# SketchEngine za JANES

kaja.dobrovoljc@trojina.si, OP FF UL, 4. 3. 2015

# Pregled tem

- Korpus
- **Iskanje**
- **Konkordance**
- **Seznami**
- **Besedne skice**
- Vprašanja
- Za navdušence

# O čem govorimo, ko govorimo o SkE

## SketchEngine

http://www.sketchengine.co.uk/

- Lexical Computing (t.i. Čehi)
- plačljivo (50+€/leto)
- hitrost
- avtonomnost
- dobra podpora
- številni viri za druge jezike
- novosti
  https://beta.sketchengine.co.uk/

## noSketchEngine

http://nl.ijs.si/noske/

- IJS
- prost dostop
- vsi referenčni slovenski korpusi
- lokalna podpora
- brez besednih skic

## Lokalna inštalacija

http://sketch.fri1.uni-lj.si/bonito/

- CJVT-FRI
- omejen dostop
- večina funkcionalnosti SkE
- najsodobnejša orodja za slovenščino (skice, GDEX)

# Spoznajmo korpus JANES

Metapodatki o besedilu

Struktura besedila

Besedilo

```
<text type="blog" author="-" date="2011" time="-" title="-"
urldomain="www.blog.nevestica.si" url="http://www.blog.nevestica.si"
id="janes.blog.062276">
<p>
<s>
Danes     danes     danes     Rgp  Rsn
bom  bom  biti Va-f1s-n  Gp-ppe-n
malo malo malo Rgp  Rsn
opisala   opisala   opisati   Vmep-sf   Ggdd-ez
tusmami   tusmami   tusma     Ncfpi     Sozmo
kanzashi  kanzashi  kanzashi  Vmep-pm   Ggdd-mm
tehniko   tehniko   tehnika   Ncfsa     Sozet
izdelave  izdelave  izdelava  Ncfsg     Sozer
rož  rož  roža Ncfpg     Sozmr
iz   iz   iz   Sg   Dr
blaga     blaga     blago     Ncnsg     Soser
s    s    z    Si   Do
pomočjo   pomočjo   pomoč     Ncfsi     Sozeo
katere    katere    kateri    Pq-fsg    Zv-zer
tudi tudi tudi Q    L
sama sama sam  Agpfsn    Ppnzei
izdelujem izdelujem izdelovati     Vmpr1s    Ggnspe
rožice    rožice    rožica    Ncfsg     Sozer
<g/>
.    .    .    Z    U
</s>
</p>
```

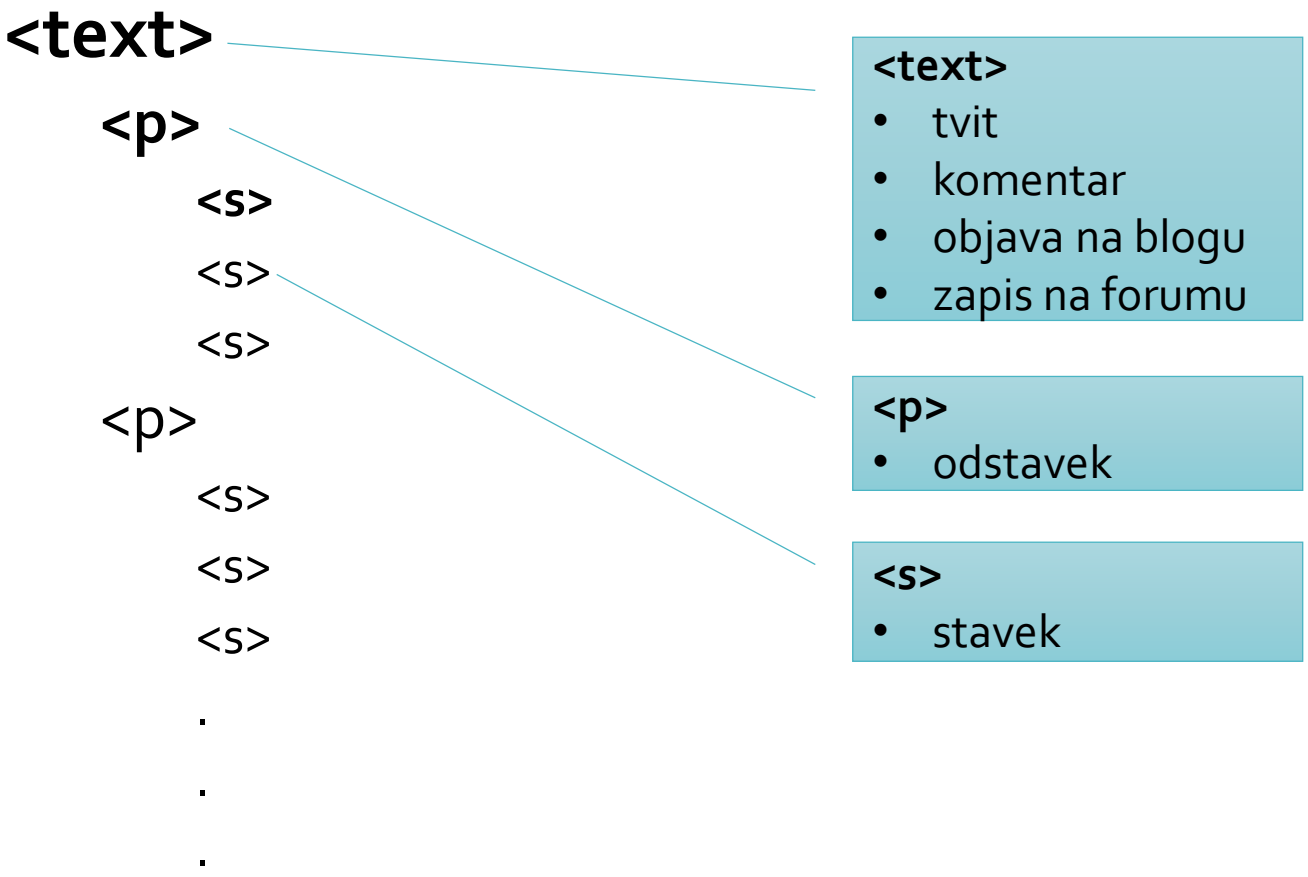
# Besedilo: atributi

Danes bom malo opisala tusmami kanzashi tehniko izdelave rož iz blaga [...].

| podatek | opisala | opisala | opisati | Vmep-sf | Ggdd-ez |
|---|---|---|---|---|---|
| tip podatka | zapisana | normalizirana | lema | msd (ang.) | msd (slo.) |
| atribut v SkE | word | lc | lemma | tag_en | tag |

NB! lc v korpusih standardnega jezika (Gigafida, Kres) označuje besedno obliko, pisano z malimi začetnicami (lowercase).

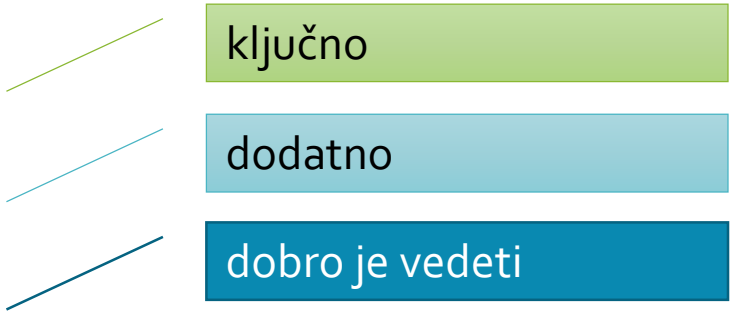
# Struktura besedila



# Metapodatki o besedilu

| | | |
|---|---|---|
| text.id | unikatna oznaka besedila | npr. `janes.tweet.362541192514256896` |
| text.type | tip besedila | `blog, comment, forum, tweet` |
| text.url | url objave, komentarja, zapisa na forumu, tvita | npr. `http://www.rtvslo.si/znanost-in-tehnologija/ko-bo-zemlje-konec-bo-na-nebu-zapisano-slovensko-ime/317001` |
| text.urldomain | url bloga, nov. portala, foruma, tviterja | npr. `piroman.blog.siol.net` |
| text.date | datum nastanka besedila | npr. `2013-07-31` |
| text.time | ura nastanka besedila | npr. `11:52:10` |
| text.author | -, avtor komentarja, forumaš, tviteraš | npr. `a3but` |
| text.title | -, naslov članka (komentarji), naslov foruma (debate), - | npr. `Ko bo Zemlje konec, bo na nebu zapisano slovensko ime` |
| text.wordcount | dolžina besedila | npr. `19` |

# Veselo na delo

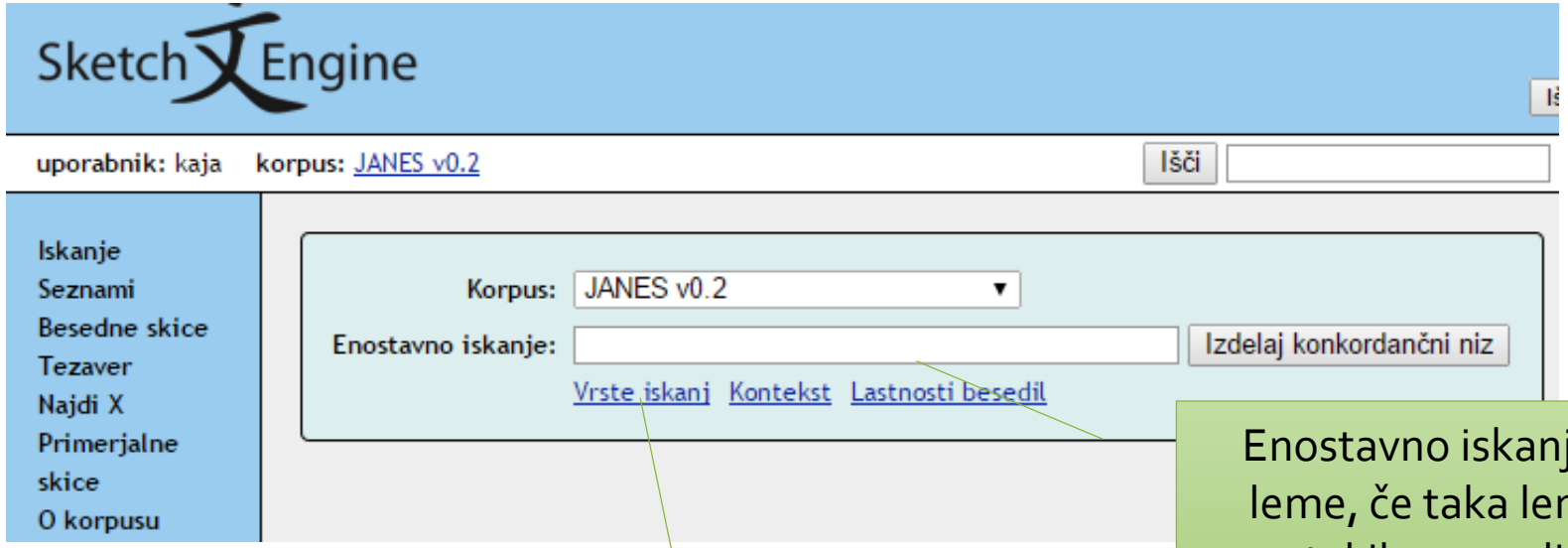



- http://sketch.fri1.uni-lj.si/bonito/
  - up. ime: janes
  - geslo: 9neztandart6

- v primeru težav:
  http://nl.ijs.si/noske/janes.cgi/first_form?corpname=janes
  - up. ime: janes
  - geslo: 9neztandart6

# ISKANJE

# Enostavno iskanje

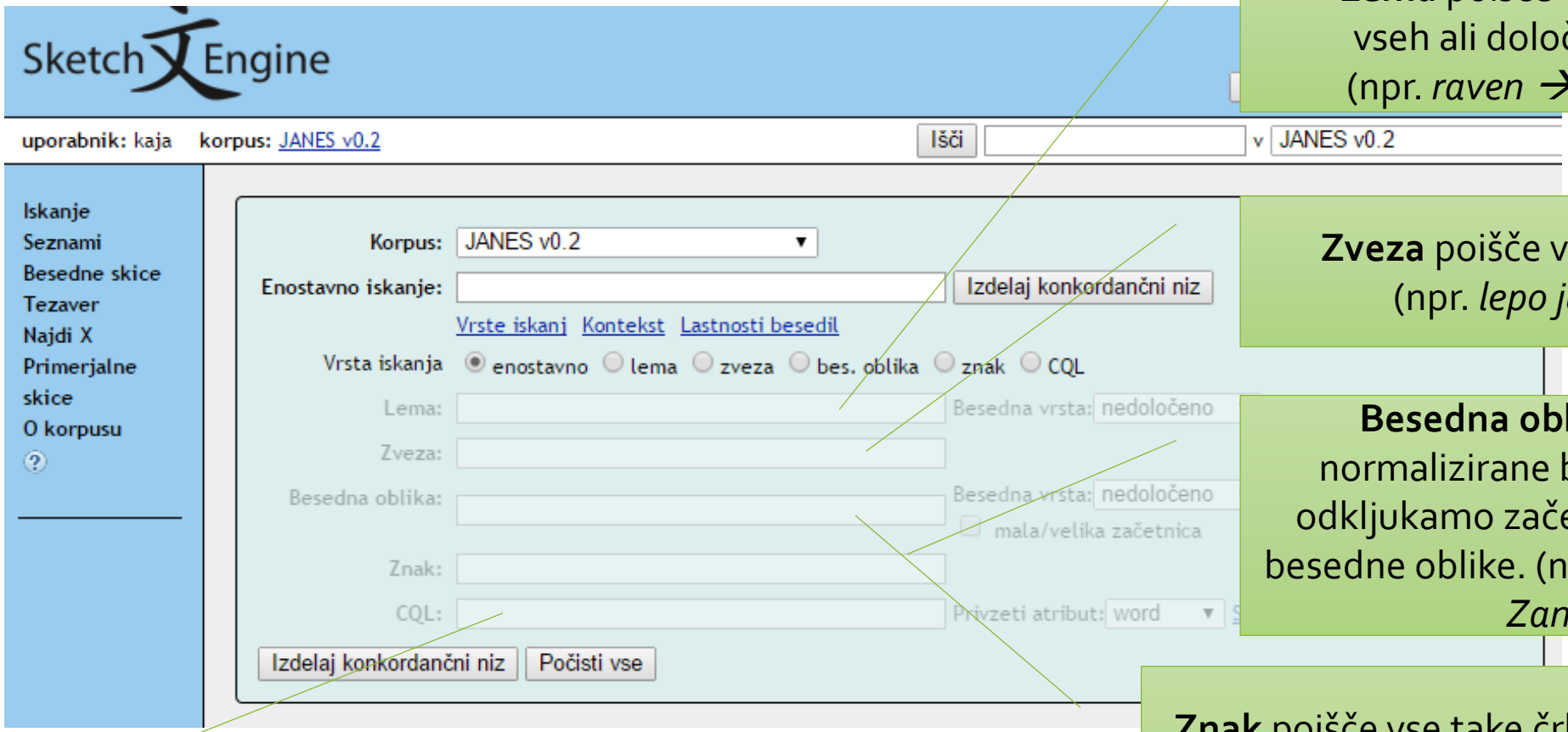



Enostavno iskanje išče po vseh oblikah leme, če taka lema obstaja, in po vseh takih normaliziranih oblikah (lc).

