Univerza v Ljubljani

KOLOS
Kolokacije
v slovenščini

cjvt

Kolokacijski slovar sodobne slovenščine

Iztok Kosem,[12] Simon Krek,[2] Polona Gantar,[1] Špela Arhar Holdt,[13]
Jaka Čibej,[123] Cyprian Laskowski,[1]

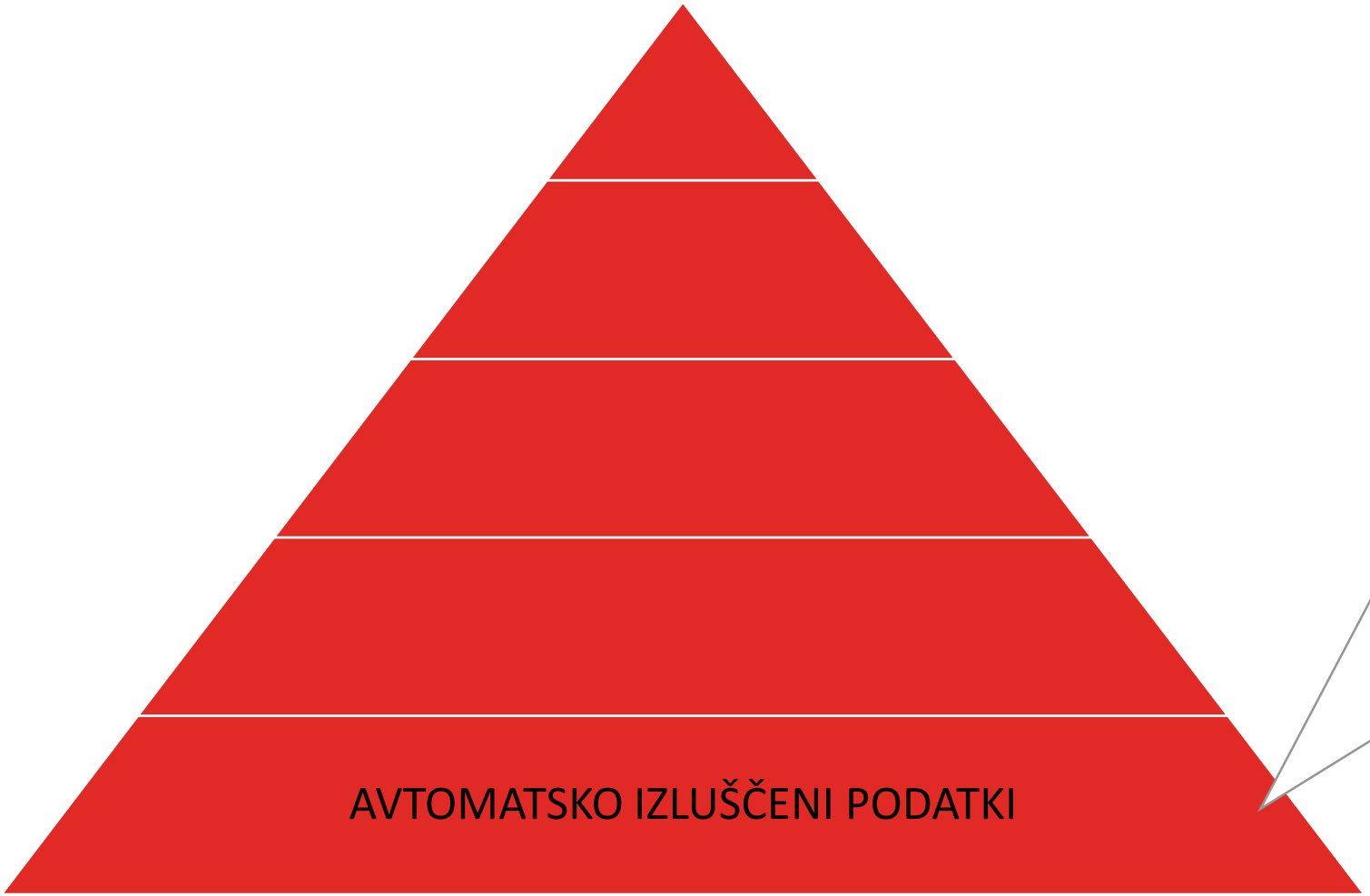
[1]Filozofska fakulteta, Univerza v Ljubljani
[2]Institut "Jožef Stefan"
[3]Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Modeli objave slovarjev

