima pisanih tekstova osoba s različitim jezičnim poremećajima, poput disleksije, afazije, traumatskog oštećenja mozga i posebnih jezičnih teškoća. U svijetu postoji nekoliko sličnih korpusa, ali znatno manjih s obzirom na broj pojavnica te užih s obzirom na uključenost ispitanika i na vrstu jezične teškoće i kronološku dob. Tako primjerice postoje dva korpusa pisanog jezika osoba s disleksijom, od čega je jedan engleski (Pedler 2007), a drugi španjolski – *DysCorpus* (Rello i Llisterri 2012). Osim što je u te korpuse uključen samo pisani jezik osoba osoba s disleksijom, i kronološka je dob ispitanika ograničena na razdoblje rane adolescencije. Oba su korpusa oblikovana sa svrhom izrade alata za pravopisnu provjeru te sadrže 12 000 pojavnica (Pedler 2007), odnosno 1 057 pojavnica (Rello i Llisterri 2012).

Izrada specijaliziranih korpusa kliničkih skupina nije važna samo za opisivanje određene jezične teškoće već i za razvoj alata za pravopisnu provjeru (Korhonen 2008, Pedler 2007), razvoj softvera za predviđanje teksta kao što je primjerice *Prenfriend XL*, razvoj edukativnih igara za djecu s disleksijom (Rello i Llisterri

2012) i obrađivača teksta koji ispravljaju najčešće pogreške u pisanju (Gregor, Dickinson, Macaffer i Andreasen 2003).

3. Metodološke nedoumice u izradi specijaliziranih korpusa

Izrada je specijaliziranih korpusa, posebice kliničkih, mnogo zahtjevnija nego izrada općih korpusa. Razlog je dijelom u uzorku ispitanika – zastupljenost je primjerice posebnih jezičnih teškoća u društvu oko 7 % pa je tim teže osigurati ispitanike. Preostali dio zahtjevnosti proizlazi iz tehničke složenosti uzorkovanja govornog jezika. Primjeri nekih nedoumica u trima koracima uzorkovanja govornih i pisanih korpusa slijede u nastavku.

3.1. Jezični uzorci i sudionici

Prvi je korak u izradi korpusa određivanje sadržaja uzorkovanja i ispitanika. Koliko je zapisa potrebno prikupiti da bi sadržaj korpusa bio dovoljan, a korpus u konačnici ujednačen i dovoljno velik? Koliko je različitih ispitanika potrebno uključiti da bi korpus bio reprezentativan? Pitanja se dodatno usložnjavaju kada se u obzir uzmu regionalni (1), društveni i funkcionalni varijeteti.

(1) Primjeri regionalnih varijeteta

Središnja Istra	Međimurje	Imotska krajina
*MAR: pegula je bijavajk, brižan. *MAR: ča god kipovida i njemu se j' [:je] inbatilo. *MAR: ne da izmišja [:izmišlja] nego je! *MAR: uvi je pa iz škalah, on je pa deset boti huje.	*INV: i devere si miela i pucnihale? *DRA: nu. *INV: koga? *DRA: Španičuvu (.) Treziku je bila jedna +... *DRA: pa Silva s Cestice.	*ANĐ: pa more [*] se nać [: naći] za dvi [: dvije] – tri iljade [: hiljade] kuna. *ANA: pa ne mogu ja sebi tliko [: toliko] priuštit [: priuštiti] da ja dvi [:dvije] -tri iljade [: hiljade] potrošin.
Prijevod		
*MAR: nesretnik je bio uvijek, siroti. *MAR: što god je tko rekao njemu se to dogodilo. *MAR: ne da izmišlja nego je! *MAR: ovaj kao da je pao niz stepenice, on je deset puta gori.	*INV: i djevere si imala i djeveruše? *DRA: da. *INV: koga? *DRA: Španičeva Trezika je bila jedna +... *DRA: i Silva iz Cestice.	*ANĐ: pa može se naći za dvije-tri tisuće kuna. *ANA: pa ne mogu ja sebi toliko priuštiti da ja dvije-tri tisuće potrošim.