Možnost vklapljanja in izklapljanja naprednega iskanja.

Če nas zanima določena besedna oblika v korpusu JANES zaradi drugačnega pomena atributa lc za ta namen raje uporabljajmo druge vrste iskanja.

# Napredno iskanje: Vrste iskanj




**Lema** poišče vse oblike neke leme vseh ali določene besedne vrste. (npr. *raven* → *raven, ravnega itd.)*

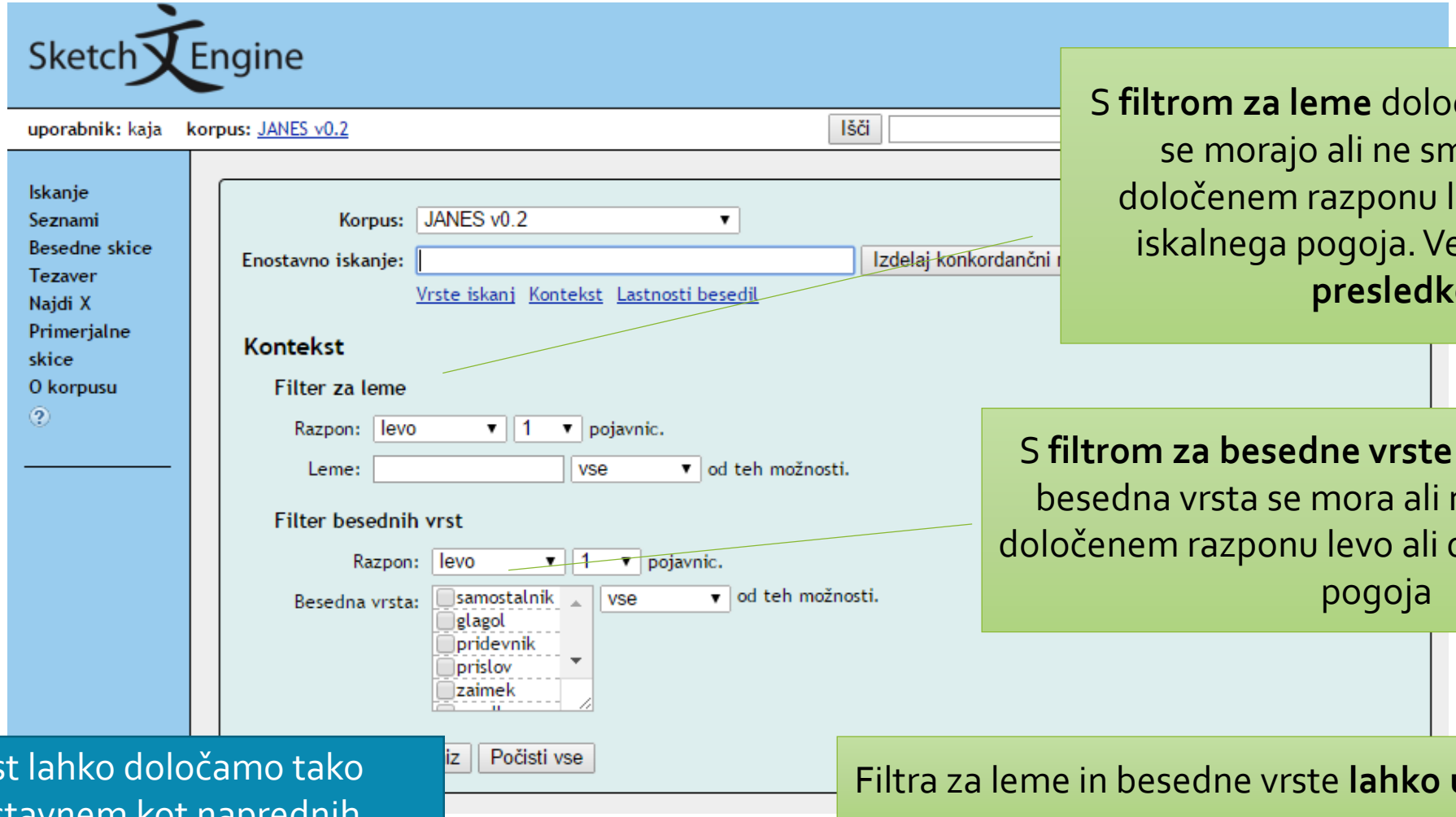
**Zveza** poišče vse take besedne nize (npr. *lepo jutro* → *lepo jutro)*

**Besedna oblika** poišče vse take normalizirane besedne oblike (lc), č odkljukamo začetnico pa take zapisa besedne oblike. (npr. *zanimivo* → *zanim Zanimivo ipd.*)

**Znak** poišče vse take črkovne nize, ki so lahko del daljših besednih oblik. (npr. *jež* → *jež, ježeš, naježeno* itd.)

**CQL**: kompleksna iskanja v jeziku CQL (Corpus Query Language)

# Napredno iskanje: Kontekst



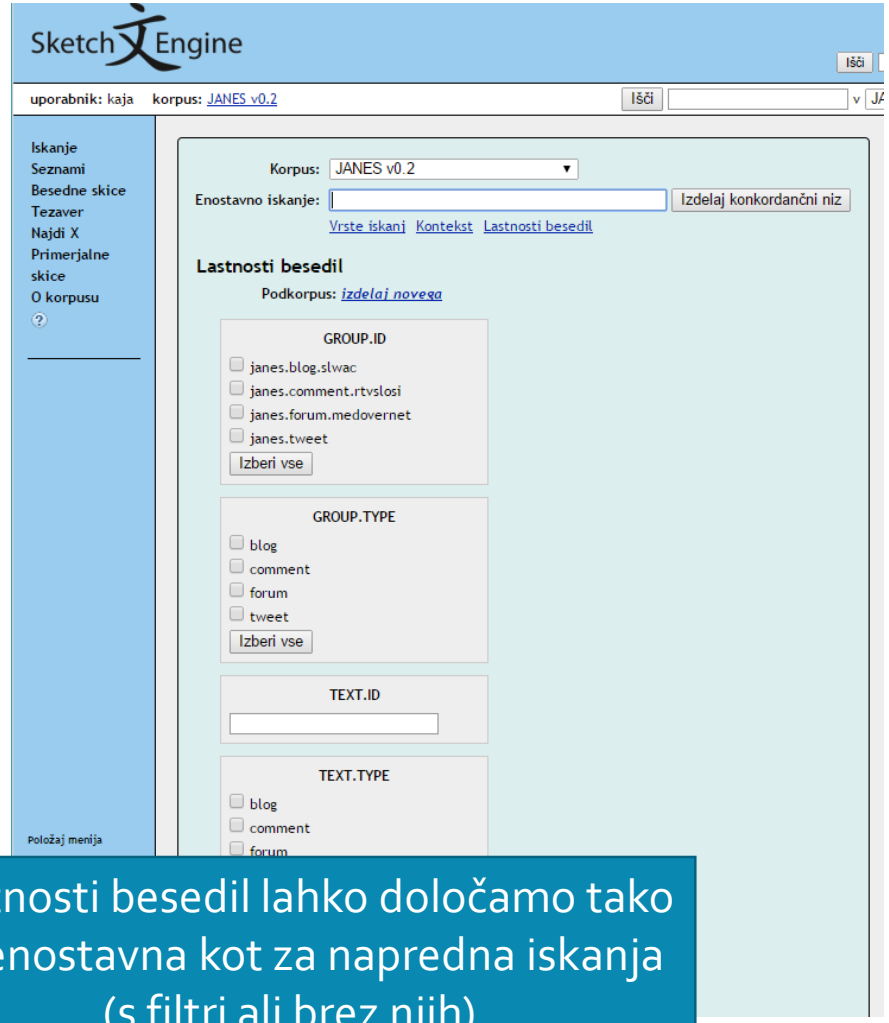

S **filtrom za leme** določimo, katere leme se morajo ali ne smejo pojaviti v določenem razponu levo ali desno od iskalnega pogoja. Več lem **ločimo s presledkom.**

S **filtrom za besedne vrste** določimo, katera besedna vrsta se mora ali ne sme pojaviti v določenem razponu levo ali desno od iskalnega pogoja

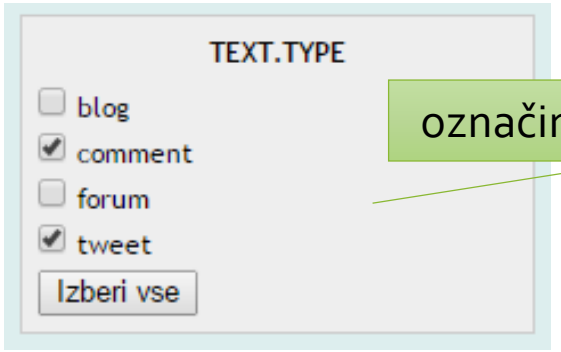
Kontekst lahko določamo tako pri enostavnem kot naprednih iskanjih.

Filtra za leme in besedne vrste **lahko uporabimo hkrati.**

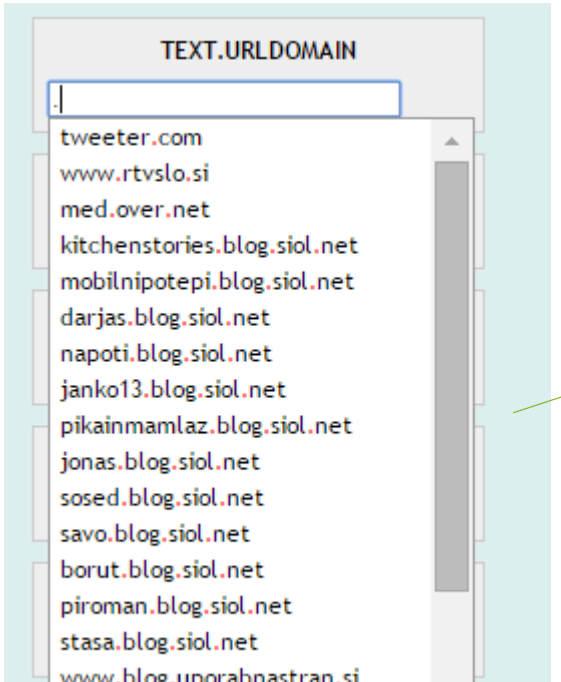
# Napredno iskanje: Lastnosti besedil




Lastnosti besedil lahko določamo tako za enostavna kot za napredna iskanja (s filtri ali brez njih).




označimo **tip besedila**




**in/ali** na spustnem seznamu izberemo **druge lastnosti besedil ,** po katerih želimo iskati

# Corpus Query Language (CQL)

- jezik za kompleksna iskanja
- zahteva nekaj učenja
- omogoča hitrejše delo
- iskanja v obliki regularnih izrazov

# CQL: osnovna iskanja

- [atribut="vrednost"]
  `[word="nauš"]` → *nauš*
  `[lemma="hiša"]` → *hiša, hiše …*
  `[tag="Rsn"]` → *zanimivo, daleč …*

- tudi za **več besed**
  `[lemma="iskati"] [tag="Sozet"]`
  →*iščem pomoč, iskati srečo …*

- **več atributov ene besede** povežemo z znakom **&**
  `[word="hiše" & tag="Sozer"]`

**Ne pozabi na oglate oklepaje in narekovaje!**
V iskalno okence lahko sicer vnesemo tudi samo vrednost atributa v narekovajih, npr. **"hiša"** namesto [word="hiša"], a moramo pri tem paziti, da smo v polju na desni izbrali ustrezen atribut (word), na kar pogosto pozabimo.

# CQL: (nekateri) posebni operatorji

| operator | pomen | primer iskanja | opis iskanja |
|---|---|---|---|
| . | katerikoli znak | `[tag="So..d"]` | poišči vse občne samostalnike v dajalniku |
| [] (znotraj besede) | katerakoli znak znotraj oglatih oklepajev | `[lemma="miš[eo]lovka"]` | poišče vse oblike lem *mišelovka* in *mišolovka* |
| * | 0 ali več pojavitev (znaka, črke, besede ...) | `[word="tvit.*"]` | poišči vse besede, ki se začnejo na tvit- (tudi *tvit*) |
| + | 1 ali več pojavitev | `[word=".+ost"]` | poišči vse besede, ki se končajo na –ost (brez *ost*) |
| ? | 0 ali 1 pojavitev | `[word="Kleme?n.*]` | poišči besede, ki se začnejo na Klemen- ali Klemn- |
| \| | disjunkcija ("ali") | `[lemma="(bloger\|blogar)"]` | poišči vse oblike lem *bloger* in *blogar* |
| ! | negacija ("ne") | `[tag="S.*" & word!="@.*"]` | poišči vse samostalnike, ki se ne začnejo z znakom @ |
| (?i) | poljubna velikost črk | `[word="(?i)ozn"]` | poišči vse besedne oblike ozn, ne glede na zapis |

# CQL: (nekateri) posebni operatorji

| operator | pomen | primer iskanja | opis iskanja |
|---|---|---|---|
| **[]** | katerakoli beseda | `[word="tistega"] []  [word="dne"]` | poišče vse zveze besed *tistega* in *dne*, med kateri je vrinjena še ena beseda |
| **{k,n}** | interval (med k in n pojavitev znaka, črke, besede) | `[tag="P.*"]{3,5} [tag="S.*"]`<br>`{0,3} – med nič in tri`<br>`{3} – točno tri`<br>`{3,} – vsaj tri` | poišči samostalnike, pred katerimi je od tri do pet zaporednih pridevnikov |
| **<s> </s>** | začetek ali konec stavka | `<s> [lemma="živjo"]` | stavki, ki se začnejo s pozdravom *živjo* |
| **within <s/>** | išči znotraj stavka | `[word="tistega"] []*`<br>`[word="dne"] within <s/>` | v stavku poišči vse zveze besed tistega in dne, med katerima je vrinjeno poljubno število besed |

# Vaja za utrditev **vrste iskanj (CQL)**

Ena izmed značilnosti spletnega jezika je tudi podaljševanje besed z nizanjem enakih črk (npr. *doooolgčas*). Poskusimo poiskati besede, ki vsebujejo vsaj 3 zaporedne samoglasnike.