- Tradicionalni model: dokončaj vse, nato objavi
  - PROBLEM: čas do objave, kdaj je nekaj dokončano?
- Rastoči model: objava dokončanih gesel sproti (npr. Veliki poljski slovar) ali v rednih intervalih
  - PROBLEM: traja več let (ali desetletij?), preden količina gesel doseže uporabno vrednost za uporabnike
- Odzivni stopenjski model: objava vseh gesel v različnih stopnjah dokončanosti (Krek et al. 2013)
  - PROBLEM: zanesljivost?

Kolokacijski slovar sodobne slovenščine

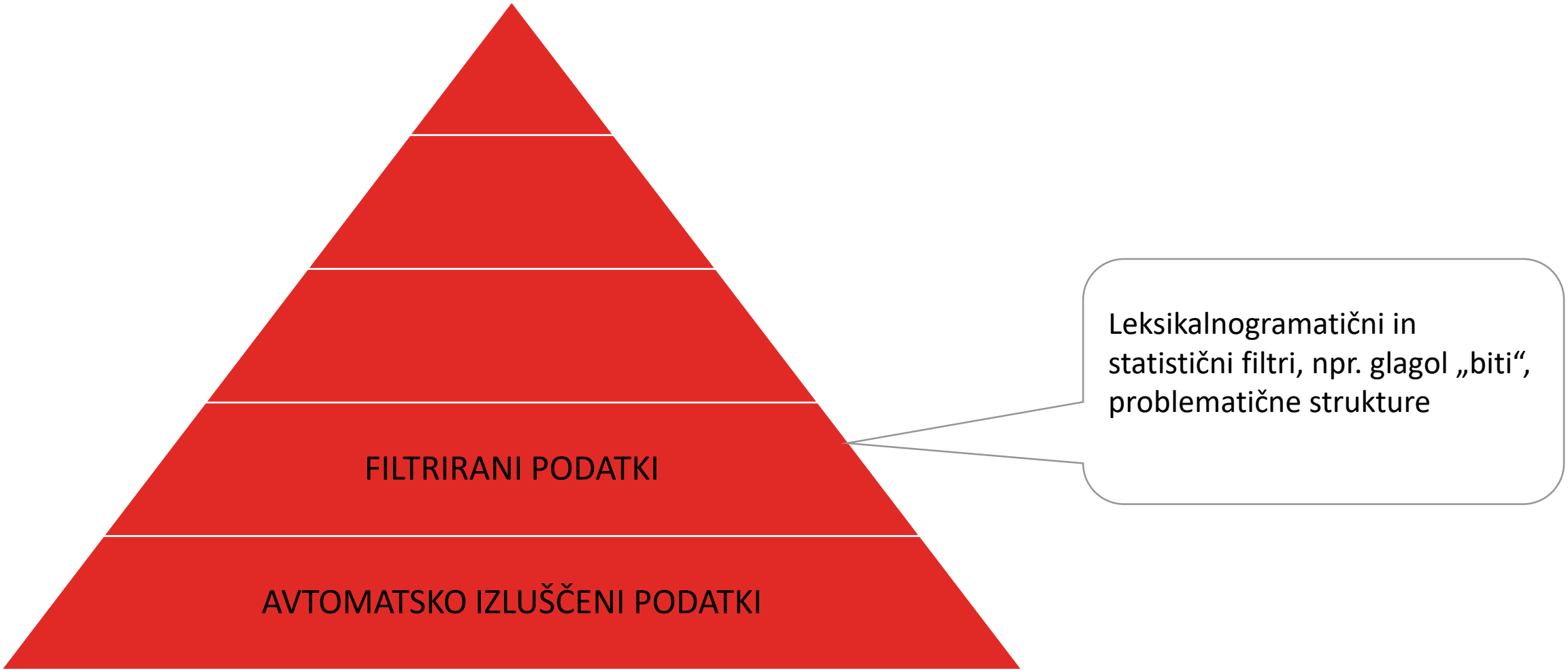
- Verzija 0.9:
  - 35.989 iztočnic
  - 7.717.561 kolokacij
  - 36.736.168 zgledov
- Podatki izluščeni iz korpusa Gigafida

