

Iztok Kosem in Eva Pori

Prvi koraki do seznama temeljnega šolskega besedišča

Povzetek: V prispevku obravnavamo izdelavo seznama temeljnega šolskega besedišča za slovenski jezik, ki bi bil primerljiv s podobnimi seznamami, kot jih poznamo za angleščino in druge jezike. Takšni seznamami ne služijo le za izboljšavo učnih gradiv in jezikovnih orodij za slovenščino, temveč lahko prispevajo tudi k nadgradnji učnih gradiv za druge predmete in k oblikovanju učnih ciljev. V uvodnem delu najprej ponudimo pregled obstoječih seznamov besed za tuje jezike, nato pa predstavimo relevantne obstoječe sezname za slovenščino, s poudarkom na referenčnem seznamu splošnega besedišča in dveh seznamih učbeniškega besedišča. Sledi podrobna predstavitev relevantnih aktualnih korpusnih virov na področju slovenščine v šolskem okolju. Nato izvedemo analizo obstoječih seznamov besed za slovenščino, v kateri jih primerjamo med sabo in s seznamami besed iz relevantnih korpusov, pri čemer preučujemo skladnosti in razlike v zajetju ključnih besed. Izkazuje se, da ima vsak seznam določene pomanjkljivosti, poleg tega pa je treba upoštevati razliko med recepcijo in produkcijo. Opravimo tudi diahrono analizo izbranega vzorca besed iz korpusov učbenikov in šolskih izdelkov, ki razkrije spremembe v rabi besed ne zgolj na časovni, temveč tudi na pomenski ravni. V zaključku povzamemo glavne ugotovitve raziskave in izpostavimo predloge za nadaljnje raziskave, ki bi lahko prispevale k izdelavi celovitega in funkcionalnega seznama temeljnega šolskega besedišča za slovenski jezik.

Ključne besede: seznamami besed, raba jezika v šoli, korpus Šolar, korpus Učbeniki

UDK: 373.3

<https://doi.org/10.63384/sptB53z794s>

Znanstveni prispevek



Dr. Iztok Kosem, višji znanstveni sodelavec, Univerza v Ljubljani, Filozofska fakulteta, Aškerčeva 2 in Fakulteta za računalništvo in informatiko, Večna pot 113, SI-1000 Ljubljana, Slovenija; elektronski naslov: iztok.kosem@ff.uni-lj.si; ©

Dr. Eva Pori, docentka, Univerza v Ljubljani, Filozofska fakulteta, Aškerčeva 2, SI-1000 Ljubljana, Slovenija; elektronski naslov: eva.pori@ff.uni-lj.si

Uvod

Seznami besed so že dolgo priljubljen in široko uporabljan vir na mnogih področjih, npr. pri poučevanju jezikov, psiholingvistiki, strojnem prevajanju ipd. V poučevanju se seznami besed uporabljajo pri izdelavi učnih načrtov, učbenikov in ostalih učnih gradiv, pri opredelitvi zahtevnosti besedil in testiranju, najdemo pa jih tudi v virih, kot so slovarji.¹ Najdaljšo tradicijo v izdelavi seznamov besed ima angleščina, za katero najdemo različne vrste seznamov, od seznamov splošnega, akademskega in različnih specializiranih besedišč do seznamov za materne in seznamov za tuje govorce.

Večina seznamov besedišča temelji na korpusnem gradivu, bodisi izključno na korpusni pogostosti bodisi na kombinaciji z drugimi metodami. Kombiniranje z drugimi metodami, od samega čiščenja seznamov do vključevanja dodatnih besed na podlagi drugih virov, je pričakovano, saj so vsi korpusi na nek način pomanjkljivi, npr. pogoste težave so neuravnoteženost gradiva, vsebnost nestandardnega in tujejezičnega besedišča, napake pri označevanju.

V Sloveniji smo v zadnjih letih dobili prve poskuse seznamov besed za slovenski jezik, nekateri od teh seznamov so bili že integrirani v jezikovne vire in orodja. Na tem področju je vseeno še precejšen manko, zlasti z vidika besedišča, s katerim se srečujejo in bi ga morali poznati šolarji, tj. besedil na receptivni ravni in pričakovani ravni produkcije. Na drugi strani imamo na voljo potrebna gradiva za izdelavo tovrstnih seznamov, saj so v zadnjih 10–15 letih nastali oz. bili posodobljeni številni korpusi oz. zbirke besedil v elektronski obliki, od referenčnih (npr. Gigafida za standardni pisni jezik, Gos za govorni jezik), razvojnih (npr. korpus šolskih pisnih besedil Šolar) do učbeniških.

Pričujoči prispevek skuša narediti prve korake k izdelavi seznama temeljnega šolskega besedišča, tj. (nestrokovnega) besedišča, ki bi ga morali poznati oz. usvojiti šolarji za razumevanje in tudi produkcijo besedil. Pregledu literature in obstoječih virov za slovenski jezik sledi analiza, v kateri primerjamo obstoječe

1 Primer rabe seznamov besed v slovarjih so iztočnice – ali (redkeje) njihovi pomeni –, ki so označene kot relevantne za določene uporabnike, recimo iztočnice, relevantne za akademsko rabo. Podobno se seznami besed uporabljajo tudi pri omejevanju razlagalnega besedišča.

sezname besed in frekvenčne sezname iz šolskih gradiv in izdelkov. Ponudimo tudi primere diahronih analiz izbranih besed. Sledi diskusija z razmisleki o nekaterih ugotovitvah analiz, prispevek pa sklenemo s predlogi o nadaljnjih korakih za izdelavo seznama šolskega besedišča.

Pregled literature in projektov

V tujini so sezname besedišča že od nekdaj deležni velike pozornosti raziskovalcev (gl. npr. Nation 2001; Nation in Waring 1997; Schmitt in McCarthy 1997; Schmitt 2000). Enega prvih seznamov temeljnega angleškega besedišča je izdelal West (1953) in vsebuje približno 2.000 besed. Temelji na predpostavki, da naj bi govorec, ki bi poznal vse te besede in njihove družine, razumel približno 80 % vsakega besedila. Seznam je kljub zastarelosti še vedno pogosto referiran, sta pa bili izdelani dve njegovi nadgradnji (Brezina in Gablasova 2015; Browne 2013), ki temeljita na sodobnih korpusnih podatkih. Precej pozornosti raziskovalcev je deležna tudi akademska angleščina, pri čemer se izdelani sezname osredotočajo na besede, ki so izrazito pogostejše zastopane v akademskem jeziku v primerjavi s splošnim jezikom. Med najbolj znanimi sezname te vrste so Academic Word List (Coxhead 2000; 570 besed), New Academic Word List (Browne idr. 2022; 957 besed) in Academic Vocabulary List (Gardner in Davies 2013; 3.000 besed).

Za učenje tujih jezikov je pomemben projekt KELLY (Kilgarrieff 2014), v okviru katerega so bili izdelani večjezični sezname besed in njihovih prevodov za devet jezikov: angleščino, arabščino, grščino, italijanščino, kitajščino, norveščino, poljščino, ruščino in švedščino. Vsak seznam je vseboval približno 9.000 besed, prevedenih v vse druge jezike projekta. Poleg seznamov je projekt zanimiv tudi zaradi uporabljene metodologije, saj raziskovalci niso vključili zgolj najpogostejših besed iz korpusa, ampak so začeli s približno 6.000 lemmi² (ločenimi po besedni vrsti), potem pa so sledili koraki čiščenja, primerjanja s pojavitvami v referenčnem korpusu in prevajanja. Dodatne kandidate so avtorji dobili z obratnim prevajanjem, tj. prevajanjem prevodov nazaj v izvirni jezik.

Ena od pogostih kritik različnih seznamov besed je, da vsebujejo zgolj besede brez pomenskih informacij. Projekt, ki je presegel to pomanjkljivost in predstavlja model drugim jezikom, je English Vocabulary Profile založbe Cambridge.³ Njegov seznam, za izdelavo katerega so uporabili Cambridge Learner Corpus⁴ (korpus izpitnih testov nematernih govorcev angleščine), sezname besed iz učbenikov in izpitnih gradiv, ima opredeljeno raven SEJO (Skupni evropski jezikovni okvir 2001) za vsak pomen angleških besed, frazemov in celo kolokacij. Pomeni so opremljeni tudi z dodatnimi informacijami, npr. kategorijo tematike; v spletni

2 Lema je osnovna oblika besede v korpusu, npr. hiša, prijazen, misliti, in predstavlja vse pojavitve besede v različnih oblikah (npr. hiša, hiše, hiši, hišo itd.). V zvezi s sezname besed so leme in besede uporabljane bolj ali manj sinonimno, glavna razlika je, da lema lahko predstavlja tudi nebesede (npr. ločila), poleg tega pa v primeru, da lem ne delimo po besednih vrstah, lahko predstavlja več besednih vrst (npr. dolg – pridevnik in samostalnik).

3 <https://www.englishprofile.org/wordlists>

4 https://www.cambridge.org/elt/corpus/learner_corpus2.htm

aplikaciji tako recimo lahko poiščemo samo pomene samostalnikov, ki se nanašajo na živali in spadajo na raven A1.

