

Naslov članka/Article:

VIRI IN ORODJA CJVT: PODPORA SLOVENŠČINI V DIGITALNEM OKOLJU

The Centre for Language Resources and Technologies: Resources and Tools to Support Slovenian in Digital Environment

Avtorja/Authors:

Dr. Špela Arhar Holdt in **dr. Iztok Kosem**

CC licenca



Priznanje avtorstva-Nekomercialno-Brez predelav



Vzgoja in izobraževanje št. 6/2024, letnik 55

ISSN 0350 -5065

Izdal in založil: Zavod Republike Slovenije za šolstvo

Kraj in leto izdaje: Ljubljana, 2024

Spletna stran revije:

<https://www.zrss.si/strokovne-revije/vzgoja-in-izobrazevanje/>

Dr. Špela Arhar Holdt in dr. Iztok Kosem

Center za jezikovne vire in tehnologije Univerze v Ljubljani (CJVT)

VIRI IN ORODJA CJVT: PODPORA SLOVENŠČINI V DIGITALNEM OKOLJU

The Centre for Language Resources and Technologies: Resources and Tools to Support Slovenian in Digital Environment

CENTER ZA JEZIKOVNE VIRE IN TEHNOLOGIJE UNIVERZE V LJUBLJANI

Center za jezikovne vire in tehnologije Univerze v Ljubljani (CJVT UL) je posebna enota Univerze v Ljubljani, ki se ukvarja z raziskovanjem in razvojem digitalne infrastrukture za sodobno slovenščino. Prostor CJVT se nahajajo na Fakulteti za računalništvo in informatiko, pri delu pa sodelujejo tudi strokovnjaki in strokovnjakinje s Filozofske fakultete, Fakultete za družbene vede, Fakultete za elektrotehniko, Pedagoške fakultete in Fakultete za upravo.

CJVT je bil ustanovljen v podporo razvoja slovenščine v digitalnem svetu. To vključuje pripravo digitalnih jezikovnih virov, kot so besedilni korpusi, slovarske baze, jezikovni modeli, vmesnikov za dostop do teh podatkov in jezikovnotehnoloških orodij. Temeljna naloga CJVT je tudi zagotavljanje odprtega dostopa do jezikovnih podatkov, zato so rezultati na voljo vsem, ki jih potrebujejo za raziskave ali razvoj novih izdelkov.

Delo na CJVT je večinoma raziskovalne narave in vključuje razvoj inovativnih rešitev. Med temi so programi, ki pospešujejo in olajšajo pripravo jezikovnih podatkov, ter vmesniki, ki omogočajo sodelovanje širše javnosti pri razvoju izdelkov. Velik poudarek je na uporabniških raziskavah, ki omogočajo boljše prilagajanje izdelkov potrebam ljudi.

Jezikovni viri in tehnologije, ki jih razvija CJVT, so koristni tudi za vzgojo in izobraževanje. Omogočajo opazovanje avtentične slovenščine, prinašajo sodobne jezikovne podatke in orodja ter spodbujajo vrednote odprte znanosti in skrb za razvoj slovenščine v digitalnem okolju. Paleta

izdelkov CJVT ponuja pregled nad gradniki digitalne jezikovne infrastrukture, njihovim namenom, prednostmi in omejitvami. Spoznavanje teh virov in tehnologij razvija digitalno pismenost in strategije za reševanje jezikovnih dilem.

V tem prispevku predstavljamo pomembnejše izdelke CJVT, ki so brezplačno na voljo na spletu. Na spletnih straneh so na tudi voljo informacije o avtorjih in znanstveni članki, ki podrobno opisujejo pripravo posameznih virov in tehnologij.

1. BESEDILNI KORPUSI

Besedilni korpusi so obsežne digitalne zbirke realnih besedil, ki so namenjene raziskovanju jezika. Besedila so skrbno izbrana, da predstavljajo vzorec določene jezikovne rabe, npr. splošni pisni jezik, šolsko pisanje, govorjena slovenščina. Besedila vključujejo različne informacije o nastanku in izvoru ter jezikoslovne informacije, kot so osnovna oblika, besedna vrsta in druge značilnosti besed.