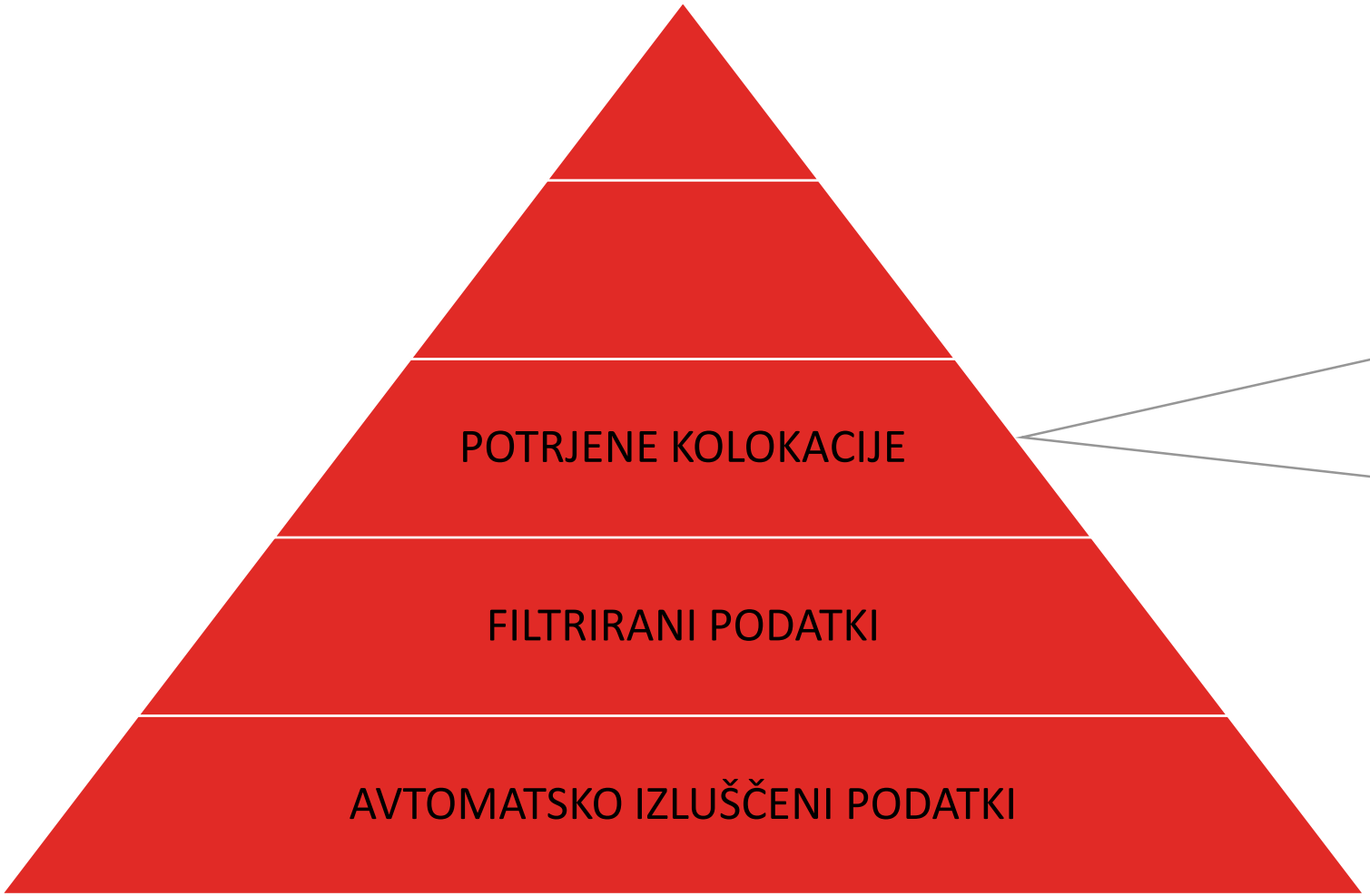


Sketch Engine API:

- Slovnica besednih skic (Krek 2011, Krek 2015)
- GDEX (Kosem et al. 2011, Kosem et al. 2013, Kosem 2015)

