Korpus srpskog kao stranog jezika (KSKS): Opis građe i plan izrade

Maja Miličević

(Univerzitet u Beogradu)

In V. Krajišnik (Ed.), *Srpski kao strani jezik u teoriji i praksi III*. Beograd: Filološki fakultet. 279-289. (2016)

Maja Miličević*
Filološki fakultet
Beograd

KORPUS SRPSKOG KAO STRANOG JEZIKA (KSKS): OPIS GRAĐE I PLAN IZRADE**

U radu je prikazan plan izrade elektronskog Korpusa srpskog kao stranog jezika. U korpus će biti uključeni različiti vidovi produkcije učenika srpskog jezika kao stranog, a najvećim delom sastavi pisani na ispitima, na času ili u okviru domaćih zadataka. U prvoj fazi izrade korpusa učenički tekstovi se skeniranjem prebacuju u digitalni format i transkribuju, uz paralelno čuvanje podataka o autorima (njihovom polu i starosti, maternjem jeziku, nivou znanja srpskog jezika). U drugoj fazi tekstovima će biti dodate lingvističke informacije poput oznaka vrsta reči i osnovnih oblika, kao i informacije o greškama, uključujući i njihovu normalizaciju. Za završnu fazu planira se postavljanje korpusa na Internet i omogućavanje njegove pretrage putem korisničkog interfejsa. Svrha izrade korpusa jeste da se objedini veća količina učeničkih tekstova i da se ti tekstovi učine dostupnima istraživačima i predavačima koji žele da proučavaju proces usvajanja srpskog jezika kao stranog ili da koriste autentičan učenički materijal u nastavi.

Ključne reči: *srpski jezik, učenički korpus, anotacija grešaka*

UVOD

U okviru brzog razvoja i napretka korpusne lingvistike tokom poslednjih nekoliko decenija značajno mesto zauzimaju i **učenički korpusi** (eng. *learner corpora*) – planski sakupljene elektronske zbirke produkcije učenika određenog jezika kao stranog (v. Leech 1998: xiv, Granger 2003: 465). Ovakve zbirke nazivaju se i korpusima usvajanja stranog jezika ili korpusima (engleskog, španskog, češkog, itd.) jezika kao stranog. Izostavljanje atributa „učenički" posebno je često u jezicima u kojima se imenica „učenik" prvobitno odnosi na decu školskog uzrasta (v. Stritar 2012: 19-20 za slovenački jezik). Međutim, u oblasti usvajanja drugog jezika ova imenica se u literaturi koristi i u značenju koje odgovara engleskom *learner* (v. npr. Kraš i Miličević u štampi, kao i Mikelić Preradović, Boras i Berać 2014), tako da se svi navedeni termini mogu smatrati odgovarajućim.

Učenički korpusi su usled složenosti izrade daleko manje brojni od opštih korpusa. Međutim, dostupni su ne samo za „velike" jezike kakvi su engleski ili španski, već i za pojedine jezike sa malim brojem govornika (i učenika), poput slovenačkog (v. Stritar 2009, 2012). [1] Na govornom području nekadašnjeg srpskohrvatskog na izradi korpusa hrvatskog jezika kao stranog radi se u centru Croaticum u Zagrebu (v. Mikelić Preradović i dr. 2014), dok u ovom radu, uz kratak pregled odlika i mogućih primena učeničkih korpusa, predstavljamo Korpus srpskog kao stranog jezika (KSKS), u čijoj izradi sarađuju Centar za srpski kao strani jezik i Katedra za opštu lingvistiku na Filološkom fakultetu Univerziteta u Beogradu.

* m.milicevic@fil.bg.ac.rs

** Rad je nastao u okviru projekta *Standardni srpski jezik: sintaksička, semantička i pragmatička istraživanja* (178004), koji finansira Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije.

[1] Spisak učeničkih korpusa različitih jezika dostupan je na stranici https://www.uclouvain.be/en-cecl-lcworld.html (poslednji pristup 10. septembra 2015).