Tudi za slovenščino imamo izdelanih že nekaj seznamov, katerih glavne značilnosti predstavljamo v nadaljevanju. Za splošni jezik so Pollak idr. (2020) izdelali referenčni seznam 4.768 pogostih splošnih besed, ki se pojavljajo med prvimi 10.000 lemmami v vseh štirih različnih korpusih: uravnoteženem referenčnem korpusu pisne slovenščine Kres⁵ (Logar idr. 2012), referenčnem korpusu govorne slovenščine Gos⁶ (Verdonik in Zwitter 2011), korpusu računalniško posredovane komunikacije Janes (Fišer idr. 2020) in korpusu šolske pisne produkcije Šolar 2.0 (Kosem idr. 2019). Seznam je bil ročno prečiščen, izločena so bila npr. lastna imena (ne pa tudi izlastnoimenski pridevniki) in tujejezične leme. Avtorji so opravili tudi analizo, ki je pokazala, da so s končnega seznama zaradi pogoja pojavitve v vseh štirih korpusih izpadle »besede, ki bi tja po intuiciji (lahko) sodile« (Arhar idr. 2020, str. 14). Seznam je bil izdelan v okviru projekta *Za kakovost slovenskih učbenikov* (KaUč),⁷ ki ga je financiralo Ministrstvo za izobraževanje, znanost in šport in je potekal v letih 2017–2022 v izvedbi Univerze v Ljubljani. V okviru projekta KaUč je bila opravljena tudi analiza 31 učbenikov različnih (naravoslovnih, družboslovnih in humanističnih) predmetov, ki je med drugim razkrila, da v učbenikih najdemo tudi redke in učencem manj znane besede, kar otežuje razumevanje podane vsebine in učenje (Košak Babuder 2021, str. 40–41).

Kot lahko ugotovimo na podlagi literature, se metodologija pri izdelavi seznamov besed večinoma opira na korpusne podatke, pri čemer sta pomembna ažurnost korpusov in upoštevanje informacij iz različnih korpusov in ostalih relevantnih virov. V nadaljevanju predstavljamo vire in projekte, ki so relevantni za izdelavo seznamov šolskega besedišča.

Korpusa Učbeniki 1.0 in ccUčbeniki 1.0

Korpus učbenikov, imenovan Učbeniki 1.0, je bil prvotno izdelan leta 2018 v sklopu projekta *Izdelava seznamov besed in n-gramov za različne ravni šolanja in različne predmete*, ki ga je financiral CLARIN.SI.⁸ Korpus, ki obsega 5.373.268 pojavnic in 4.302.857 besed, vsebuje 127 osnovnošolskih in srednješolskih učbenikov za 16 različnih predmetov (Preglednica 1).

5 <https://www.clarin.si/ske/#dashboard?corpname=kres10>

6 <https://viri.cjvt.si/gos/>

7 Spletna stran projekta: <https://kauc.splet.arnes.si/>

8 Več o projektu in viru: <https://www.cjvt.si/infrastruktura-podpora/seznami-za-ravni-solanja/>; <http://hdl.handle.net/11356/1215>

Predmet	Pojavnice	Besede
Slovenščina	1.826.630	1.397.666
Matematika	996.194	762.249
Biologija	352.128	269.434
Kemija	344.205	263.372
Glasbena umetnost	282.807	216.392
Naravoslovje	269.881	206.502
Geografija	241.000	184.403
Naravoslovje in tehnika	225.816	172.785
Fizika	221.169	169.229
Zgodovina	205.802	157.471
Likovna umetnost	175.480	134.270
Družba	79.134	60.550
Spoznavanje okolja	53.731	41.112
Gospodinjstvo	40.925	31.314
Tehnika in tehnologija	29.818	22.815
Domovinska in državljanska kultura in etika	28.548	21.843

Preglednica 1: Velikost podkorpusev učbeniških besedil posameznih predmetov v korpusu Učbeniki 1.0

Gledano po obdobjih šolanja je osnovnošolski del korpusa večji, pri čemer učbeniki za prvo triado obsegajo 3,83 % (205.801 pojavnice), učbeniki za drugo triado 23,7 % (1.273.970 pojavnice), učbeniki za tretjo triado 44,06 % korpusa (2.367.824 pojavnice). Učbeniki za srednjo šolo predstavljajo 28,39 % korpusa (1.525.673 pojavnice).

Na podlagi korpusa Učbeniki 1.0 smo leta 2022 izdelali in objavili podkorpus ccUčbeniki 1.0 (Kosem idr. 2022), ki obsega 2.181.602 pojavnice in vsebuje 32 učbenikov, ki so na voljo pod odprtim dostopom. Učbeniki so nastali v okviru projekta Zavoda Republike Slovenije za šolstvo v letih 2014–2015⁹ in pokrivajo predmete slovenščina, matematika, kemija, naravoslovje, naravoslovje in tehnika, fizika, geografija, glasbena umetnost in likovna umetnost.¹⁰ Korpus je označen s cevovodom CLASSLA v. 1.1.1 na ravni tokenizacije, stavčne segmentacije, lematizacije in oblikoslovnih oznak. Pomemben je predvsem zaradi možnosti uporabe v različne namene, tako raziskovalne kot komercialne. Korpus je bil izdelan v projektu *Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti* (PROP) (2021–2024).¹¹

⁹ Objavljeni so na strani <https://eucbeniki.sio.si/>.

¹⁰ Učbenik za šport, ki je tudi na spletni strani, takrat v podkorpus ni bil vključen, ker še ni potrjen.

¹¹ Več o projektu: <https://www.cjvt.si/prop/>

Korpus Šolar 3.0

V okviru projekta PROP je bila razvita tudi različica 3.0 razvojnega korpusa Šolar – jezikovnega vira, specializiranega za pedagoško rabo, ki že več kot desetletje omogoča raziskovanje razvoja pisnih zmožnosti učencev (Arhar Holdt idr. 2022; Arhar Holdt in Kosem 2024)¹². Gre za korpus šolskih pisnih izdelkov, nastalih na slovenskih osnovnih in srednjih šolah, ki ponuja pomemben vpogled v pisno jezikovno zmožnost šolske populacije. Trenutno dostopna različica 3.0 vključuje 5.485 besedil, večinoma esejev, pa tudi nekatera praktičnosporazumevalna besedila učencev slovenskih osnovnih in srednjih šol. Korpus zajema tudi 36.570 avtentičnih učiteljskih jezikovnih popravkov, ki so del izvornih pisnih izdelkov učencev in podajajo realno sliko popravljanja šolskih spisov v izobraževalnem procesu.

Seznami besed

Za šolsko rabo so predvsem relevantni seznami, ki so nastali na podlagi korpusa Učbeniki 1.0 (za več o korpusu glej spodaj).

- Seznam širšega učbeniškega besedišča (Š-UČB; Kosem idr. 2023) vsebuje 11.906 besed z vsaj 20 pojavitvami (4,65 na milijon besed) v korpusu. Seznam ni bil prečiščen, so pa bili izločeni vsi števniki, neznane leme in ločila. Na seznamu je največ samostalnikov (6.534), pridevnikov (2.665), glagolov (1.778) in prislovov (615), ostale besedne vrste so zastopane manj kot stokrat.
- Seznam splošnega učbeniškega besedišča (S-UČB; Kosem idr. 2019) vsebuje 5.977 besed, ki se pojavljajo v besedilih vsaj osmih od 16 različnih šolskih predmetov. Ravno ta kriterij je ključna razlika glede na seznam širšega besedišča, saj tako dobimo bolj splošne besede, ne pa terminov, ki so mogoče vezani na nekaj predmetov ali celo samo enega. Seznam je bil ročno prečiščen – odstranili smo morebitne napake pri označevanju besedil (npr. napačna osnovna oblika, zatipki, simboli). Tudi na tem seznamu je največ samostalnikov (2.739), sledijo pridevniki (1.184), glagoli (1.121) in prislovi (416).
- 16 frekvenčnih seznamov ključnih besed po predmetih (Kosem idr. 2019), za vsak predmet po en seznam, pri čemer so na seznamih ključne besede z vsaj petimi pojavitvami (skupaj 23.270 ključnih besed).
- Seznam n-gramov (zaporednih nizov besed), ki vsebuje 9.177 n-gramov, od tega 7.310 bigramov, 1.600 trigramov, 184 štirigramov in 83 petgramov, z vsaj desetimi pojavitvami v korpusu (Kosem idr. 2019).
- Seznam kolokacij (Munda idr. 2025a), ki vsebuje 701.943 kolokacij iz 82 skladenjskih struktur.
- Seznam skladenjskih struktur z dvema do desetimi pojavnicami (Munda idr. 2025b).

¹² Prva različica korpusa Šolar je nastala kot eden izmed produktov aktivnosti Nove didaktike poučevanja slovenskega jezika v okviru projekta Sporazumevanje v slovenskem jeziku (2008–2013). Vsebovala je skoraj milijon besed oz. 967.477 besed.

Za slovenščino kot tuji jezik imamo na voljo seznam 5.273 lem (Klemen idr. 2022a, 2022b), ki so razdeljene v naslednje skupine glede na raven jezikovnega znanja (SEJO): 350 lem z oznako A1-jedro, 864 z oznako A1-širše, 1.451 z oznako A2 in 2.608 z oznako B1. Kriterij za vključitev besede v A1-jedro je bil, da se je pojavila v vseh petih učbenikih stopnje A1, v skupino A1-širše pa so bile vključene besede, ki so se pojavile v štirih, treh ali dveh učbenikih iste stopnje. Beseda je bila uvrščena v skupino A2, če se je pojavila v petih, štirih, treh ali dveh učbenikih stopnje A2, največ pa v enem učbeniku stopnje A1. V skupino B1 pa je bila vključena beseda, ki se v učbenikih stopnje A1 sploh ni pojavila, na stopnji A2 največ v enem učbeniku, na stopnji B1 pa v enem ali dveh učbenikih. Seznam in oznake temeljijo na korpusu učbenikov za učenje slovenščine kot drugega in tujega jezika (KUUS), ki zajema 17 učbenikov. Vsak učbenik ima pripisano raven oz. ravni SEJO A1 (A1-jedro, A1-širše), A2 ali B1. Seznam besed je bil ročno prečiščen, odstranjeni so bili šumi, napačno označene leme, lastna imena in števila.

