Luščenje in jezikoslovna analiza kolokacij iz korpusa Šolar

Tadeja Rozman,*† Špela Arhar Holdt,*† Senja Pollak,§ Iztok Kosem*†

* Zavod za uporabno slovenistiko Trojina,
Trg republike 3, 1000 Ljubljana
spela.arhar@trojina.si, tadeja.rozman@trojina.si, iztok.kosem@trojina.si
† Filozofska fakulteta Univerze v Ljubljani,
Aškerčeva 2, 1000 Ljubljana
§ Inštitut »Jožef Stefan«,
Jamova cesta 39, 1000 Ljubljana
senja.pollak@ijs.si

1 Uvod

Korpus šolskih pisnih izdelkov Šolar (Kosem et al., 2012) predstavlja pomemben vir za raziskovanje pisne jezikovne zmožnosti učencev in je tako dragocen vir informacij za pripravo didaktičnih gradiv in priročnikov, namenjenih šolski populaciji. Čeprav je bil Šolar zgrajen že leta 2011, pa je bilo zaenkrat opravljenih malo korpusnih analiz, ki bi bile usmerjene v pridobivanje podatkov, pomembnih za izdelavo tovrstnih gradiv. V pričujočem prispevku bomo zato predstavili raziskavo, s katero deloma zapolnjujemo vrzel na področju leksikalnih analiz, ki omogočajo uvid v proces usvajanja besedišča v šolskem kontekstu.

2 Predhodne raziskave

Raziskava, ki jo predstavljamo, je nadgradnja raziskave Arhar Holdt in Rozman (2015). Čeprav je korpus Šolar s približno milijonom besed za analize leksike razmeroma majhen,[1] je omenjena raziskava med drugim pokazala,

a) da Šolar omogoča pridobitev novih spoznanj o rabi besedišča pri učencih ter tako predstavlja pomemben vir relevantnih informacij za izdelavo priročnikov, gradiv ter prenovo didaktičnih načel in praks na področju šolske obravnave besedišča,

b) da je za zanesljivejše zaključke potrebno analizirati večji del korpusnega gradiva, za kar moramo bolj avtomatizirati postopke luščenja informacij.

Ker so se kot gradivno zanimivejši pokazali učiteljski popravki besed, vezanih na kolokacijske omejitve rabe, bomo v pričujočem prispevku predstavili metodo luščenja kolokacij ter analizirali izbrane primere s stališča uporabnosti tako pridobljenih podatkov za pripravo šolskih priročnikov in gradiv.

3 Metoda in rezultati

Za luščenje kolokacij bomo uporabili postopek, ki je bil predhodno preizkušen za primerjavo kolokacij v korpusih GOS in Kres in za luščenje korpusnospecifičnih kolokacij korpusa uporabniških vsebin Janes (Pollak, 2015). Z metodo bomo primerjali kolokacije v korpusu Šolar in uravnoteženem referenčnem korpusu Kres (Logar et al., 2012) .

Za ponazoritev navajamo nekaj rezultatov metode luščenja kolokacij, ki se opira na orodje SketchEngine (Kilgariff et al., 2004) in z avtomatskim izvozom preko API-ja (Pollak, 2015) za določen seznam besed izvozi kolokacije, njihove frekvence in kolokacijske vrednosti ter povezavo na korpusni zgled. Luščili smo kolokatorje, ki se pojavljajo ob najpogostejših stotih samostalniških lemah v korpusu Šolar (*človek, življenje, ljubezen, otrok, čas* itd.), in sicer tiste, ki se pojavljajo na mestu pred lemo in so označeni bodisi kot pridevnik, glagol ali samostalnik. Med kolokacijami, specifičnimi za korpus Šolar, ki imajo vsaj 5 pojavitev in vrednost logDice nad 3, je po pričakovanjih najti precej primerov z lastnoimenskimi kolokatorji (*Bogomilina ljubezen, hlapec Jernej*) in primere, ki vsebujejo jezikovne popravke (o teh gl. spodaj). Za razumevanje procesa usvajanja besedišča v šolskem kontekstu je zanimiva ugotovitev, da so preostali primeri – čeprav na pogled vsebinsko širši – pogosto tesno vezani na specifično obravnavano delo, npr. *pretentati barona* (Ta veseli dan ali Matiček se ženi), *večinska vera* (Krst pri Savici), *absurdno dejanje* (Tujec). Na drugi strani je najti kolokacije, ki se pojavljajo kot terminologija v šolskih testih, npr. *aplikativni cilj, fobični človek*.

V raziskavi bomo metodo luščenja razširili s primerjalno metodo (Pollak in Arhar Holdt, 2015) in prilagodili analizo, da bo ustrezala specifikam korpusa Šolar in namenu raziskave. Za razliko od predhodnih raziskav, v katerih je bilo v središču pozornosti predvsem besedišče, ki je glede na referenčni korpus novo in

[1] Šolar se bo do leta 2018 povečal na predvidoma dva milijona besed, saj trenutno poteka projekt »Nadgradnja korpusa Šolar«, ki ga financira Ministrstvo za kulturo RS.

drugačno, nas namreč pri primerjavi s korpusom Šolar zanima tudi besedišče, ki je v korpusih enako oz. primerljivo. Podatke bomo rangirali v različne skupine: (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. Ob (kritično ovrednoteni) predpostavki, da korpus Šolar predstavlja pisanje mladostnikov, ki pisno kompetenco šele razvijajo, Kres pa vzorec odraslih, izkušenih piscev, je mogoče podatke nadalje kategorizirati in interpretirati v iskanju zadreg in močnih točk šolskega pisanja. Druga sprememba v zornem kotu je premik od tipičnih, pogostih kolokacij do zvez, ki se pojavljajo redko, vendar v širšem naboru podobnih primerov lahko ponudijo uvid v usvajanje večbesednih enot (npr. korpusnospecifične zveze [*kovati, opredeliti, povprašati, izvedeti, dopustiti*] *mnenje* ali [*predati, upreti*] *se mnenju*).