Korpusi se uporabljajo za jezikoslovne raziskave, pripravo slovarjev in slovnice ter za razvoj jezikovnih tehnologij. Med uporabniki so tudi lektorji, prevajalci, učitelji, učenci osnovnih in srednjih šol, tisti, ki se učijo slovenščine kot drugega ali tujega jezika, in drugi podobni uporabniki. Raba korpusov v vmesnikih CJVT je zelo preprosta, saj spominja na spletne iskalnike: besedo ali besedno zvezo, ki nas zanima, vtiskamo v iskalno okence, nato program poišče in prikaže vsa besedila, kjer se beseda pojavlja, ter ponudi različne informacije, ki pomagajo pri raziskovanju rezultatov.

1.1

Gigafida 2.0 (<https://viri.cjvt.si/gigafida/>) zajema vzorec pisne standardne slovenščine. Vključuje dnevne časopise, revije, izbrane spletne vire in knjižne publikacije različnih vrst (leposlovje, učbeniki, stvarna literatura). Vsebuje 1.134.693.933 besed iz 38.310 besedil, ki so nastala med letoma 1990 in 2018. Novejša besedila, zlasti z novičarskih portalov, se zbirajo v spremljevalni korpus **Trendi**, s pomočjo katerega delujejo tudi nekatera orodja, ki jih predstavljamo v nadaljevanju.

Zanimivost iz korpusa Gigafida: najpogostejša beseda v slovenskih pisnih besedilih je glagol *biti*, sledita predlog *v* in veznik *in*. Med najpogostejšimi glagoli so še *imeti*, *morati*, *iti*, med samostalniki *leto*, *čas*, *dan*, *Slovenija*, med pridevniki pa *velik*, *nov*, *slovenski*, *dober*. V korpusu lahko raziskujemo, kako se besede in zveze uporabljajo v realnih besedilih, zato je tudi bogat vir raznolikih in avtentičnih jezikovnih zgledov.

1.2 GOS 2.0 (<https://viri.cjvt.si/gos/>) vključuje transkripcije približno 320 ur posnetkov pogovorov v različnih vsakdanjih situacijah, kot so radijske in televizijske oddaje, šolske ure, predavanja, pogovori med prijatelji, družinski pogovori, delovni sestanki, svetovanja in podobno. Posnetki so v različnih slovenskih regijah. Govor je transkribiran, kar omogoča iskanje po korpusu prek spletnega vmesnika, za vsak izpis pa je mogoče tudi poslušati pripadajoči del posnetka.

Korpus GOS vsebuje posnetke govorcev iz različnih narečnih skupin in s tem ponuja možnost zabavnega interaktivnega raziskovanja slovenskih narečij, razlik med njimi ter standardnim govorjenim jezikom. Tako lahko npr. poslušamo, kako v nekaterih primorskih narečjih zveza *ne upam* pomeni *ne morem* ali pa opazujemo, kakšne različice za besedo *kako* se pojavljajo v različnih narečjih (*kok, kuk, koko, keku* ...).

1.3 Šolar 3.0 (<https://solar.cjvt.si/>) je korpus, ki vključuje besedila, ki so jih učenci slovenskih osnovnih in srednjih šol napisali pri pouku. Vsebuje 5.485 besedil, ki so jih ustvarili dijaki srednjih šol in učenci od 7. do 9. razreda osnovne šole, nekaj malega tudi iz 6. razreda. Za vsako besedilo so navedeni podatki o šoli, predmetu, stopnji (razred ali letnik), vrsti besedila, regiji in letu nastanka. Nekatera korpusna besedila vsebujejo tudi učiteljske popravke.

V korpusu Šolar je mogoče raziskovati jezikovne popravke glede na vrsto jezikovnega problema, na primer različne težave s črkovanjem, oblikoslovjem, skladnjo itd. S pomočjo korpusa lahko denimo najdemo najpogostejše napačno črkovane besede v šolskih spisih, kot so *življenski*, *bol*, *use*, *vrjeti*, *mišljenje*, *bljižen*.