Metodologija

Namen pričujoče analize je predvsem identificirati pomanjkljivosti obstoječih seznamov in ugotoviti, kako bi jih lahko kombinirali oz. dopolnjevali, da bi prišli bližje seznamu temeljnega šolskega besedišča. Uporabili smo naslednje sezname in vire: referenčni seznam pogostih splošnih besed (REFS), seznam širšega učbeniškega besedišča (Š-UČB), seznam splošnega učbeniškega besedišča (S-UČB), frekvenčni seznam iz korpusa Šolar 3.0 (Šolar) in frekvenčni seznam iz korpusa Učbeniki 1.0 (Učbenik). Za vključitev obeh učbeniških seznamov smo se odločili, ker S-UČB predstavlja splošno besedišče, ki ga najdemo pri raznoraznih predmetih in bi moral pokrivati osnovno potrebno (receptivno) znanje šolske populacije, Š-UČB pa prinaša neko nadgradnjo, in sicer kombinacijo splošnega in terminološkega oz. področno zamejenega besedišča, pri čemer slednje lahko kasneje prehaja v splošno jezikovno rabo.

Opravili smo dve vrsti analiz: a) primerjalno analizo seznamov Š-UČB, S-UČB, REFS, Šolar in Učbeniki, pri čemer preučujemo skladnosti in razlike v zajetju ključnih leksikalnih enot; b) diahrono analizo izbranega vzorca besed iz korpusa Učbeniki 1.0, s čimer osvetlimo spremembe v učbeniškem besedišču skozi čas.

Primerjava REFS s Š-UČB in S-UČB ter korpusom Šolar

V prvi primerjavi smo želeli preveriti, kolikšen delež lem z REFS najdemo na obeh seznamih učbeniškega besedišča. Rezultati, prikazani v Preglednici 2, kažejo, da 4.336 (91 %) pogostih splošnih lem najdemo na seznamu Š-UČB, ko jih primerjamo s S-UČB, je prekrivnost pričakovano nekoliko nižja (71,7 %). Gledano po besednih vrstah je prekrivnost REFS pri polnopolninskih besednih vrstah

dokaj stabilna pri primerjavi z obema učbeniškima seznamoma, pri funkcijskih besednih vrstah pa so odstopanja nekoliko večja, zlasti pri medmetih.

	Š-UČB	S-UČB
REFS	91 %	71,7 %
Samostalniki (2.038)	91,9 %	71 %
Pridevniki (914)	90,8 %	70,9 %
Glagoli (1.174)	89,9 %	69,5 %
Prislovi (427)	91,8 %	75,2 %
Predlogi (43)	97,8 %	90,7 %
Členki (37)	100 %	83,8 %
Vezniki (39)	100 %	97,4 %
Zaimki (59)	98,3 %	84,8 %
Števniki (30)	/ ¹³	93,3 %
Medmeti (7)	100 %	14,3 %

Preglednica 2: Zastopanost lem z REFS na obeh učbeniških seznamih

V naslednjem koraku smo natančneje analizirali razlike med seznamami. V Preglednici 3 navajamo po 20 najpogostejših polnopomenskih besed in vse besede izbranih funkcijskih besednih vrst s seznama REFS oz. Š-UČB in S-UČB, ki jih ne najdemo na primerjanem seznamu oz. seznamih. Vidimo lahko, da po eni strani seznama učbeniškega besedišča ne vsebujeta marsikatero pogoste splošne besede (npr. *ponedeljek*, *vikend*, *vodja*, *čuden*, *nesrečen*, *prijaviti*, *zavrniti*), za katere bi (lahko) pričakovali, da bi jih šolarji morali poznati in usvojiti. Po drugi strani lahko med besedami, ki jih najdemo na učbeniških seznamih, jih pa ni na seznamu REFS, opazimo veliko področno specifičnih besed oz. besed s prevladujočo rabo na določenih področjih (npr. *poved*, *molekula*, *ogljikov*, *polaren*). Medtem ko bi za te besede še lahko rekli, da mogoče zaradi omejenosti rabe ne sodijo na potencialni seznam temeljnega šolskega besedišča, pa zlasti glagoli, kot so *zapisovati*, *preoblikovati*, *razporediti*, *poenostaviti* ipd., kažejo, da tudi učbeniška seznama ponujata precej relevantnih kandidatov za takšen seznam.

13 Števniki so bili s seznama učbeniškega besedišča izločeni.

	<i>REFS – DA Š-UČB – NE S-UČB – NE</i>	<i>REFS – DA Š-UČB – NE</i>	<i>REFS – NE Š-UČB – DA S-UČB – DA</i>
Samostalniki (2.038)	<i>gospod, stranka, politika, sodišče, plača, punca, usoda, kriza, minister, upanje, ponedeljek, vikend, vodja, volitev, kolega, klub, sezona, samomor, reč, letnik</i>	<i>šef, firma, umor, veza, panika, konkurenca, rit, zavarovanje, finance, kvaliteta, faks, tolar, sramota, norec, kontrola, izgled, seks, morilec, poslovanje, upokojenec</i>	<i>poved, molekula, cm, stranica, masa, kislina, raztopina, spojina, prostornina, graf, preglednica, kvadrat, ploskev, model, tlak, formula, pravokotnik, ravnina, tekočina, km</i>
Pridevniki (914)	<i>super, čuden, žalosten, spoštovan, jezen, zaljubljen, praven, letošnji, priden, grd, ubog, bivši, važen, božji, pogumen, razočaran, nesrečen, pozdravljen, konkreten, ljubezenski</i>	<i>navajen, pijan, sporen, mariborski, kvaliteten, nesramen, včerašnji, kazniv, sovražen, mišljen, pokojen, noseč, trmast, nesposoben, ljubosumen, profesionalen, prodajen, razumen, nemočen, prisilen</i>	<i>ogljikov, kemijski, pridobljen, pravokoten, celičen, rastlinski, skladien, organski, poljuben, vzporeden, manjkajoč, vrsten, geometrijski, nepopoln, naštet, svetloben, merski, krožen, polaren, kovinski</i>
Glagoli (1.174)	<i>priznati, poročiti, izgledati, klicati, tikati, zmeniti, zaljubiti, naročiti, smejati, zgledati, prijaviti, obljubiti, pogrešati, zavrnuti, udeležiti, dvomiti, pozdraviti, oprostiti, splačati, sanjati</i>	<i>zgledati, splačati, javiti, odpustiti, dojeti, obetati, obupati, obljubljeni, brigati, ukiniti, žaliti, obžalovati, zadevati, tvegati, pretepati, obsojati, preteptati, staviti, varati, razočarati</i>	<i>izračunati, razvrstiti, zapisovati, izpisati, pretvoriti, zaokrožiti, preoblikovati, ponazoriti, izločiti, prerisati, krožiti, pobarvati, prehranjevati, segrevati, seštevati, razporediti, poganjati, razstaviti, poenostaviti, prešteti</i>
Prislovi (427)	<i>noter, notri, zdajle, zadnjič, skratka, osebno, tiho, srečno, slučajno, zaenkrat, nekam, normalno, nikjer, mar, vsakič, zastoj, absolutno, čudno, pošteno, prepozno</i>	<i>zdajle, slučajno, absolutno, prepozno, fizično, možno, iskreno, konkretno, direktno, grozno, kakorkoli, navsezadnje, noro, čimprej, drago, nakar, psihično, pridno, tretjič, brezplačno</i>	<i>desno, enakomerno, levo, praviloma, grafično, gosto, poljubno, vzporedno, gospodarsko, postopno, navpično, pravokotno, zmerno, sčasoma, razmeroma, sočasno, kopno, naravnost, ročno, spretno</i>
Predlogi (43)	<i>zoper, izpred, zavoljo, nasproti</i>	<i>zavoljo</i>	<i>vzdolž, izpod</i>
Členki (37)	<i>itak, menda, češ, kajne, morebiti, nikar</i>	/	/
Vezniki (39)	<i>četudi</i>	/	/
Zaimki (59)	<i>oni, karkoli, marsikdo, nobeden, najin, nikakršen, čigav, vajin</i>	<i>nikakršen</i>	<i>kolikšen, medme, katerikoli, skozme, nase, čigar, vsakršen, kakršenkoli</i>

Preglednica 3: Do 20 besed po besednih vrstah, ki manjkajo na vsaj enem od primerjanih seznamov

Za zaključek smo primerjali seznam REFS s korpusom Šolar in ugotovili, da se samo sedem besed s seznama ne pojavlja v Šolarju: samostalniki *druga, nova, reklo* in *testo*, pridevnik *noseč*, glagol *besediti* in prislov *zabavno*.

Primerjava seznamov učbeniškega besedišča in splošnega učbeniškega besedišča s korpusom Šolar

Primerjava učbeniškega besedišča s korpusom Šolar 3.0 je relevantna, ker primerjamo šolsko recepcijo (učbeniki) s produkcijo (Šolar). Pri primerjavi nas je najprej zanimalo, koliko besed z obeh seznamov učbeniškega besedišča najdemo v korpusu Šolar. Kot kažeta Preglednici 4 in 5, se precejšen delež besedišča iz korpusa Učbeniki pojavlja tudi v korpusu Šolar, funkcijske besede v celoti, nekoliko manjša sta le deleža ujemanja samostalnikov in pridevnikov. Med besedami, ki jih ne najdemo v korpusu Šolar, prevladuje področno specifično izrazje, pri S-UČB v manjši meri.