Dodaten izziv luščenja je upoštevanje posebnih oznak korpusa Šolar, tj. oznak učiteljskih jezikovnih popravkov v besedilih. Trenutno je postopek na te oznake neobčutljiv, zato med rezultati korpusnospecifičnih kolokacij dobimo tudi pare napaka-popravek na ravni posamezne besede (npr. *dogotek dogodek; zakonik zakon*). Te primere je smiselno ločevati od primerov, kjer se učiteljska oznaka nanaša na kolokacijsko raven. V prispevku raziščemo možnosti za avtomatsko ločevanje enih in drugih primerov.

Primerjalno analizo kolokacij v korpusu Šolar in korpusu Kres, ki jo omogoča postopek luščenja kolokacij po zgornji metodi, bomo dopolnili z analizo jezikovnih popravkov v korpusu Šolar. V ta namen smo opravili postopek luščenja kolokacij, ki morajo zadostiti dvema pogojema: vsaj en del kolokacije mora biti označen kot napaka in biti hkrati popravljen, jezikovni popravek pa je uvrščen v tip 'napaka besedišča'. Podatke smo izluščili za šest tipov kolokacij (del s popravkom je pisan z velikimi začetnicami, v oklepaju je podano število najdenih zadetkov): pridevnik + SAMOSTALNIK (147), PRIDEVNIK + samostalnik (131), SAMOSTALNIK + samostalnik (176), samostalnik + SAMOSTALNIK (95), glagol + SAMOSTALNIK (171), GLAGOL + samostalnik (103). Ker pri večjem kolokacijskem razponu, npr. -5 +5, zaradi načina označenosti korpusa z jezikovnimi popravki dobimo precej šuma, smo iskanje kolokacij omejili na zaporedne kombinacije besed. Sledila je kategorizacija, podrobna analiza in primerjava s podatki prvega luščenja.

4 Sklep

V prispevku bomo predstavljeno metodo luščenja kolokacij evalvirali, na izbranih primerih opravili kvalitativno analizo in ovrednotili korpus Šolar kot vir (relevantnih) podatkov o specifičnosti rabe kolokacij pri šolski populaciji. S tem nadaljujemo prizadevanja po pridobivanju empiričnih podatkov o rabi besedišča med učenci ter želimo spodbuditi nadaljnje empirične raziskave leksikalne problematike. Te so pomembne tudi v luči prizadevanj za izdelavo sodobnih slovarjev, ki bodo upoštevali zmožnosti in potrebe ciljnih uporabnikov (Gorjanc et al., 2015), med drugim tudi mladih v procesu izobraževanja (Rozman et al., 2015), ter pozivov po spremembah poučevanja slovenščine v šolah, ki so mu deloma botrovali tudi rezultati raziskav (PISA, PIRLS, NPZ) o upadu bralne pismenosti in neučinkovitosti pouka pri razvijanju znanj na višjih taksonomskih ravneh (npr. Rozman et al., 2012; Stabej, 2011).

Literatura

Špela Arhar Holdt in Tadeja Rozman. 2015. Možnosti uporabe podatkov iz korpusa Šolar za pripravo slovarskih priročnikov. V: M. Smolej, ur., Obdobja 34: Slovnica in slovar - aktualni jezikovni opis, 1. del, str. 67–74. Znanstvena založba Filozofske fakultete UL. http://centerslo.si/wp-content/uploads/2015/11/34_1-Arhar-Hol-Roz.pdf

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

Adam Kilgarriff, Pavel Rychly, Pavel Smrz in David Tugwell. 2004. The Sketch Engine. V: G. Williams in S. Vessier, ur., *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004 Lorient, France July 6–10, 2004*, str. 105–116. Universite de Bretagne-sud.

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

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

Senja Pollak. 2015. Luščenje kolokacij iz korpusa uporabniških spletnih vsebin. V: M. Smolej, ur., *Obdobja 34: Slovnica in slovar - aktualni jezikovni opis*, 2. del, str. 601–607. Znanstvena založba Filozofske fakultete UL. http://centerslo.si/wp-content/uploads/2015/11/34_2-Pollak.pdf

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

Tadeja Rozman, Iztok Kosem, Nataša Pirih Svetina in Ina Ferbežar. 2015. Slovarji in učenje slovenščine. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 150–167. Znanstvena založba Filozofske fakultete UL.

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

Marko Stabej. 2011. Jezikovni potrošnik in potrošnica. *Sodobna pedagogika*, 62=128(2), 102–113. Zveza društev pedagoških delavcev Slovenije. http://www.dlib.si/details/URN:NBN:SI:doc-RF8JNPLQ