trojína

CENTER ZA UPORABNO JEZIKOSLOVJE

Šolar 2.0

http://solar.trojina.si/

nadgradnja korpusa šolskih pisnih izdelkov

IZTOK KOSEM

TADEJA ROZMAN

ŠPELA ARHAR HOLDT

POLONCA KOCJANČIČ

CYPRIAN LASKOWSKI

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

trojína

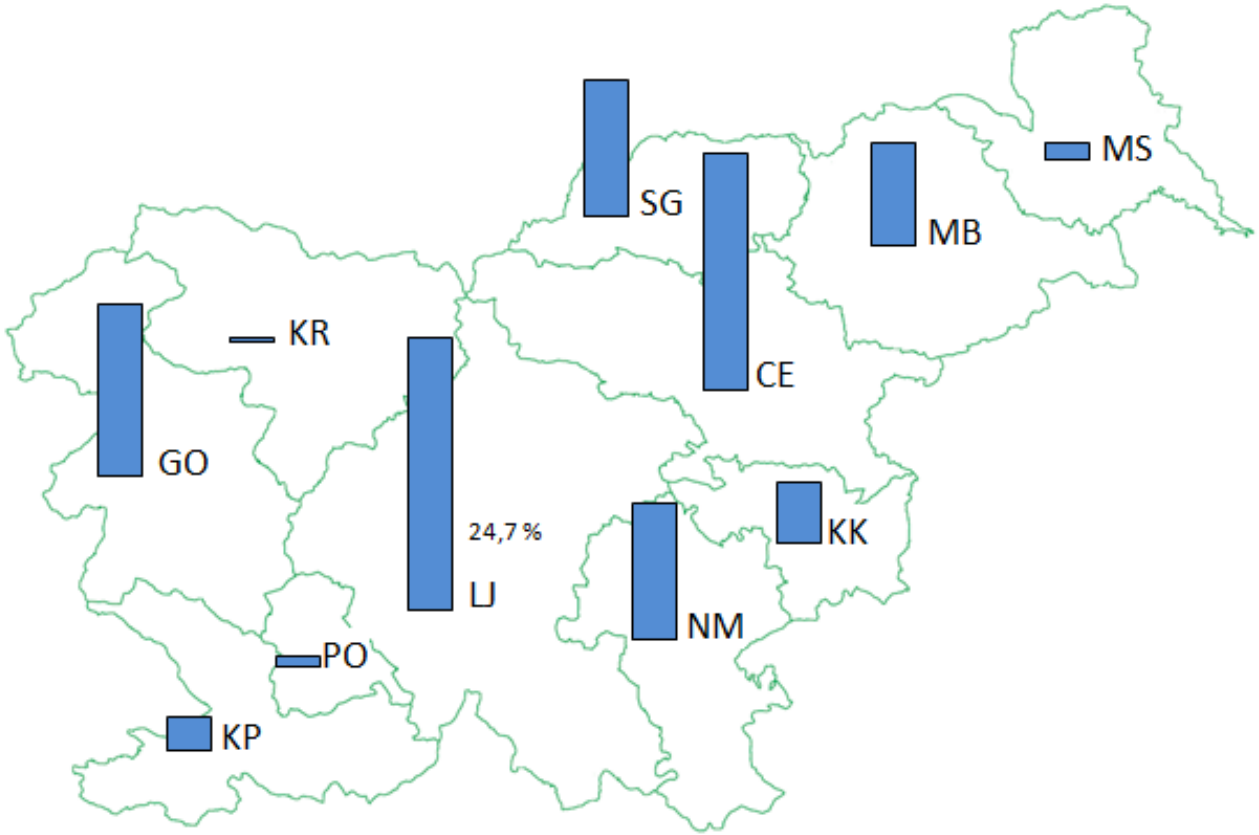
Šolar (1.0)