1.4 Korpusnik (<https://korpusnik.cjvt.si/>) je orodje, ki na enem mestu združuje podatke iz petih slovenskih korpusov. Poleg že omenjenih Gigafide, Trendov in GOS-a vključuje tudi korpus **OSS 1.0**, ki zajema znanstvena besedila, in **JANES 1.0**, ki obsega besedila s slovenskih družbenih omrežij. V zavihku *Poudarki* uporabniki najprej vidijo povzetek glavnih značilnosti iskane besede, na primer v katerem korpusu se najpogostejše pojavlja in s katerimi drugimi besedami tipično nastopa. Vsak korpus ima tudi svoj zavihek, kjer so prikazani grafi in primeri rabe.

S Korpusnikom lahko pridobimo zanimive statistične podatke o besedah, npr. da se je beseda *kajtar* v v korpusnih virih prvič pojavila leta 2002, da beseda *majica* prevladuje v besedilih s športno tematiko, da je besedi *računalnik* v obdobju 1991–2024 raba upadla za 72 %.

The screenshot displays the Šolar 3.0 web application interface. At the top, there is a navigation bar with the logo 'cjvt šolar 3.0', search and filter buttons, and a language dropdown set to 'Slovenščina'. Below the navigation bar, a search bar contains the text '(*)' and a note 'Iskanje po izvornih besedilih / velikost korpusa: 28.953 pojavnic'. A filter bar shows 'Besedila s popravki' and 'Vokali/Odvečni vokali + 26'. The main content area is divided into two panels. The left panel, titled 'Filtri', contains filters for 'Osnovne oblike' (with a list of words and their counts) and 'Tip popravka' (with a list of correction types and counts). The right panel displays search results for the word 'mišljenje', showing various contexts where it appears, such as 'on prestajal vojno. Ko se podaš v vojno, nesmeš imeti lastnih vrednot in lastnega mišljenja, ampak moraš misliti in vrednotiti vojno kot partizani, ki pobijajo tudi ne'.

► SLIKA 1: Korpus Šolar

Sopomenke				
1 za službo ali funkcijo				
prosilec	:	aspirant	:	prijavljenec
prijavljeni	↻		:	interesent ekonomija
2 na izpitu				
prijavljenec	:	prijavljeni	↻	udeleženeec
3 za nagrado				
predlaganec	↻	nominiranec	:	

► SLIKA 2: Slovar sopomenk sodobne slovenščine

2. LEKSIKON IN SLOVARJI

Med pomembnimi izdelki CJVT so oblikoslovni leksikon Sloleks ter Slovar sopomenk sodobne slovenščine in Kolo kacijski slovar sodobne slovenščine. Gre za jezikovne vire, ki prinašajo informacije o sodobni slovenščini, in sicer: kako se besede pregibajo (*agent, agenta, agentu ... agenti / agentje* itd.), katere sopomenke imajo (npr. *agent – informator; posrednik*) in s katerimi besedami se v jeziku tipično pojavljajo – t. i. kolokacije (npr. *tajni, obveščevalni, skrivni, nepremičninski, zavarovalni, potovalni agent*).

Pomembna in za izobraževanje zanimiva lastnost slovarskih virov CJVT je njihova "odzivnost": uporabnicam in uporabnikom omogočajo aktivno sodelovanje pri razvoju vsebine, npr. snemanje posnetkov izgovora besed v leksikonu, dodajanje novih sopomenk in protipomenk, pregledovanje in urejanje kolokacij. Sodelovanje po principu občanske znanosti spodbuja demokratično vključevanje širše javnosti v razvoj slovarjev in hitro posodabljanje slovarskih podatkov, ki so nato odprto na voljo celotni skupnosti.

2.1

Sloleks 2.0 (<https://viri.cjvt.si/sloleks/>) je oblikoslovni leksikon in vsebuje osnovne podatke o 100.802 slovenskih besedah, predvsem v katero besedno vrsto spadajo, kakšne so njihove slovnične lastnosti in kako se pregibajo. V različici 2.0 so pri vsaki besedi zabeležene vse njene pregibne oblike z morebitnimi variantami, matko pripisanimi naglasi, zapisi v mednarodni fonetični pisavi (IPA) in posnetki izgovarjave, ki so bili generirani s pomočjo sistema eBralec.

