Tadeja Rozman
Univerza v Ljubljani
Fakulteta za upravo

Špela Arhar Holdt
Univerza v Ljubljani
Center za jezikovne vire in tehnologije

Senja Pollak
Institut "Jožef Stefan"
Univerza v Edinburgu
Institut Usher

Iztok Kosem
Univerza v Ljubljani
Center za jezikovne vire in tehnologije
Institut "Jožef Stefan"

UDK 811.163.6'373.7:004.56

KOLOKACIJE V KORPUSU *ŠOLAR*

Prispevek se ukvarja z možnostmi uporabe korpusa šolskih pisnih izdelkov *Šolar* za namene ugotavljanja procesov usvajanja kolokacij v slovenskem jeziku ter s potencialnim vplivom korpusno pridobljenih podatkov o kolokacijah na jezikovni pouk slovenščine. Na številčno omejenem gradivu zvez *pridevnik + samostalnik* in *samostalnik + samostalnik* sta bili preizkušeni dve metodi: a) s kvantitativno primerjalno analizo korpusov *Šolar* in *Kres* smo preverjali podobnosti in razlike v rabi kolokacij v šolskih spisih in splošni pisni slovenščini, b) s kvalitativno analizo učiteljskih popravkov pa smo želeli ugotoviti, s katerimi kolokacijami imajo učenke in učenci težave. Rezultati kažejo, da je v rabi kolokacij med mladimi zaslediti marsikatere atipičnosti v primerjavi z običajno rabo v standardnem jeziku, ter potrjujejo uporabnost korpusnih raziskav na področju usvajanja leksike.

Ključne besede: kolokacije, korpus *Šolar*, korpusna analiza, usvajanje besedišča, jezikovna didaktika

1 Uvod

Pojma kolokacija in kolokacijskost sta v jezikoslovju znana že več kot pol stoletja, če ne drugače, vsaj po slavni in večkrat citirani izjavi J. R. Firtha (1957): »Besedo spoznaš po njeni okolici«.[1] Ključni poudarek izjave je v tem, da sta pomen in pojavnost besede odvisna od konteksta, takšen metodološki pristop pri raziskovanju leksike pa zasledimo npr. v Hallidayevi funkcijski slovnici (1966), Mel'čukovih leksikalnih funkcijah (1996, 1998), Sinclairjevih na korpusu temelječih kolokacijskih študijah (1987, 1991), slovnici vzorcev (Hunston in Francis 2000) in teoriji leksikalnega proženja (Hoey 2005).

[1] Angl. *You shall know a word by the company it keeps.*

Številne raziskave (npr. Kjellmer 1991, James 1998, Nation 2001) poudarjajo pomen kolokacij pri učenju jezika, s poudarkom na učenju tujega jezika, saj naj bi kolokacije predstavljale enega ključnih korakov tujejezičnih govorcev k doseganju ravni maternih govorcev jezika. Dejansko je razmerje med raziskavami, ki se ukvarjajo z vlogo kolokacij pri učenju tujega jezika in pri usvajanju maternega jezika, močno v prid prvih. Prav tako je večina raziskav kolokacij opravljena na angleškem jeziku, medtem ko je raziskav kolokacij za mnoge druge, zlasti manjše jezike, zelo malo.

V Sloveniji je bil v zadnjih letih narejen velik napredek pri prepoznavi kolokacij, zlasti v okviru izdelave Leksikalne baze za slovenščino[2] (Gantar 2015) in z razvojem postopkov za avtomatsko luščenje leksikalnih podatkov iz korpusov (Gantar idr. 2016, Kosem idr. 2013, Pollak 2015). Pomemben del avtomatskega luščenja je podrobna slovnica besednih skic (zbirka poizvedb za avtomatsko prepoznavo različnih slovničnih relacij v korpusu), ki je omogočila podrobno razporeditev kolokacij po slovničnih relacijah (Krek in Kilgarriff 2006). Omenjeni postopek je tudi eden izmed temeljev pri izdelavi Kolokacijskega slovarja slovenskega jezika,[3] ki bo namenjen predvsem maternim govorcem, in pa z njim povezane Baze kolokacijskega slovarja slovenskega jezika (Krek idr. 2016). Poleg tega velja omeniti temeljni raziskovalni projekt KOLOS[4] (*Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki*, J6-8255), ki se je začel leta 2017 in je posvečen preučevanju različnih vidikov kolokacij v slovenskem jeziku, ter raziskave kolokacij spletne slovenščine, ki so bile opravljene v okviru projekta JANES[5] (*Jezikoslovna analiza nestandardne slovenščine*, J6-6842).

Večjih raziskav, ki bi preučevale usvajanje in učenje kolokacij v slovenskem jeziku, pa zaenkrat ni, in analiza, ki jo predstavljamo v pričujočem članku, je poskus, da bi to vrzel vsaj deloma zapolnili. Analizo smo zasnovali empirično in jo lahko umestimo v okvir korpusnih leksikalnih analiz s področja usvajanja jezika. Tovrstnih raziskav je pri nas zaenkrat sicer malo (Arhar Holdt in Rozman 2015, Kosem idr. 2012), kljub temu da je že od leta 2011 na voljo prosto dostopen korpus *Šolar*[6] (gl. poglavje 2), tj. korpus šolskih pisnih izdelkov, ki omogoča empirične raziskave pisne jezikovne zmožnosti slovenskih učencev.