UČENIČI KORPUSI I NJIHOVE MOGUĆE PRIMENE

Učenički korpusi predstavljaju dragocen izvor podataka za istraživanje procesa usvajanja stranog jezika. Oni istraživačima stavljaju na raspolaganje planski sakupljen autentični materijal pogodan za proučavanje različitih aspekata učeničkog međujezika (eng. *interlanguage*), a od posebnog je metodološkog značaja to što znatno olakšavaju kvantitativne analize učeničke produkcije. Gries (2008) kao osnovne formate upotrebe učeničkih korpusa izdvaja: (1) frekvencijske liste i liste kolokata, korisne za izdvajanje najučestalijih leksičkih elemenata u tekstovima učenika (i njihovo poređenje sa produkcijom izvornih govornika); (2) liste koligacija (kombinacija leksičkih i gramatičkih obrazaca), korisne za proučavanje leksičko-sintaksičkih pojava poput glagolske valentnosti; (3) konkordance, odnosno prikaze ključnih reči u kontekstu, putem kojih se mogu izučavati kako leksičke tako i gramatičke odlike tekstova. Zatim, ukoliko učenički korpus uz tekstove sadrži i detaljne informacije o učenicima (njihovom maternjem jeziku, uzrastu, nivou znanja drugog jezika i sl.), te se informacije mogu upotrebiti kao filteri prilikom pretrage, što čini korpus pogodnim za proučavanje različitih faktora koji utiču na proces usvajanja stranog jezika. Najzad, aspekt proučavanja učeničkog jezika koji posebno privlači pažnju (s obzirom na odstupanja od standarda ciljnog jezika), a koji elektronski korpusi mogu značajno olakšati, jeste analiza grešaka.

Da bi ovakvi napredni načini upotrebe učeničkih korpusa bili mogući, potrebno je da korpus ima odgovarajuće odlike. Sa jedne strane, potrebno je da sadrži relevantne **metapodatke**, odnosno podatke o maternjem jeziku učenika, njihovom nivou poznavanja drugog jezika, starosti, polu, kontekstu u kome usvajaju jezik i drugim faktorima koji se ocenjuju kao relevantni (pregled metapodataka unetih u različite učeničke korpuse može se naći u Stritar 2012: 61). Metapodaci omogućuju grupisanje rezultata pretrage korpusa prema potencijalno značajnim kriterijumima, kao i ograničavanje pretrage na produkciju učenika sa određenim karakteristikama (v. npr. opcije pretrage u češkom korpusu CzeSL, dostupnom na adresi https://kontext.korpus.cz/first_form). Sa druge strane, upotrebljivost svakog korpusa, pa i učeničkog, u značajnoj meri je određena nivoom njegove **lingvističke anotacije** (v. Leech 2005). U slučaju učeničkih korpusa bitna je ne samo morfosintaksička anotacija (unošenje podataka o osnovnim oblicima reči, vrstama kojima reči pripadaju i gramatičkim kategorijama poput roda, broja ili padeža), koja je relevantna za sve tipove korpusa, već i **anotacija grešaka**, karakteristična upravo za učeničke korpuse (v. Gries and Berez u štampi). Anotacija grešaka sastoji se u (najčešće ručnom) označavanju grešaka koje se javljaju u učeničkoj produkciji, a kao izuzetno bitan korak prethodi joj razrada klasifikacije grešaka koje će biti označene. Česti pristupi jesu odvajanje grešaka koje se tiču pojedinačnih reči i onih koje se tiču većih celina, uz dalje izdvajanje podvrsta unutar svake kategorije (v. npr. Hana et al. 2010:16 za češki, gde su ortografske i morfološke greške odvojene od onih koje se tiču većih celina, poput valentnosti, reda reči, leksike i frazeologije), i klasifikacija grešaka prema nivou jezičke strukture (v. Stritar 2012: 154-155 za podelu na ortografske, leksičke, morfološke i strukturne greške u učeničkom korpusu slovenačkog jezika).