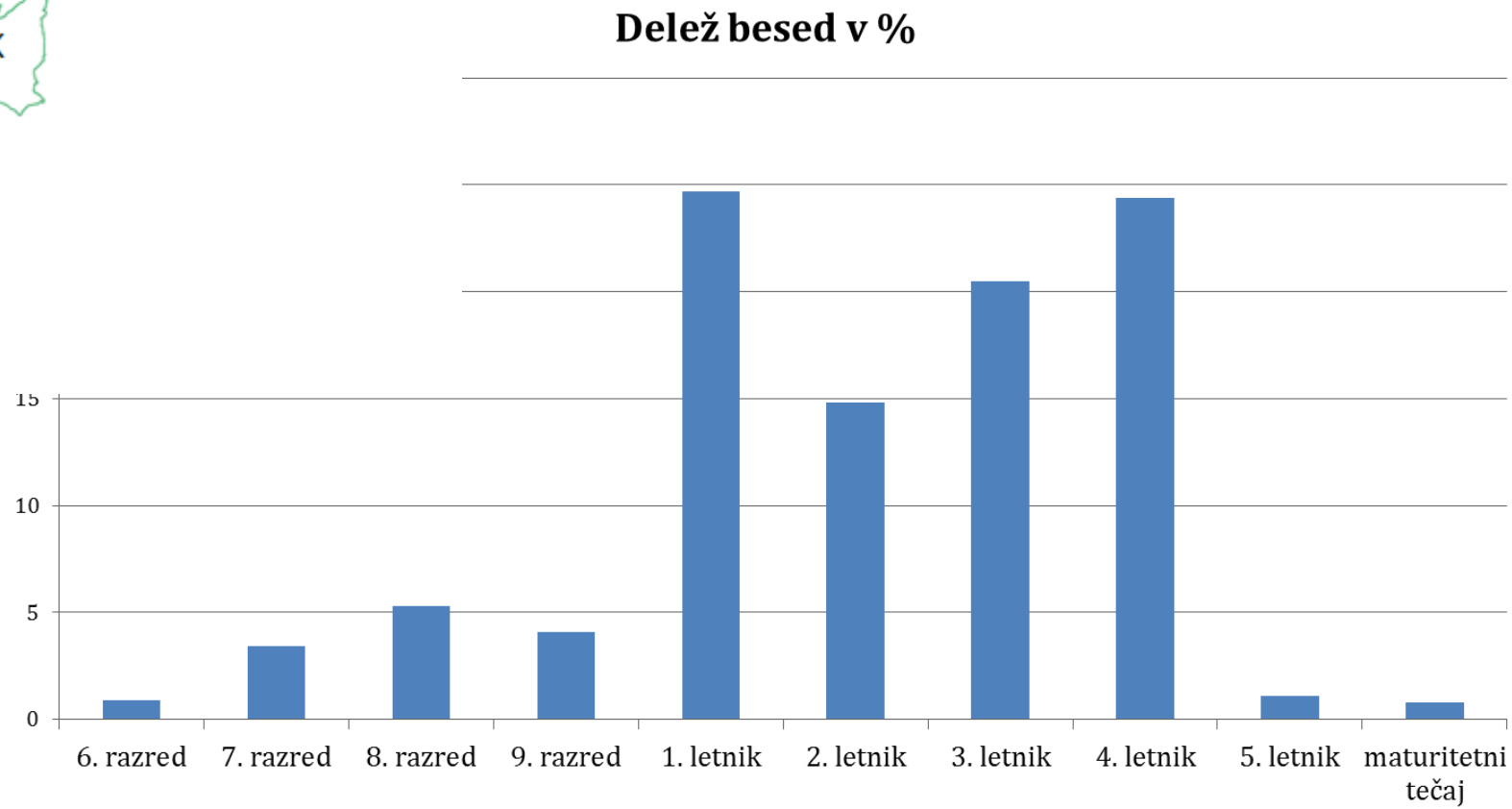
- Nastal v okviru projekta Sporazumevanje v slovenskem jeziku
- Šolsko leto 2009/2010
- 2703 besedil (od 8594)
- 967.477 besed
- Pisna besedila, samostojno tvorjena pri pouku (zadnje triletje OŠ in srednja šola)
- 39 šol (60 % JZ Slovenije, 40 % SV)
- 56 % besedil vsebuje jezikovne popravke učiteljev
- Dostopen na: http://www.korpus-solar.net/ (+repozitorij CLARIN.SI)
- Licenca: CC BY-NC-SA 2.5 SI (Creative Commons Priznanje avtorstva-Nekomercialno-Deljenje pod enakimi pogoji 2.5 Slovenija)