1. Najprej poiščimo besede, ki **vsebujejo črko a.**
   `[word=".*a.*"]`
2. Najprej poiščimo besede, ki vsebujejo **tri ali več zaporednih črk a**.
   `[word=".*a{3,}.*"]`
3. Seveda nas zanimajo **tako male kot velike črke**.
   `[word="(?i).*a{3,}.*"]`
4. **Izločimo še imena** tviterašev.
   `[word="(?i).*a{3,}.*" & word!="@.*"]`
5. Zdaj pa iskanje razširimo še na vse **druge samoglasnike.**
   `[word="(?i).*(a{3,}|e{3,}|i{3,}|o{3,}|u{3,}).*" & word!="@.*"]`
   tudi: `[word="(?i).*a{3,}.*" & word!="@.*"]|[word="(?i).*e{3,}.*" & word!="@.*"]` itd.
6. Ali se take besede pogosteje pojavljajo v **tvitih** ali na **blogih**?
   `Lastnosti besedil: blog = 89.8/milj.; tviter = 258.8/milj.`

# KONKORDANCE

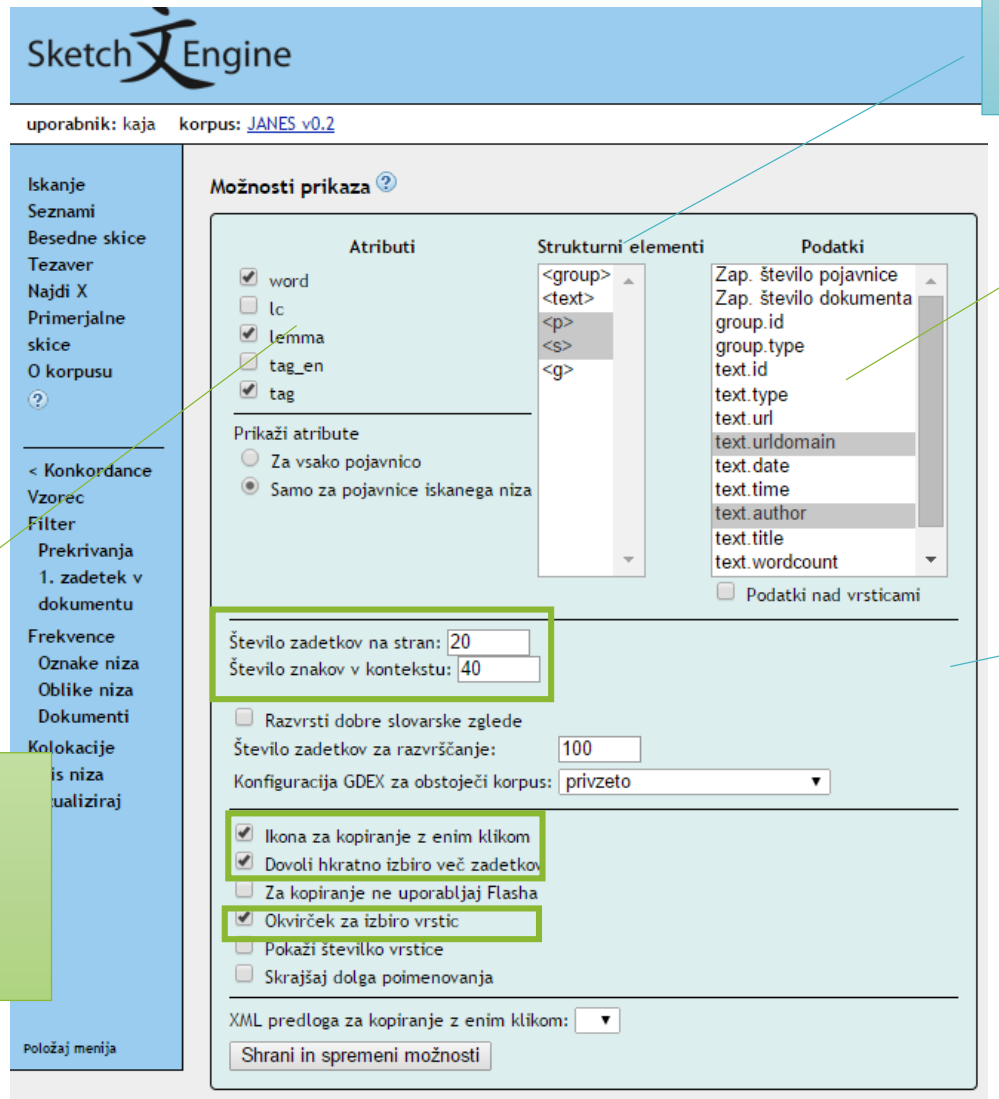
# Konkordančni niz

# Konkordance: Možnosti prikaza

Možnosti prikaza
Usredinjeno
Stavek

Privzete funkcije omogočajo **preklapljanje med usredinjenim in stavčnim prikazom** konkordanc.

V polju **atributi** določimo prikazovanje normalizirane oblike, osnovne oblike ali oblikoskladenjskih oznak.




Strukturni elementi: prikazovanje oznak za stavke in odstavke

V polju **podatki** določimo prikazovanje lastnosti besedil (za več lastnosti držimo CTRL)

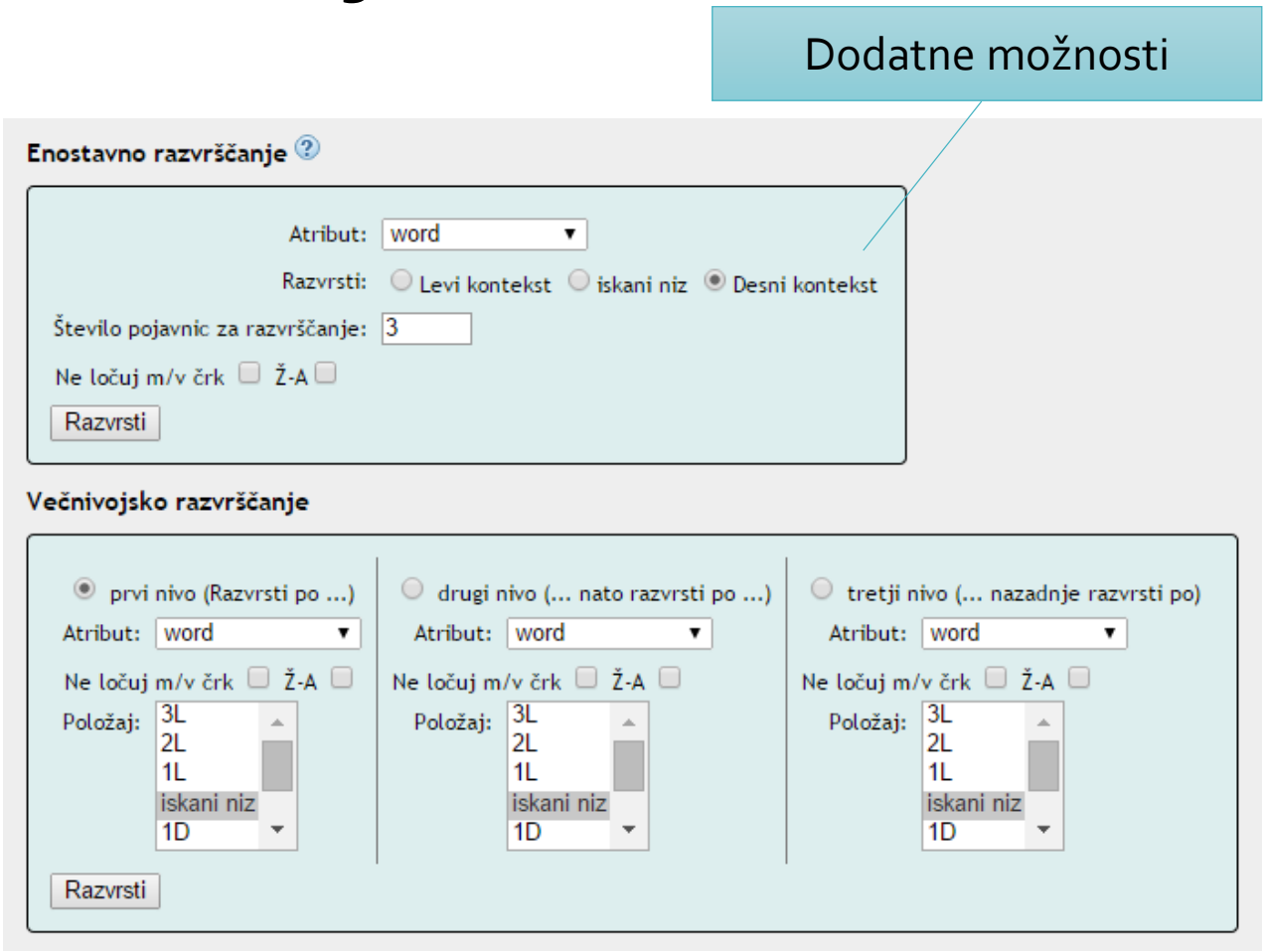
Druge možnosti

# Konkordance: Razvrščanje

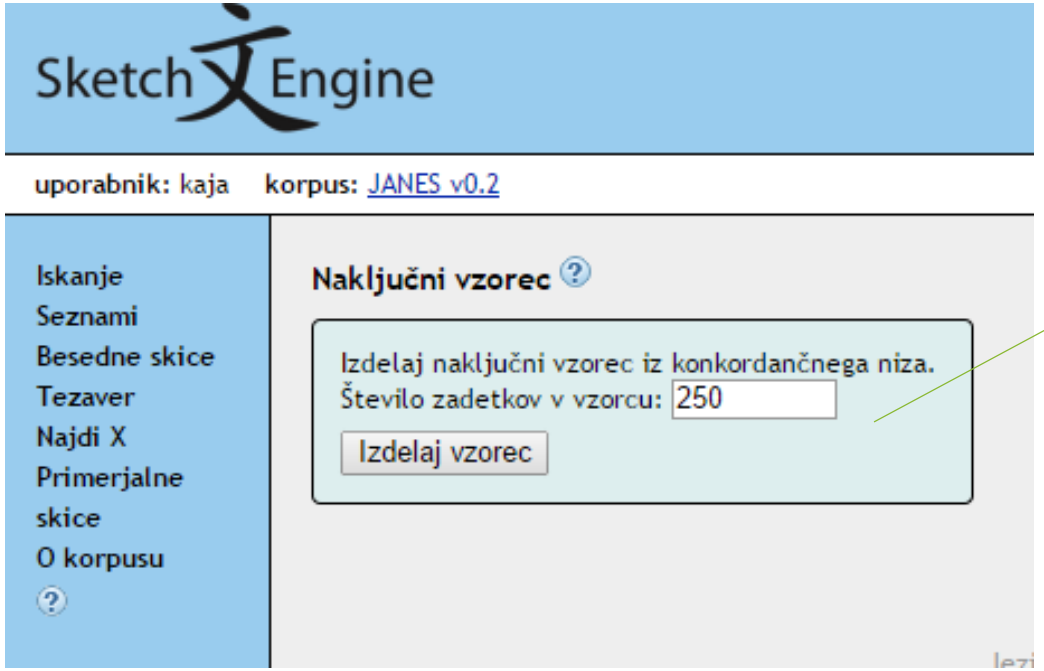
Razvrščanje
- po levi
- po desni
- iskani niz
- Podatki
- Premešaj

Privzete funkcije omogočajo razvrščanje glede na:
- prvo besedo na levi
- prvo besedo na desni
- obliko iskanega niza
- lastnosti besedila
- naključno premešaj

Dodatne možnosti



# Konkordance: Vzorec




Določimo **število konkordanc v naključnem vzorcu** zadetkov iz konkordančnega niza.

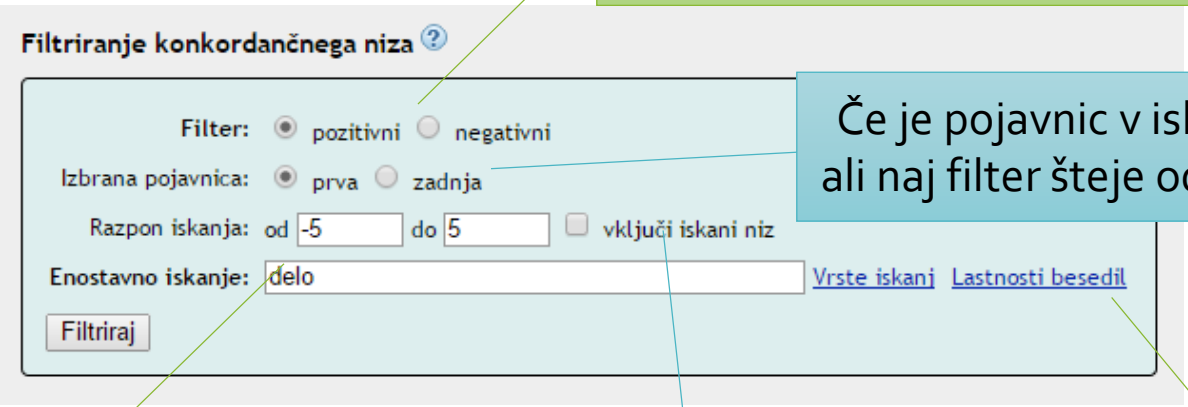
Naključni vzorec danega konkordančnega niza je **vedno enak.**

# Konkordance: Filter

Filter
Prekrivanja
1. zadetek v dokumentu

Privzeti funkciji filtrirata:
- ujemajoče se vrstice
- prvi zadetek v vsakem dokumentu

V polju **filter** določimo, se iskani pogoj mora (pozitivni filter) ali ne sme pojaviti (negativni filter) v danem kontekstu.




Če je pojavnic v iskanem nizu več, določimo, ali naj filter šteje od prve ali zadnje pojavnice.

Vnesemo besedo ali besedno zvezo, po kateri želimo filtrirati zadetke.

Pri filtriranju lahko tako kot pri možnostih iskanja uporabimo **enostavno** ali **napredno** iskanje ter **lastnosti besedil**.

Običajno iskalnega niza ne vključujemo (filtriramo zgolj glede na okolico).