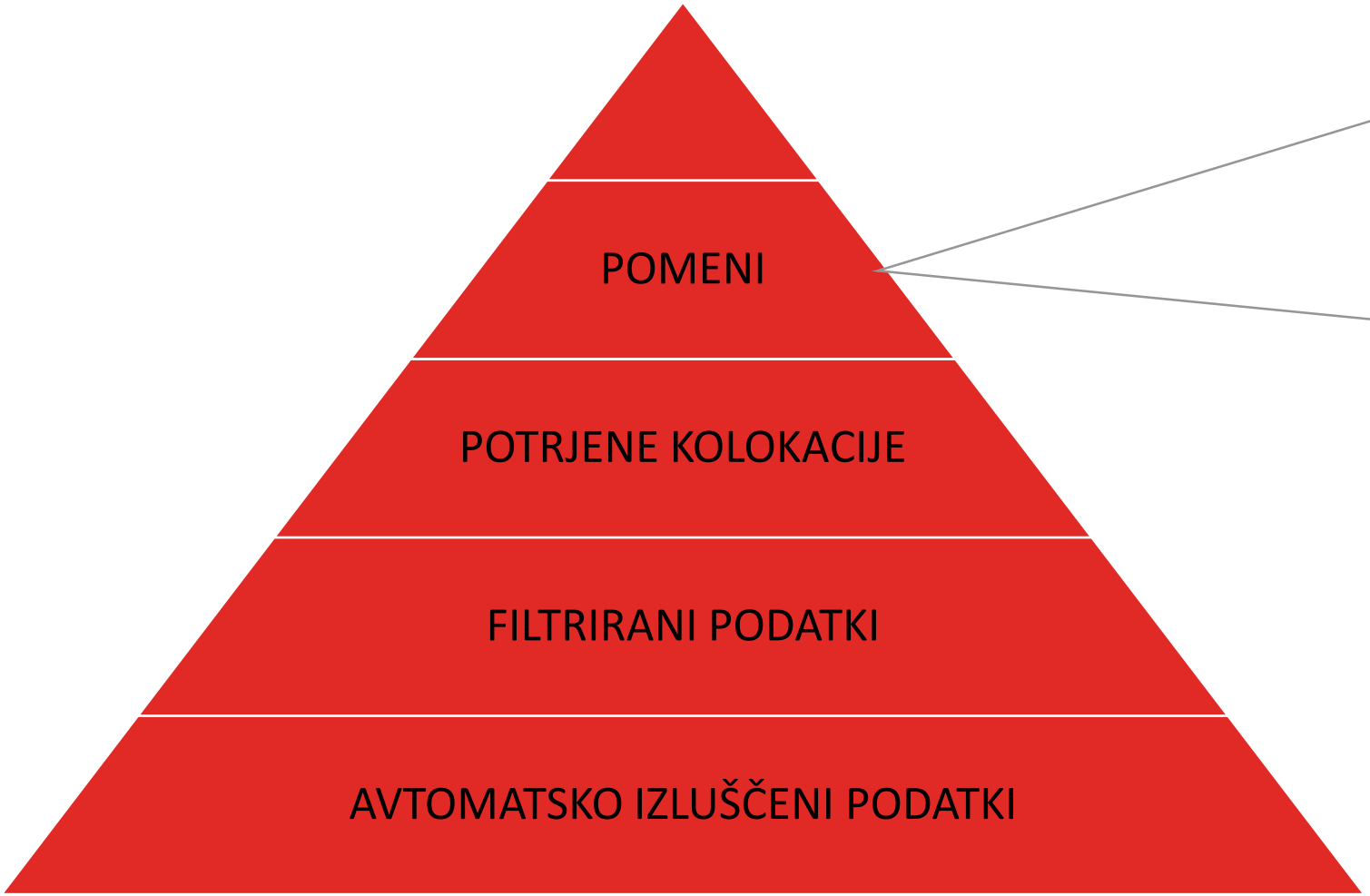
Leksikografski postopek:
Krek et al. 2013
Gantar et al. 2016
Gorjanc et al. 2015, 2016


FILTRIRANI PODATKI
AVTOMATSKO IZLUŠČENI PODATKI
Leksikalnogramatični in statistični filtri, npr. glagol „biti“, problematične strukture