trojína

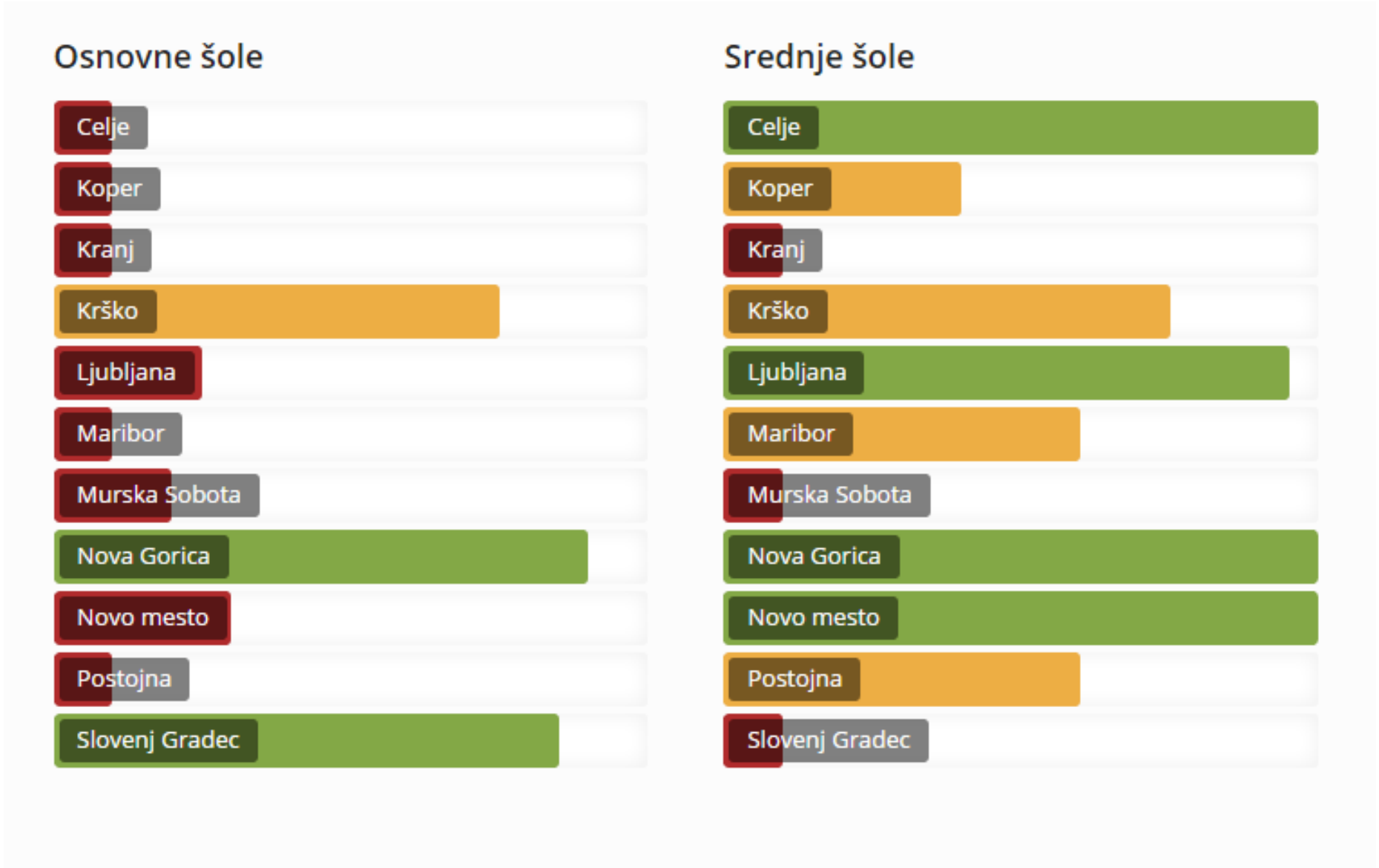
Nadgradnja korpusa Šolar

- projekt pri Ministrstvu za kulturo (2015-2018):
  - Z načrtnim dodajanjem novo zbranih (in deloma že zbranih besedil) izboljšati regijsko uravnoteženost korpusa Šolar in uravnoteženost po stopnji šolanja.
  - Vnesti informacijo o podkategoriji napake učenca v obstoječih besedilih in na podlagi teh informacij izdelati učni korpus za jezikovnotehnološke namene.
  - Izdelati podkorpus besedil šolarjev dislektikov kot primer korpusa besedil učencev s posebnimi potrebami. Podkorpus naj bi kasneje služil kot model za izdelavo vseh nadaljnjih podobnih podkorpusov.