# Konkordance: Frekvence

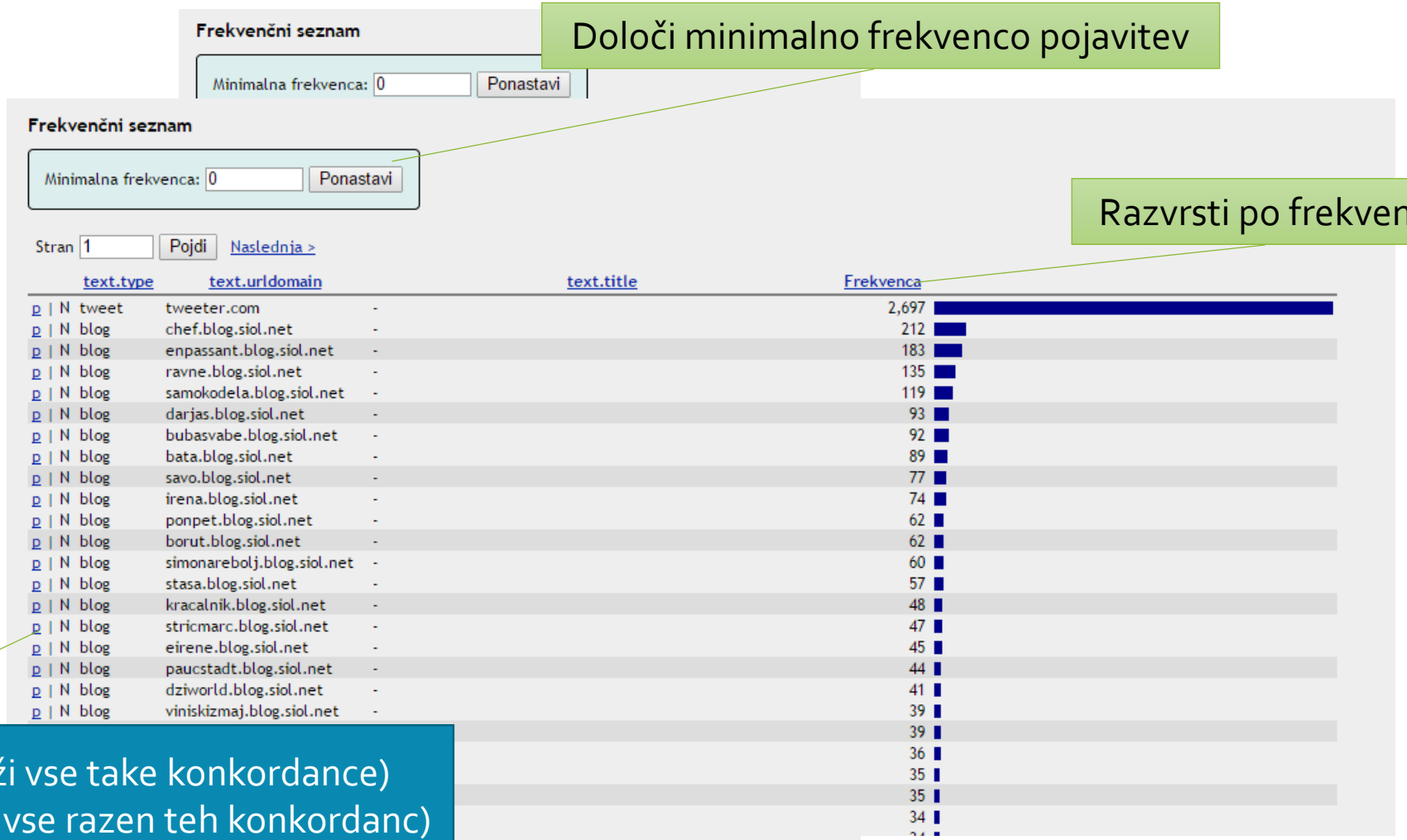
Frekvence
- Oznake niza
- Oblike niza
- Dokumenti

Določi minimalno frekvenco pojavitev

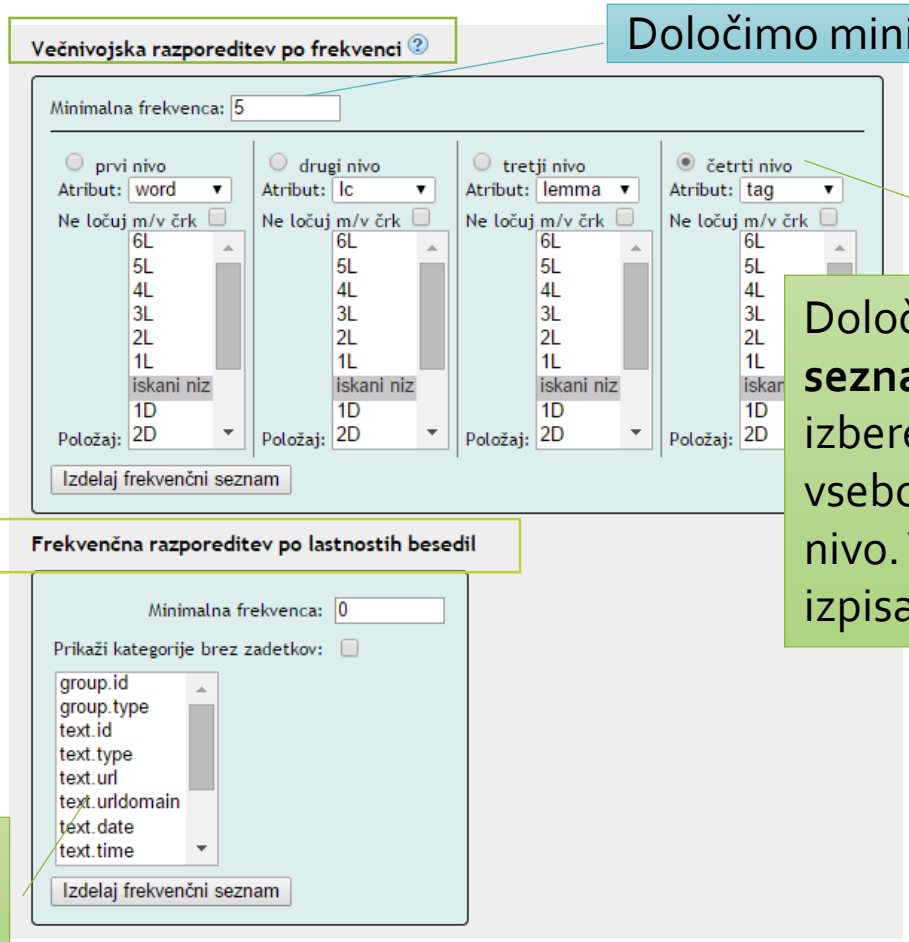
Razvrsti po frekvenci

klik do konkordanc

p: pozitivni filter (pokaži vse take konkordance)
N: negativni filter (pokaži vse razen teh konkordanc)



# Konkordance: Frekvence

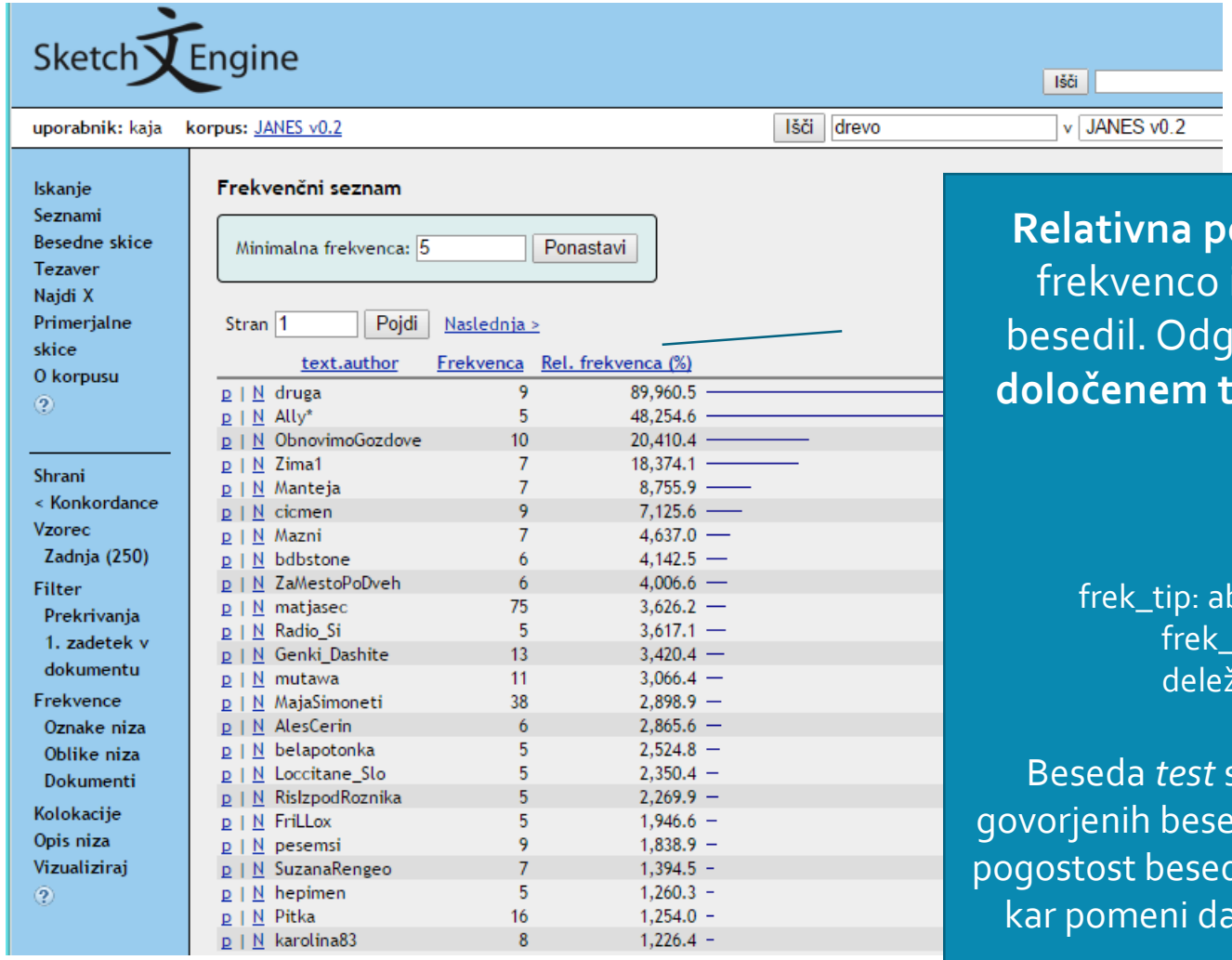



Določimo minimalno frekvenco.

Določimo do štiri **nivoje izpisa seznama** in nastavitve za izpis. Če izberemo npr. drugi nivo, bo seznam vseboval podatke za prvi in drugi nivo. Vsak nivo bo na seznamu izpisan v svojem stolpcu.

Frekvence lahko opazujemo tudi glede na **vrsto besedila.**

# Konkordance: Frekvence




**Relativna pogostost [%]** označuje razmerje med absolutno frekvenco iskanega niza in velikostjo obravnavanega tipa besedil. Odgovarja na vprašanje, **kako pogost je iskani niz v določenem tipu besedila v primerjavi s celotnim korpusom**.

Izračunamo jo tako:

**(frek_tip / frek_korpus) / delež**

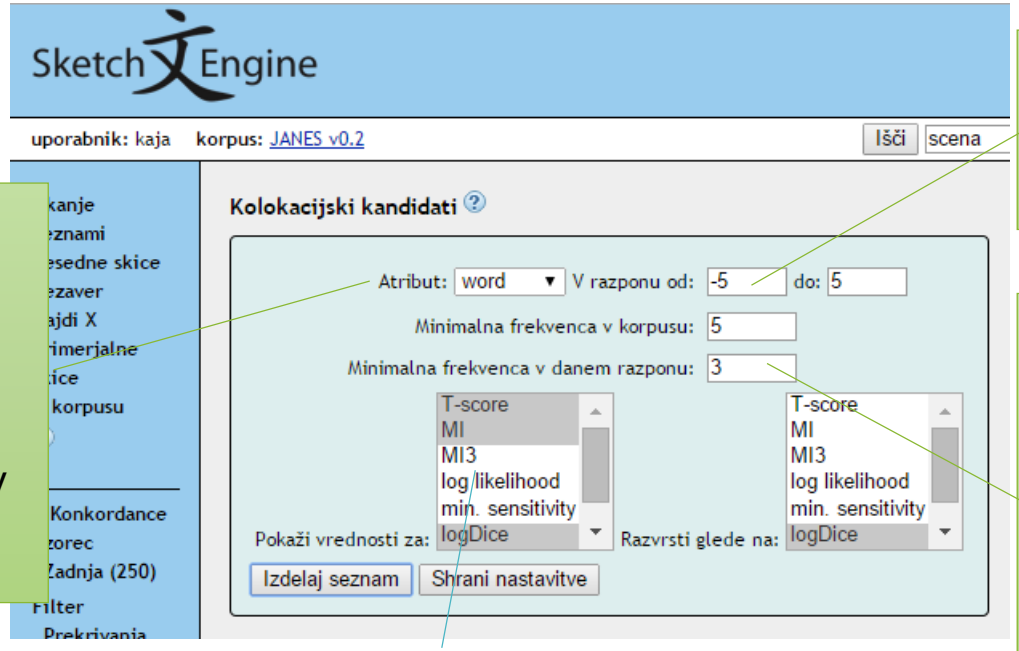
frek_tip: absolutna pogostost iskanega niza v izbranem tipu besedil
frek_korp: absolutna pogostost iskanega niza v korpusu
delež = delež izbranega tipa besedil v celotnem korpusu

Beseda *test* se npr. v korpusu pojavi 2000-krat, od tega 400-krat v govorjenih besedilih, ki predstavljajo 10 % celotnega korpusa. Relativna pogostost besede *test* v tem tipu besedil je torej (400/2000)/0.1 = 200 %, kar pomeni da se beseda *test* v govorjenih besedilih pojavlja dvakrat pogosteje kot v celotnem korpusu.

# Konkordance: Kolokacije

Funkcijo **Kolokacije** uporabimo za izdelavo seznama besed, katerih frekvenca pojavljanja v okolici iskanega niza je statistično pomembna.

**Izberemo atribut** kolokacijskih kandidatov glede na to, ali nas zanimajo zapisane besedne oblike, standardizirane besedne oblike, leme ali oblikoskladenjske oznake v okolici našega iskanega niza.




Določimo **razpon okolice,** v kateri iščemo kolokacijske kandidate.

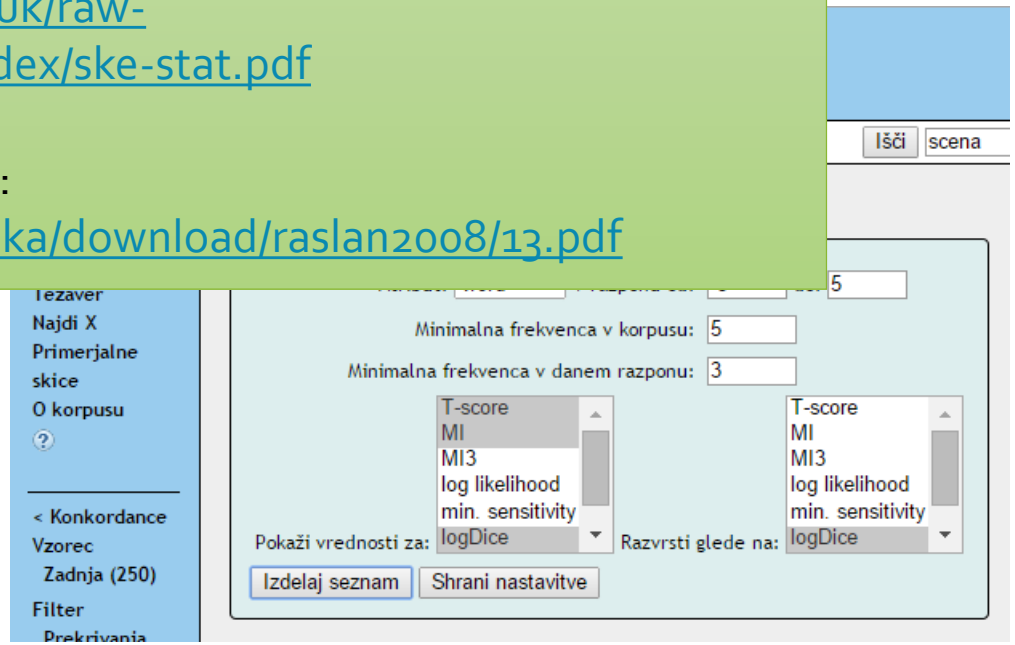
Določimo **minimalno pogostost kolokacijskega kandidata** (kolokatorja) v korpusu in minimalno pogostost kolokacijskega kandidata v danem razponu (relaciji).

Izberemo lahko različne statistike za izpis ter statistiko, po kateri naj bodo rezultati razvrščeni (privzeto logDice).