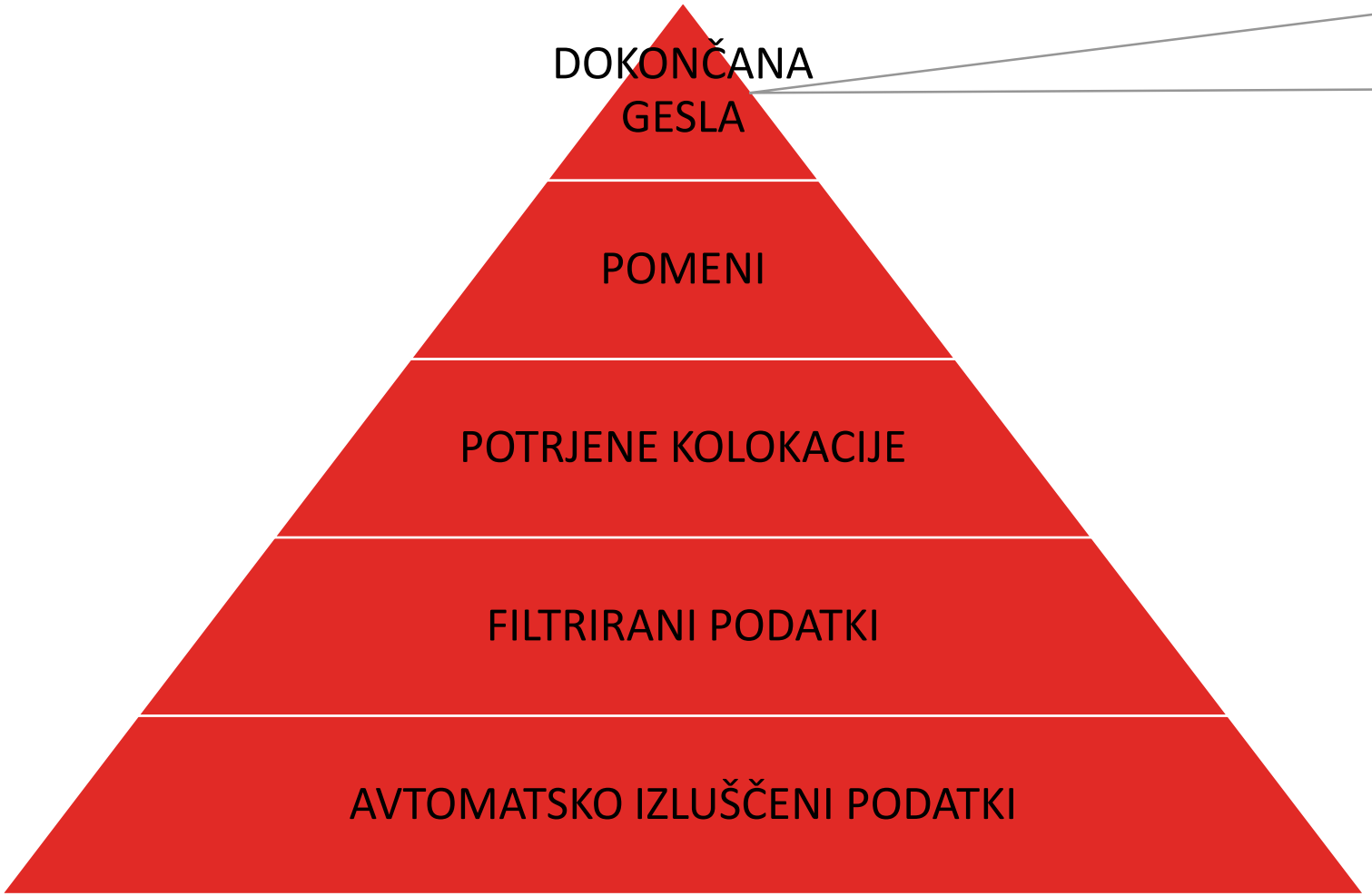
IZLOČAMO V CELOTI:
- nekolokacije
- napake strukture