Učenički korpusi nude brojne mogućnosti i za nastavnu praksu. Na primer, u učionici i prilikom pripreme nastavnog materijala mogu biti korisni za izdvajanje leksičkih jedinica ili gramatičkih pojava koje su posebno problematične (za govornike određenog maternjeg jezika, učenike na određenom nivou znanja, ili uopšte), a mogu poslužiti i kao osnova za izradu vežbanja. Zanimljive primene mogu se pronaći i za

greške, koje se mogu upotrebiti pri izradi negramatičnih distraktora u pitanjima višestrukog izbora, ili mogu biti zadate za ispravljanje (up. Stritar 2009: 146-147).

Neki od poznatijih postojećih učeničkih korpusa jesu korpus engleskog jezika ICLE (*International Corpus of Learner English*, Granger et al. 2009), korpus nemačkog jezika FALKO (*Fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache*, Lüdeling et al. 2008) i korpus španskog jezika CEDEL2 (*Corpus Escrito del Español L2*, Lozano 2009). U domenu slovenskih jezika posebno treba istaći već pomenute korpuse češkog (CzeSL – *Czech as a Second Language*, Hana et al. 2010) i slovenačkog (KUST – *Korpus usvajanja slovenščine kot tujega jezika*, Stritar 2012). Zajednička osobina svih navedenih korpusa jeste da uključuju pisanu produkciju (na različite teme i različitog stepena spontanosti), a većina korpusa deli i odliku da obuhvata učenike različitih maternjih jezika (izuzetak je CEDEL2, fokusiran na maternje govornike engleskog jezika). Veličina ovih korpusa kreće se od nekoliko desetina hiljada (KUST) do nekoliko miliona reči (ICLE). Svi korpusi su u elektronskom obliku, dostupni putem namenskih sučelja ili kao datoteke (na zahtev), a anotirani su u različitoj meri i na različite načine.

KORPUS SRPSKOG KAO STRANOG JEZIKA

Izrada Korpusa srpskog kao stranog jezika započeta je 2014. godine. Prilikom osmišljavanja njegovog nacrta pored teorijsko-metodoloških razmatranja značajnih za sve učeničke korpuse bilo je potrebno uzeti u obzir i nekoliko specifičnih praktičnih činjenica, a pre svega to da je srpski jezik sa malim brojem govornika i srazmerno malim brojem učenika koji ga uče kao strani, što se nužno odražava na ukupno moguću veličinu korpusa i dovodi u pitanje mogućnost primene usko definisanih kriterijuma u izboru građe. Među učenicima su uz to najbrojniji oni koji su na nižim nivoima znanja, sa jedne strane kao posledica relativno niskog broja učenika koji se trajno ili na duži period naseljavaju u Srbiji, a sa druge usled toga što na univerzitetima van Srbije srpski u najvećem broju slučajeva učestvuje u nastavnim programima kao jedan od nekoliko slovenskih jezika, i to najčešće ne glavni.[2] Ova činjenica je značajna ne samo za planiranje nivoa znanja koji će biti uključeni u korpus, već i za razradu načina anotacije grešaka, budući da se na različitim nivoima znanja očekuju različiti tipovi i različita količina grešaka. U narednim odeljcima opisujemo odluke i planove formulisane imajući u vidu ova praktična ograničenja.

Izbor i prikupljanje građe

Jedan od osnovnih zahteva pri izradi elektronskih korpusa uopšte jeste precizno definisanje kriterijuma izbora građe koja će biti uključena u korpus (v. Sinclair 2005). Kod učeničkih korpusa potrebno je odlučiti da li će biti sakupljana pisana ili usmena produkcija (ili produkcija u oba modaliteta), da li će prema tipu produkcija biti spontana ili elicitirana (u drugom slučaju – na koji način će biti elicitirana, kakve teme će biti pokrivene i sl.), čija produkcija će biti uključena (s obzirom na maternji jezik/jezike učenika, nivo/nivoe njihovog znanja drugog jezika, i sl.), kao i da li će biti korišteni već postojeći materijali ili će se tek pristupiti namenskom prikupljanju tekstova (up. Stritar 2012). Prilikom donošenja odluka vezanih za ove kriterijume uvek se moraju uzeti u obzir i praktične okolnosti, odnosno dostpunost građe.