Za raziskovalno izhodišče smo si zastavili naslednje vprašanje: Kaj nam lahko korpus *Šolar* pove o usvajanju kolokacij in kako te informacije uporabiti pri načrtovanju pouka leksike ter pripravi didaktičnih gradiv in jezikovnih priročnikov, namenjenih šolski populaciji? Da bi na to vprašanje lahko vsaj delno odgovorili, smo se odločili,

---

[2] <http://www.slovenscina.eu/spletni-slovar/leksikalna-baza>. (Dostop 28. 8. 2018.) Leksikalna baza in v nadaljevanju navedeni korpusi so dostopni preko repozitorija Slovenske raziskovalne infrastrukture za jezikovne vire in tehnologije Clarin.si. <https://www.clarin.si/repository/xmlui/>. (Dostop 28. 8. 2018.)

[3] <https://viri.cjvt.si/kolokacije/slv/#>. (Dostop 19. 11. 2018.)

[4] <https://www.cjvt.si/kolos/>. (Dostop 28. 8. 2018.)

[5] <http://nl.ijs.si/janes/>. (Dostop 28. 8. 2018.)

[6] <http://www.slovenscina.eu/korpusi/solar>. (Dostop 28. 8. 2018.)

da preizkusimo dve metodi: a) s primerjalno analizo korpusov *Šolar* in *Kres*[7] smo preverjali, ali obstajajo specifike v rabi kolokacij pri pisanju mladih v primerjavi z običajno rabo v pisnih besedilih odraslih; b) z analizo učiteljskih popravkov v korpusu *Šolar* smo želeli detektirati težave v rabi kolokacij pri mladini.

2 Opis uporabljenih korpusov

Korpus šolskih pisnih izdelkov *Šolar* vsebuje približno milijon besed oz. 2.703 pisna besedila srednješolcev in učencev zadnjega triletja osnovnih šol, ki so jih napisali pri pouku v šolskem letu 2009/2010. Besedila so avtentična in niso bila ustvarjena za namene gradnje korpusa, so besedilnozvrstno različna (eseji in spisi, testi in različni pisni sestavki, napisani med poukom), nastala pa so pri različnih predmetih, čeprav večji del pri pouku slovenščine (82,3 % vseh besedil). Korpus je bil zgrajen za namene empiričnega raziskovanja pisne jezikovne zmožnosti šolajoče se populacije, saj omogoča detekcijo tipičnih jezikovnih težav, ki jih imajo učenke in učenci pri pisanju. Približno polovica korpusnega gradiva je tudi opremljena z informacijami o avtentičnih jezikovnih (pa tudi vsebinskih) popravkih in komentarjih, ki so jih pri pregledovanju besedil naredili oziroma dopisali učitelji. Več o korpusu gl. Rozman idr. 2012.

Korpus *Kres* je uravnoteženi referenčni korpus pisne slovenščine. Vsebuje skoraj 100 milijonov besed, sestavljajo pa ga besedila, ki so izšla med letoma 1990 in 2010. Besedilnovrstno je pester in uravnotežen, saj vsebuje primerljive, približno 20-odstotne deleže različnih besedil (časopisi, revije, stvarna besedila, leposlovje, internet). Korpus tako predstavlja merodajen vir o sodobni pisni slovenščini in je zato prvenstveno namenjen jezikoslovnim raziskavam. Več o korpusu gl. Logar idr. 2012.

3 Primerjava kolokacij v korpusih *Šolar* in *Kres*

V prispevku razvijamo metodologijo, ki omogoča ugotovitve, kakšne oz. katere kolokacije učenci tipično usvajajo in »vadijo« pri pripravi šolskih nalog; kakšne oz. katere kolokacije se pri šolskem pisanju tipično ne pojavljajo, čeprav so v časopisih, revijah, knjigah zelo pogoste; kakšen oz. kolikšen je presek; kaj nam o usvajanju jezika povedo redke, atipične sopojavitve v šolskem pisanju; in kako je mogoče primerjalne podatke uporabiti za izboljšanje jezikovne infrastrukture in učnih pristopov pri pouku slovenščine.

Primerjava kolokacij v korpusih *Šolar* in *Kres* nam pomeni primerjavo jezika, ki ga mladi govorci in govorke pišejo v sklopu šolskih aktivnosti, z jezikom, ki ga berejo – in naj bi ga kot jezikovno usposobljeni in opremljeni člani družbe razumeli – odrasli govorci in govorke. Pri primerjavi je pomembno upoštevati strukturo obeh korpusov. Na eni strani *Šolar* prinaša besedila enega samega šolskega leta, pri čemer

[7] <http://www.slovenscina.eu/korpusi/kres>. (Dostop 28. 8. 2018.)

gre večinoma za šolske eseje (64,2 % vseh besedil); teme, ki se v slednjih pojavljajo, so vezane na leposlovna dela, ki so se v letu zbiranja korpusa obravnavala v izbranih razredih sodelujočih šol. Čeprav se v rezultatih pojavljajo tudi terminološke kolokacije iz testov pri drugih predmetih, podatki torej odslikavajo predvsem pisanje pri pouku slovenščine. Na drugi strani *Kres* zajema širše časovno obdobje in različne besedilne vrste, ki seveda tudi prinašajo določene značilnosti, pogojene s tipičnimi vsebinami in ubeseditvenimi načini. Rezultate je treba razumeti v luči navedenih značilnosti.