IZLOČAMO,
A OBDRŽIMO V BAZI:
- semantično manj relevantne kolokacije



Kolokacije in pripadajoče zglede razvrstimo pod pomene

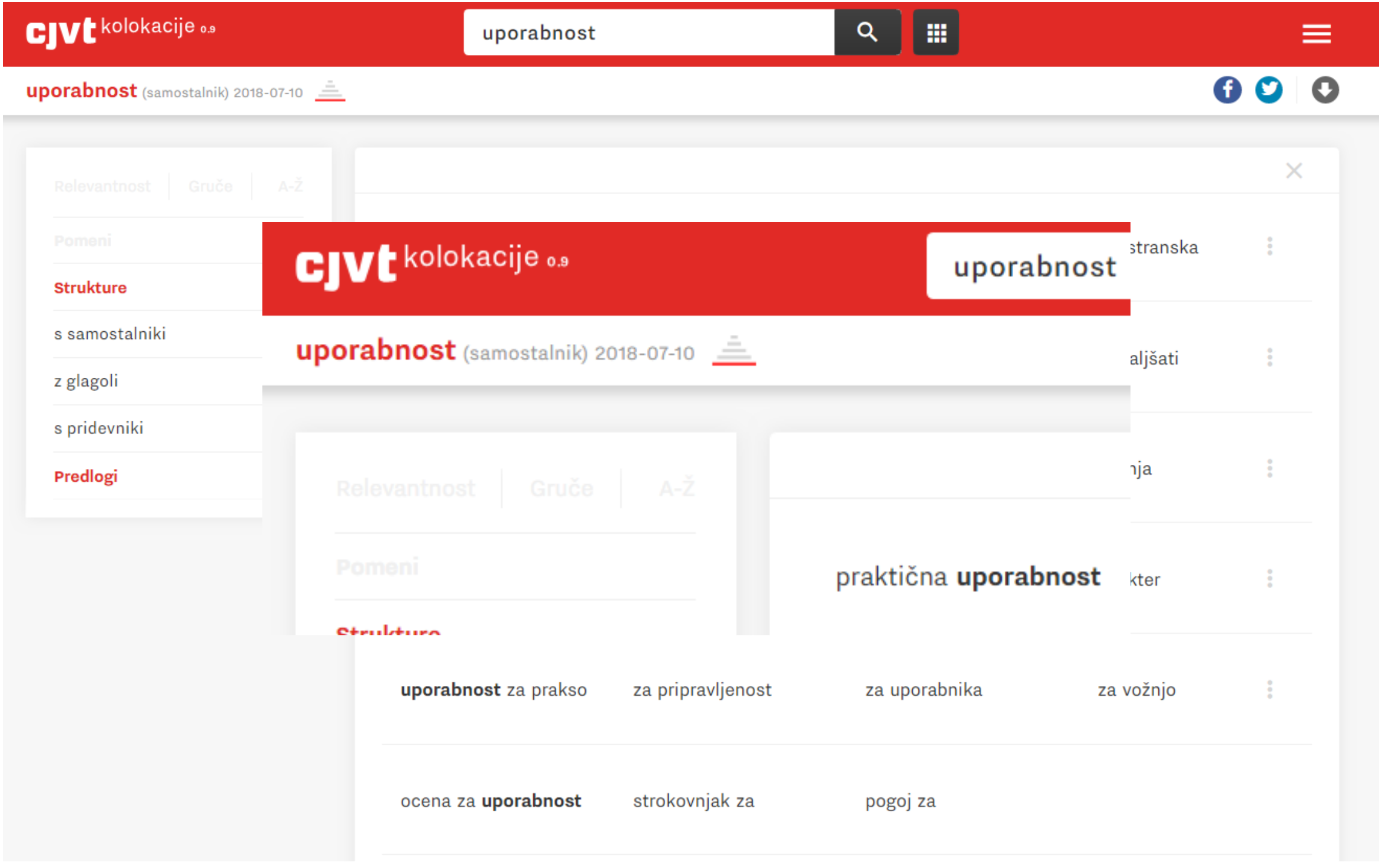
Uporaba množičenja (deluje!)