[2] U periodu od akademske 2010/2011. godine izuzetak od ovog trenda predstavljaju polaznici programa Svet u Srbiji, u okviru koga stipendisti iz država članica i posmatrača Pokreta nesvrstanih zemalja po završetku intenzivnog kursa srpskog jezika nastavljaju studije na Univerzitetu u Beogradu.

U slučaju Korpusa srpskog kao stranog jezika iz praktičnih razloga je odlučeno da se prikupljanje građe započne od materijala neposredno dostupnog istraživačima uključenim u izradu korpusa, da se zatim uputi poziv drugim istraživačima i nastavnicima za proširenje materijala, a da se kriterijumi vezani za osobine učenika definišu što je šire moguće, odnosno da u korpus bude uključena produkcija učenika različite starosti i pola, različitih maternjih jezika i svih dostupnih nivoa znanja, uz kodiranje metapodataka vezanih za ove faktore. U tabeli 1 dat je pregled odlika KSKS-a s obzirom na gornje kriterijume.

Kriterijum	**KSKS**
Modalitet produkcije	Pisana produkcija[3]
Tip produkcije	Autentična elicitirana produkcija[4]
Osobine učenika	Bez ograničenja
Vreme nastanka materijala	Postojeći i novi tekstovi

Tabela 1. Kriterijumi izbora građe za KSKS.

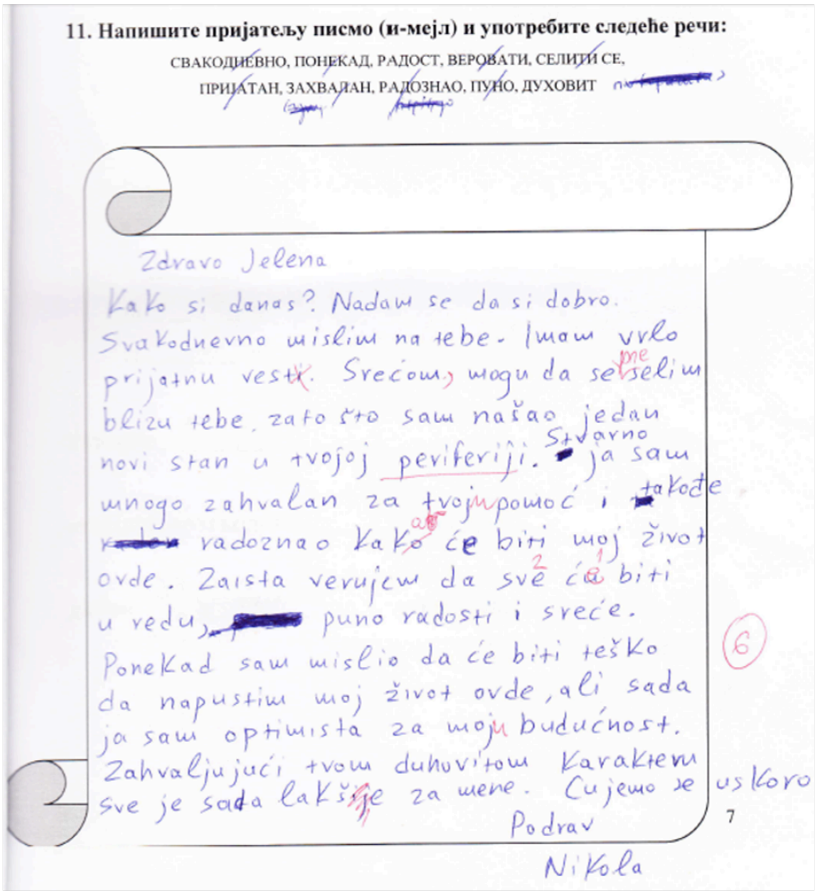
Najveću količinu neposredno dostupnih tekstova čine učenički sastavi pisani u okviru ispita i nastavnih aktivnosti u Centru za srpski kao strani jezik na Filološkom fakultetu u Beogradu (i institucijama sa kojima Centar sarađuje prilikom organizacije ispita). Ispitni radovi dostupni su za ispitne rokove od 2006. godine do danas. Tekstovi koji se mogu uključiti u korpus predstavljaju jedan deo ispita i prisutni su od nivoa A2 (u nekim ispitnim rokovima B1) naviše, budući da se na najnižem nivou (A1) ne zadaje pisanje sastava. U pitanju je produkcija koja je uvek u nekoj meri vođena, odnosno elicitirana – sastavi su napisani na zadatu temu, a zadaci su često zasnovani na opisivanju slika (v. primer na slici 1), ili su uz temu zadate i reči koje je potrebno upotrebiti (slika 2). Najčešće su u pitanju sastavi deskriptivnog tipa (u učeničkim korpusima uopšte vrlo su zastupljeni i argumentativni eseji, v. Stritar 2012: 67). Dužina pojedinačnih sastava u proseku iznosi 100-120 reči (dužina sastava je uglavnom srazmerna nivou), a procenjena ukupna veličina dela KSKS-a koji će činiti ovi tekstovi jeste približno 50.000 reči. Uz većinu tekstova postoje i osnovni podaci o autorima, koji imaju različite maternje jezike i različite su starosti (najvećim delom je u pitanju studentska populacija starosti 19-23 godine), dok strukturiranost kurseva i ispita prema nivoima Zajedničkog evropskog okvira za žive jezike olakšava kodiranje nivoa znanja srpskog jezika.[5]