3.1 Metodologija

Za kvantitativni del analize smo na korpusih *Šolar* in *Kres* uporabili metodo za primerjalno luščenje kolokacij, ki je bila predhodno že preizkušena na podatkih korpusa govorjene slovenščine *GOS*[8] (Pollak in Arhar Holdt 2015) in korpusa računalniško posredovane slovenščine *Janes*[9] (Pollak 2015a). Metoda je v tehničnem smislu natančno opisana v navedenih prispevkih, zato na tem mestu navajamo le glavne značilnosti in prilagoditve, pripravljene za korpus *Šolar*.

Luščenje podatkov se opira na orodje SketchEngine (Kilgariff idr. 2004), iz katerega za določen seznam besed izpiše kolokacije, njihove frekvence, kolokacijske vrednosti, povezavo na korpusni zgled v konkordančniku NoSketchEngine ter izračuna primerjalno kolokacijsko vrednost LD-diff. Izhodiščni seznam besed za primerjavo kolokacij v korpusih *Šolar* in *Kres* je najpogostejših sto samostalniških lem v korpusu *Šolar*, kot npr. *človek*, *življenje*, *ljubezen*, *otrok*, *čas*. Za te besede smo pridobili kolokatorje, ki se v korpusnih besedilih pojavljajo na mestu tik pred lemo in so oblikoskladenjsko označeni kot pridevnik ali samostalnik.

V predhodnih primerjalnih raziskavah kolokacij so bili v središču podatki, ki so glede na referenčni korpus novi oz. drugačni. Pri šolskem pisanju pa nas zanimajo tudi kolokacije, ki so v obeh korpusih prekrivne. Širši zajem podatkov omogoča rangiranje podatkov na: (I) kolokacije, ki se pojavljajo zgolj v korpusu *Šolar*; (II) kolokacije, ki se pojavljajo v obeh korpusih; (III) kolokacije, ki se pojavljajo samo v korpusu *Kres* (razdelek 3.2). Obenem nas v korpusu *Šolar* ne zanimajo samo tipične, statistično relevantne kolokacije, ampak tudi redke, atipične sopojavitve. Ta del podatkov ob ustrezni organizaciji v sopojavitvene nize omogoča uvid v močne in šibke točke usvajanja večbesednih enot v sklopu šolskega pisanja (razdelek 3.3). Skladno z navedenimi prioritetami so bili parametri prilagojeni na primerjalno luščenje brez navedbe minimalne meje za pogostost pojavitev v korpusu in vrednost logDice. Po preizkusnem luščenju (Rozman idr. 2016), ki je potekalo na izvirni različici korpusa *Šolar*, smo za raziskavo uporabili prilagojeno različico korpusa, ki vsebuje besedila brez učiteljskih popravkov. Z opisanim postopkom je bilo pridobljenih 221.855 besednih sopojavitev. Po pričakovanjih (zaradi velike razlike v obsegu korpusov) je večina podatkov specifičnih za korpus *Kres* (213.559), za korpus *Šolar* je specifičnih 2061 sopojavitev in 6235 se jih pojavlja v obeh korpusih.

[8] <http://www.slovenscina.eu/korpusi/gos>. (Dostop 28. 8. 2018.)

[9] <http://nl.ijs.si/janes/>. (Dostop 28. 8. 2018.)

3.2 Najpogostejše kolokacije v obeh korpusih

Za ponazoritev v prispevku navajamo po 30 najpogostejših kolokacij, združeno za oba oblikoskladenjska vzorca. Rezultati so preoblikovani v besednozvezno obliko, da so lažje berljivi. S seznama so bili ročno odstranjeni nerelevantni rezultati, ki so posledica težav na ravni označevanja korpusnih besedil.[10]

Samo korpus *Kres*	**Korpusa *Šolar* in *Kres*[11]**	**Samo korpus *Šolar***
pravna oseba, člen zakona, prihodnje leto, predlog zakona, nadzorni svet, socialno delo, državni svet, delovni čas, fizična oseba, združene države, občinski svet, nadaljnje besedilo, gospodarska družba, dopolnitev zakona, terensko delo, nov zakon, leva stran, desna stran, posebne potrebe, začetek leta, raziskovalno delo, sprememba zakona, določba zakona, varstvo okolja, odgovorna oseba, zvezna država, delniška družba, uradna oseba, zemljiška knjiga	šolski spis, šolska naloga, današnji čas, nezakonska mati, družbene razmere, osnovna šola, dobro življenje, glavni junak, naslednji dan, srednja šola, glavni lik, današnji svet, prosti čas, prava ljubezen, dober prijatelj, športni dan, glavna oseba, materni jezik, izgubljeni sin, pravi prijatelj, celo življenje, svetovna vojna, krščanska vera, posmrtno življenje, spolni odnos, velik vpliv, organska potreba, kulturni dom, ljubljena oseba, nezakonski otrok	lik žene, Kreonov brat, Lojzkini starši, Rebulov roman, Antigonin brat, Antigonina sestra, Kreonova žena, Črtomirjev vojak, Bronjin mož, zavedni del, Hamletova mati, Kreonov zakon, Polikarpov odnos, kiparska naloga, Jazonova žena, Antigonina ljubezen, Kovačičev roman, Hamletova težava, Descartova misel, mrtvaški pot, Ožbejev oče, umrli brat, Bubijeva družina, pomembnost razuma, vrnitev očeta, Odisejevo mnenje, Ofelijin oče, Hamletova ljubezen, Polikarpovo dejanje, Salomin odnos

Tabela 1: Najpogostejše kolokacije kot rezultat primerjalnega luščenja podatkov.