	Šolar 3.0	Primeri besed, ki jih ni v korpusu Šolar 3.0
Š-UČB (11.906)	76,5 %	
Samostalniki (6.534)	71,6 %	<i>molekula, stranica, raztopina, spojina, ploščina, prostornina, preglednica, ploskev, daljica, elektron, krožnica, ion, pravokotnik, polinom, polmer, naslovnik, oklepaj, oglišče, vodik, diagonala</i>
Pridevniki (2.665)	72,3 %	<i>številski, celičen, decimalen, koordinaten, računski, geometrijski, natrijev, vodikov, merski, perioden, raztopljen, deljiv, desetiški, toploten, molekulski, biotski, agregaten, karboksilen, atomski, navpičen</i>
Glagoli (1.778)	93,3 %	<i>prerisati, besediti, segreti, izrezati, prišteti, mirovati, razgraditi, razvrščati, pretakati, odmeriti, prezrcaliti, skrčiti, ohlajati, absorbirati, stikati, izhlapati, konstruirati, skicirati, oksidirati, zmnožiti</i>
Prislovi (615)	91 %	<i>enakomerno, slogovno, poljubno, navpično, pravokotno, premo, zmerno, kopno, podredno, energijsko, oblikovno, priredno</i>
Predlogi (53)	100 %	/
Členki (41)	100 %	/
Vezniki (42)	100 %	/
Zaimki (71)	98,6 %	<i>kolik</i>
Medmeti (19)	68,4 %	<i>ho, čiv, hi, hoj, hov, uf</i>

Preglednica 4: Primerjava seznama Š-UČB s korpusom Šolar 3.0

	Šolar 3.0	Primeri besed, ki jih ni v korpusu Šolar 3.0
S-UČB (5.977)	87,8 %	
Samostalniki (2.739)	90,7 %	<i>molekula, stranica, raztopina, spojina, prostornina, preglednica, ploskev, pravokotnik, oklepaj, kota, fotosinteza, ogljik</i>
Pridevniki (1.184)	89,8 %	<i>celičen, geometrijski, merski, raztopljen, toploten, atomski, navpičen, Zemljin, poševen, pritrjen, temperaturn, zaokrožen, višinski, žveplov, pripadajoč</i>
Glagoli (1.121)	96 %	<i>prerisati, segreti, izrezati, mirovati, razgraditi, razvrščati, pretakati, skrčiti, ohlajati, absorbirati, stikati, izhlapeti, skicirati, podvojiti, raztegniti</i>
Prislovi (416)	95,2 %	<i>enakomerno, poljubno, navpično, pravokotno, zmerno, kopno, tem, rdeče, rumeno, zaporedno, prostorsko, pregledno, industrijsko, trajno, vodoravno, izmenično, koristno, šestkrat, mehansko, škodljivo¹⁴</i>
Predlogi (42)	100 %	/
Členki (31)	100 %	/
Vezniki (39)	100 %	/
Zaimki (58)	100 %	/
Medmeti (3)	100 %	/

Preglednica 5: Primerjava seznama S-UČB s korpusom Šolar 3.0

Opravili smo tudi primerjavo lem v korpusu Šolar 3.0 z vsemi lemmi v učbeniškem korpusu in izkazalo se je, da od 33.304 lem (56,8 %) iz Šolarja v korpusu učbenikov ne najdemo 18.899 lem. Od teh jih je sicer 12.192 takih s samo eno pojavitvijo v korpusu Šolar. Med pogostejšimi skupinami lem, ki se pojavljajo samo v korpusu Šolar, so napake zapisa (npr. *utemeljitel, vznemerljivo, izobroženost, individealist*), nestandardno, neformalno in grobo besedje (npr. *zjeban, seueda, reskirati, štadijon*) ter (anonimizirana) lastna imena (npr. *Polikarp, Hasan, Saloma, Polinejk, XImeX¹⁵, XKrajX*); zlasti slednja so sicer bolj odraz tematik esejev¹⁶ v času zbiranja besedil za korpus Šolar. Za občnoimenske samostalnike in ostale besedne vrste naštevamo po nekaj (pogostejših) primerov iz korpusa Šolar, ki jih ni v učbeniškem korpusu:

- samostalniki: *fobija, ekstraspekcija, intraspekcija, partnerka, uporništvo, baronica, samodejavnost, pomočnica, vice, halucinacija, prokreacija, absurd, behaviorizem, očim, klerikalec, izdajalka, bolničar, ločenost, daritev, plebejec,*

¹⁴ Pri prislovih smo navedli vseh 20 prislovov, ki jih ni v korpusu Šolar.

¹⁵ XImeX smo uporabili za anonimizacijo imen. Nekateri drugi korpusi uporabljajo pristop za menjave imen z generičnim pogostim imenom, npr. Janez, vendar pa je potem težko slediti, katero ime je dejansko anonimizirano in katero pravo.

¹⁶ Gre za tip besedila v korpusu Šolar, za katerega je v šolskem kontekstu uporabljen tudi izraz spis.

- pridevniki: *noseč, ljubosumen, briljanten, usmiljen, razpravljaljen, legalen, nepravilčen, absurden, hraber, nuklearen, nelegalen, dramatski, dopustniški, zagnan, klerikalen, poskusen, aplikativen, objokan, posvetilen, ponižan,*
- glagoli: *pokristjaniti, zaničevati, spreobrniti, krstiti, osvobajati, privoliti, poniževati, pobotati, zanemarjati, kesati, zalesketati, zanositi, podleči, odrea-girati, razočarati, pokesati, kloniti, zblazneti, ignorirati, popivati,*
- prislovi: *psihično, zadovoljivo, hudobno, protinaravno, legalno, definitivno, nepremišljeno, brezpogojno, zavedno, prevečkrat, direktno, intraspektivno, progresivno, častno, optimistično, ekstraspektivno, hladnokrvno, obupano,*
- predlogi: *napram, vpricho, tekom, širom,*
- členki: *bržkone, malodane, edinole,*
- zaimki: *medse, redkokdo, taisti, mnogokaj, prenekateri, marsikakšen.*

Kot lahko vidimo, je med pogostejšimi primeri kar nekaj besed tematsko specifičnih, vseeno pa najdemo tudi precej splošno razširjenih in uporabljenih besed.

Vendar pa zgolj identifikacija besed, ki jih najdemo na enem seznamu, ne pa tudi na drugem, ne razkrije tudi ostalih razlik v rabi besed. Tako smo se osredotočili še na besede, ki se pojavljajo tako v korpusu Šolar kot v korpusu Učbeniki, a izkazujejo največje razlike v pogostosti rabe. S primerjavo relativnih frekvenc smo identificirali besede, ki so primerjalno precej pogostejše v korpusu Šolar, in besede, ki so precej pogostejše v korpusu Učbeniki. V preglednicah v prilogi podajamo prvih sto lem z vsakega od seznamov.

Poleg glagola *biti* na vrhu seznama besed, ki so precej pogostejše uporabljene v korpusu Šolar kot v korpusu Učbeniki, najdemo tu predvsem zaimke in veznike, izstopajo tudi členki (npr. *ne, tudi, še*). Med samostalniki je precej besed, povezanih s sorodniki (npr. *oče, otrok, starš, mama*), pri glagolih pa prevladujejo glagoli kognitivnih procesov (npr. *hoteti, imeti, misliti, želeti*), premikanja (npr. *priti, oditi*) in sporazumevanja (npr. *govoriti, povedati*). Po drugi strani so v učbeniškem korpusu v primerjavi s Šolarjem izrazito pogostejše uporabljeni predlogi (npr. *v, z, na*), potem pa že sledijo polnopomenske besedne vrste in področno zamejene besede (npr. *število, besedilo, beseda, snov*). Večina besed na obeh seznamih je rabljena dokaj pogosto v obeh korpusih, najdemo pa izjeme: s seznama izrazito pogostejše rabljenih besed v korpusu Šolar so v korpusu Učbeniki dokaj redko rabljeni strogo tematska lastnoimenska samostalniki *Antigona* in Črtomir ter glagola *poročiti* in *ubiti*; s seznama izrazito pogostejše rabljenih besed v korpusu Učbeniki je redko rabljenih v korpusu Šolar precej več, npr. *izračunati, enačba, njihov, sila* (prislov), *naš, njen*.

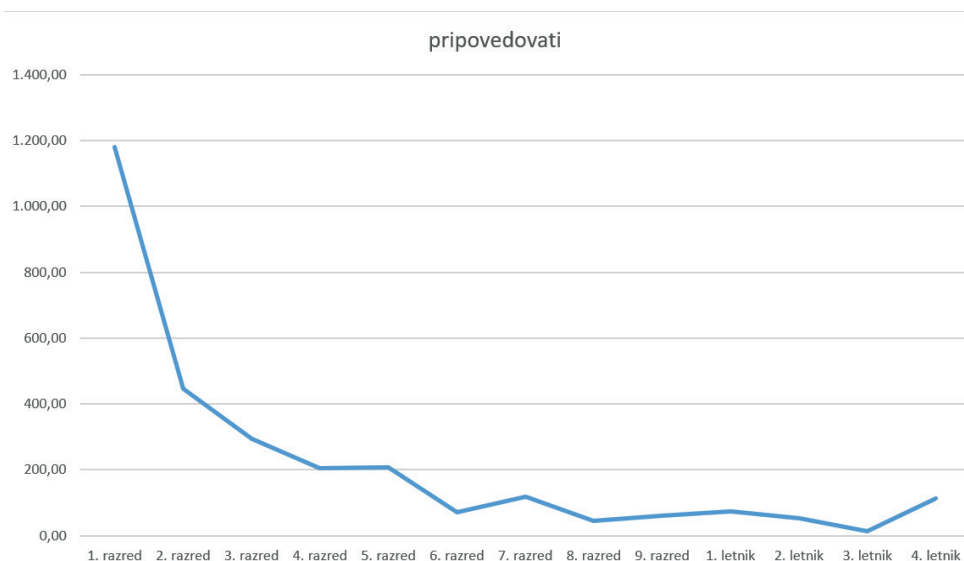
Diahrona analiza izbranega besedišča

V diahronem pregledu (upo)rabe besed nas je zanimalo, kako se spreminja (upo)raba učbeniškega besedišča v učbeniškem gradivu in/ali pri šolski populaciji skozi čas. Na izbranih primerih s seznamov učbeniškega besedišča smo spremljali spremembe v rabi besed skozi različne ravni šolanja (osnovnošolsko in srednje-

šolsko obdobje) po posameznih razredih oz. letnikih, od 1. razreda osnovne šole (OŠ) do 4. letnika srednje šole (SŠ). Zanimalo nas je predvsem, ali se določeno besedišče, ki se pojavlja v učbenikih, dejansko uporablja/pojavlja v pisni produkciji učencev ter ali se skozi različne ravni šolanja v rabi besed kažejo tudi spremembe na ravni pomena ipd.