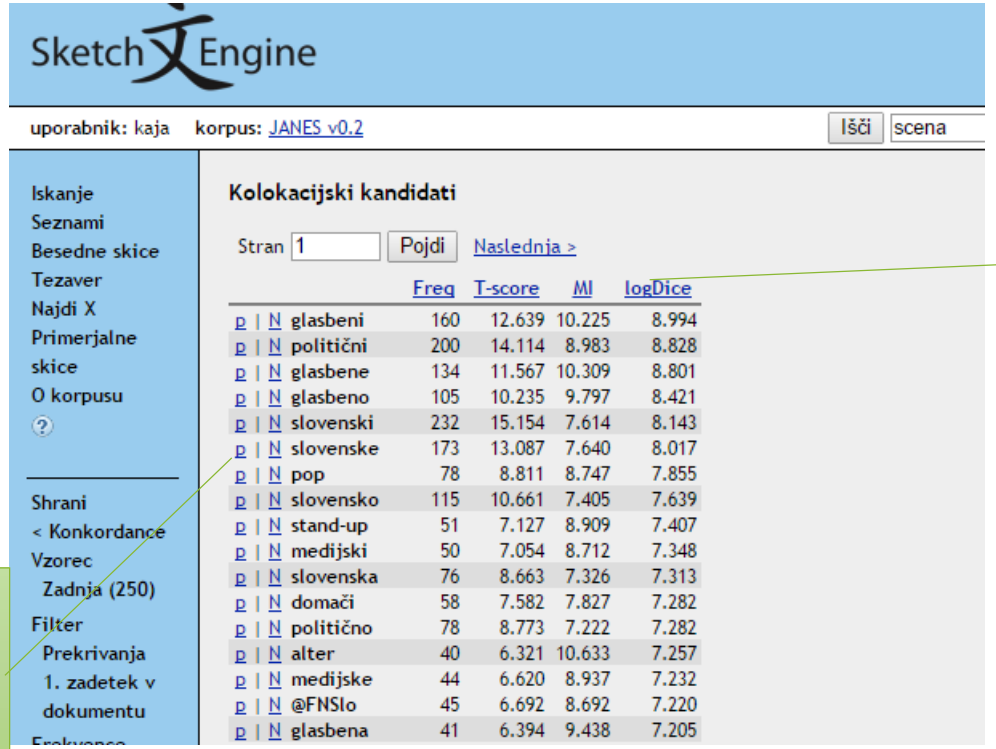
# Konkordance: Kolokacije

Več o …
- statističnih vrednostih:
https://trac.sketchengine.co.uk/raw-attachment/wiki/SkE/DocsIndex/ske-stat.pdf

- statistični vrednosti logDice:
http://www.fi.muni.cz/usr/sojka/download/raslan2008/13.pdf



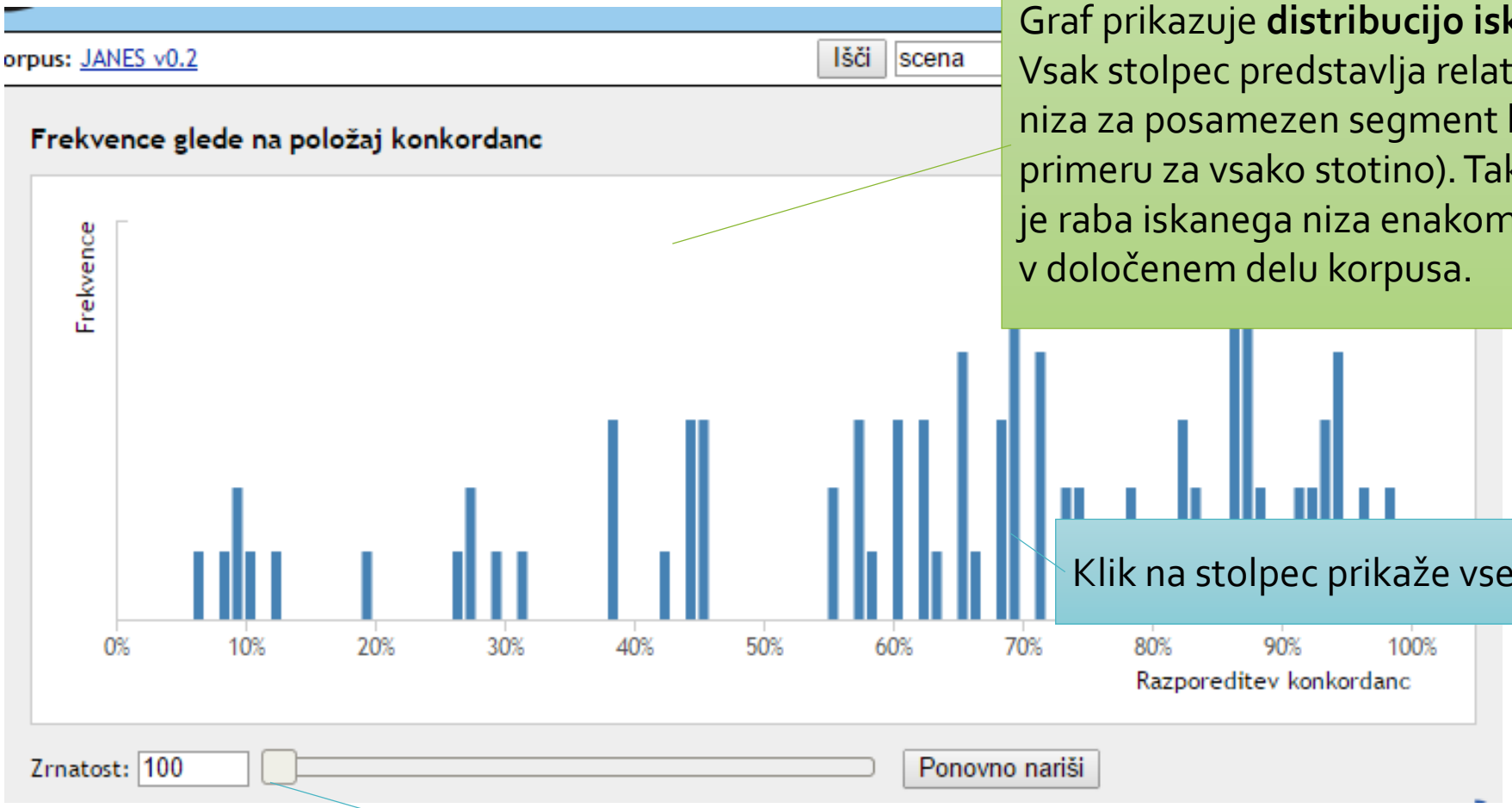
# Konkordance: Kolokacije




S klikom na naslov stolpca **razvrščamo** po poljubni statistični vrednosti.

S klikom na p si ogledamo **primere rabe** kolokacij.

# Konkordance: Vizualiziraj




Graf prikazuje **distribucijo iskanega niza v korpusu**. Vsak stolpec predstavlja relativno frekvenco iskanega niza za posamezen segment korpusa (v prikazanem primeru za vsako stotino). Tak graf nam takoj pokaže, ali je raba iskanega niza enakomerna ali pa se pojavlja samo v določenem delu korpusa.

Klik na stolpec prikaže vse konkordance tega segmenta.

Poljubno določamo število segmentov. Korpus je privzeto razdeljen na 100 enakomernih delov (100 stolpcev).

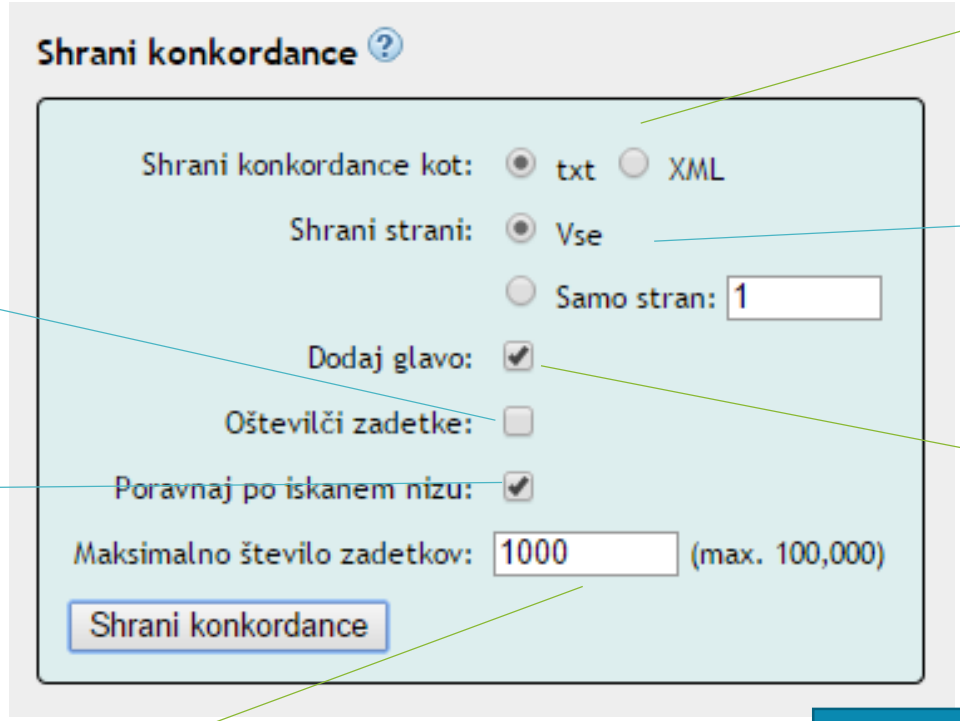
# Konkordance: Shranjevanje

Poleg konkordančnih nizov lahko na podoben način shranjujemo tudi sezname, kolokacije ipd.

Vsaki konkordanci pripiši zaporedno številko.

Usredinjen ali stavčni izpis.

**Število shranjenih zadetkov**. Za vse vpiši čim večjo številko, npr. 1000000.




Izbira formata: golo besedilo (**txt**) ali **XML**

Shrani vse konkordance ali samo določeno stran

Shrani tudi **dodatne informacije** (o korpusu, iskalnem pogoju ipd.)

Poleg besedila v konkordancah se shranijo še tisti podatki, ki so glede na izbrane možnosti prikaza prikazani v konkordančnem nizu (npr. podatki o besedilu, lema ipd.)

# Vaje za utrditev konkordanc

1. Za poimenovanje spletne predmetnosti se je v slovenščini v zadnjem času pojavilo veliko novih besed in njihovih tvorjenk. Poišči vse besede, ki se začnejo na *blog-*, in izdelaj seznam njihovih lem.

```
Iskanje: [word="blog.*"]
Frekvence >> Prvi nivo: lemma+iskani niz >> Izdelaj frekvenčni seznam
```

2. Izberi si poljubno besedo ali besedno zvezo.

- Ali se značilneje uporablja v katerem izmed tipov besedil?

```
Frekvence >> text.type >> Izdelaj frekvenčni seznam
```

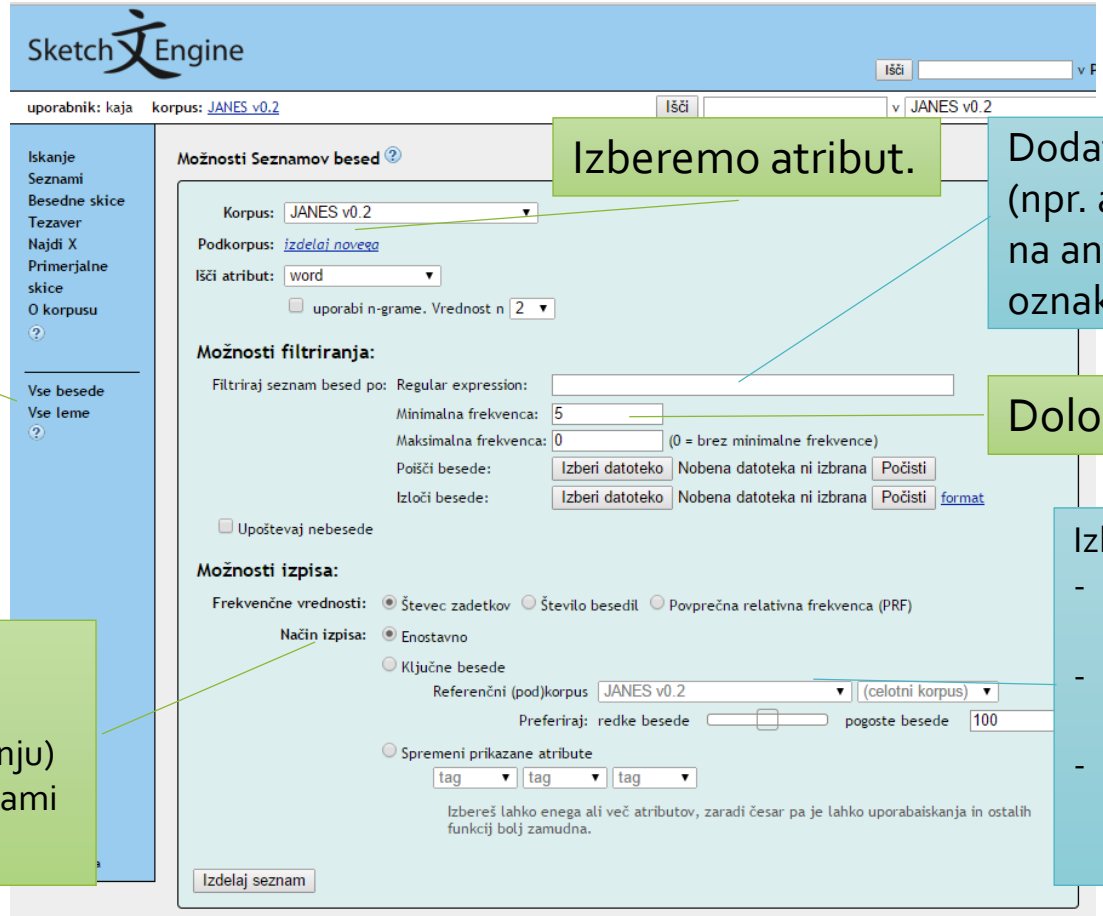
- Katera je najpogostejša lema na njeni levi?

```
Frekvence >> Prvi nivo: lema + 1L >> Izdelaj frekvenčni seznam
```

# SEZNAMI

# Seznami

Že **pripravljena seznama vseh besed in vseh lem,** ki se v korpusu pojavijo vsaj 5-krat.




Izberemo atribut.