Prvi stolpec Tabele 1 prinaša besedišče, ki se v korpusu *Kres* pojavlja pogosto, v korpusu *Šolar* pa se ne pojavi. Kot je razvidno, gre predvsem za besedišče, ki opisuje družbeni sistem in njegovo delovanje (*nadzorni svet*, *delovni čas*, *gospodarska družba*, *varstvo okolja*, *socialno delo*, *posebne potrebe*). Razumevanje tovrstnega besedišča je predpogoj za aktivno državljanstvo v demokratični družbi, zato je ključno zagotoviti njegovo spoznavanje in usvajanje. Pridobljeni korpusni podatki so v sklopu jezikovnega pouka uporabni za izbiro neumetnostnih besedil ter za pripravo strategij za medpredmetno povezovanje pouka slovenščine z relevantnimi drugimi predmeti.

Drugi stolpec prinaša besedišče, ki je v korpusih prekrivno. Na eni strani je opaziti besedišče, vezano na kontekst nastanka in vsebino šolskih besedil (*šolski spis*, *šolska naloga*, *osnovna šola*, *srednja šola*, *glavni junak*). Na drugi strani kolokacije zrcalijo glavne probleme, o katerih učenci pišejo (*materni jezik*, *prosti čas*, *dober prijatelj*, *prava ljubezen*, *posmrtno življenje*, *krščanska vera*, *svetovna vojna*, *spolni odnos*, *nezakonski otrok*), kar

[10] Odstranjenih je bilo 13 primerov: v podatkih iz korpusa *Kres* se najde težava z zapisom šumnika (*drzaven svet*), ostali primeri izvirajo iz korpusa *Šolar*, kjer 8 primerov napačne segmentacije naslovov eseja vodi v vtis zaporednosti besed, ki v besedilu v resnici ne sodijo skupaj (npr. *interpretacija oče* pri naslovu *Vodena interpretacija / Oče in sin*), ter 4 primeri razlik v postopku lematizacije, npr. pripis leme *edini* za razliko od korpusa *Kres*, kjer se pripiše lema *edin*.

[11] Urejeno po pogostosti v korpusu *Šolar*.

omogoča empirični uvid v neposredno povezavo med šolsko obravnavo literarnih besedil in usvajanjem besedišča pri pouku slovenščine.

V tretjem stolpcu so kolokacije, specifične za korpus *Šolar* (teh torej ni v *Kresu*). Po pričakovanjih gre predvsem za besedne zveze z lastnimi imeni, ki se v referenčnem korpusu ne pojavljajo. Kolokacije, ki se osredotočajo na družinske odnose in osebna razmerja književnih likov (*Kreonov brat, Jazonova žena, Antigonina ljubezen*), sopostavljene ob družbeno relevantno besedišče iz korpusa *Kres*, spodbujajo razmislek o možnih pozitivnih vplivih širitve šolskega pisanja na neliterarne kritične eseje, ki bi, usmerjeni k izbrani sodobni družbenopolitični temi, spodbujali usvajanje relevantnega besedišča kot tudi kritičnega mišljenja in argumentacije.

Pridobljene podatke (njihovo celoto, ne samo predstavljeni del) je zato mogoče uporabiti za pripravo infrastrukture za usmerjeno usvajanje besedišča v sklopu šolskega pouka, primarno za: (I) določevanje temeljnega besedišča, ki naj bi ga učenci (s)poznali na določeni stopnji šolanja, ter (II) šolski slovar. Predvsem podatki iz korpusa *Šolar* pa so uporabni tudi za razkrivanje težav, ki jih imajo učenci pri usvajanju besedišča, k čemur se vračamo v poglavju 4, v katerem predstavljamo analizo učiteljskih popravkov – celotni nabor kolokacij iz korpusa bi predstavljeno kvalitativno analizo pomembno dopolnil.

3.3 Razvrstitev kolokacij glede na tipičnost v jezikovni rabi

Ob predpostavki, da korpus *Šolar* predstavlja pisanje mladostnikov, ki pisno kompetenco šele razvijajo, *Kres* pa vzorec odraslih, izkušenih piscev, je mogoče podatke preučevati tudi na ravni posamezne leme. Premik od tipičnih, pogostih kolokacij do zvez, ki se pojavljajo redko, v širšem naboru podobnih primerov ponudi uvid v usvajanje večbesednih enot v sklopu šolskega pisanja. Razvrstitev je mogoče pripraviti na osnovi podatkov o frekvenci in statistični jakosti kolokacij v obeh korpusih. V nadaljevanju za ponazoritev navajamo del podatkov za besedo *odnos*.

Samostalnik *odnos* se pojavlja v 2258 kolokacijah oz. sopojavitvah, od tega 2167 primerov samo v korpusu *Kres* in 23 primerov samo v korpusu *Šolar*. Slika 1 prikazuje po 10 primerov za posamezno skupino, pri čemer lastnoimenske primere tipa *Polikarpov odnos* izpuščamo iz obravnave.