Primer 1

Glagol *pripovedovati* najdemo na seznamu REFS in obeh učbeniških seznamih, pa tudi v korpusu Šolar. Raba po razredih in letnikih v korpusu Učbeniki kaže (Slika 1), da se glagol na receptivni ravni pojavlja že v zelo zgodnjem obdobju šolanja, saj doseže najvišjo pogostost v 1. razredu OŠ. V prvih treh razredih se pojavlja predvsem kot del učbeniških navodil (npr. *pripoveduj; opiši in pripoveduj*) pri matematiki in slovenščini, medtem ko od 4. razreda naprej že kot kombinacija navodil in poročanja (npr. o liku v knjigi, ki pripoveduje o čem). Pogostost glagola nato hitro upada skozi drugo in tretjo triado OŠ, pri čemer se, sicer v nižjih vrednostih, kontinuirano pojavlja tudi v SŠ.

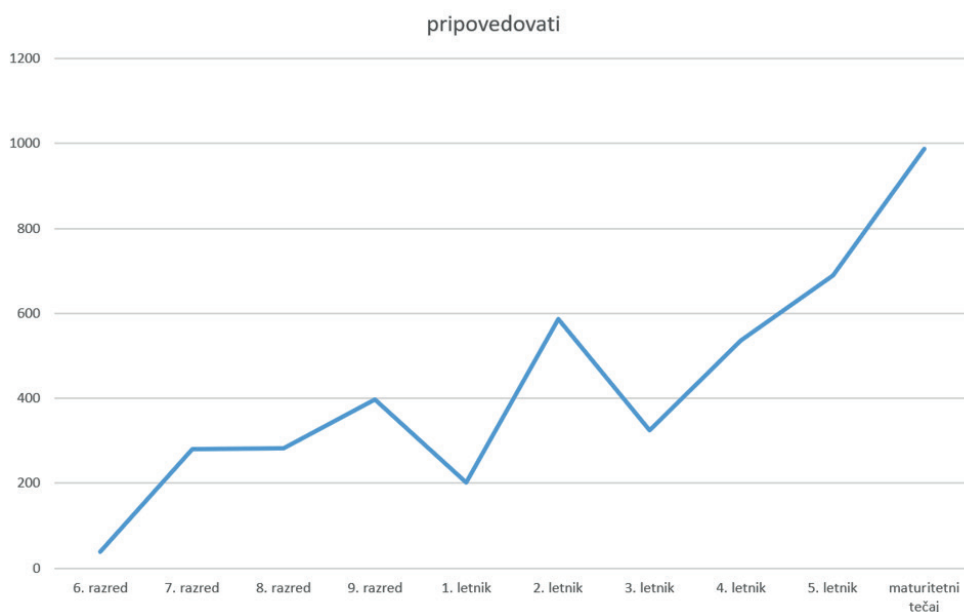


Slika 1: Prikaz relativne besedilne pogostosti¹⁷ glagola *pripovedovati* po razredih in letnikih v korpusu Učbeniki

Čeprav je v učbenikih v nekaterih razredih opazna manjša raba, kar je povezano z različno tematiko posameznih vsebin, pa podatki iz korpusa Šolar (Slika 2)

¹⁷ Relativna besedilna pogostost primerja pogostost iskanega niza v določenem podkorpusu (recimo v podkorpusu posameznega razreda) s pogostostjo niza v celotnem korpusu. Za več glej spletno stran: <https://www.sketchengine.eu/glossary/relative-text-type-frequency/>.

kažejo, da se v kasnejših letih šolanja (od 8. razreda naprej) razvija produktivna raba glagola, zlasti v kontekstu pisnih izdelkov pri pouku slovenščine (eseji, spisi in drugi samostojni pisni sestavki). Uporaba se v višjih letnikih naglo povečuje in doseže vrh v okviru intenzivnejših priprav na maturo (Slika 2), ko se od učencev pričakuje kompleksnejša pisna produkcija v smislu ustvarjalnega pisanja in analize besedil (maturitetni esej).



Slika 2: Prikaz relativne besedilne pogostosti rabe glagola pripovedovati po razredih in letnikih v korpusu Šolar¹⁸

Primer 2

Zanimiv kazalnik trenda uporabe besed je skupina besed, ki se pojavljajo izključno v učbenikih, medtem ko niso del REFS in jih tudi ne zasledimo v pisni produkciji (v korpusu Šolar). Pojav nakazuje, da so določeni termini, specifični za učbeniški jezik (npr. *molekula*, *stranica*, *raztopina*; glej Preglednico 4), uporabljeni predvsem na receptivni ravni. Gre za izraze, ki jih učenci morajo poznati in jih obvladujejo v kontekstu učenja in razumevanja pri posameznih predmetih, vendar se ne izkazujejo za del aktivne rabe oz. se najverjetneje le redko pojavljajo v njihovem samostojnem pisnem izražanju.

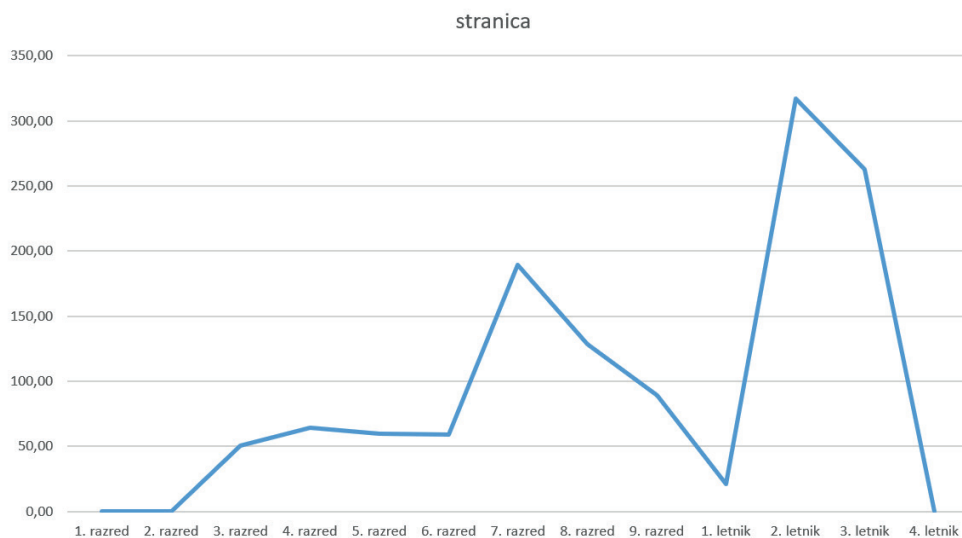
Izračunane pogostosti rabe izbranega samostalnika *stranica* v učbenikih za posamezni razred/letnik kažejo, da se beseda začne pojavljati konec prve triade OŠ (v 3. razredu) in v drugi triadi OŠ (Preglednica 6), kar kaže na uvajanje termino-

¹⁸ Maturitetni tečaj je enoletni program, ki omogoča pripravo na maturo.

logije, specifične za predmet matematike, predvsem na področju geometrije (npr. *dolžina stranice, kocka s stranico; Trikotnik ima tri stranice.*). Korpus namreč pokaže, da se tako v OŠ kot SŠ beseda večinsko uporablja pri matematiki, v manjši meri pa še pri fiziki (8.–9. razred in 1. letnik) ali pri predmetu tehnika in tehnologija (6. razred) (Preglednica 6).

Predmet	Absolutna pogostost	Relativna besedilna pogostost
Matematika	2.106	598,4
Fizika	269	27,3
Tehnika in tehnologija	13	104,1
Naravoslovje in tehnika	12	12,5
Likovna umetnost	4	5,2
Naravoslovje	6	5,1
Spoznavanje okolja	1	5,1
Kemija	4	2,7
Družba	1	3
Zgodovina	1	1,1
Slovenščina	4	0,5

Preglednica 6: Relativna besedilna pogostost samostalnika stranica po šolskih predmetih v korpusu Učbeniki



Slika 3: Prikaz relativne besedilne pogostosti samostalnika stranica po razredih (OŠ) in letnikih (SŠ) v korpusu Učbeniki

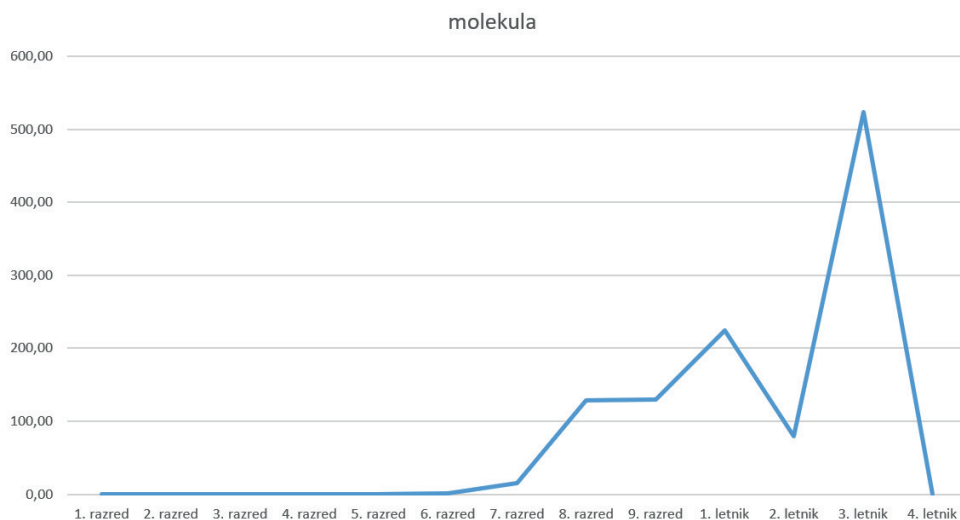
V zadnji triadi pogostost rabe samostalnika *stranica* naglo naraste, pri čemer doseže vrh v 7. razredu (Slika 3), kar sovpada z učnimi vsebinami pri matematiki, kjer je poudarek na razumevanju in uporabi geometrijskih pojmov. Po izrazitem porastu v 7. razredu opazimo upad v rabi v zadnjem delu tretje triade OŠ in na začetku SŠ (1. letnik) (Slika 3). Upad prevladujoče rabe v posameznih razredih ali letnikih je, sledeč učnemu načrtu, predvsem odraz preusmeritve poudarka vsebinskih sklopov in ciljev na druge matematične vsebine (npr. z geometrije na algebro ali aritmetiko).