Končni pregled gesel, brisanje slabih zgledov, dodajanje informacij (npr. oznak)

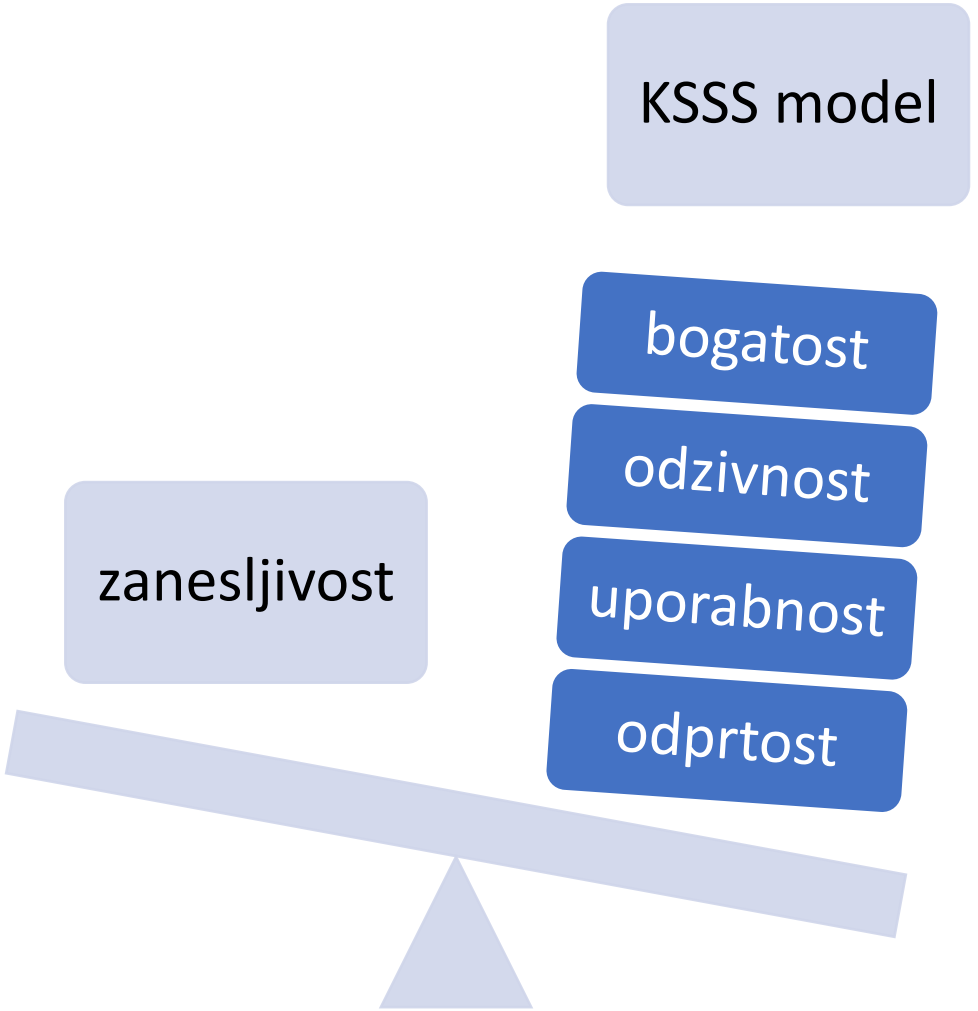
Vmesnik

- Glavne značilnosti:
  - Izhodišče so kolokacije in ne pomeni
  - Jasna informacija o stanju gesla in datumu zadnje posodobitve


cjvt kolokacije 0.9
uporabnost
uporabnost (samostalnik) 2018-07-10
Relevantnost
Gruče
A-Ž
Pomeni
Strukture
s samostalniki
z glagoli
s pridevniki
Predlogi
praktična uporabnost
uporabnost za prakso
za pripravljenost
za uporabnika
za vožnjo
ocena za uporabnost
strokovnjak za
pogoj za

Vmesnik

- Glavne značilnosti:
  - Izhodišče so kolokacije in ne pomeni
  - Jasna informacija o stanju gesla in datumu zadnje posodobitve
  - Veliko možnosti razvrščanja in filtriranja
  - Prilagojenost za različne digitalne medije


KSSS model
bogatost
odzivnost
uporabnost
odprtost
zanesljivost

Projekt Kolokacije v slovenščini (KOLOS)