[3] Prikupljaju se i snimci usmenih materijala, ali se oni čuvaju za kasnije dodavanje, budući da je ovakva građa manje dostupna od pisane, a ujedno i teža za transkribovanje.

[4] Brojni autori ističu da je u domenu usvajanja drugog jezika u kontekstu učionice, odakle potiče najveći deo korpusnih materijala, gotovo nemoguće govoriti o potpuno autentičnoj spontanoj produkciji, pa se autentičnim u ovom domenu smatraju tekstovi nastali u pedagoške svrhe, ili namenski za korpus, pri čijem nastanku van okvira zadavanja zadatka nije postojala značajna kontrola od strane nastavnika (v. Nesselhauf 2004).

[5] Budući da je dostupan podatak o tome da li je određeni učenik položio ispit za dati nivo, moguće je klasifikovati ga/je u taj ili niži nivo znanja.

Slika 1. Primer (dela) ispitnog zadatka zasnovanog na slikama (nivo A2).



Slika 2. Primer tekstualnog ispitnog zadatka (nivo B2).

Pored ispitnih tekstova, Centar za srpski kao strani jezik od početka izrade korpusa prikuplja i sastave koje učenici pišu u okviru domaćih zadataka, a određena količina materijala sakupljena je, i sakuplja se i dalje, uz pomoć lektora koji predaju srpski jezik van Srbije. Pri sakupljanju novih tekstova beleže se i propratni podaci o njihovim autorima: pol, starost, maternji jezik, nivo znanja srpskog jezika, dužina

učenja srpskog jezika, poznavanje drugih stranih jezika.[6] U tu svrhu sastavljeni su uputstvo za prikupljanje tekstova i upitnik za podatke o učenicima (dostupni od autorke ovog rada). Upitnici za učenike sadrže i stavku u kojoj se daje saglasnost za upotrebu tekstova u svrhe istraživanja, a posebno treba podvući da se u procesu izrade korpusa nigde ne otkriva identitet ma kog od autora tekstova.

Po završetku početne faze prikupljanja građe biće sprovedena evaluacija ostvarenih rezultata, nakon čega će biti razmotrene i dodatne opcije sakupljanja. Jedna od mogućnosti jeste izrada kontrolisanijeg potkorpusa koji bi bio zasnovan na grupi tema planski zadatih većem broju učenika; up. proceduru izrade korpusa CEDEL2 (Lozano 2009), gde učenici pišu sastav minimalne dužine od 500 reči na jednu od 12 ponuđenih tema, ili Valico (Corino i Marello 2009), gde je učenicima zadato da napišu sastav dužine preko 100 reči na osnovu jednog od pet setova slika.