Podoben trend lahko opazujemo na primeru samostalnika *molekula*. Raba samostalnika je vezana na področje naravoslovja, večinsko se pojavlja v učbenikih za kemijo, manj pogosto v učbenikih za biologijo ali fiziko (Preglednica 7).

Predmet	Absolutna pogostost	Relativna besedilna pogostost
Kemija	2.602	1.110,9
Biologija	709	298,7
Fizika	99	65
Naravoslovje	67	35,4
Matematika	4	0,7
Geografija	2	1,2
Tehnika in tehnologija	1	5
Slovenščina	1	0,1
Likovna umetnost	1	0,8

Preglednica 7: Relativna besedilna pogostost samostalnika *molekula* po šolskih predmetih v korpusu Učbeniki

Iz učbeniških podatkov je razvidno, da se samostalnik *molekula* v prvih petih razredih osnovne šole ne pojavlja, kar se sklada z dejstvom, da se terminologija, povezana s kemijo, v tem obdobju še ne uvaja (Slika 4). Prvič se beseda pojavi v 6. razredu z nizko frekvenco, nato pa ta v 7. razredu močno naraste in doseže vrh v 8. in 9. razredu (Slika 4), kar sovpada z učnim načrtom, saj se kemijski pojmi začnejo sistematično uvajati v 6. in 7. razredu v okviru predmeta naravoslovje, od 8. razreda naprej pa v okviru kemije kot samostojnega predmeta.

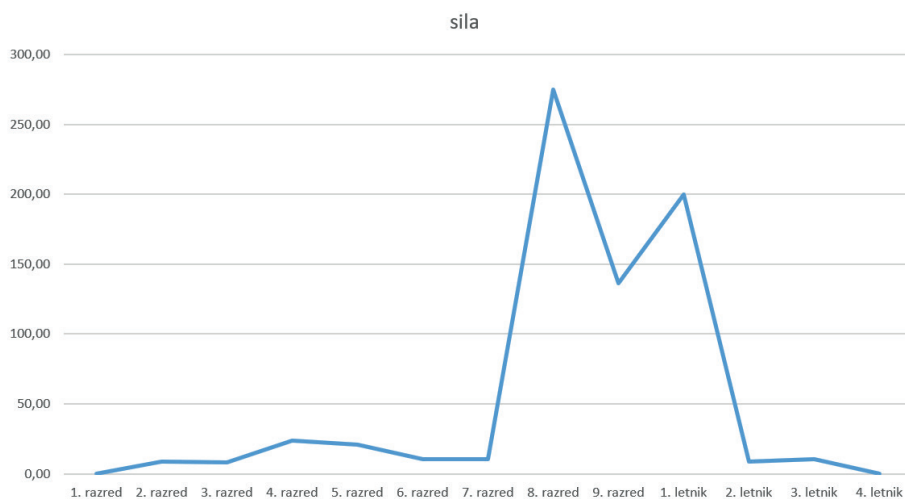


Slika 4: Prikaz relativne besedilne pogostosti samostalnika molekula po razredih (OŠ) in letnikih (SŠ) v korpusu Učbeniki

V srednji šoli frekvenca še dodatno naraste (Slika 4), kar kaže na intenzivnejšo obravnavo kemijskih vsebin, pri čemer je najvišja v 3. letniku, v 4. letniku pa sledi nagel upad in se beseda ne pojavlja več.

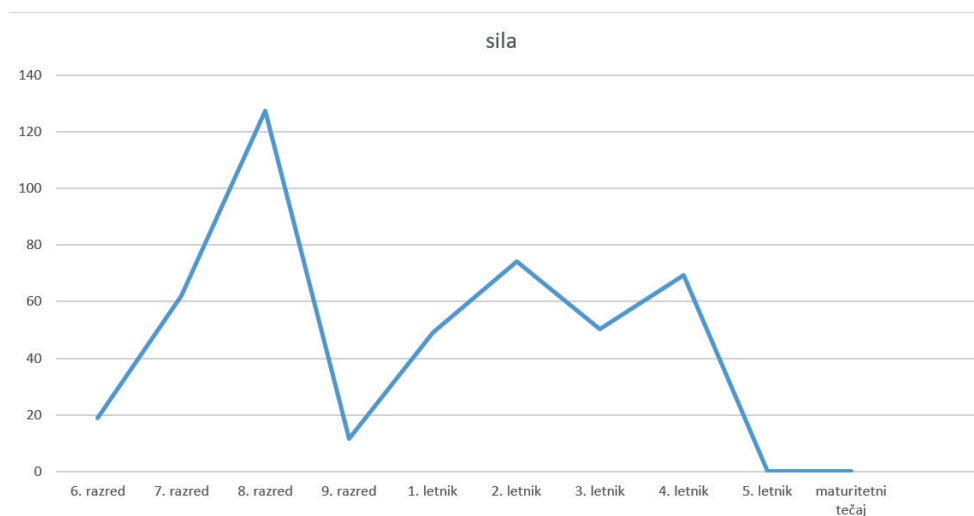
Primer 3

Na naslednjem primeru smo želeli preveriti še vidik spreminjanja rabe besede na ravni pomena skozi različne ravni in razrede/letnike šolanja. Raba samostalnika *sila* je v učbenikih pričakovano najpogostejša pri predmetu fizika, kjer izraz predstavlja enega izmed temeljnih pojmov. Beseda se v učbenikih pojavlja predvsem v višjih razredih osnovne šole in srednji šoli. Prva omemba vredna uporaba je v 6. razredu, medtem ko v 8. in 9. razredu OŠ ter 1. letniku SŠ opazimo izrazit porast pogostosti (Slika 5), kar sovпада z uvajanjem fizikalnih konceptov sile v učnih načrtih. Konkretni korpusni primeri (raba besede v širšem besedilnem okolju) pokažejo, da se pri prehodu v srednjo šolo (v 1. letniku) zgodi premik na ravni pomena – od konkretnega k abstraktnemu. Čeprav redkeje, se beseda ne uporablja več zgolj v osnovnem, fizikalnem pomenu, predvsem v učbenikih za slovenščino lahko opazimo, da se njen pomenski obseg (raz)širi in prehaja na raven figurativne oz. metaforične rabe (*ljubezen je najvišja sila; so krive sile znotraj njega; oblast in sila ipd.*), kar se sklada tudi z večjo zmožnostjo abstraktnega mišljenja pri starejših učencih oz. srednješolcih.



Slika 5: Prikaz relativne besedilne pogostosti samostalnika *sila* po razredih (OŠ) in letnikih (SŠ) v korpusu Učbeniki

Na drugi strani podatki v korpusu Šolar kažejo močan porast rabe v 8. razredu, kjer frekvenca doseže vrh, nato pa v 9. razredu občutno upade (Slika 6). Prehod v srednjo šolo spremlja manjša raba v 1. letniku, nato pa postopno naraščanje do 3. letnika, kjer se vrednosti stabilizirajo (Slika 6). Opazovanje besede v pisni produkciji in konkretnem besedilnem okolju pokaže, da porasti in upadi niso neposredno vezani na obravnavo fizikalnega pojma *sila* v skladu z učnim načrtom. Beseda se tako v OŠ kot SŠ v šolskih pisnih izdelkih, zlasti esejih in pisnih izdelkih pri slovenščini, sociologiji, filozofiji, državljski vzgoji in etiki, uporablja v drugih, nefizikalnih kontekstih in pomenih (*pa smo se ognili Bogu, višji sili; nadnaravna sila; delovna sila; pri tem Hasan uporabi silo; saj je zanj sila ničvredna; pride na oblast s silo; jo skuša na silo izničiti ipd.*).



Slika 6: Prikaz relativne besedilne pogostosti rabe samostalnika sila po razredih (OŠ) in letnikih (SŠ) v korpusu Šolar

Diskusija in zaključki

Primerjalne analize seznamov so razkrile, da nobeden od obstoječih seznam sam na sebi ne zadošča za izdelavo seznama temeljnega šolskega besedišča. Resda v šolskih besedilih najdemo veliko večino besed s seznama REFS in tako ta predstavlja dobro izhodišče, ki pa ga je treba nadgraditi z manjkajočimi (pogostimi) besedami iz (predvsem) učbenikov in korpusa Šolar. Primerjava med učbeniki in Šolarjem ter diahrona analiza rabe izbranih primerov sta pokazali, da se besedišče v šolskem okolju oblikuje v dveh smereh: recepcijsko (sprejemanje in branje učnih vsebin) in produkcijsko (aktivna uporaba pri govornem in pisnem izražanju). Učbeniki pogosto vključujejo specifično strokovno terminologijo, ki ostaja večinoma na receptivni ravni oz. je le redko vključena v pisno produkcijo. Da je določeno strokovno izrazje relevantno tudi za sezname šolskega besedišča, priča dejstvo, da ga najdemo tudi na seznamu S-UČB, ki vključuje besede iz učbenikov vsaj osmih različnih šolskih predmetov.