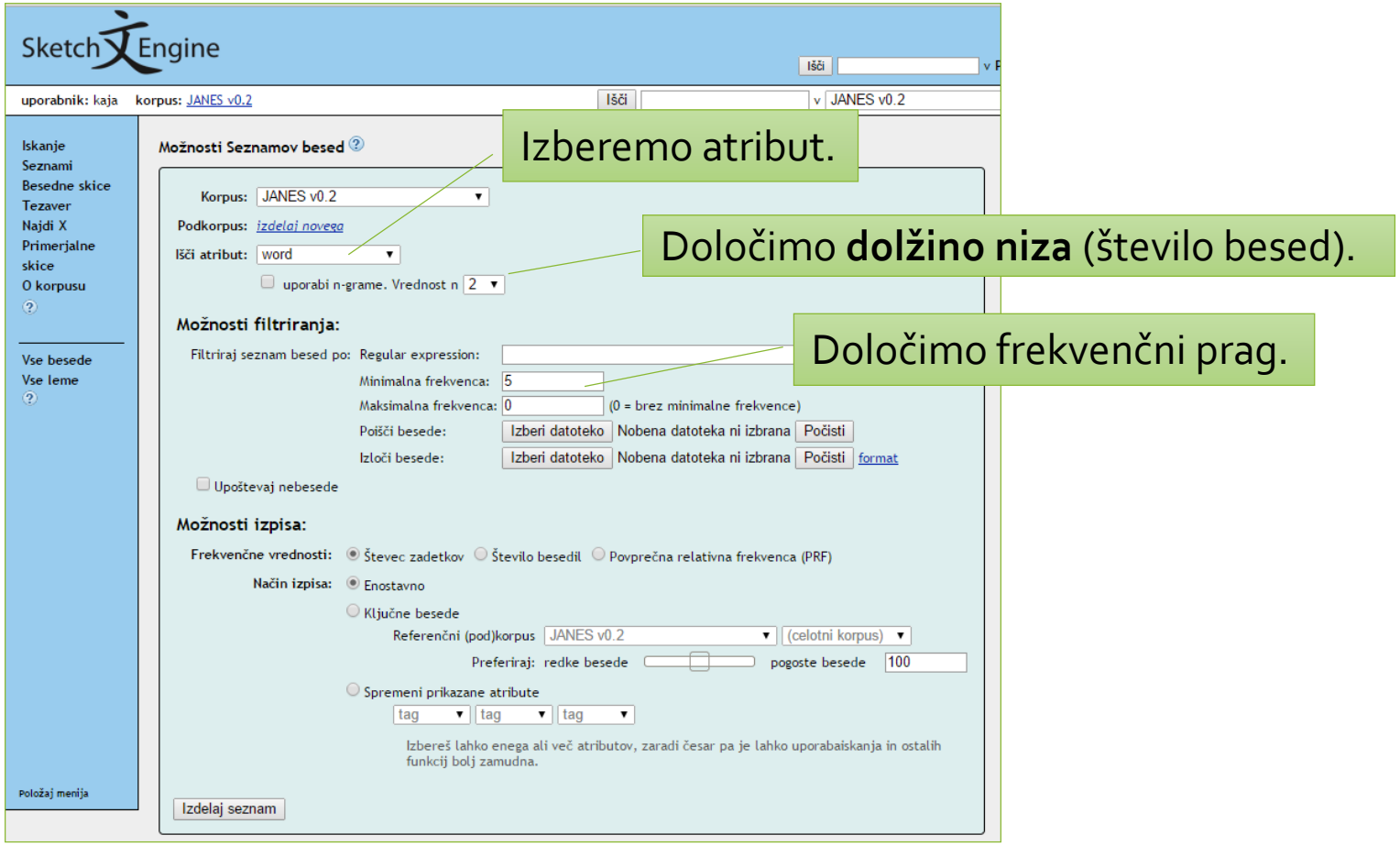
Dodatni pogoji v obliki regularnega izraza (npr. anti.* za seznam besed, ki se začnejo na anti-, S.* za seznam vseh glagolskih oznak itd.).

Določimo frekvenčni prag.

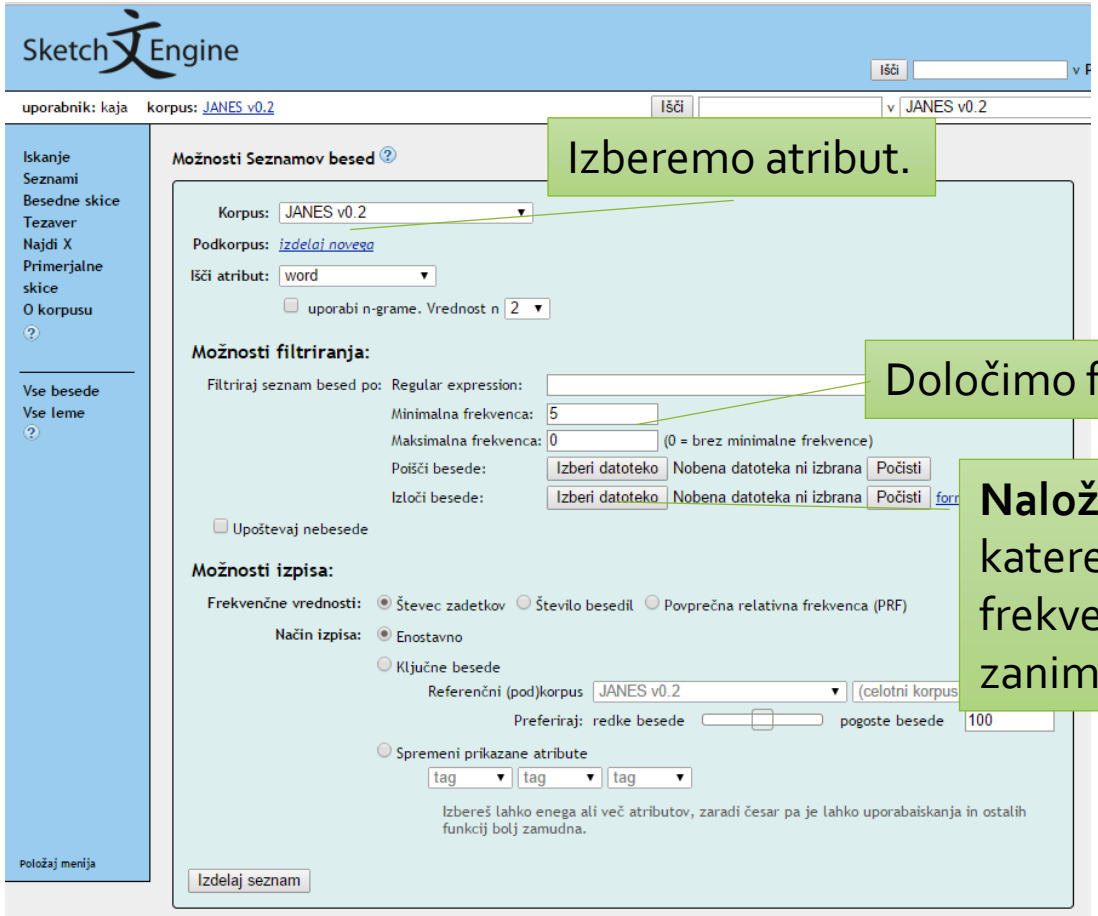
Izberemo želeno frekvenčno vrednost:
- <u>števec zadetkov</u>: kolikokrat se besede pojavijo v korpusu
- <u>število besedil</u>: v koliko različnih besedilih se besede pojavijo
- <u>PRF:</u> varianta frekvenčnega seznama, ki ne šteje pojavitev iste besede, ki se pojavljajo skupaj, npr. v istem dokumentu

Izberemo način izpisa:
- <u>enostavno</u>: iskani atribut s frekvenco
- <u>ključne besede</u>: (podrobneje v nadaljevanju)
- <u>večnivojsko:</u> izpis z dodatnimi informacijami o iskanem atributu (npr. lema in msd)

# Seznami: besedni nizi



# Seznami: poišči ali izloči besede

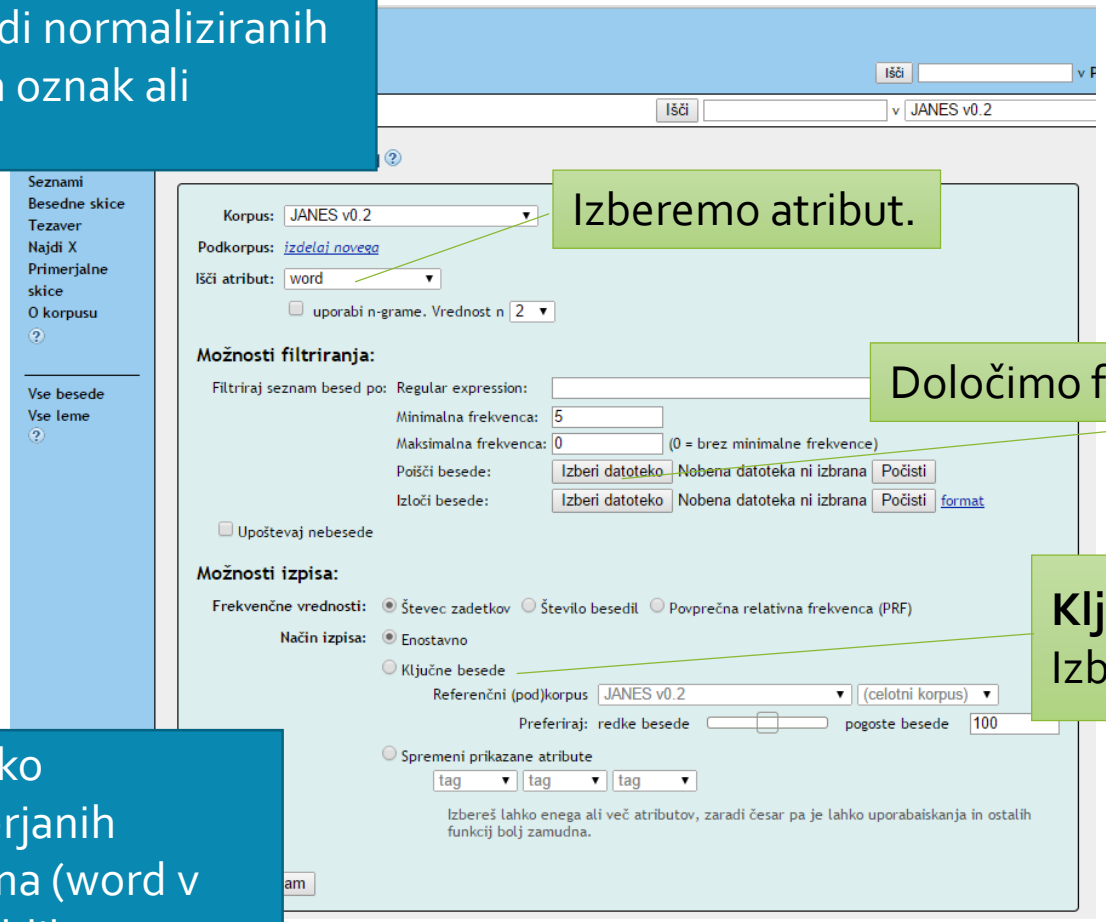



Seznam besed mora biti v navadni tekstovni datoteki (.txt) v kodiranju UTF-8, pri čemer je vsaka beseda v svoji vrstici. Besede na seznamu se morajo ujemati z izbranim atributom, npr. če je na seznamu atributov izbrana 'lema', mora seznam vsebovati leme.

**Naložimo datoteko** z besedami, za katere želimo dobiti podatek o frekvencah (poišči) ali pa nas ne zanimajo (izloči).

# Seznami: ključno besedišče

S funkcijo ključnih besed lahko izdelamo seznam ključnih besed kot tudi normaliziranih oblik, lem, oblikoskladenjskih oznak ali besednih nizov.




Izberemo atribut.

Določimo frekvenčni prag.

**Ključne besede:**
Izberemo referenčni korpus (npr. Kres)

Pozor! Orodje vedno primerja enako poimenovani atribut v obeh primerjanih korpusih, ki pa nista nujno prekrivna (word v JANES-u npr. označuje zapisano obliko, v korpusu GOS pa pogovorno obliko).

# Seznami: ključno besedišče

**Možnost preklopa** (ključnost besedišča v KRES-u glede na JANES)

**Seznam besed**

Korpus: JANES v0.2

Referenčni korpus: Kres
Spremeni cilj in referenčni (pod)korpus

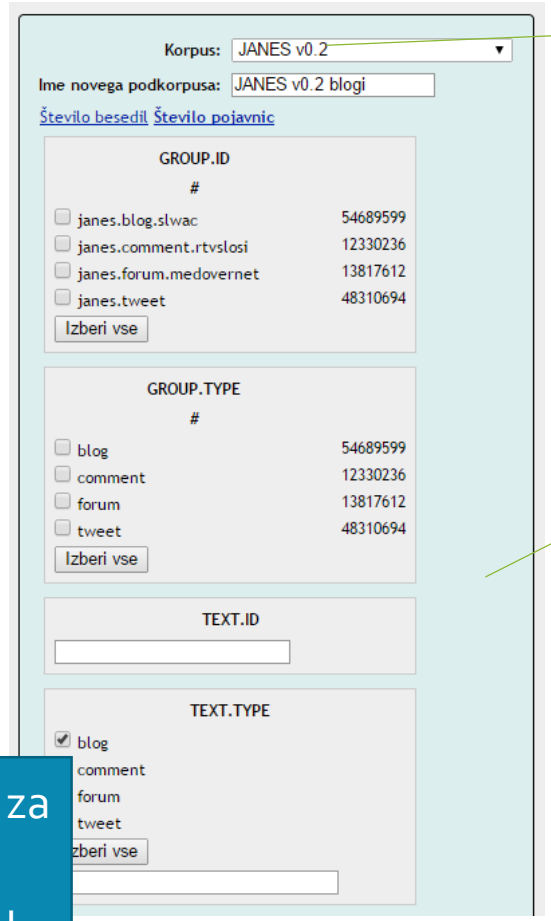
Stran 1 Pojdi Naslednja >

| | *JANES v0.2* | | *Kres* | | |
|---|---|---|---|---|---|
| lemma | Frekvenca | Frekvenca/mil | Frekvenca | Frekvenca/mil | Rezultat |
| slovenija | 129,490 | 1002.6 | 1,862 | 15.5 | 9.6 |
| ljubljana | 46,228 | 357.9 | 284 | 2.4 | 4.5 |
| janša | 39,366 | 304.8 | 267 | 2.2 | 4.0 |
| edini | 34,487 | 267.0 | 0 | 0.0 | 3.7 |
| tale | 79,971 | 619.2 | 12,145 | 100.8 | 3.6 |
| slovenec | 35,640 | 276.0 | 950 | 7.9 | 3.5 |
| blog | 32,889 | 254.7 | 1,155 | 9.6 | 3.2 |
| mogoče | 62,146 | 481.2 | 9,981 | 82.9 | 3.2 |
| via | 28,345 | 219.5 | 143 | 1.2 | 3.2 |
| hvala | 77,058 | 596.7 | 15,133 | 125.6 | 3.1 |
| sds | 26,652 | 206.4 | 78 | 0.6 | 3.0 |
| itak | 37,507 | 290.4 | 3,701 | 30.7 | 3.0 |
| slo | 26,020 | 201.5 | 314 | 2.6 | 2.9 |
| rabiti | 35,734 | 276.7 | 3,705 | 30.8 | 2.9 |
| levi | 22,490 | 174.1 | 61 | 0.5 | 2.7 |
| super | 34,191 | 264.7 | 4,080 | 33.9 | 2.7 |
| pozdravljen | 26,126 | 202.3 | 1,614 | 13.4 | 2.7 |
| maribor | 22,114 | 171.2 | 211 | 1.8 | 2.7 |
| veliko | 244,634 | 1894.2 | 78,131 | 648.7 | 2.7 |
| bravo | 24,111 | 186.7 | 1,036 | 8.6 | 2.6 |
| res | 207,559 | 1607.1 | 66,282 | 550.3 | 2.6 |
| ampak | 214,071 | 1657.6 | 69,091 | 573.6 | 2.6 |
| danes | 164,189 | 1271.3 | 51,454 | 427.2 | 2.6 |
| evropa | 21,357 | 165.4 | 467 | 3.9 | 2.6 |
| janez | 20,023 | 155.0 | 144 | 1.2 | 2.5 |
| desni | 19,444 | 150.6 | 0 | 0.0 | 2.5 |
| pač | 93,142 | 721.2 | 27,681 | 229.8 | 2.5 |
| http | 19,988 | 154.8 | 506 | 4.2 | 2.4 |
| lej | 19,647 | 152.1 | 411 | 3.4 | 2.4 |
| malo | 153,800 | 1190.9 | 51,892 | 430.8 | 2.4 |

Absolutna in **relativna frekvenca** v obeh korpusih.