Slika 1: Kolokacije in sopojavitve s samostalnikom *odnos*.

Podatki, ki jih je mogoče nadalje členiti in natančneje analizirati, izkazujejo uporabno vrednost za različne namene. Njihov ključni potencial je za pripravo didaktičnega gradiva za usvajanje kolokacij pri pouku slovenščine, pri čemer je prekrivni del besedišča mogoče razumeti kot izhodišče, na osnovi katerega se pripravi gradivo iz korpusa *Kres*. V zgornjem primeru prekrivni del besedišča razkriva kolokacije, ki se vežejo na vrsto odnosov, kamor lahko ob *prijateljskega* in *medčloveškega* dodamo še [*mednarodni*, *ekonomski*, *delovni*, *poslovni*, *ljubezenski*, *sosedski*] *odnos* in podobno. Na drugi strani so za nadaljnjo analizo zanimive tudi sopojavitve v korpusu *Šolar*, ki so v primerjavi s tipično jezikovno rabo, izkazano v korpusu *Kres*, redke, npr. [*neodkrit*, *osovražen*, *zvit*] *odnos*. Takšni podatki so uporabni za izboljšanje razumevanja asociativnosti kolokacij oz. določanja težavnosti pri njihovem usvajanju ter kot vir za pripravo didaktičnih gradiv za učenje leksike.

4 Analiza učiteljskih popravkov v korpusu *Šolar*

Kvalitativno analizo učiteljskih popravkov smo naredili s predpostavko, da bodo popravki večbesednih zvez razkrili, s katerimi kolokacijami imajo učenci in učenke pri pisanju težave, ter bomo tako pridobili pomembne podatke za načrtovanje učenja besedišča, ki bi preseglo običajne vzorce šolske obravnave – ta pogosto ostaja dekontekstualizirana, kar se med drugim kaže tudi v izrazitem ločevanju obravnave besedišča in slovnice (Stabej idr. 2008), koncept kolokacije pa je pri pouku slovenščine kot prvega jezika malodane prezrt. Predvidevamo lahko, da bi sistematična vključitev leksikogramatike v pouk slovenščine omogočila funkcionalnejše in uspešnejše usvajanje besedišča.[12]

4.1 Metodologija

Tako kot v primerjalni analizi, opisani v poglavju 3, smo se tudi tukaj osredotočili le na dvobesedne samostalniške zveze tipa *samostalnik/pridevnik + samostalnik*. Iz korpusa smo izluščili vse take zveze, kjer je bil vsaj en del zveze popravljen, popravek pa je bil uvrščen v tip »napaka besedišča« (o tipih napak v *Šolarju* gl. Rozman idr. 2012). Od 549 izluščenih zvez je bilo za analizo relevantnih 354 (64,5 %), od tega 228 zvez s pridevniško in 126 s samostalniško sestavino.[13] Vse relevantne zveze smo nato s pregledom konkordanc in odstavkov analizirali v besedilnem kontekstu.

4.2 Rezultati

Analiza je pokazala, da je razmeroma velik delež popravkov vsebinske narave in so torej za leksikalne analize manj relevantni (npr. *znan* ***grški > francoski*** *pisatelj Descartes*;

[12] S tem želimo poudariti, da je pri obravnavi besedišča pomembna tudi sintagmatika besed, kamor uvrščamo tudi pomensko družljivost.

[13] Izločili smo vse tiste primere, ki so se v izpisih ponavljali, ter tiste, ki zaradi napak pri lematizaciji niso predstavljali oblikoslovno ustreznih zvez.

Odisej se ***v času vojne > po vojni*** *odpravi s svojimi možmi v boj na Trojo*). Manj relevantni so tudi dokaj pogosti stilistično-skladenjski (besedilni) popravki, med katerimi so najpogostejši popravki zaradi ponavljanja (npr. *Potem me mama odpelje v Dobjo vas, kjer imam pevske vaje.* ***Pevske vaje > Te*** *trajajo do šestih*), čeprav so nekateri popravki, ki jih lahko uvrstimo v to kategorijo, pomembni tudi za učenje rabe leksemov v širšem kontekstu, bodisi s stališča stilistične ustreznosti kombinacije besed bodisi zaradi leksikogramatičnih omejitev (npr. *Zgodba se konča* ***z tragičnim koncem > tragično***; *Kreon je v bolečinah in* ***v občutku > z občutkom*** *krivde moral živeti naprej*).

V izluščenem gradivu je (predvidevamo, da zaradi velikosti korpusa, morda pa tudi zaradi metode luščenja) malo primerov z več pojavitvami – največkrat se pojavi zveza *materni jezik* (16-krat, popravljena iz zvez *materin/materinski jezik*) in beseda *knjiga* (9-krat) v različnih zvezah. Vendar je analiza pokazala, da se v tej raznolikosti skrivajo vzorci, tako da posamezne primere lahko uvrstimo v omejeno število skupin podobnih popravkov. To smo tudi naredili in tako dobili zasnovo tipologije popravkov (razdelek 4.3). Tipologija zaradi majhnega vzorca ne more biti popolna, dodaten izziv pa je, da so med rezultati tudi popravki posameznih besed (npr. *Celotna* ***knjiga > drama*** *je dokaj grozna*), ki jih ne moremo tolmačiti kot napake zveze. Teh primerov pred analizo nismo izločili, saj se v gradivu po eni strani pojavljajo primeri, ko je težko določiti, ali je popravek relevanten le s stališča besede, po drugi strani pa je tipologizacija pokazala, da se v skoraj vse skupine uvrščajo tako enobesedni kot večbesedni popravki, celostna slika pa je za interpretacijo oz. razumevanje popravkov kolokacij pomembna, saj lažje izpostavi specifike v rabi zvez. Zaradi pomanjkanja predhodnih leksikalnih raziskav se zato vnaprejšnje izločanje ni zdelo ustrezno.