Odluka o veličini korpusa, kao i kriterij odabira ispitanika, ovise o sastavljaču korpusa. Nijedan korpus nije dovoljno velik da obuhvati sve riječi jednog jezika. Mali jezični uzorci predstavljaju problem jer vjerojatno neće obuhvatiti jezične pojave koje se rjeđe javljaju. Stoga, što je manji uzorak, to je veća mogućnost da će se u lingvističkoj analizi precijeniti omjer češćih pojava ili podcijeniti omjer rjeđih pojava (Hržica 2011). Hunston (2002) smatra da je, kako bi korpus bio ujednačen i reprezentativan, poželjno da sadrži ujednačenu količinu zapisa svih relevantnih društvenih skupina. No jezik postoji u nepoznatim i nedokučivim kvantitetama i u nepoznatom rasponu varijeteta te je sve relevantne grupe nemoguće obuhvatiti. Prema Hunston (2002) postoje dva pristupa u rješavanju tog problema. Jedan je da se pri kreiranju plana uzorkovanja sastavi lista relevantnih varijabli poput dobi, spola, socijalnog statusa i rodnog mjesta ispitanika, kao i varijable o govornom ili pisanom kontekstu. Druga mogućnost jest uzeti u obzir sve varijable i sve podatke koje je moguće prepoznati, što omogućava sveobuhvatniji korpus, ali manju ujednačenost. Također, potrebno je promišljati i o pitanjima koliko korpus treba sadržavati ispitanika i koliko je iskaza nužno da svaki ispitanik proizvede kako bi se njegovi iskazi uvrstili u korpus. Većina autora smatra da je za pouzdane jezične analize potrebno stotinjak iskaza po ispitaniku (Cameron 2001), dok neki smatraju da je i pedeset iskaza dovoljno (Shipley i McAfee 2015).

3.2. Uvjeti uzorkovanja

Nakon određivanja veličine i raznolikosti uzorka koji će predstavljati korpus potrebno je osmisliti najbolje uvjete u kojima će se prikupljati zapisi spontanoga govora i pisanog jezika te pri tome uvažavati sve propisane etičke kriterije. Jezični korpus koji se temelji na podatcima govornog jezika sastoji se od dva dijela: zvučnog ili videozapisa te prijepisa ili transkripta (Kuvač i Palmović 2007), dok se pisani korpusi sastoje samo od pisanih uzoraka koje piše sam ispitanik. Prilikom izrade govornog korpusa prvo pitanje koje se nameće jest je li bolje govorne uzorke prikupljati uređajem za snimanje zvuka (diktafonom) ili videokamerom. Premda videozapisi sadrže mnogo više informacija o obilježjima komunikacije, jezika i govora ispitanika, transkripcija takvog uzorka traje mnogo duže od transkripcije zvučnog zapisa, a javljaju se i drugi problemi jer takav zapis zahtijeva veće kapacitete za pohranu, a pretraživanje i obrada mogu biti složeniji. Tako transkripcija i obrada jednog sata videozapisa traje i do dvadeset sati, a jednog sata zvučnog zapisa do šest sati (Rowland, Fletcher i Freudenthal 2008). No videozapisi pružaju i mnogo više podataka pa se sve češće rabe (Kuvač i Palmović 2007), a posebice su pogodni za prikupljanje specijaliziranih govornih korpusa gdje ispitanici s, primjerice, afazijom proizvode vrlo malo govornog jezika pa svaki način komunikacije koji se vidi može biti važan za opisivanje obilježja jezika osoba s takvim poremećajem.

Prikupljanje uzoraka u neuobičajenom kontekstu kao što je kabinet nekog stručnjaka može narušiti spontanost, a time i autentičnost i kvalitetu uzorka. Laboratorijski uvjeti čine posebnu društvenu situaciju pa govornik prilagođava svoj jezik (Kuvač i Palmović 2007) sukladno poznatom paradoksu promatrača koji objašnjava kako se očekuje da se osoba spontano ponaša premda se nalazi u nespontanim uvjetima (Labov 1972). Spontanost govora posebno je osjetljiva varijabla na koju mogu utjecati još neki vanjski čimbenici. Primjerice, spontanost govora odraslih ispitanika moguće je narušiti traženjem informiranog pristanka za sudjelovanje u prikupljanju govornog jezika ili zbog prisutnosti uređaja za snimanje zvučnog zapisa. I dok je uređaj za snimanje zvuka donekle moguće ukloniti iz vidokruga ispitanika stavljajući mikrofon na skriveno mjesto, klasičnu je kameru gotovo nemoguće sakriti (2).