Kot se je pokazalo na izbranih primerih, pogostost uporabe posameznih besed odraža ciklične spremembe v učnem procesu: besede se uvajajo, v posameznih razredih ali letnikih raba doseže vrh, ko se vsebinski poudarki preusmerijo na nove koncepte, pa pogosto upade. Jezikovna kompetenca učencev je tako tesno povezana z razvojem predmetnospecifičnega znanja, pri čemer terminološka raba sledi zahtevam posameznih učnih ciljev in predmetnih področij.

Besedišče pa se skozi različna obdobja šolanja ne spreminja le po pogostosti rabe, temveč tudi glede na funkcijo in pomen, kar se najizraziteje kaže prav pri šolskem pisanju. Ista beseda se lahko skozi šolanje uporablja v različnih pomenih,

njena funkcija pa se spreminja – od receptivne rabe v navodilih do receptivne in produktivne rabe pri samostojnem tvorjenju besedil, npr. pri poročanju, obnavljanju ali argumentiranju. Različne analize, ki smo jih opravili za ta prispevek, zato predstavljajo dobro izhodišče za nadaljnje raziskave in kažejo na to, da je v zvezi s šolskim besediščem treba razmišljati dlje kot zgolj o izdelavi seznama besed, saj je nujno vključiti tudi pomenske informacije ter upoštevati tako receptivne kot produktivne vidike rabe. Smiselno je, da seznam služi pedagoški rabi: sledijo učnim ciljem posameznih predmetov in znanju, ki ga morajo učenci usvojiti, vključujejo primere rabe v kontekstu, omogočajo medpredmetne povezave ter spodbujajo aktivno delo učencev, ki novo besedišče samostojno uporabljajo pri lastni analizi, sintezi in vrednotenju. Poleg tega bi bilo smiselno seznam razširiti z večbesednimi enotami (stalne zveze, frazemi), kar je tudi ena pogostih kritik obstoječih seznamov besed. Kot vzorčni primer pri snovanju temeljnega šolskega besedišča bi lahko vzeli projekt English Vocabulary Profile in vsak pomen izbranih besed ali besednih zvez označili z vidika stopnje šolanja, na kateri bi ga učenci morali usvojiti ali že znati aktivno uporabljati.

Pri načrtovanju seznama temeljnega šolskega besedišča bo treba upoštevati tudi nekatere pomanjkljivosti obstoječih korpusov. Kot prvo sta korpusa Učbeniki 1.0 in Šolar 3.0 po obsegu dokaj majhna in bi ju bilo dobro povečati. Poleg tega bi bilo potrebno korpus učbenikov dopolniti z najnovejšimi učbeniki različnih predmetov, pa tudi povečati deleže posameznih predmetov, ki so trenutno slabše zastopani. Podobno bi bilo smiselno nadgraditi korpus Šolar z besedili različnih predmetov (trenutno dominira slovenski jezik), da bi lahko bolje preučili tudi produkcijo strokovnega izrazja. Koristno pa bi bilo razmisliti tudi o analizi drugih korpusov relevantnih besedil (ali njihovih delov), kot je npr. korpus Maks (MLAdinski KorpuS; Verdonik idr. 2022), ki med drugim vsebuje besedila iz mladinskega leposlovja.

Literatura in viri

- Arhar Holdt, Š., Pollak, S., Robnik Šikonja, M. in Krek, S. (2020). Referenčni seznam pogostih splošnih besed za slovenščino. V: D. Fišer in T. Erjavec (ur.). *Proceedings of the Conference on language technologies and digital humanities*. Ljubljana: Institute of Contemporary History, str. 10–15.
- Arhar Holdt, Š., Rozman, T., Stritar Kučuk, M., Krek, S., Krapš Vodopivec, I., Stabej, M., Pori, E., Goli, T., Lavrič, P., Laskowski, C., Kocjančič, P., Klemenc, B., Krsnik, L. in Kosem, I. (2022). *Developmental corpus Šolar 3.0* (Slovenian language resource repository CLARIN.SI). Dostopno na: <http://hdl.handle.net/11356/1589> (pridobljeno 3. 6. 2025).
- Arhar Holdt, Š. in Kosem, I. (2023). *Šolar, the developmental corpus of Slovene* (preprint). Dostopno na: <https://doi.org/10.21203/rs.3.rs-3274669/v1> (pridobljeno 3. 6. 2025).
- Arhar Holdt, Š. in Kosem, I. (2024). *Šolar, the developmental corpus of Slovene. Language Resources and Evaluation*. Dostopno na: <https://doi.org/10.1007/s10579-024-09758-4> (pridobljeno 3. 6. 2025).

- Brezina, V. in Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, 36, št. 1, str. 1–22.
- Browne, C. (2013). The new general service list: Celebrating 60 years of vocabulary learning. *The Language Teacher*, 34, št. 7, str. 13–15.
- Browne, C., Culligan, B. in Phillips, J. (2022). *New academic wordlist*. Dostopno na: <https://www.newgeneralservicelist.com/new-general-service-list-1> (pridobljeno 3. 6. 2025).
- Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Strasbourg: Council of Europe Publishing. Dostopno na: www.coe.int/lang-cefr (pridobljeno 3. 6. 2025).
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34, št. 2, str. 213–238.
- Fišer, D., Ljubešić, N. in Erjavec, T. (2020). The Janes project: Language resources and tools for Slovene user generated content. *Language Resources and Evaluation*, 54, str. 223–246.
- Gardner, D. in Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, 35, št. 3, str. 305–327.
- Kilgariff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J. B., Khalil, S., Kokkinakis, S. J., Lew, R., Sharoff, S., Vadlapudi, R. in Volodina, E. (2014). Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48, št. 1, str. 121–163.
- Klemen, M., Kosem, I., Arhar Holdt, Š., Pollak, S., Huber, D. in Lutar, M. (2022a). *Corpus of textbooks for learning Slovenian as L2 KUUS 1.0* (Slovenian language resource repository CLARIN.SI). Dostopno na: <http://hdl.handle.net/11356/1696> (pridobljeno 3. 6. 2025).
- Klemen, M., Arhar Holdt, Š. in Pollak, S. (2022b). *Core vocabulary for Slovenian as L2 1.0* (Slovenian language resource repository CLARIN.SI). Dostopno na: <http://hdl.handle.net/11356/1697> (pridobljeno 15. 6. 2025).
- Kosem, I. idr. (2019). *Developmental corpus Šolar 2.0*. (Slovenian language resource repository CLARIN.SI). Dostopno na: <http://hdl.handle.net/11356/1214> (pridobljeno 15. 6. 2025).
- Kosem, I., Pori, E. in Arhar Holdt, Š. (2019). *Keywords and n-grams from a textbook corpus* (Slovenian language resource repository CLARIN.SI). Dostopno na: <http://hdl.handle.net/11356/1215> (pridobljeno 15. 6. 2025).
- Kosem, I., Stritar Kučuk, M., Može, S., Zwitter Vitez, A., Arhar Holdt, Š. in Rozman, T. (2020). *Analiza jezikovnih težav učencev: korpusni pristop*. Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno na: <https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/229/329/5311-1> (pridobljeno 15. 6. 2025).
- Kosem, I., Pori, E., Žagar, A. in Arhar Holdt, Š. (2022). *Corpus of Slovenian textbooks ccUčbeniki 1.0* (Slovenian language resource repository CLARIN.SI). Dostopno na: <http://hdl.handle.net/11356/1693> (pridobljeno 15. 6. 2025).
- Kosem, I., Pori, E. in Arhar Holdt, Š. (2023). *Frequency list of textbook vocabulary by level of education in elementary and secondary schools* (Slovenian language resource repository CLARIN.SI). Dostopno na: <http://hdl.handle.net/11356/1719> (pridobljeno 15. 6. 2025).
- Košak Babuder, M., Mavrič Gavez, I., Ivanuš Grmek, M. in Torkar, G. (2021). Kako dobri so slovenski učbeniki? Kazalniki kakovosti in ocenjevanje učbenikov. *Sodobna pedagogika*, 72(138), št. 3, str. 26–44.
- Logar, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. in Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina in Fakulteta za družbene vede.

- Munda, T., Arhar Holdt, Š., Kosem, I., Pori, E. in Krek, S. (2025a). *Frequency list of collocations from the Učbeniki 1.0 corpus* (Slovenian language resource repository CLARIN.SI). Dostopno na: <http://hdl.handle.net/11356/2012> (pridobljeno 15. 6. 2025).
- Munda, Tina idr. (2025b). *Frequency lists of syntactic structures from the Šolar 3.0 corpus* (Slovenian language resource repository CLARIN.SI). Dostopno na: <http://hdl.handle.net/11356/2009> (pridobljeno 15. 6. 2025).
- Nation, P. in Waring, R. (1997). Vocabulary size, text coverage and word lists. V: N. Schmitt in M. McCarthy (ur.). *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press, str. 6–19.
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Pollak, S., Arhar Holdt, Š., Krek, S. in Robnik-Šikonja, M. (2020). *Reference list of Slovene frequent common words* (Slovenian language resource repository CLARIN.SI). Dostopno na: <http://hdl.handle.net/11356/1346> (pridobljeno 15. 6. 2025).
- Schmitt, N. in McCarthy, M. (1997). *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press.
- Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.
- Verdonik, D. in Zwitter Vitez, A. (2011). *Slovenski govorni korpus Gos*. Ljubljana: Trojina, zavod za uporabno slovenistiko.
- Verdonik, D., Majninger, S., Dobrovoljc, K., Antologa, Š., Zögling Markuš, A., Voršič, I., Zemljak Jontes, M., Koletnik, M., Valh Lopert, A., Šek, P., Kosem, I., Majhenič, S., Ferme, M., Žagar, A. in Arhar Holdt, Š. (2022). *Corpus of Slovenian texts for pedagogical purposes ccMAKS 1.0* (Slovenian language resource repository CLARIN.SI). Dostopno na: <http://hdl.handle.net/11356/1692> (pridobljeno 15. 6. 2025).
- West, M. (1953). *A general service list of English words*. London: Longman, Green and Co.