SG
MS
MB
KR
CE
GO
24,7 %
LJ
KK
NM
PO
KP

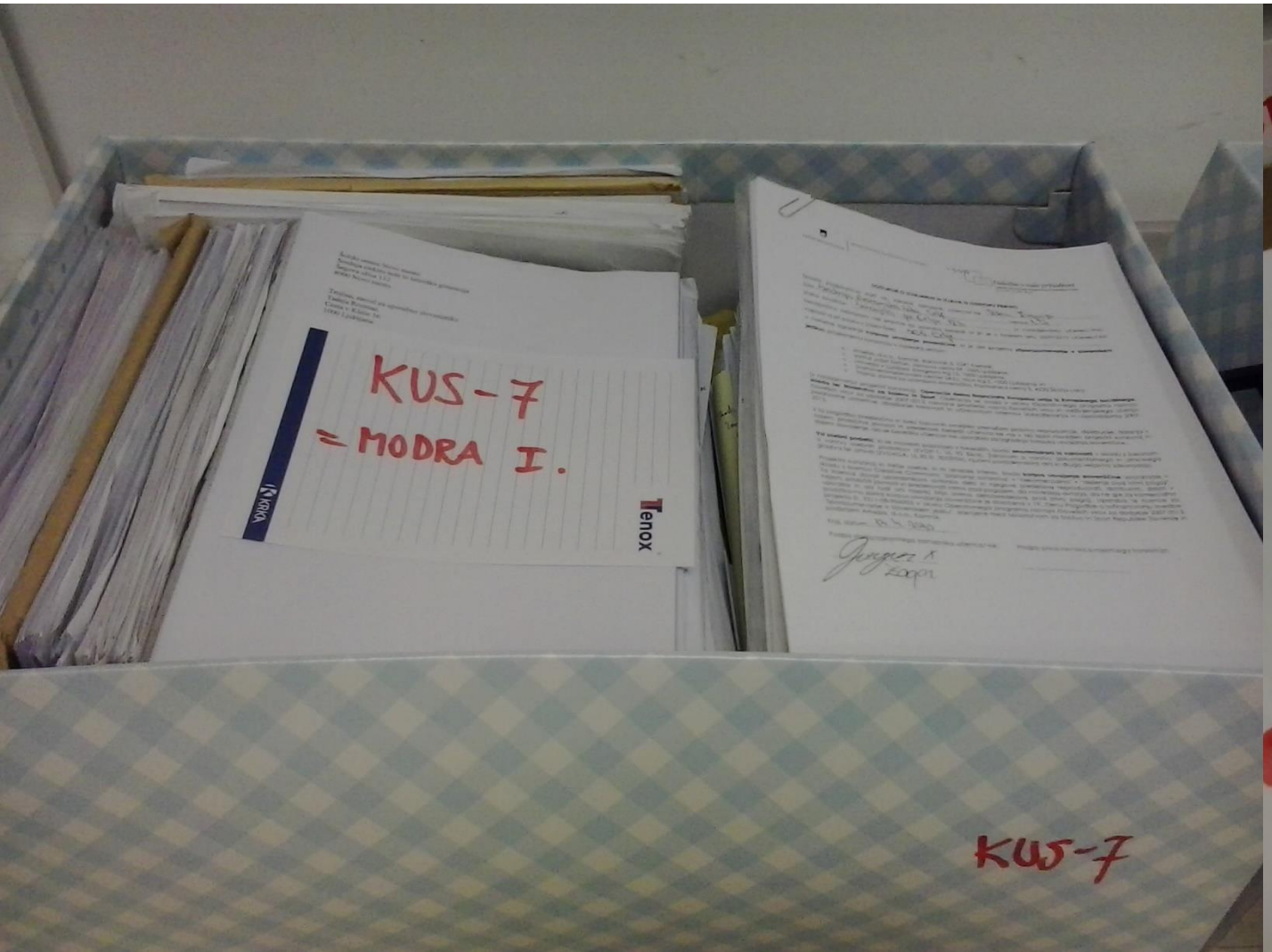

Delež besed v %
15
10
5
0
6. razred
7. razred
8. razred
9. razred
1. letnik
2. letnik
3. letnik
4. letnik
5. letnik
maturitetni tečaj