(2) Utjecaj prisutnosti snimača na spontanost govora

*MAT:	neman ja šta [: što] tu (.).
*JEL:	neš [: nećeš] da radiš ili?
*ANA:	anđelko molin te (.) ...
*ANA:	nemoj me snimat [: snimati] s mobitelon i kameron.
%com:	JEL se smije.

Premda u prikupljanju pisanih uzoraka problem, barem u većoj mjeri, ne predstavlja spontanost, problemi uzorkovanja pisanog jezika tiču se u prvom redu sveobuhvatnosti mogućih struktura pisanog teksta. Kako bi se obuhvatilo što više razina strukturiranosti i stilova pisanja u pisanome korpusu, materijal za uzorkovanje koji je pripremljen za ispitanike opsežan je. Primjerice, uzorkovanje pisanog jezika u sklopu *Hrvatskog korpusa neprofesionalnog pisanog jezika* sastoji se od pisanja dvaju eseja, odgovaranja na deset pitanja (pet odgovora namijenjenih pisanju rukom, pet na računalu), pisanja dviju priča prema slikovnom predlošku, dvaju pisama te dviju formalnih poruka. Iako izlaze iz okvira spontanosti jer imaju najveću razinu strukturiranosti, pišu se još i dva diktata: jedan se piše ručno, drugi na računalu. Međutim, pisanje tolikog broja zadataka odjednom dugotrajno je i zahtjevno, što povećava vjerojatnost pogrešaka i kod ispitanika urednog jezičnog statusa, a osobito kada je riječ o djeci, starijim osobama i(li) osobama s jezičnim poremećajima. Ako se ispitivanje provodi u više navrata, uvjeti za sve ispitanike nisu ujednačeni, što također može utjecati na konačan pisani proizvod.

3.3. Transkripcija

Pitanje transkripcije posebno je važno pri izradi korpusa spontanoga govora, no na probleme se nailazi i u korpusu pisanog jezika.

Transkript spontanoga govornog jezika trebao bi biti takav da se iskaze doživi upravo onakvima kako ih je ispitanik proizveo te je ključno kodirati pojedina

obilježja poput ponavljanja, oklijevanja, zastajkivanja ili pogrešnih početaka (Cameron 2001). Problem predstavlja i pitanje o načinu zapisivanja iskaza, to jest određivanje kako će se govorni niz razlomiti na sastavne dijelove (3). Određivanje kraja iskaza oduvijek je predstavljalo problem ispitivačima. Granica se iskaza najčešće određuje na temelju izgovornih karakteristika kao što su pauza ili intonacija koje su podložne subjektivnoj procjeni, ali i nisu nužno vezane uz strukturu izrečenoga. Kao alternativa osmišljen je i način odvajanja govornog niza koji se temelji na lingvističkim jedinicama kao što je komunikacijska jedinica (engl. *Communication Unit*, *CU*; Logan 1966, vidi Crookes 1990). No, kriterij je razdiobe govornog niza u tom slučaju sasvim različit od onoga koji se primjenjuje za razdvajanje iskaza.

(3) Primjer nedoumica u označavanju kraja iskaza

*PAT: i tak da: ono: (.) a i nekak [: nekako] ta (.) em: (.) obiteljska atmosfera te baš nekak [: nekako] (.) ono: (.) kao da te potiče da: sve drugo radiš samo ne ono kaj si odrediš ... na primjer, ako želiš učiti, onak [:onako], stalno te neko, ono, na neki način, pod navodnicima uznemirava da ne možeš to napraviti nego je ta atmosfera da te stalno onak [: onako], nekako vuče da se usmjeriš na sve drugo samo ne na to, kaj ti je zapravo naj [//] najvažnije

Bez obzira na područje istraživanja kojim se istraživač bavi, smatra se da svaki prijepis, kako bi bio pouzdan, treba biti autentičan, odnosno mora u potpunosti odgovarati interakciji koja se opisuje, i praktičan, što znači da mora biti primjenjiv i drugim korisnicima (Edwards 1993). Da bi se ta dva kriterija zadovoljila, prijepis treba biti strukturiran u skladu sa sljedeća tri načela: 1) načelom kategorije dizajna – svaki slučaj u prijepisu mora imati jedinstvenu oznaku kojom se razlikuje od ostalih slučajeva, 2) načelom čitljivosti – informacije moraju biti lako dostupne istraživaču te vizualno i prostorno zamjetljive i 3) načelom računalne provodljivosti – sustavan i predvidljiv prijepis.