Zaradi značilnosti korpusa (gl. poglavji 2 in 3) se precejšen del popravkov nanaša na izraze s področja slovenščine (npr. *Dandanes imamo vse preveč tujk in* ***prenesenih > prevzetih*** *besed iz naše bljižnje Avstrije, Nemčije*; *Delo Nezakonska mati spada v liriko in je* ***vložena > vložna*** *pesem*), kar bi bilo v prihodnje smiselno podrobneje raziskati, saj bi primeri neustrezno rabljenega strokovnega besedišča[14] lahko pomagali bolje razumeti procese usvajanja besedišča pri mladostnikih. Kot pomembno dopolnitev primerjalni analizi, opisani v poglavju 3, pa bi bilo popravke zanimivo analizirati tudi s stališča tipologije besedišča, torej raziskati pomenska/pojmovna polja besedišča, s katerim imajo pišoči težave, poleg strokovnega besedišča preveriti, ali se popravki nanašajo tudi na abstraktno besedišče, kako je s težavnostjo rabe v jeziku bolj/manj frekventnih besed in zvez ipd., s čimer pa se v prispevku ne ukvarjamo.

4.3 Tipi popravkov

Po izločitvi za leksikalne analize manj relevantnih primerov smo dobili nekaj tipov popravkov, ki razkrivajo nekatere tipične leksikalne težave mladih pišočih, pa tudi

[14] V korpusu *Šolar* je trenutno največ izrazja s področja slovenščine, raziskave pa bi bilo zagotovo smiselno razširiti tudi na druga strokovna področja.

šolska pričakovanja oz. vsebinske zahteve na področju jezikovnega pouka, učiteljske poglede na normo in taktike za širjenje besednega zaklada učenk in učencev.

Zelo pogosti so bili popravki zaradi slogovne zaznamovanosti (*A ti si imel še vedno **prejšnjo punco > prejšnje dekle***), rabe tujk (*Za naše normalno **funkcioniranje > delovanje** potrebujemo drug drugega*) in rabe besed oz. zvez, ki niso napačne oz. neustrezne, ampak so po mnenju učitelja oz. učiteljice pomensko preširoke ali ne dovolj natančne[15] (*Ker je bil **pomemben partizan** > **partizanski komandant**, Celotna **knjiga > drama** je dokaj grozna*).

Na težave z razumevanjem besedišča in razvijajočo se produktivno leksikalno zmožnost pa kažejo pogosti primeri uporabe v kontekstu pomensko neustrezne besede oz. zveze (npr. ***Opredeljeni > Dani** odlomek spada v epiko*; *Antigona je imela zelo veliko dobrih **vrednot > lastnosti***), pri čemer pišoči (v zvezi) pogosto uporabijo besedo, ki je oblikovno blizu, a pomensko drugačna (*Dandanes se družba ali bi raje rekli država ponekod še vedno obnaša na tak **diskriminanten > diskriminatoren** način*; *Zamoti se le s pisanjem o asimulantu, **izdajatelju > izdajalcu** slovenstva Jerneju Jerobniku*) ali pa pride do uporabe besede, ki je pomensko blizu ustrezni (npr. *Poved s katere je viden **dosežek > učinek** Čedermaca*).

Če podrobneje pogledamo, kako se napake odražajo v rabi kolokacij, dobimo naslednje skupine:

- uporaba kolokacije, ki ni slogovno ustrezna (*Oni pripravijo **kupe denarja > denar** in že imajo prostost*; *Zato me je pol **cel čas > ves čas** tišalo, kaj bojo pa rekli*),
- uporaba kolokacije, ki ni pomensko ustrezna (*Krščanska vera govori o **posvetnem > posmrtnem** življenju*; *čeprav je bil le ta bolj izobražen kot Volodja* in je na družbeni ***ravni > lestvici** on veljal več*),
- manj običajna kombinacija besed (*Gregor je partizan, ki na rutinskem **sprehodu > obhodu** opazi mladenko*; *Odločil bi se za pot, ki bi bila v dobro meni in ljudem, ki so v mojem krogu **življenja > ljudi***),
- mešanje[16] (*Črna je kraj z 575 m nadmorske **gladine > višine***; *Kot pravim najprej je pomembno **dokončanje izobrazbe > dokončati šolo***).[17]

Učitelji popravljajo tudi manj pogoste kolokacije s pogosteje rabljenimi ali pomensko natančnejšimi[18] (*delanje > pisanje nalog*, *zdravstveni problemi > zdravstvene*

[15] Pa tudi vsebinsko manj natančne (npr. *V času pred vojno je živel v **slovenskih deželah > notranjosti Slovenije***), a ti primeri so za analizo manj zanimivi.