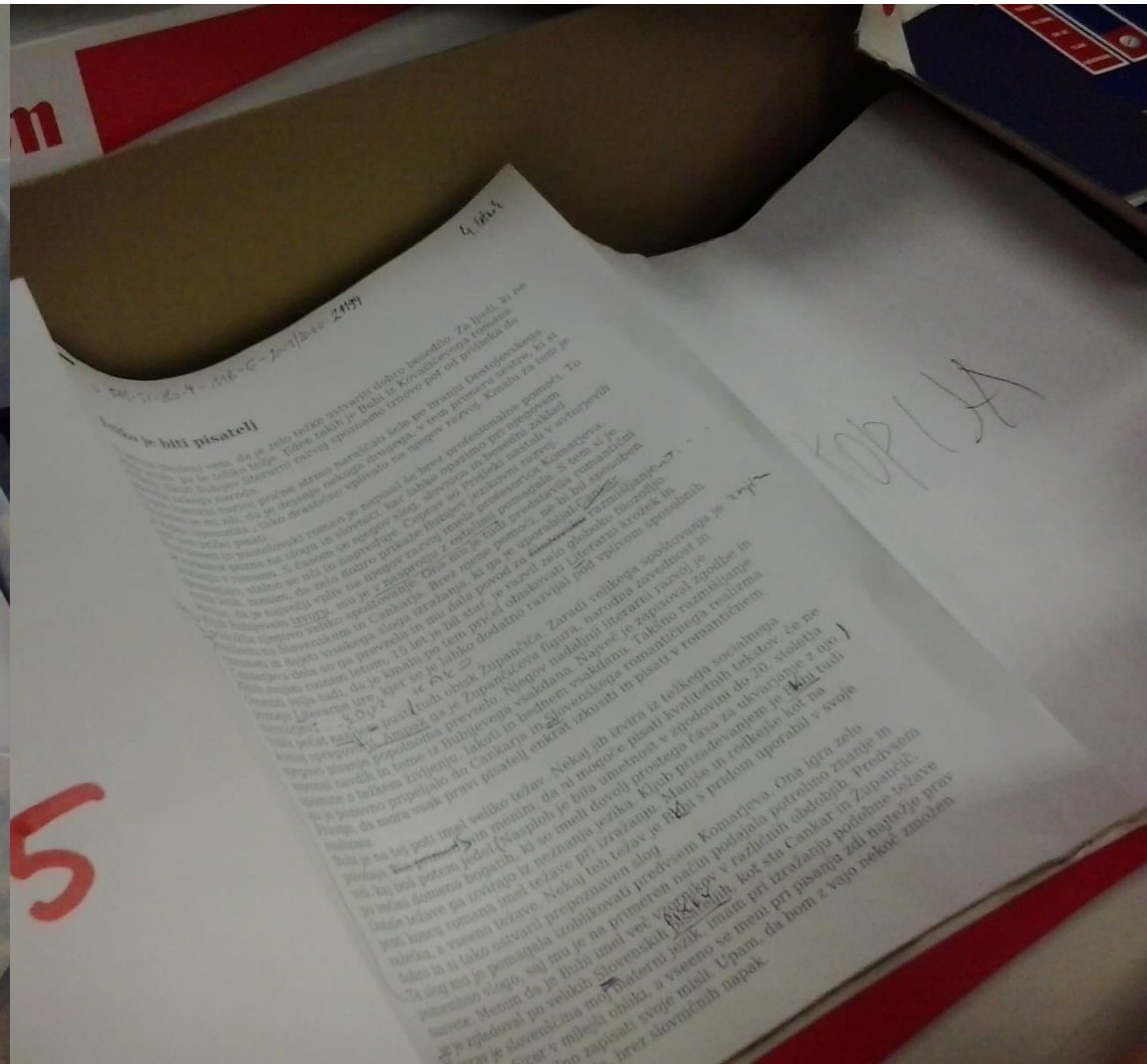
trojína


Osnovne šole
Celje
Koper
Kranj
Krško
Ljubljana
Maribor
Murska Sobota
Nova Gorica
Novo mesto
Postojna
Slovenj Gradec
Srednje šole
Celje
Koper
Kranj
Krško
Ljubljana
Maribor
Murska Sobota
Nova Gorica
Novo mesto
Postojna
Slovenj Gradec

trojína

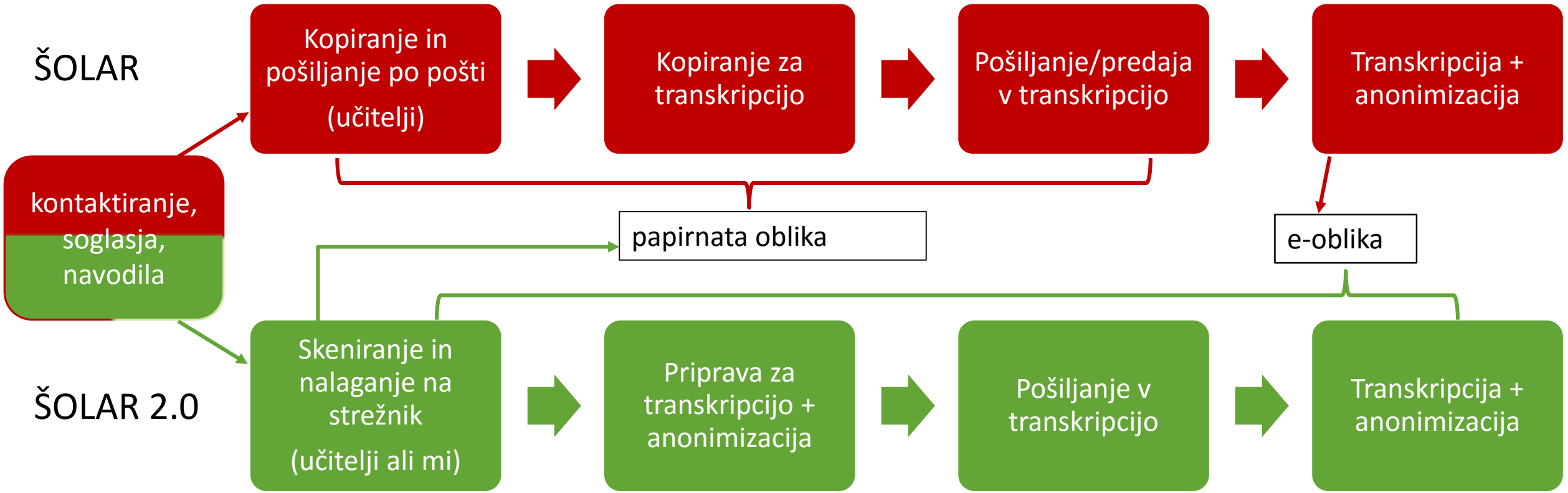
Zbiranje in digitalizacija





trojína

Zbiranje in digitalizacija



trojína

Digitalizacija že zbranih besedil

- 5891 že zbranih besedil, ki v prvotni korpus niso bila vključena
- Digitalizacija:
  - Digitaliziramo VSA besedila (e-arhiv)
  - Prednost imajo besedila, relevantna za Šolar 2.0
  - Formata datotek TIFF in PDF
  - Spletni repozitorij za shranjevanje (OwnCloud)
  - En izdelek – ena datoteka
  - Skeniranje:
    - Ni učiteljskih popravkov → sivinsko
    - Učiteljski popravki → barvno