Poradi lakše čitljivosti i praktičnosti pri transkripciji se obično bira jedan od sustava transkripcije i kodiranja. Takvi se sustavi sastoje od popisa pravila transkripcije, kodova za označavanje različitih jezičnih karakteristika (na primjer samoispravljanje, ponavljanje riječi, pauza...) te računalnog programa koji omogućuje transkripciju i kasniju obradu podataka. Neki su od takvih sustava javno dostupni (uz bazu *TalkBank* tako se vezuje sustav za transkripciju *Codes for Human Analysis of Transcripts (CHAT)* i računalni program *Computerized Language Analysis (CLAN)*), a neki isključivo komercijalno (sustav *Systematic Analysis of Language Transcripts (SALT)*). Odabir prikladnog sustava transkripcije važan je jer svaki sustav nužno ima neka ograničenja vezana uz transkripciju i uz kasniju automatiziranu ili poluautomatiziranu obradu. Također, o formatu pohrane ovisi i koliko će jezični uzorci biti prikladni za obradu drugim jezičnim tehnologijama (na primjer automatsko morfološko označavanje ili lematizacija).

Problem prilikom transkripcije predstavlja i postupak transkripcijske selekcije koji se odnosi na pitanja o tome što uključiti, a što isključiti iz prijepisa (Miller i Klee 1995). Istraživač odlučuje o tome koja će obilježja promatrati i kodirati, a koja će ispustiti, ali samo ona jezična obilježja koja se kodiraju mogu biti analizirana. Problem nastaje kada istraživač odluči svoj korpus učiniti javno dostupnim, a prilikom prijepisa pretjerano se usmjerava samo na vlastiti nacrt istraživanja pa ne kodira velik broj jezičnih pojava. S druge strane kodiranje velikog broja jezičnih pojava predstavlja dodatni posao koji je ekonomski i vremenski zahtjevniji, a prijepis s previše detalja pun je distraktora koji ometaju istraživača da se usredotoči na ciljane sastavnice analize (DuBois, Schuetze-Coburn, Cumming i Paolino 1993).

Kada je riječ o govornom jeziku, redovito se nameće i odluka o tome hoće li se prijepis oblikovati fonetski ili fonološki. Pri fonetskoj transkripciji čuvaju se izgovorne razlike (na primjer, *ovo*, *'vo* i *'o'*), ali se gubi preciznost u obradi. Računalni će program takve elemente zbrojiti kao tri različnice iako je riječ o istom obliku. Taj je problem moguće riješiti kombinacijom fonetskog i fonološkog načela. Na sličan je način moguće tretirati i dijalektalne oblike, kao i pogreške u govorenju i pisanju (4). No, takve su pojave još uvijek poseban izazov za jezične alate koji su nastali na temelju kontroliranih tekstova velikih nacionalnih korpusa te je njihova primjena na specijaliziranim korpusima znatno složenija.

(4) Kombinacija fonetskog i fonološkog načela u kodiranju

*JOS:	mi doli [:dolje] čekamo kavu bokte [: bog te] mazo [: mazao].
*LUC:	xxx.
%com:	zvuk televizije.
*JEL:	kakav je ovo divan film, a?
%com:	kraća pauza.
*INV:	nenan [: ne znam] ćać [:ćaća] koi [: koji] je o: [:ovo] film?

Kao i govorni, i korpus pisanog jezika ima svoje probleme transkripcije, osobito ako je riječ o spontanom pisanju. Jedan od prvih problema koji se ističe problem je točnosti prijepisa. Ako ispitanik piše primjerice esej na zadanu temu, zbog nečitkog rukopisa ponekad je teško biti siguran da je učinjen točan prijepis. On će ovisiti isključivo o prepisivaču i njegovoj sposobnosti da dešifrira rukopis u kompletnom tekstu ili pak samo pojedina slova, vizualno slična jedna drugima, kao što su vokali. Primjerice, ako rečenicu *Sjedila sam s Tanjom*. ispitanik napiše nečitko, osoba koja vrši prijepis ne može iščitati je li ispitanik napisao *Sjedila sam s Tanjom* ili *Sjedila sam s Sanjom.* jer ispitanik na sličan način piše velika pisana slova T i S. Ako je ispitanik napisao *Sjedila sam s Sanjom.*, u rečenici koju je napisao nalazi se pogreška (*sa Sanjom* umjesto *s Sanjom*), a ako je napisao *Sjedila sam s Tanjom.* – pogreške nema. Ta razlika čini kvantitativnu razliku u ukupnom broju pogrešaka, kao i kvalitativnu u analizi istih – a ovisi isključivo o prepisivaču.