Obrada tekstova

Kao što se može videti na slikama 1 i 2, originali tekstova koji ulaze u KSKS najvećim delom su rukopisni radovi. Prvi korak u daljoj obradi stoga nužno predstavlja čuvanje u digitalnom formatu, odnosno skeniranje tekstova, a zatim njihovo ručno prekucavanje.[7] Iako naizgled jednostavno, prekucavanje tekstova je aktivnost koja zahteva ogromne vremenske resurse, posebno u slučaju kada je materijal rukopisni i neizbežno sadrži nečitke delove; Stritar (2009: 138) navodi da je pri izradi učeničkog korpusa slovenačkog jezika za prekucavanje teksta od 200 reči u proseku bilo potrebno približno 20 minuta. Učenički tekstovi su ujedno specifični po greškama na koje je neophodno obratiti posebnu pažnju kako bi bile verno prenete u transkriptu, a ne ispravljene u pravilan oblik, odnosno normalizovane od strane prepisivača. Među dodatnim pitanjima koja se nameću u vezi sa srpskim jezikom svakako je bitno i ono koje se tiče pisma, naime da li je bolje rešenje sve tekstove transkribovati na jedno odabrano pismo ili svuda zadržati pismo originala. Odabrano je zadržavanje pisma originala, uz kasnije omogućavanje upotrebe pisma kao filtera za pretragu (što je posebno bitno za istraživanja ortografije), ali i omogućavanje jedinstvene pretrage u kojoj se razlika u pismu zanemaruje.

Sa tehničke strane, u izradi svakog korpusa izuzetno je značajan format u kome će tekstovi biti sačuvani. Budući da je u transkripciju uključen veći broj studenata Filološkog fakulteta, upućeni su da transkripte radi jednostavnosti čuvaju kao dokumente u nekom od formata koje nudi Microsoft Word. Po završnoj proveri transkripcije dokumenti će biti prebačeni u XML format, gde će svakom tekstu u zaglavlju biti pridruženi raspoloživi metapodaci: identifikaciona oznaka teksta, datum nastanka teksta, poreklo teksta, tip aktivnosti u okviru koje nastao, identifikaciona oznaka učenika, starost učenika, njegov/njen maternji jezik itd. Još jednom napominjemo da će imena i prezimena učenika radi postizanja anonimnosti tekstova biti izostavljena.

Za fazu koja će slediti posle transkripcije predviđa se lingvistička obrada tekstova, u vidu dodeljivana oznaka za vrste reči i osnovni oblik; drugim rečima, korpus će biti tagiran i lematizovan (v. Leech 2005). Procedura koja će pri tome biti

[6] U slučaju nedostatka određenih podataka za pojedine učenike predviđena je upotreba oznake „nepoznato“, koja će omogućiti pretragu tekstova sa nepotpunim podacima u slučajevima kada istraživaču podaci o učeniku nisu bitni, a njihovo isključivanje u slučaju kada jesu. Ovakva praksa je uobičajena pri izradi učeničkih korpusa (up. npr. opcije u korpusu italijanskog jezika Valico, http://www.valico.org/valico_b_CORPUS.html, poslednji pristup 10. septembra 2015.).

[7] S obzirom na broj različitih rukopisa, nije moguće koristiti neki od programa za automatsku transkripciju.

sleđena najverovatnije će, usled relativno male ukupne količine materijala, biti ručna ili zasnivana na normalizaciji i naknadnoj upotrebi alatki za automatsku anotaciju razvijenih za standardni jezik.[8] Posebna pažnja biće posvećena učeničkim greškama.

Označavanje grešaka

Lingvistička obrada KSKS-a obuhvatiće i označavanje učeničkih grešaka, koje predstavlja jedan od najkorisnijih aspekata korpusa stranih jezika. Ovaj korak u obradi korpusa biće sproveden ručno, a za njega će od posebnog značaja biti klasifikacija grešaka i osmišljavanje oznaka za njihove različite tipove, na čemu se rad upravo započinje. Iz prakse je poznato da u veće probleme u srpskom kao stranom jeziku spadaju, na primer, savladavanje upotrebe dvaju pisama, bogata flektivna i derivaciona morfologija, glagolski vid, složena pravila kongruencije i relativno slobodan red reči. Stoga će se sasvim sigurno pribeći nekom vidu višeslojne anotacije u kojoj će na različitim nivoima biti predstavljeni različiti tipovi grešaka (up. primere iz češkog i slovenačkog korpusa u gornjem tekstu), imajući na umu preporuku da sistemi anotacije grešaka treba istovremeno da budu fleksibilni i konzistentni (Granger 2003: 467).