trojína

Nadgradnja korpusa Šolar

- projekt pri Ministrstvu za kulturo (2015-2018):
  - Z načrtnim dodajanjem novo zbranih (in deloma že zbranih besedil) izboljšati regijsko uravnoteženost korpusa Šolar in uravnoteženost po stopnji šolanja.
  - Vnesti informacijo o podkategoriji napake učenca v obstoječih besedilih in na podlagi teh informacij izdelati učni korpus za jezikovnotehnološke namene.
  - Izdelati podkorpus besedil šolarjev dislektikov kot primer korpusa besedil učencev s posebnimi potrebami. Podkorpus naj bi kasneje služil kot model za izdelavo vseh nadaljnjih podobnih podkorpusov.

trojína

Jezikovni popravki učiteljev

- 56 % besedil
- 35.035 popravkov: Zapis (61,1 %), Besedišče (10,9 %), Oblika (10,3 %), Skladnja (17,7 %)
- Analiza jezikovnih težav učiteljev: korpusni pristop (Kosem et al., 2012)
  - Pripisovanje kategorij jezikovnim popravkom
    - Zapis (68), Besedišče (207), Oblika (120), Skladnja (334)
- Analiza opravljena v orodju WordSmith Tools
- Pomanjkljivosti:
  - Nekateri tipi popravkov manj podrobno obravnavani kot drugi
  - Več različnih sistemov kategorizacije jezikovnih popravkov
  - Kategorije še niso pripisane v korpus
  - Omejitev en popravek → ena kategorija