[16] Gre pravzaprav za podtip kategorije »manj običajna kombinacija besed«, kjer lahko opazujemo neustrezno kombinacijo besed iz več obstoječih kolokacij, ki so pomensko povezane oz. pomensko blizu. Gl. naslednjo opombo.

[17] Nadmorska višina – morska gladina, dokončati šolo/dokončanje šole – pridobiti izobrazbo/pridobitev izobrazbe.

[18] Včasih gre za večplastne primere, ki jih ne moremo zlahka uvrstiti v eno samo skupino.

težave, *mimoidoči* > *mimovozeči avti*), nekatere kolokacijske zveze pa zamenjajo z enobesednim izrazom (*v takratnem času* > *tedaj*).

4.4 Uporabnost rezultatov

Predstavljena kvalitativna analiza popravkov zajema le manjši delež relevantnih kolokacijskih zvez iz korpusa *Šolar*, a so predstavljeni rezultati kljub temu pomembni, saj gotovo prispevajo vsaj majhen delček k razumevanju procesov usvajanja »aktivnega« besedišča, predvsem kolokacij, ki do sedaj v slovenskem prostoru v kontekstu usvajanja in učenja jezika empirično še niso bile raziskane. Rezultati kažejo, da imajo učenke in učenci kar nekaj težav na ravni ustreznega (pomenskega) povezovanja besed, pa tudi pri ustrezni rabi zvez (ki posredno kažejo tudi na nerazumevanje besedišča). Rezultati so zato pomembni tako za pripravo šolskih jezikovnih priročnikov, saj prinašajo veliko avtentičnega gradiva, primernega za vključitev v priročnike, kot tudi za načrtovanje pouka leksike, ki bi ga bilo smiselno razširiti oz. nadgraditi tudi z vsebinami o kolokativnosti besed in sistematičnimi vajami za usvajanje kolokacij.[19] Konkretneje to pomeni, da učenci kolokacijskosti ne bi usvajali le intuitivno, npr. z branjem ali preko učiteljskih popravkov, ampak bi (tudi) s ciljno zastavljenimi dejavnostmi med poukom ozaveščali, kaj so bolj in manj običajne kombinacije besed, stilistične razlike med pomensko enakovrednimi izbirami, spoznavali, kdaj in zakaj določene povezave besed niso ustrezne ... Na podlagi trenutnih delnih analiz posploševanje rezultatov sicer še ni zanesljivo, zato bi bilo empirično raziskovanje treba nadaljevati in ga razširiti na različne, tudi daljše zveze, korpusno metodo luščenja podatkov pa še izboljšati, saj je predstavljena metoda zaradi razmeroma velikega korpusnega šuma dokaj zamudna.

5 Sklep

Če se za konec vrnemo k vprašanju, ki smo si ga na začetku prispevka zastavili kot izhodišče naše analize, lahko zaključimo, da se korpus *Šolar* pokaže kot dober vir podatkov o rabi kolokacij, predstavljeni metodi pa sta se izkazali za učinkoviti, saj njuna uporaba razmeroma dobro odgovarja na zastavljena raziskovalna vprašanja in potrjuje smiselnost preučevanja tako tipičnosti kot atipičnosti v povezovanju besed, tudi pri manj frekventnih pojavitvah. Za zanesljivejše rezultate bi bilo metodi sicer treba uporabiti na obširnejšem gradivu in ju kombinirati z drugimi, ne samo korpusnimi raziskovalnimi metodami. Za razumevanje procesov usvajanja kolokacij je nujno tudi nadaljevati z empiričnimi raziskavami razumevanja in rabe besedišča na sploh, ki jih v slovenskem prostoru primanjkuje, saj se bomo tako uspešneje spopadli z izzivom poučevanja leksike (tudi) na področju slovenščine kot maternega jezika. Predstavljeni rezultati kažejo potencial za razvoj novih šolskih metod ter za pripravo didaktičnih gradiv in priročnikov, pa tudi jezikovnotehnoloških orodij, s

[19] O pomembnosti nalog za spodbujanje učenja besedišča gl. npr. Rozman 2010.

katerimi bi na osnovi tako pridobljenih podatkov lahko npr. v besedilih avtomatsko predoznačili potencialne težave pri povezovanju besed in tako razvili sodobne metode, ki bi lahko prispevale tudi k individualizaciji razvoja pisne kompetence. Ampak to je le ena izmed idej in ker dobre ideje tlakujejo svetlo prihodnost, se zdi prav, da prispevek zaključimo z njo – in z mislijo na uspešen razvoj področja.

Literatura

Arhar Holdt, Špela, in Rozman, Tadeja 2015: Možnosti uporabe podatkov iz korpusa Šolar za pripravo slovarskih priročnikov. Smolej, Mojca (ur.): *Obdobja 34: Slovnica in slovar - aktualni jezikovni opis, 1. del*. Ljubljana: Znanstvena založba Filozofske fakultete. 67–74.

Firth, John R., 1957: *Papers in Linguistics: 1934–51.* London: Oxford University Press.

Gantar, Polona, 2015: *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete.

Gantar, Polona, Kosem, Iztok in Krek, Simon 2016: Discovering Automated Lexicography: The Case of the Slovene Lexical Database. *International Journal of Lexicography* 29/2. 200–225.

Gorjanc, Vojko, Gantar, Polona, Kosem, Iztok, in Krek, Simon (ur.), 2015: *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete.