Iztok KOSEM (University of Ljubljana, Faculty of Arts and Faculty of Computer and Information Science, Slovenia)

Eva PORI (University of Ljubljana, Faculty of Arts, Slovenia)

FIRST STEPS TOWARDS A LIST OF BASIC SCHOOL VOCABULARY

Abstract: In this paper, we address the issue of creating a list of basic school vocabulary for the Slovene language that would be comparable to similar word lists we know for English and other languages. Such word lists not only serve to improve teaching materials and language tools for Slovene, but can also contribute to the upgrading of teaching materials for other subjects and to the development of teaching standards. In the introductory section, we first provide an overview of existing word lists for foreign languages, and then present the relevant existing word lists for Slovene, focusing on a reference list of general vocabulary and two lists of textbook vocabulary. This is followed by a detailed presentation of relevant current corpus resources for Slovene in the school environment. We then conduct an analysis of the existing Slovene vocabulary lists, comparing them with each other and with word lists from relevant corpora, examining consistencies and differences in coverage of key lexical units. It turns out that each list has certain shortcomings and that the difference between reception and production has to be taken into account. We also perform a diachronic analysis of a selected sample of words from the corpora of textbooks and school written products, which reveals changes in word usage not only at the temporal level but also at the semantic level. In conclusion, we summarise the main findings of the study and highlight suggestions for further research that could contribute to the production of a comprehensive and functional list of basic school vocabulary for the Slovene language.

Keywords: word lists, language use in schools, corpus Šolar, corpus of textbooks

Email for correspondence: iztok.kosem@ff.uni-lj.si

Priloga 1:

Lema	Besedna vrsta	Rel. frekvenca Šolar	Rel. frekvenca Učbeniki
biti	g	91431,89	60396,15
da	v	17003,38	6883,8
on	z	20190,69	10740,77
se	z	27041,32	17652,23
ta	z	12958,43	7195,68
jaz	z	6503,59	2277,56
pa	v	9425,64	5600
saj	v	4412,44	975,4
ko	v	5037,32	1816,47
zelo	r	3746,67	1038,84
svoj	z	5506,51	3020,09
ker	v	3240,26	1181,77
človek	s	3671,18	1680,51
ljubezen	s	2111,07	141,53
ne	l	6481,57	4526,53
oče	s	2198,09	308,4
življenje	s	2601,75	993,53
tudi	l	6509,88	4904,65
družina	s	1900,86	335,36
ves	z	4645,72	3111,89
njegov	z	2697,68	1189,67
vendar	v	1934,93	434,59
otrok	s	2031,91	604,25
moj	z	1819,60	426,46
hoteti	g	1521,31	193,36
imeti	g	5874,51	4583
pri	g	1857,34	575,66
sam	p	1961,67	686,06
misliti	g	1454,74	236,82

do	d	3133,84	1962,88
veliko	r	1762,98	641,9
dober	p	1684,87	597,74
mati	s	1235,08	167,56
starš	s	1307,43	257,04
želeti	g	1550,15	511,52
šola	s	1697,98	676,53
tako	r	3393,86	2381,91
še	l	3378,13	2411,42
nek	z	1308,48	350,93
videti	g	1603,62	657,7
iti	g	1597,85	681,18
sin	s	1046,89	135,49
ampak	v	1135,48	228,92
oditi	g	981,88	110,39
kar	z	1626,68	767,16
morati	g	2198,09	1339,11
lep	p	1129,19	297,48
prijatelj	s	1085,15	267,73
vedno	r	1351,46	556,14
stvar	s	943,61	166,63
njen	z	1484,09	726,73
odlomek	s	1148,59	392,3
zgodba	s	1064,19	321,88
govoriti	g	1308,48	581,71
rad	r	998,66	284,46
samo	l	1304,81	590,77
dan	s	1693,26	998,64
odločiti	g	864,98	179,18
zaradi	d	1681,73	1016,07
mama	s	912,16	252,39
čas	s	2091,15	1432,54
ženska	s	805,22	159,43

Črtomir	s	644,28	8,13
kot	v	4182,83	3548,8
že	l	1955,38	1323,07
zato	v	1787,62	1162,02
Romeo	s	649,52	32,54
zdeti	g	888,04	272,14
družba	s	804,17	190,57
bolj	r	1276,50	674,2
začeti	g	1311,10	731,61
umreti	g	688,31	115,74
smrt	s	732,35	161,29
konec	s	1017,53	451,1
narediti	g	1026,44	461,09
odnos	s	791,59	237,05
mnenje	s	826,19	284,23
vera	s	615,97	79,48
žena	s	635,37	105,05
slab	p	657,91	129,68
povedati	g	1200,49	676,06
živeti	g	1264,44	743,23
moči	g	999,71	480,61
reči	g	1059,47	540,57
problem	s	630,12	112,25
brat	s	664,20	153,39
mož	s	653,19	147,34
poročiti	g	514,79	33,47
delati	g	708,76	237,28
takrat	r	683,07	226,59
čeprav	v	695,65	239,38
ter	v	1752,50	1303,09
nikoli	r	591,85	142,93
vedeti	g	1326,82	878,49
tisti	z	1062,61	615,87

Antigona	s	455,56	9,3
storiti	g	555,68	109,46
le	l	1677,01	1243,36
moški	s	538,91	107,14
ubiti	g	461,32	31,14

Preglednica 1: Seznam prvih sto lem z izrazito višjo relativno pogostostjo rabe v korpusu Šolar kot v korpusu Učbeniki

Lema	Besedna vrsta	Rel. frekvenca Šolar	Rel. frekvenca Učbeniki
v	d	19760,83	27346,71
z	d	9638,48	15792,3
ali	v	1857,34	6022,97
na	d	10736,22	14762,05
kaj	r	88,07	3885,79
kateri	z	1872,55	5437,78
število	s	51,37	2853,22
a	s	82,83	2508,1
kaj	z	1603,62	3885,79
voda	s	243,77	2414,91
iz	d	2435,05	4388,01
besedilo	s	592,38	2272,68
pri	d	2350,12	4013,38
beseda	s	644,80	2197,61
slika	s	111,66	1586,62
snov	s	67,10	1521,55
zapisati	g	34,60	1451,83
napačno	r	27,26	1408,13
oblika	s	131,06	1436,49
različen	p	461,32	1709,1
telo	s	120,05	1339,81
velik	p	1431,15	2635,23
koliko	r	115,33	1287,52

med	d	1898,23	3068,43
izračunati	g	2,62	1157,84
od	d	2457,59	3604,35
stran	r	39,32	1171,08
ki	v	7917,44	9017,96
točka	s	72,34	1116,7
del	s	619,64	1655,18
odgovor	s	185,05	1220,58
podatek	s	68,67	1096,25
držati	g	181,91	1205,94
vrsta	s	214,41	1237,09
zvezek	s	16,25	1038,38
enak	z	206,55	1222,44
lahko	r	3750,86	4673,87
npr.	o	379,02	1294,72
pravilen	p	90,17	994,46
a	v	1621,97	2508,1
enačba	s	1,05	884,76
vrednost	s	22,54	902,19
dva	k	830,38	1702,36
celica	s	11,53	881,51
rastlina	s	34,60	904,05
pravilno	r	88,59	953,78
za	d	6991,12	7835,72
imenovati	g	112,71	949,83
C	s	61,33	888,25
njihov	p	1,57	825,73
zakaj	r	394,74	1215,7
izraz	s	45,61	850,83
sila	r	1,05	803,19
poved	s	223,32	1022,58
energija	s	48,75	833,4
en	k	1149,63	1928,25

majhen	p	361,19	1129,95
Nato	s	5,24	770,42
atom	s	1,05	756,94
naš	p	0,52	753,22
prebrati	g	191,87	942,16
nastati	g	215,46	964,01
skupina	s	169,85	917,3
sila	s	60,29	803,19
svet	p	108,52	848,04
njen	p	1,05	726,73
narisati	g	9,96	729,28
enota	s	21,49	727,66
dopolniti	g	23,07	715,8
prvi	k	885,95	1564,54
žival	s	176,67	855,25
uporabljati	g	305,63	982,84
trikotnik	s	15,73	681,18
dolžina	s	7,34	664,44
ogledati	g	144,16	793,66
pod	s	2,62	647,71
izbrati	g	252,15	894,06
spodnji	p	11,01	643,99
barva	s	89,12	716,5
cm	s	19,92	645,85
zato	r	547,82	1162,02
več	l	756,99	1369,56
slovenski	p	628,55	1236,62
ustrezen	p	23,59	631,44
zrak	s	62,38	661,65
primer	s	787,39	1385,59
trije	k	325,02	917,76
količina	s	36,17	627,95
temperatura	s	13,11	604,71

Slovenija	s	286,75	875,46
poiskati	g	115,33	699,77
organizem	s	8,39	589,84
ime	s	411,00	992,36
po	d	3690,05	4265,31
dolg	s	14,68	589,14
premica	s	0,52	568,46
element	s	66,05	630,51
stran	s	617,54	1171,08
glasba	s	72,87	619,36
uporabiti	g	112,19	653,05

Preglednica 2: Seznam prvih sto lem z izrazito višjo relativno pogostostjo rabe v korpusu Učbeniki kot v korpusu Šolar