Anotacija grešaka tesno je povezana i sa pitanjem njhivog ispravljanja, odnosno normalizacije građe (v. napomenu 8 za razjašnjenje pojma normalizacije); za KSKS planira se i ovaj tip anotacije. Naime, normalizacija je korisna ne samo u svrhu jednostavnijeg tagiranja i lematizacije korpusa, o čemu je bilo reči iznad, već i da bi se korisnicima korpusa omogućila pretraga prema standardnim oblicima koja bi kao rezultat dala i ispravne oblike i različite greške. Na primer, upitom kojim bi se tražio oblik *lakše* mogli bi se istovremeno dobiti i pogrešno izvedeni komparativi poput *lakšije* (v. Miličević i Vuković 2015 za slične primere iz dijalekatske građe). Drugim rečima, anotacija grešaka koja uključuje i njihovu normalizaciju olakšava ne samo analizu grešaka prema različitim tipovima, već i paralelno pronalaženje ispravnih i neispravnih rešenja u učeničkim tekstovima, što može dovesti do otkrivanja faktora koji dovode do varijabilnosti u učeničkoj produkciji.

Dostupnost korpusa

Za završnu fazu izrade korpusa planira se njegovo postavljanje na Internet i omogućavanje pretrage putem korisničkog interfejsa. Detalji vezani za ovaj korak biće naknadno utvrđeni.

ZAKLJUČAK

Korpus srpskog kao stranog jezika u velikoj meri će automatizovati pronalaženje reči i kombinacija reči u učeničkim tekstovima, omogućiće pretragu prema vrstama reči i osnovnim oblicima reči, uz prikaz rezultata u kontekstu i automatsko dobijanje podataka o učestalosti. Takođe će pružiti mogućnost filtriranja pretrage prema kriterijumima poput maternjeg jezika i nivoa znanja učenika. Najzad, mogućnošću pretrage prema tipovima grešaka i normalizovanim oblicima značajno će olakšati analizu grešaka. KSKS će moći da se koristi kako u istraživanjima, tako i u nastavi srpskog jezika kao stranog i u izradi nastavnog materijala. Najveći problem u izradi korpusa ostaje mala količina dostupne građe, zbog čega u zaključku želimo da pozovemo predavače koji imaju ili mogu da sakupe odgovarajuće tekstove da se

[8] Normalizacija u ovom slučaju ne predstavlja intervenciju na originalnom transkriptu, već dodavanje standardnih oblika kao novog sloja informacija, slično postupku razrađenom za druge tipove nastandardne građe (up. višeslojnu normalizaciju dijalekatske građe opisanu u Vuković 2015).

uključe u izradu korpusa. Više informacija mogu dobiti kontaktiranjem autorke ovog rada ili preko Centra za srpski kao strani jezik na Filološkom fakultetu u Beogradu (http://learnserbian.fil.bg.ac.rs/cooperation.php?id=f).

LITERATURA

Corino, E. and C. Marello (2009). Elicitare scritti a partire da storie disegnate: il corpus di apprendenti VALICO. In: *Corpora di Italiano L2: Tecnologie, metodi, spunti teorici* (C. Andorno and S. Rastelli, eds), Perugia: Guerra, 113-138.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal* 20/3: 465-480.

Granger, S., E. Dagneaux, F. Meunier, and M. Paquot (2009). *The International Corpus of Learner English. Version 2. Handbook and CD-Rom*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Gries, S. Th. (2008). Corpus-based methods in analyses of SLA data. In: *Handbook of Cognitive Linguistics and Second Language Acquisition* (P. Robinson and N. C. Ellis, eds), New York: Routledge, 406-431.