Kako jezični alati za morfološku i sintaktičku obradu rade na rečeničnoj razini i ne uzimaju u obzir odnose koji nadilaze rečenicu, potrebno je odijeliti rečenice u novi redak. Neki ispitanici, osobito oni s jezičnim teškoćama, nerijetko izostavljaju rečenične znakove i na ispitivaču je da procijeni granicu kod izrazito distorziranog teksta (npr. kod fragmentiranih rečenica bez interpunkcije). Ponekad je jako teško procijeniti gdje je granica kod takvih tekstova.

4. Umjesto zaključka

Korpusi već desetljećima predstavljaju vrijedan izvor jezičnih podataka, a ekološka vrijednost korpusnih istraživanja sastoji se u dohvatu osnovnih spoznaja o jeziku u stvarnoj uporabi, preciznije, mogućnosti dohvata dokumentiranih podataka o čestotnosti i distribuciji riječi i struktura, varijacijama i kontekstualnim obrascima (Pusch 2006). Opisane nedoumice u izradi korpusa pisanog i govornog jezika ni u kojem slučaju ne umanjuju tu važnost i vrijednost. Dapače, one predstavljaju izazov za sve istraživače koji su interesno usmjereni prema korpusima da broj tih nedoumica umanje kako bi se dodatno povećala objektivnost i pouzdanost riznica jezičnih podataka.

Literatura

Cameron, Deborah. 2001. *Working with Spoken Discourse*. London: Sage.

Chomsky, Noam. 1965. *Aspect of Theory of Syntax*. Cambridge Massachusetts: MIT Press.

Crookes, Graham. 1990. „The Utterance and Other Basic Units for Second Language Discourse Analysis." *Applied Linguistics* 11: 183–199.

DuBois, John, Schuetze-Coburn, Stephan, Cumming, Susanna i Danae Paolino. 1993. „Outline of discourse transcription." U *Talking data: Transcription and coding in discourse research, uredili* Edwards, Jane A. i Lampert, Martin D., 45–89. Hillsdale, NJ: Lawrence Erlbaum Associates.

Edwards, Jane. 1993. „Principles and contrasting systems of discourse transcription." U *Talking data: Transcription and coding in discourse research, uredili* Edwards, Jane A. i Lampert, Martin D., 3–32. Hillsdale, NJ: Lawrence Erlbaum Associates.

Francis, Nelson W. 1992. „Language corpora B.C." U *Directions in Corpus Linguistics. Trends in Linguistics. Studies and Monographs, uredio* Svartvik, Jan, 17–32. Berlin: Mounton de Gruyer.

Gregor, Peter, Dickinson, Anna, Macaffer, Alison i Andreasen, Peter. 2003. „Seeword: a personal word processing environment for dyslexic computer users." *British Journal of Educational Technology* 34 (3): 341–355.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hržica, Gordana. 2011. *Glagolske kategorije aspekta, vremena i akcionalnosti u usvajanju hrvatskog jezika.* Doktorska disertacija. Sveučilište u Zagrebu: Filozofski fakultet.

Klobučar Srbić, Iva. 2008. „Obol korpusne lingvistike suvremenoj leksikografiji." *Studia lexicographica* 2 (3): 39–51.

Korhonen, Tuomas. 2008. „Adaptive spellchecker for dyslexic writers." U *Proceedings of the 11th international conference on Computers Helping People with Special Needs, ICCHP '08,* 733–741. Berlin, Heidelberg: Springer-Verlag.

Kovačević, Melita. 2002. Hrvatski korpus dječjeg jezika. http://childes.psy.cmu.edu/ Pristupljeno: 20. svibnja 2015.

Kuvač, Jelena i Palmović, Marijan. 2007. *Metodologija istraživanja dječjeg jezika.* Jastrebarsko: Naklada Slap.

Labov, William. 1972. *Sociolinguistic patterns*. Pennsylvania: University of Pennsylvania Press.

Ljubešić, Nikola, Erjavec, Tomaž. 2011. „hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene." *Text, Speech and Dialogue. Lecture Notes in Computer Science*, Springer. 395–402.