trojína

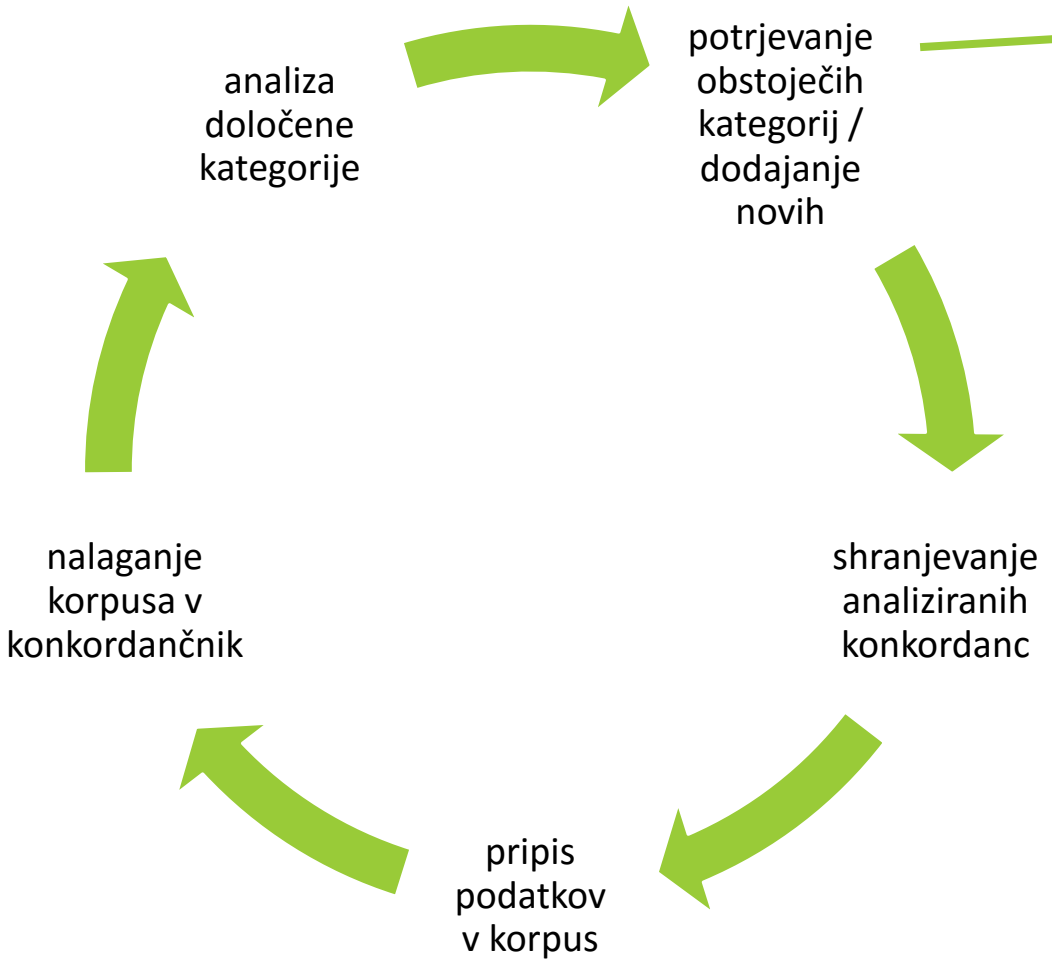
A	B	C
N	Concordance	Set
1	p1>je Simon poljski oficir</p1></u1> Drohojewski. Simonu ni jasno<u1 tip="Z-LOC"><p1>,</p1></u1> <u1 tip="B">zakaj<	zč
2	arečja dovoljena tudi v šolah. Mladi <u1 tip="B">večina<p1>večinoma</p1></u1> uporabljamo sleng, <u1 tip="B">privzete<	zv
3	l njihove ukaze. Ljudi, ki so karkoli poskušali narediti proti oblasti v državi, so ujeli in jih <u1 tip="B">interveniral<p1>interni	zv
4	je že dober mesec po moževi smrti poročila z drugim <u1 n=""/> <u1 n=""/>. Zato Hamlet poroke ne <u1 tip="B">obravnav	zv
5	štirideset in petdeset let, prste ima skoraj brez nohtov in zaudarja po mesu, po mesnici, in po <u1 tip="B">prepoznanem<	zv
6	/u1><u1 tip="S-IZP"><p1>, kar imajo</p1></u1> po materi. Pri otrocih je bila mati vse, oče pa le <u1 tip="B">medlel<p1>	zv
7	idno in pazljivo reševati, saj nam spremenijo način življenja. Knjiga mi je bila zelo všeč in jo <u1 tip="B">posvečam<p1>prip	zv
8	ovej nam enkrat po pravici," je rekla socialna delavka, ki ji je bilo ime Darja. „Zakaj si ušla? <u1 tip="B">Kaj<p1>Saj</p1><	zv
9	<u1 tip="Z-LOC">,<p1></p1></u1> sama <u1 tip="B">potegovati<p1>boriti</p1></u1> zase in ni imela <u2 tip="B">zavet	zv
10	ip="Z-SN">nebo<p1>ne bo</p1></u1> nič manjkalo<u1 tip="Z-LOC"><p1>,</p1></u1> če se spreobrne in <u1 tip="B">va	zv
11	lovek tako mirno joka. Iztopi iz avtobusa in naprej živi svoje življenje. Pripovedovalec je zelo <u1 tip="B">nadzoren<p1>naz	zv
12	ala od leta 1830 do 1849. Je lirsko-epska pesnitev, ki je sestavljena iz treh delov. Prvi del je <u1 tip="B">posvetni<p1>posv	zv
13	: 2009/2010 ID: KUS-Sl-slo-4-KP-E-2009/2010-20983 </head> <body> Književna zvrst tega odlomka je <u1 tip="B"><u2 tip	zv
14	<u1 tip="S-IZP"><p1>njen</p1></u1> gib posebej. Vsaka stvar se mu zdi zelo zanimiva, saj jo tako <u1 tip="B">nadzornc	zv
15	a</p2></u2> rada. Otroke tudi prevladajo emocije, od </u1> sreče so začeli kar jokati. "Bili smo <u2 tip="B">presrečni<p2	zv
16	<u1 tip="Z-LOC" pov="1">,<p1>.</p1></u1> <u1 tip="Z-MV" pov="1">pred<p1>Pred</p1></u1> tem se ni <u1 tip="B">pre	zv
17	</p1></u1> bo <u1 tip="S-STR">tudi on storil samomor<p1>še nekdo umrl.</p1></u1>" Kreon še vedno <u1 tip="B">obst	zv
18	u1 tip="O" pov="2">svojem<p1>svojim</p1></u1> <u1 tip="O" pov="2">vojaškem<p1>vojaškim</p1></u1> <u1 tip="B" pc	zv
19	ti<p1>služit</p1></u1> denar, da so imeli za kruh, da so se lahko preživljali. Meta se je steška <u1 tip="B">prebivala<p1>	zv
20	ada v socialni <u1 tip="B">razred<p1>realizem</p1></u1>. Zvrst<u1 tip="S-IZP">:<p1> je</p1></u1> <u1 tip="B">etika<p	zv
21	mrtve. To je bil zame srečen in hkrati žalosten dan. Zjutraj, ko sem prišel iz Ljubljane, nisem <u1 tip="B">okreval<p1>okle	zv
22	a, da sta <u1 tip="B">ga<p1>Tomažka</p1></u1> prebudila in še dolgo v noč Boga hvalila in hudiča <u1 tip="B">prekinjal	zv
23	kovana, ampak polna preprek, preizkušenj, ki ti <u1 tip="B" pov="2">stojijo<p1>pridejo</p1></u1> <u1 tip="B" pov="2">na	zv
24	pov="1">prikazuje<p1>prikazujejo</p1></u1>, kako so si bogati in revni različni. Spada v obdobje<u1 tip="B">etike<p1>e	zv

trojína

p1>je Simon poljski oficir</p1></u1> Drohojewski. Simonu ni jasno<u1 tip="Z-LOC"><p1>,</p1></u1> **<u1 tip="B">zakaj<p1>čemu</p1></u1>** so ga sprejeli v zavod. Zaposleni<u2 tip="Z-LOC"><p2>,</p2></u2> na