Halliday, M. A. K., 1966: Lexis as a linguistic level. Bazell, C., Catford, J., Halliday, M. A. K., in Robins, R. (ur.): *In Memory of J. R. Firth*. London: Longman. 148–162.

Hoey, Michael, 2005: *Lexical Priming: A new Theory of Words and Language*. London: Routledge.

Hunston, Susan, in Francis, Gill, 2000: *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins.

James, Carl, 1998: *Errors in Language Learning and Use: Exploring Error Analysis*. London: Longman.

Kilgarriff, Adam, Rychly, Pavel, Smrz, Pavel, in Tugwell, David, 2004: The Sketch Engine. Williams, G., in Vessier, S. (ur.): *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004*. Lorient: Universite de Bretagne-sud. 105–116.

Kjellmer, Göran, 1991: A mint of phrases. Aijmer, Karin, in Altenberg, Bengt (ur.): *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman. 111–127.

Kosem, Iztok, Stritar Kučuk, Mojca, Može, Sara, Zwitter Vitez, Ana, Arhar Holdt, Špela, in Rozman, Tadeja, 2012: *Analiza jezikovnih težav učencev: korpusni pristop*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Kosem, Iztok, Gantar, Polona, in Krek, Simon, 2013: Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., in Tuulik, M. (ur.): *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference*. Ljubljana/Tallinn: Trojina, zavod za uporabno slovenistiko/Eesti Keele Instituut. 32–48.

Krek, Simon, in Kilgarriff, Adam, 2006: Slovene Word Sketches. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Zbornik 5. slovenske in 1. mednarodne konference Jezikovne tehnologije 2006.* Ljubljana: Institut Jožef Stefan. 62–67.

Krek, Simon, Gantar, Polona, Kosem, Iztok, Gorjanc, Vojko, in Laskowski, Cyprian, 2016: Baza kolokacijskega slovarja slovenskega jezika. Erjavec, Tomaž, in Fišer, Darja (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika*. Ljubljana: Institut Jožef Stefan, Oddelek za prevajalstvo. 101–105.

Logar, Nataša, Grčar, Miha, Brakus, Marko, Erjavec, Tomaž, Arhar Holdt, Špela, in Krek, Simon, 2012: *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, Fakulteta za družbene vede.

Mel'čuk, Igor, 1998: Collocations and Lexical Functions. Cowie, A. P. (ur.): *Phraseology. Theory, Analysis, and Applications*. Oxford: Clarendon Press. 23–53.

Nation, I. S. P., 2001: *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Pollak, Senja, in Arhar Holdt, Špela, 2015: Identifying corpus-specific collocations: the case of spoken Slovene. Gajdošová, K., in Žáková, A. (ur.): *Natural language processing, corpus linguistics, lexicography: proceedings*. RAM-Verlag. 117–125.

Pollak, Senja, 2015: Identifikacija spletno specifičnih kolokacij pogostega besedišča. Fišer, Darja (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete. 57–62.

Pollak, Senja, 2015a: Luščenje kolokacij iz korpusa uporabniških spletnih vsebin. Smolej, Mojca (ur.): *Slovnica in slovar - aktualni jezikovni opis, 2. del*. Obdobja 34. Ljubljana: Znanstvena založba Filozofske fakultete. 601–607.

Rozman, Tadeja, 2010: *Vloga enojezičnega razlagalnega slovarja slovenščine pri razvoju jezikovne zmožnosti: doktorska disertacija*. Ljubljana: Filozofska fakulteta.

Rozman, Tadeja, Krapš Vodopivec, Irena, Stritar Kučuk, Mojca, in Kosem, Iztok, 2012: *Empirični pogled na pouk slovenskega jezika*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Rozman, Tadeja, Kosem, Iztok, Pirih Svetina, Nataša, in Ferbežar, Ina, 2015: Slovarji in učenje slovenščine. Gorjanc, Vojko, Gantar, Polona, Kosem, Iztok, in Krek, Simon (ur.): *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete. 150–167.

Rozman, Tadeja, Arhar Holdt, Špela, Pollak, Senja, in Kosem, Iztok, 2016: Luščenje in jezikoslovna analiza kolokacij iz korpusa Šolar. Erjavec, Tomaž, in Fišer, Darja (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika*. Ljubljana: Institut Jožef Stefan, Oddelek za prevajalstvo. 222–224.

Sinclair, John, 1987: *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London and Glasgow: Collins ELT.

Sinclair, John, 1991: *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stabej, Marko, 2011: Jezikovni potrošnik in potrošnica. *Sodobna pedagogika* 62=128/2. 102–113.

Stabej, Marko, Rozman, Tadeja, Pirih Svetina, Nataša, Modrijan, Nina, in Bajec, Boštjan, 2008: *Jezikovni viri pri jezikovnem pouku v osnovni in srednji šoli: končno poročilo z rezultati dela*. Ljubljana: Pedagoški inštitut.

Zahvala

Prispevek je rezultat znanstvenoraziskovalnega dela pri temeljnih raziskovalnih projektih Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki (J6-8255), Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256) ter programov Center za uporabno jezikoslovje (I0-0051), Tehnologije znanja (P2-0103) ter Slovenski jezik – bazične, kontrastivne in aplikativne raziskave (P6-0215). Vse naštete projekte in programe financira Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.