Ljubešić, Nikola, Klubička, Filip. 2014. „{bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian." U *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, uredili Felix Bildhauer, Felix i Schäfer, Roland, 1-24.

Gothenburg, Sweden: Association for Computational Linguistics.

Miller, Jon F. i Klee, Thomas. 1995. „Quantifying language disorders in children." U *Handbook of Child Language, uredili* Fletcher, Paul i MacWhinney, Brian, 545–572. UK: Blackwell.

Pedler, Jennifer. 2007. *Computer correction of real-word spelling errors in dyslexic text.* Unpublished PhD thesis. Birkbeck: University of London.

Pusch, Claus D. 2006. „Corpora of Spoken Discourse." U *Encyclopedia of Language & Linguistics, uredio* Brown, Keith, 226–230. Oxford: Elsevier.

Ramus, Franck. 2014. „Should there really be a „Dyslexia debate"?" *Brain* 137: 3371–3374.

Rello, Luz i Llisterri, Joaquim. 2012. „There are phonetic patterns in vowel substitution errors in texts written by persons with dyslexia." U *21st Annual World Congress on Learning Disabilities (LDW 2012)*, Vol. 7. Oviedo, Spain.

Rowland, Caroline F., Fletcher, Sara L. i Freudenthal, Daniel. 2008. „How big is enough? Assessing the reliability of data from naturalistic samples." U *Corpora in Language Acquisition Research: History, Methods, Perspectives. Trends in Language Acquisition Research 6., uredila* Behrens, Heike, 1–24. Amsterdam/ Philadelphia: John Benjamins Publishing Company.

Salmelin, Riitta, Kiesilä, Kimmo Uutela, Elisabet Service i Oili Salonen 1996. „Impaired visual word processing in dyslexia revealed with magnetoencephalography." *Annals of neurology* 40 (2): 157–162.

Shipley, Kenneth G. i McAfee, Julie G. 2015. *Assessment in Speech-LanguagePathology: A resource manual.* 5. izdanje. USA, Boston: Cengage Learning. https://books.google.hr/books?id=NWOOBAAAQBAJ&pg=PR2&dq=kenneth+shipley+5th+edition&hl=hr&sa=X&ei=qAxhVa6ZCsfoUu7pgZgI&ved=0CBsQ6AEwAA#v=onepage&q&f=false Pristupljeno: 23. lipnja 2015.

Tadić, Marko. 2003. *Jezične tehnologije i hrvatski jezik*. Zagreb: Ex libris.

Tadić, Marko. 2009. „New version of the Croatian National Corpus." U *After Half a Century of Slavonic Natural Language Processing, uredili* Hlaváčková, Dana, Horák, Aleš, Osolsobě, Klara i Pavel Rychlý, 199–205. Brno: Masaryk University.

Westerhout, Eline. 2006. *A corpus of Dutch aphasic speech: sketching the design and Performing a pilot study.* Master Thesis Utrecht: Faculty of Humanities Theses, Utrect University.

Sampling challenges of specialized spoken and written adult speakers corpora

The importance of language corpora is highly acknowledged and well-documented in contemporary linguistics. The majority of them are based on written texts of different genres. However, there is a tendency to include 10% of spoken language into the national corpora, as well. These data consist of samples collected during informal conversations. There is a need to build such a corpus of spoken language in Croatian, too. In addition, as the interest in special population groups increases, the need for specialized corpora grows as well. Specialized language corpora are formed by constraining the sample to a relevant group (e.g. children, bilingual speakers, persons with aphasia...) or to a relevant kind of samples (e.g. samples of different levels of complexity). Generally, these corpora are divided into spoken and written corpora. In the sampling process some methodological questions are put forward such as the sort of the texts chosen, input options, orthographic type, communication context. Two specialized corpora are being built within two projects that have been carried out in the Laboratory for Psycholinguistic Research: *Adult Language Processing* (Croatian Science Foundation UIP-11-2013-2421) and *Text input computer assistant for persons with language disorders*. The first one is the Croatian Adult Spoken Language Corpus and the second one is the Croatian Written Language Corpus. Healthy speakers and speakers with different language disorders are included in both corpora. In this paper some methodological questions which need to be addressed prior to sampling will be discussed, primarily regarding the choice of the sampling method, the control of the extra-linguistic factors and the materials used for sampling.

Keywords: specialized corpus, Croatian Adult Spoken Language Corpus, Croatian written language corpus, clinical corpus