Gries, S. Th. and A. L. Berez (u štampi). Linguistic annotation in/for corpus linguistics. In: *Handbook of Linguistic Annotation* (N. Ide and J. Pustejovsky, eds), Berlin & New York: Springer.

Hana, J., S. Škodová, and B. Štindlová (2010). Error-tagged learner corpus of Czech. In: *Proceedings of the Fourth Linguistic Annotation Workshop at ACL 2010*, Association for Computational Linguistics, 11-19.

Kraš, T. i M. Miličević (u štampi). *Eksperimentalne metode u istraživanjima usvajanja drugoga jezika*. Rijeka: Filozofski fakultet.

Leech, G. (1998). Preface. In: *Learner English on Computer* (S. Granger, ed.), London: Longman, xiv-xx.

Leech, G. (2005). Adding linguistic annotation. In: *Developing Linguistic Corpora: A Guide to Good Practice* (M. Wynne, ed.), Oxford: Oxbow Books, 17-29. Dostupno na adresi http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter2.htm [poslednji pristup 10. septembra 2015.].

Lozano, C. (2009). CEDEL2: Corpus Escrito del Español L2. In: *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente* (C. M. Bretones Callejas, ed.), Almería: Universidad de Almería, 197-212.

Lüdeling, A., S. Doolittle, H. Hirschmann, K. Schmidt, and M. Walter (2008). Das Lernerkorpus Falko. *Deutsch Als Fremdsprache* 45: 67-73.

Mikelić Preradović, N., D. Boras i M. Berać (2014). Učenički korpus hrvatskog kao stranog jezika. Rad izložen na 28. međunarodnom skupu Hrvatskog društva za primjenjenu lingvistiku „Višejezičnost kao predmet multidisciplinarnih istraživanja“, Zagreb, 25-27. aprila 2014.

Miličević, M. i T. Vuković (2015). Creation and some ideas for classroom use of an electronic corpus of the dialect of Bunjevci. Rad izložen na XV International Conference on Minority Languages, Beograd, 28-30. maja 2015.

Nesselhauf, N. (2004). Learner corpora and their potential in language teaching. In: *How to Use Corpora in Language Teaching* (J. Sinclair, ed.), Amsterdam/Philadelphia: John Benjamins, 125-152.

Sinclair, J. (2005). Corpus and text – basic principles. In: *Developing Linguistic*

Corpora: A Guide to Good Practice (M. Wynne, ed.), Oxford: Oxbow Books, 1-16. Dostupno na adresi http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm [poslednji pristup 10. septembra 2015.].
Stritar, M. (2009). Slovene as a foreign language: The Pilot Learner Corpus perspective. *Slovenski jezik – Slovene Linguistic Studies* 7: 135–152.
Stritar, M. (2012). *Korpusi usvajanja tujega jezika*. Ljubljana: Zveza društev Slavistično društvo Slovenije.
Vuković, T. (2015). *Izrada modela dijalekatskog korpusa bunjevačkog govora*. Neobjavljeni master rad. Beograd: Filološki fakultet.

Maja Miličević

BUILDING A CORPUS OF SERBIAN AS A FOREIGN LANGUAGE

Summary

This paper outlines the procedure envisaged for the creation of KSKS (*Korpus srpskog kao stranog jezika*), the first corpus of Serbian as a foreign language. Available texts are described first: KSKS will comprise different types of learner production, for the most part short essays written during exams, as part of classroom activities or for homework. Within the first stage of corpus construction, learner texts are currently being digitised – each text undergoes scanning and manual transcription; information about the learners is also coded (e.g. their age and gender, mother tongue, and proficiency level in Serbian). The second stage will consist in annotating texts with linguistics information (part-of-speech tags and lemmas), as well as in implementing an error annotation scheme that is currently being developed and in adding normalised forms to enable standard language queries that will simultaneously provide results complying with the standard and those departing from it. In the final stage of its development the corpus will be made available online via a graphical user interface. The purpose of KSKS is to assemble a substantial amount of learner production data and to put that data at the disposal of researchers and teachers interested in studying the acquisition of Serbian as a foreign language, or in using authentic learner material in their teaching activities.

Key words: Serbian, learner corpus, error annotation