**Izračun ključnosti.**

# Seznami: izdelava podkorpusa




Vnesemo **ime podkorpusa**.

Določimo **lastnosti besedil** v podkorpusu (npr. tip, avtor, domena ipd.)

Podkorpus se shrani v bazo in je na voljo tudi za poznejše delo. Poleg izdelave seznamov ga lahko uporabimo tudi pri iskanju oz. filtriranju konkordanc in primerjalnih skicah.

# Vaje za utrditev **seznamov**

1. Tako v govorjeni kot v spletni slovenščini je pogost izpust končnega -i pri glagolskih deležnikih na -l v moškem spolu množine (npr. *našli* → *našl*), ki smo jih zbrali v datoteki http://goo.gl/ofN1t8.* Preverimo njihovo frekvenco v korpusu JANES.

```
Seznami: atribut: word; min. frekvenca: 5; whitelist: delezniki.txt
```

2. Izberi si priljubljenega tviteraša ali blogerja. Katero je njegovo tipično besedišče v primerjavi s korpusom Kres? In katero v primerjavi z drugimi spletnimi besedili?

```
Ustvari podkorpus (filter: blog.siol.net ali @miljonar)
Seznami: atribut: lema; ključne besede: JANES
```

* Datoteka vključuje take deležnike v leksikon Sloleks, ki smo jim s funkcijo poišči-in-zamenjaj odstranili končni –i. Zaradi lažje obdelave so v resnici vključeni samo tisti, ki imajo pred končnim -li soglasnik (razen r).

# BESEDNE SKICE

# Besedne skice

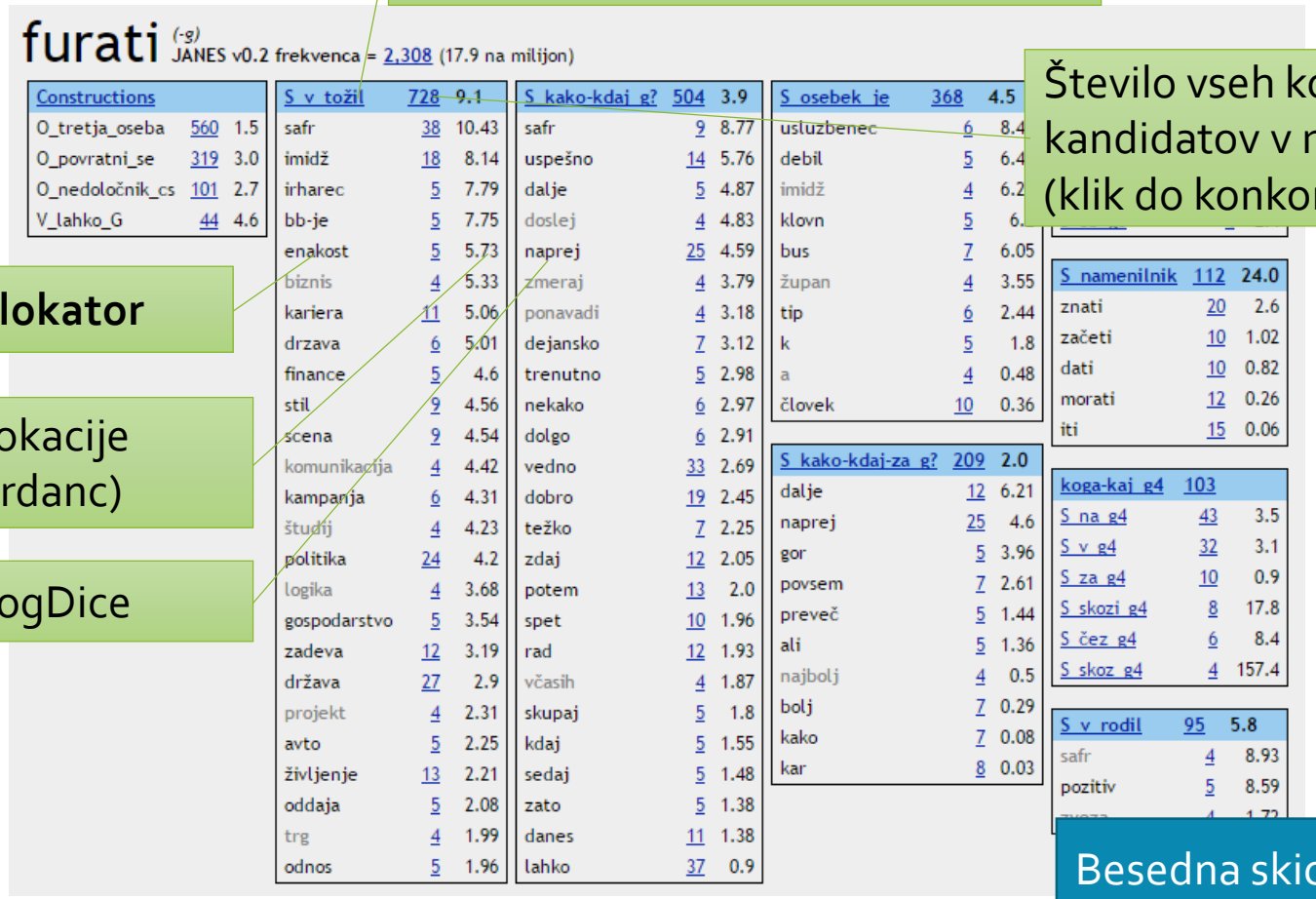
**Ime slovnične relacije**
(klik do opisa poizvedbe)

Število vseh kolokacijskih kandidatov v relaciji
(klik do konkordanc)

**Kolokator**

Pogostost kolokacije
(klik do konkordanc)

Vrednost logDice




Besedna skica je povzetek slovničnega in kolokacijskega obnašanja besede.

# Slovnica besednih skic

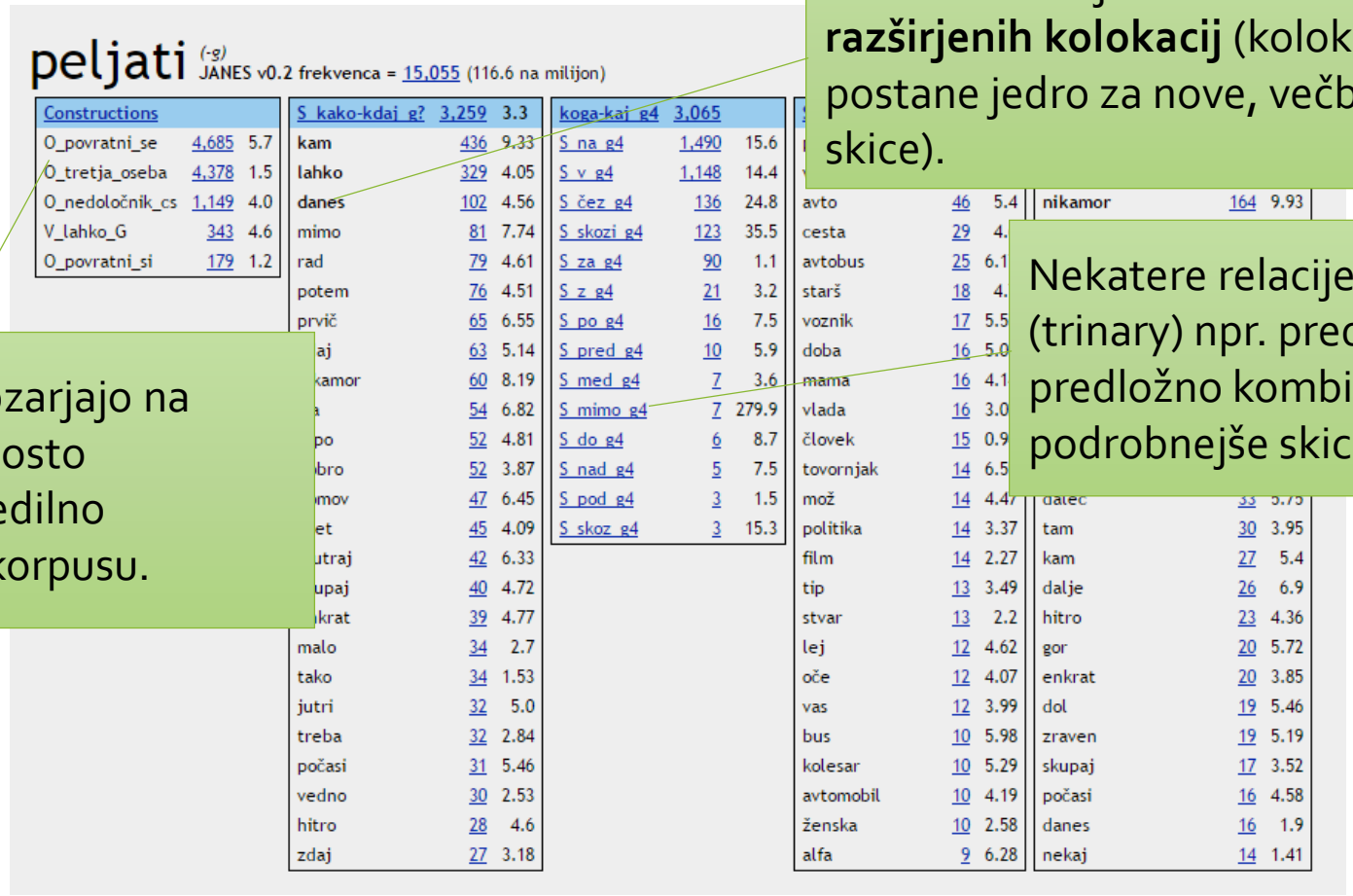
- besedna skica pokaže, katere besede (kolokatorji) se v definiranem besednem okolju povezujejo z izhodiščno lemo
- besedilno okolje je definirano s **slovnično relacijo**
- slovnično relacijo sestavljajo trije elementi:
  - ime relacije
  - tip odnosa med besedami v relaciji (direktiva)
  - poizvedba (regularni izraz)
- v.08 = **103 slovnične relacije**

*DUAL
=S_v_rodil-s/S_s-koga-česa
1:samostalnik brez_GSVD{0,2} 2:samost_rod
npr. *delovanje motorja*

| direktiva | število |
|---|---|
| SEPARATEPAGE + TRINARY | 36 |
| DUAL | 23 |
| UNARY | 2 |
| CONSTRUCTION | 13 |
| CONSTRUCTION +UNARY | 6 |
| COLLOC | 3 |
| SYMMETRIC | 2 |
| brez | 18 |
| skupaj | 103 |

Tabela 1: slovnične relacije po direktivah

# Besedne skice: posebne direktive




Klik na odebeljene kolokatorje vodi do **razširjenih kolokacij** (kolokacija postane jedro za nove, večbesedne skice).

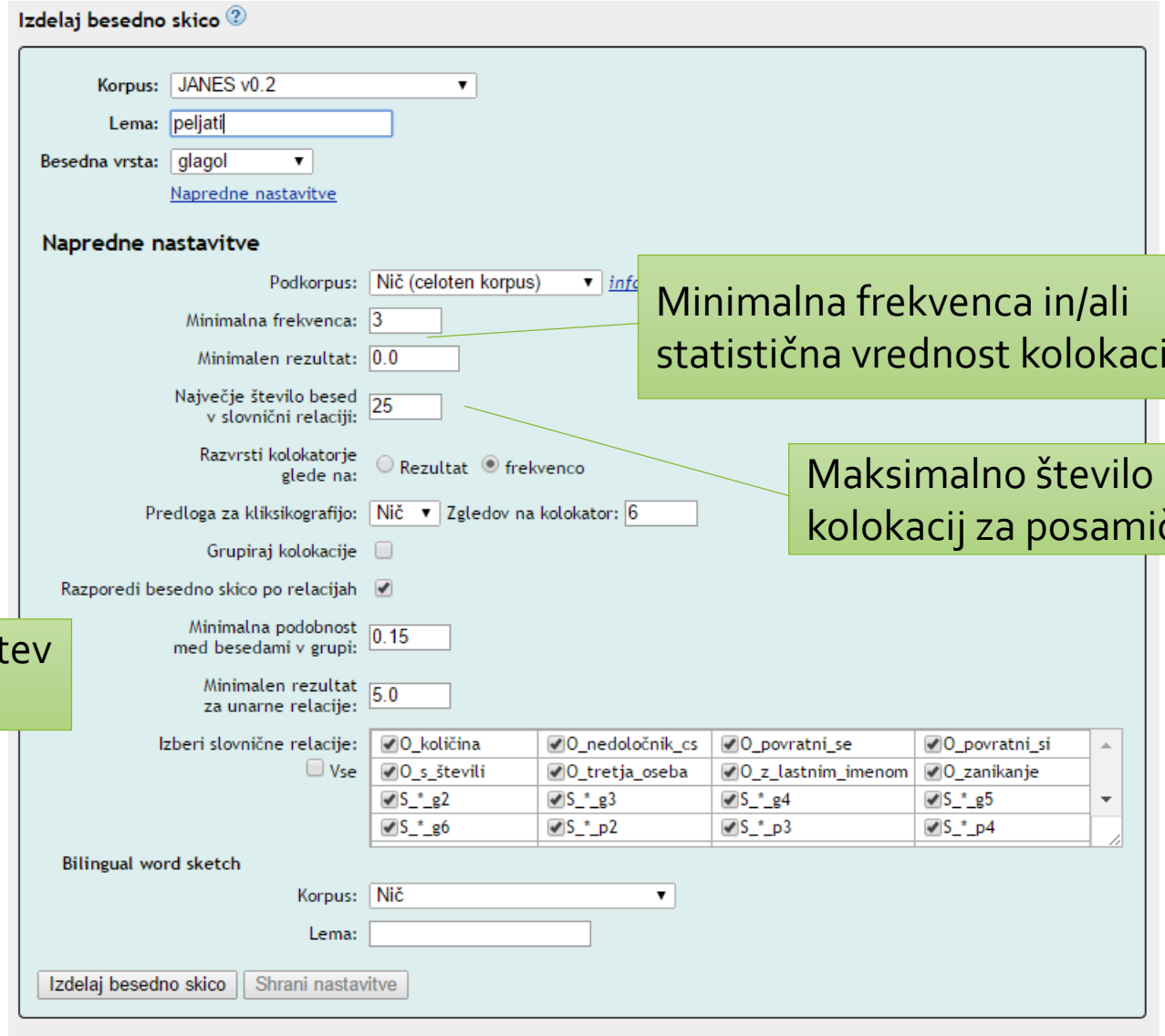
**Constructions** opozarjajo na nadpovprečno pogosto skladenjsko in besedilno obnašanje leme v korpusu.