1. Identifikacija teksta iz konkordance v korpusu
   - Neupoštevanje tipov in podtipov, ker so bili za potrebe analize skrajšani (tip=Z podtip=LOC → tip=Z-LOC
2. identifikacija pravega popravka v konkordanci/korpusu
3. pripis podkategorije

trojína


analiza določene kategorije
potrjevanje obstoječih kategorij / dodajanje novih
shranjevanje analiziranih konkordanc
pripis podatkov v korpus
nalaganje korpusa v konkordančnik


Filter Sort by label Bootstrap Finish New label: Add Batch annotation: selection
13 > Label filter 13
coinvolgere dalla storia e l' ho letteralmente DIVORATO 2 . Questo libro dura quasi 50 anni e ci
coinvolto e lo consiglio a chi ha voglia di divorare 2 un romanzo , e sottolineo romanzo , in
scrittura è nato nell' adolescenza , quando " divoravo 2 " un libro dietro l' altro e da sola leggevo
sono barricata in casa , mangio e studio . Divoro 2 libri , trascrivo appunti , le mani nei
quotidianamente dai giornali , che venivano divorati 2 avidamente e appassionatamente ; ma adesso
arina . Argomento : MANGA (e , visto che ne divoro 2 miliardi , chiedo una TOP 5) La mia personale
sfigato " quattrocchi " sempre immerso a divorare 2 romanzi e saggi ormai sia roba da naftalina
poi gli avrei reso la cortesia ! Mentre divoravamo 2 libri-game e provavamo tutti i giochi che
a chi ancora non lo ha letto , è di non divorare 2 questo libro in poche ore come verrebbe
avea profondamente studiato , compreso , e divorato 2 un gran libro divino : " D. Gaspare Bertoni
infoiati di Romero , mi butto su un libro che divoro 2 in troppo poco tempo . È una storia che
Non le importava di balli o teatri , ma divorava 2 letteralmente libri , e sia detto a suo
L' anima cerca calore tra gli antichi e divorati 2 libri , nel passaggio fuggitivo tra l'

trojína

Nadgradnja korpusa Šolar

- projekt pri Ministrstvu za kulturo (2015-2018):
  - Z načrtnim dodajanjem novo zbranih (in deloma že zbranih besedil) izboljšati regijsko uravnoteženost korpusa Šolar in uravnoteženost po stopnji šolanja.
  - Vnesti informacijo o podkategoriji napake učenca v obstoječih besedilih in na podlagi teh informacij izdelati učni korpus za jezikovnotehnološke namene.
  - Izdelati podkorpus besedil šolarjev dislektikov kot primer korpusa besedil učencev s posebnimi potrebami. Podkorpus naj bi kasneje služil kot model za izdelavo vseh nadaljnjih podobnih podkorpusov.

trojína

Podkorpus besedil šolarjev dislektikov

- Pridobivanje informacije o odločbah učencev in soglasja
- Kompleksnost pridobivanja informacije, označevanja besedil učencev in dosledno upoštevanje anonimnosti
- (Pod)korpus s tovrstnimi metapodatki dostopen samo raziskovalcem (potreben podpis pogodbe)
- Možne raziskave:
  - Primerjave jezikovnih težav in zmožnosti učencev s primanjkljaji v primerjavi s preostalo šolsko populacijo
  - Vpogled v posredovane povratne informacije učencem s primanjkljaji

trojína

Pričakovani glavni rezultati projekta

- Korpus Šolar 2.0
  - 5400 besedil
  - 2 milijona besed
- Učni korpus z označenimi kategorijami jezikovnih popravkov
  - cca 550.000 besed
- Sistem kategorizacije jezikovnih popravkov
- Dokumentacija!

- Dostopnost:
  - Šolar 2.0 in učni korpus: CC BY-NC-SA 2.5 SI
  - Novo zbrana besedila: CC BY 4.0

trojína

Projekt Nadgradnja korpusa Šolar financira Ministrstvo za kulturo, pogodba 3340-15-141006, v okviru razpisa Digitalizacija na področju kulture 2015-2018.

Infrastrukturna podpora projektu, zlasti digitalizaciji obstoječega gradiva, se izvaja v okviru infrastrukturnega programa Center za uporabno jezikoslovje (I0-0051), ki ga sofinancira Javna agencija za raziskovalno dejavnost Republike Slovenije.