V Sloleksu je mogoče preveriti redke oblike, npr. roditeljski dvojniki samostalnika *dno* (*dveh dan* ali *dveh dnov*). Pomagajo tudi podatki o frekvenci besedne oblike v korpusu Gigafida, ki denimo povedo, da je oblika *agenti* v jeziku izrazito pogostejša od *agentje*.

2.2

Slovar sopomenk sodobne slovenščine 2.0 (<https://viri.cjvt.si/sopomenke/>) vsebuje 100.837 iztočnic in 362.828 sopomenk in je najobsežnejša odprto dostopna zbirka sopomenk za slovenščino. Slovar je bil v prvem koraku pripravljenih povsem strojno z naprednimi metodami, sedaj pa se postopoma ročno pregleduje in dopolnjuje, pomočjo skupnosti. Slovar vsebuje tudi protipomenke, ki so za zdaj na voljo pri skoraj 3.600 iztočnicah.

Največ sopomenk v Slovarju sopomenk sodobne slovenščine ima beseda *hud*, in sicer več kot sto. Beseda *hud* ima več pomenov, vsi pa so sopomensko bogati, npr. *grozen*, *pretresljiv*, *travmatičen*; *ogorčen*, *jezen*, *strog*; *mega*, *nor*, *carski* ...

2.3 Kolokacijski slovar sodobne slovenščine 2.0 (<https://viri.cjvt.si/kolokacije/>) vsebuje 81.443 iztočnic in 4.491.958 kolokacij. Tudi ta slovar nastaja z naprednimi računalniškimi metodami, ki se redno izboljšujejo in nadgrajujejo. V slovarju je mogoče opazovati, katere besede tipično nastopajo skupaj, kar lahko pomaga pri pisanju, pripravi učnega gradiva in podobno. Tudi Kolokacijski slovar podpira sodelovanje skupnosti.

Kolokacijski slovar je uporaben tudi za primerjavo dveh sorodnih besed, saj besedna okolica pregledno prikaže razlike v njuni rabi. Primerjava pridevnikov *kakovosten* in *kvaliteten* npr. razkrije, da se s *kakovosten* pojavljajo predvsem samostalniki *razred*, *raven*, *razlika*, *boben*, *lestvica*, s *kvaliteten* pa *nivo*, *nasprotnik*, *tujec*, *arhitektura*, *postrežba*, medtem ko se z obema pridevnikoma pojavljajo *izdelek*, *življenje*, *ponudba*, *ekipa* in *hrana*.

3. JEZIKOVNA ORODJA IN TEHNOLOGIJE

Na CJVT nastajajo tudi številna jezikovna orodja in tehnologije. Na eni strani se na slovenščino prenašajo in preizkušajo metode, ki so na voljo za druge jezike, razvijajo pa se tudi povsem inovativne rešitve. Med izdelki, ki so zanimivi za široko rabo, v tem prispevku omenjamo Jezikovni sledilnik in orodje za avtomatsko postavljanje vejic.

3.1 Jezikovni sledilnik (<https://viri.cjvt.si/sledilnik/slv>)

ponuja vpogled v časovne trende slovenskega besedišča. Orodje prikazuje različne sezname besed, npr. tistih, ki so zaznamovale predhodni mesec ali leto, ter novih besed določenega obdobja. Zanimiva je tudi možnost primerjave rabe izbranih sorodnih besed (npr. rabe besed *neurje* in *ujma*). Vsaka beseda je opremljena z grafičnim prikazom rabe v izbranem časovnem obdobju, omogočeno pa je tudi primerjanje časovnih trendov več besed.

Podrobnejšo analizo besed meseca z zanimivimi podatki iz različnih naših virov objavljamo tudi v mesečnem novičniku.