Kolokativnost

struktura	% of good	% of noisy	Agreement	same answers	% nestrinjanja ali dva Ne vem
PRID + SAM	88.9	3.1	0.42	0.78	11,8
SAM + SAM_2	84.9	7.8	0.45	0.63	13,3
GLAG + SAM_4	87.0	10.0	0.46	0.73	16,1
PRISL + GLAG	87.7	1.3	0.37	0.63	19,5
PRISL + PRID	63.6	20.7	0.46	0.64	15,9
SAM + v + SAM_5	64.2	0.5	0.35	0.50	39,0
GLAG + PRISL	59.6	7.3	0.47	0.61	19,4
GLAG + v + SAM_5	86.0	2.1	0.33	0.46	26,4
SAM + s/z + SAM_6	74.7	0.2	0.42	0.56	23,3
GLAG + s/z + SAM_6	92.7	1.3	0.46	0.61	10,2
GLAG + SAM_2	8.6	90.6	0.40	0.79	20,6

satelit *(samostalnik)*

KOLIKO KOLOKACIJ IZLUŠČITI?

Kriterij:

- jakost = 6,00 jakost?
- Prvih 20 kolokacij?

salience

S_kakšen?		43.95
vohunski +	357	9.81
Zemljin +	217	9.72
geostacionaren +	179	9.72
Jupitrov +	130	9.09
meteorološki +	111	8.33
umeten +	510	8.13
komunikacijski +	246	7.96
Saturnov	53	7.84
telekomunikacijski +	160	7.78
Nasin	50	7.61
opazovalen	40	7.14
Marsov	33	7.09
izstreljen	28	6.90
navigacijski	47	6.57
Galilejev	20	6.54
sovjetski +	123	6.46
astronomski	34	6.37
Uranov	17	6.33
odslužen	26	6.19
infrardeč	24	6.10
istrski	34	5.88
nedelujoč	13	5.88
utirjen	11	5.74
vremenski	83	5.67
izvidniški	13	5.67

frequency

S_kakšen?		43.95
umeten +	510	8.13
vohunski +	357	9.81
ameriški +	308	4.28
naraven +	249	5.52
komunikacijski +	246	7.96
Zemljin +	217	9.72
geostacionaren +	179	9.72
vojaški +	171	5.01
telekomunikacijski +	160	7.78
velik +	150	1.07
Jupitrov +	130	9.09
sovjetski +	123	6.46
nov +	116	0.60
meteorološki +	111	8.33
ruski	97	4.38
majhen	88	2.55
evropski	84	1.77
vremenski	83	5.67
komercialen	65	5.67
nekdanji	61	2.39
slovenski	61	0.09
teniški	57	5.07
zemeljski	55	5.45
raziskovalen	54	4.60
edin	54	2.97

Distribucijska semantika

struktura	baseline	SkE SVM	vložitve SVM	SkE + vložitve SVM
PRID + SAM	0,5	0,563	0,715	**0,721**
SAM + SAM_2	0,5	0,615	0,841	**0,843**
GLAG + SAM_4	0,5	0,644	0,816	**0,822**
PRISL + GLAG	0,5	0,716	**0,879**	**0,879**
PRISL + PRID	0,5	0,794	0,852	**0,856**
SAM + v + SAM_5	0,5	0,651	0,740	**0,746**
GLAG + PRISL	0,5	0,696	0,801	**0,806**
GLAG + v + SAM_5	0,5	0,510	0,666	**0,668**
SAM + s/z + SAM_6	0,5	0,533	**0,786**	0,785
GLAG + s/z + SAM_6	0,5	0,638	0,660	**0,683**
GLAG + SAM_2	0,5	0,862	0,795	**0,810**

Zaključki

- Uporabnost
  - KSSS je odzivni stopenjski slovar z bogatim naborom podatkov o kolokacijah
- Odprtost
  - Kolokacijski podatki iz KSSS bodo na voljo kot baza podatkov v repozitoriju CLARIN.SI pod licenco Creative Commons 4.0 CC-BY
- Podprtost
  - Povezani projekti nudijo možnost za izboljšanje metodološkega, vsebinskega in predstavitvenega dela

Zahvala

- Predstavitev je podprl projekt Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki, ki ga financira Javna agencija za raziskovalno dejavnost Republike Slovenije (KOLOS, J6-8255).

- Projekt sta podprla infrastrukturna programa:
  - Center za jezikovne vire in tehnologije, Univerza v Ljubljani
  - Centre za uporabno jezikoslovje pri zavodu Trojina (I0-0051)
- Projekt je podprl raziskovalni program “Slovenski jezik – bazične in kontrastivne raziskave” (P6-0215)