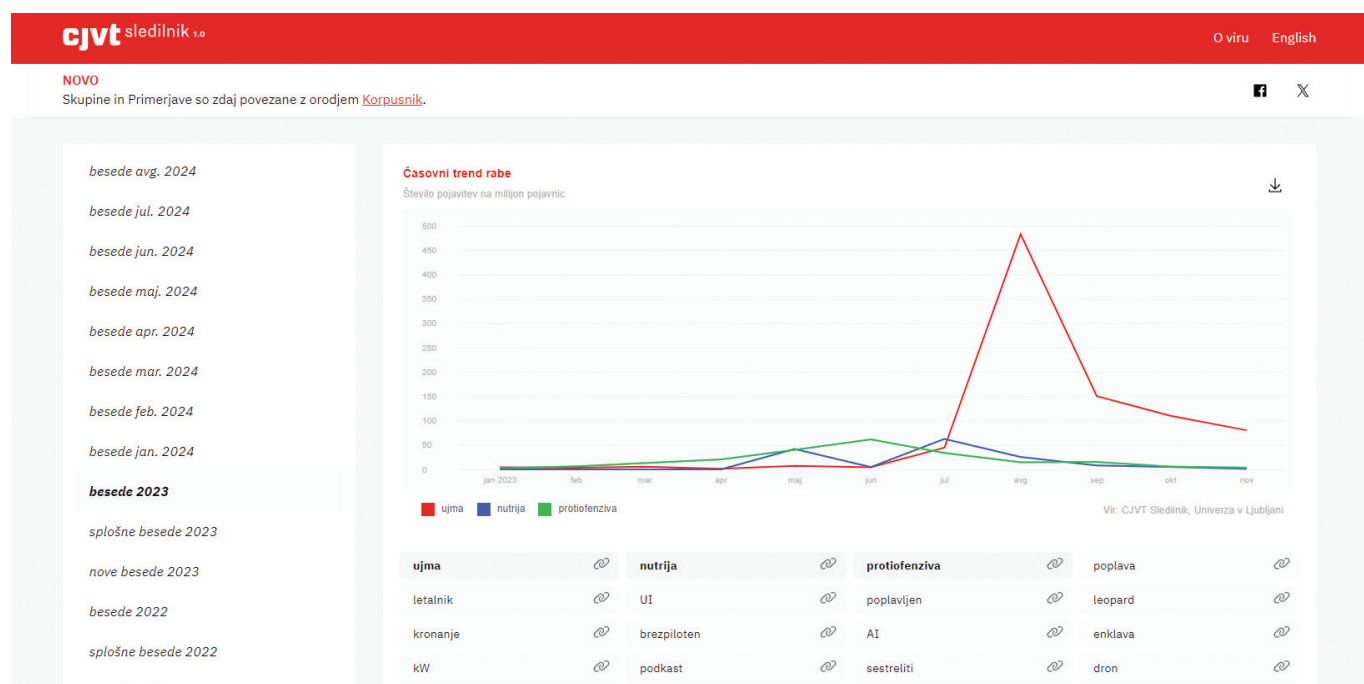
3.2 Vejice (<https://orodja.cjvt.si/vejice/>) so spletno orodje, ki na temelju metode nevronske mreže preverja postavitev vejic v slovenskih besedilih. Orodje, ki je po pričakovanjih naletelo na velik in spodbuden odziv širše javnosti, opozarja na manjkajoče vejice s sivo barvo in na odvečne vejice z modro barvo. Glede na teste program trenutno deluje uspešno v 94 odstotkih primerov.

KAJ SE ŠE PRIPRAVLJA?

Poleg rednega posodabljanja vsebin v Korpusniku in Jezikovnem sledilniku se v prihajajočem letu obetajo posodobitve Kolokacijskega slovarja, Slovarja sopomenk in Vejic, objavljene pa bodo tudi nove jezikovne igre. Če bi želeli prejemati mesečni novičnik in druga obvestila o aktivnostih CJVT, se prijavite na novice na spletni strani ali pa CVJT spremljajte na družbenih omrežjih Facebook ali X.



QR KODA spletne strani CJVT



► SLIKA 3: Jezikovni sledilnik