Nekatere relacije so **tročlenske** (trinary) npr. predložne zveze. Klik na predložno kombinacijo nas pripelje do podrobnejše skice.

# Besedne skice: napredne nastavitve

Spremeni nastavitve
Grupiranje
Razvrsti
Relacije
Več podatkov
Manj podatkov

Privzete možnosti nastavitev v meniju na levi.




Minimalna frekvenca in/ali statistična vrednost kolokacije.

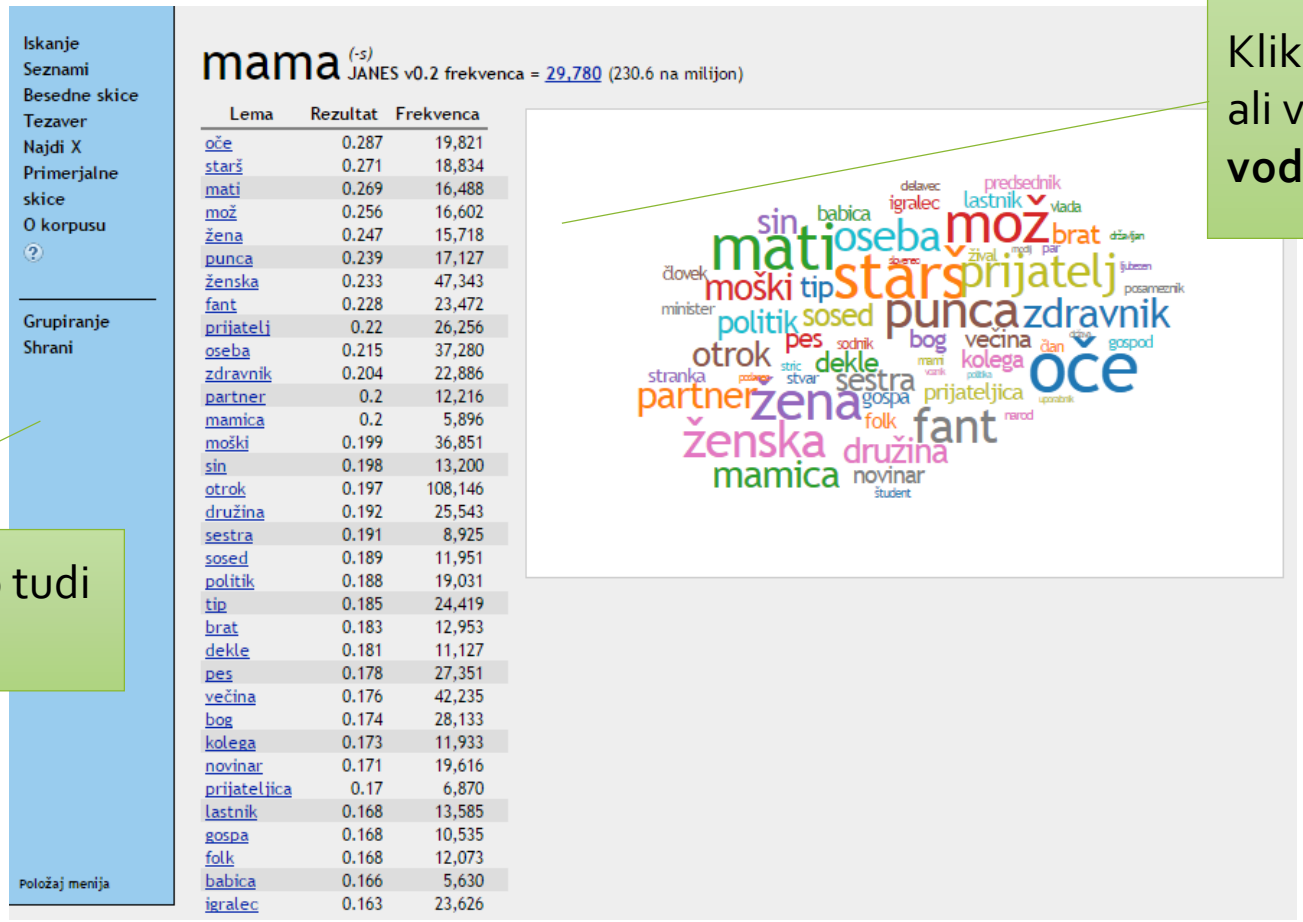
Maksimalno število prikazanih kolokacij za posamično relacijo.

# Tezaver


Določimo lemo in besedno vrsto.

Avtomatsko izdelani tezaver poišče izhodiščni lemi najbolj podobne leme glede na njihove slovnične in kolokacijske lastnosti (podoben kontekst).

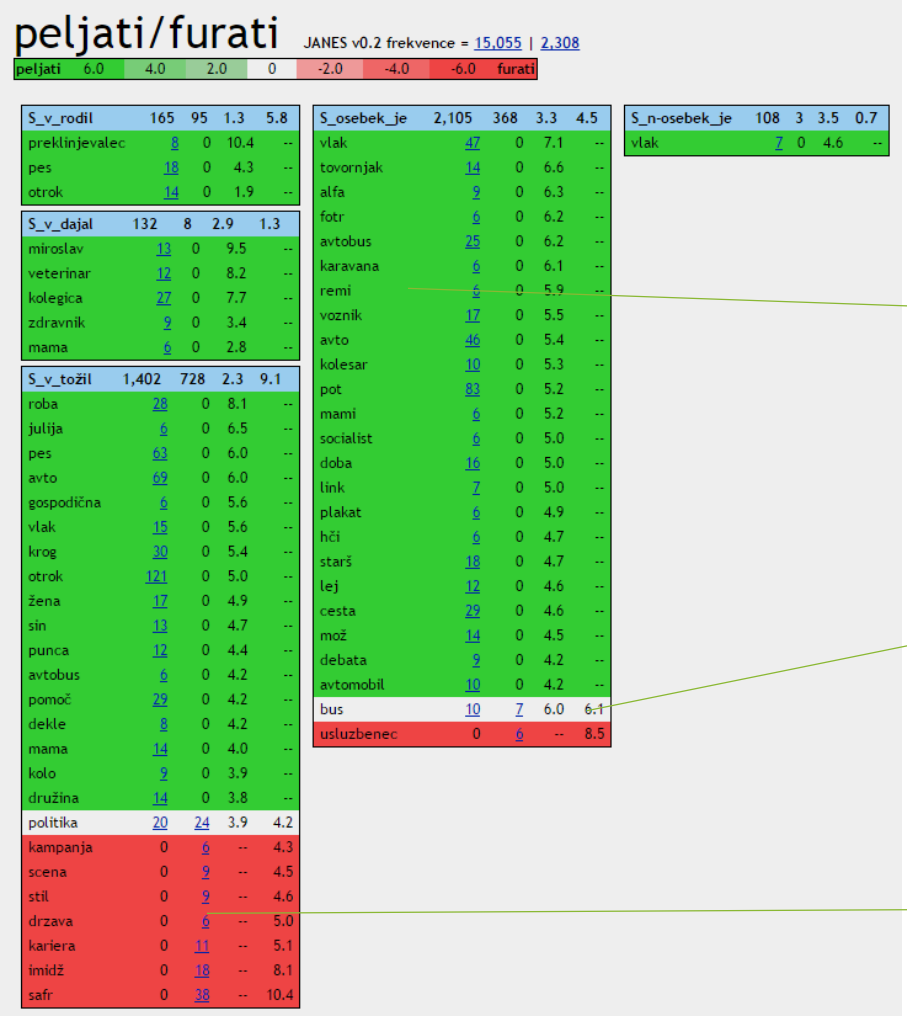
# Tezaver




Klik na besedo na seznamu ali v besednem oblaku **vodi do primerjalne skice.**

Seznam lahko razvrstimo tudi **po skupinah** podobnosti.

# Primerjalne skice

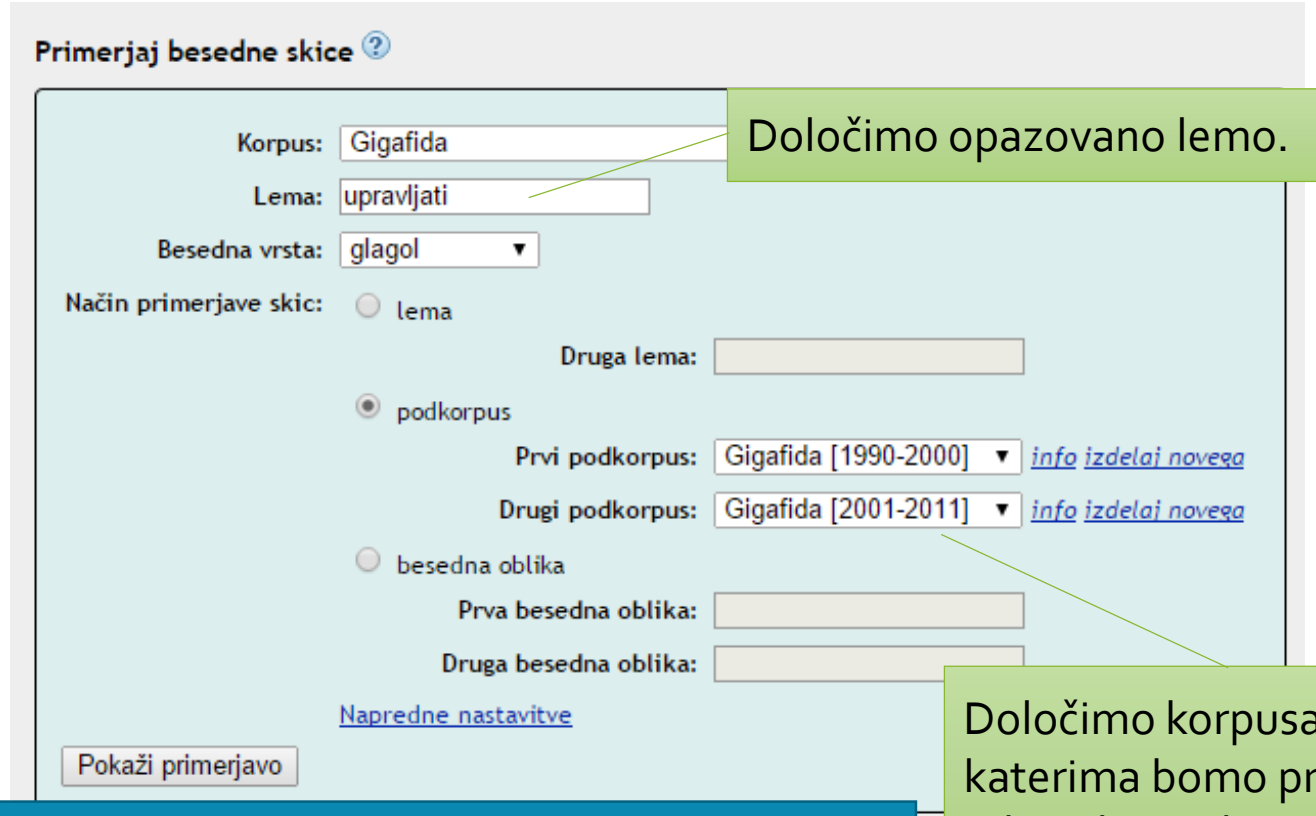



**Zeleno**: kolokacije, ki se tipično pojavljajo s prvo lemo (*peljati*).

**Neobarvano**: kolokacije, ki so značilne za obe lemi.

**Rdeče**: kolokacije, ki se tipično pojavljajo z drugo lemo (*furati*).

# Primerjalne skice




Določimo opazovano lemo.

Določimo korpusa, med katerima bomo primerjali rabo izbrane leme.

S funkcijo primerjalnih skic lahko poleg pomenske sorodnosti dveh besed detektiramo tudi pomenske spremembe besedišča v različnih časovnih obdobjih ali žanrih.

Tipične kombinacije pred letom 2000.

| S_kako-kdaj-za_g? | 367 | 1,472 | -0.8 | 1.2 |
|---|---|---|---|---|
| mehansko | 7 | 0 | 6.0 | -- |
| dzu | 21 | 26 | 6.9 | 7.2 |
| daljinsko | 6 | 8 | 5.5 | 5.9 |
| samostojno | 5 | 11 | 3.0 | 4.2 |
| ročno | 9 | 30 | 3.7 | 5.4 |
| odgovorno | 0 | 5 | -- | 4.0 |
| ločeno | 0 | 9 | -- | 4.3 |
| povečini | 0 | 9 | -- | 4.5 |
| elektronsko | 0 | 6 | -- | 4.7 |
| optimalno | 0 | 6 | -- | 4.8 |
| glasovno | 0 | 5 | -- | 5.8 |
| občinsko | 0 | 5 | -- | 5.9 |
| centralno | 0 | 6 | -- | 6.0 |
| gospodarno | 0 | 8 | -- | 6.2 |
| zpo | 0 | 6 | -- | 6.5 |

Tipične kombinacije po letu 2000.

# Vaje za utrditev besednih skic

Kaj furamo?

safr, imidž, irharce …

Skica za *furati*, relacija S_v_tožil

O čem nakladamo?

metaforah, članstvu, avtomobilih …

Skica za *nakladati*, relacija kom-čem_g5 >> skica za S_o_g5

Kakšna je glasbena scena?

slovenska, notranja, britanska, cirkusantska …

Skica za *scena*, relacija S_kakšen? >> Skica za glasben, relacija S_kakšen?

*Kako nekomu povedati, da je nadležen?

zoprn, neprijeten, moteč …

Tezaver za *nadležen*

*Ali sta duet in duo popolna sinonima?

Primerjalne skice >> Lema: duet, Druga lema: duo

# VPRAŠANJA

# Pomoč

- v angleščini:
  http://www.sketchengine.co.uk/documentation
- v slovenščini:
  http://www.trojina.org/sketchengine/
- video tutoriali:
  http://www.sketchengine.co.uk/documentation/wiki/SkE/Help/VideoHowTo

Najkoristnejši je klik na znak  , ki se običajno nahaja ob imenu funkcije.

# ZA NAVDUŠENCE

# Besedne skice: ALLP API

- API: Application Programming Interface
- ALLP: postopek avtomatskega luščenja leksikalnih podatkov (Kosem et al. 2013)
- python **skripta za luščenje kolokacij in dobrih slovarskih zgledov**
- izvoz v obliki LBS-XML (primeren za uvoz v iLex)
- trenutno za polnopomenske besede [GPRS]
- pred luščenjem določiti parametre

# ALLP API: določanje parametrov

- korpus
- lema
- slovnična relacija
- GDEX konfiguracija
- število zgledov na kolokator
- število kolokatorjev na slovnično relacijo
- minimalna frekvenca kolokatorja
- minimalna frekvenca slovnične relacije
- minimalna izpostavljenost kolokatorja (salience)
- minimalna izpostavljenost slovnične relacije (salience)

# ALLP API: osnovni parametri

```
python tblscript.py janes -U http://sketch.fri1.uni-lj.si/bonito/ -u janes -p 9neztandart6 -l lemmalist.txt -r gramrellist.txt
```

Seznam lempos-ov (.txt)

```
dan-s
priden-p
hitro-r
imeti-g
```
itd.

Seznam slovničnih relacij s parametri (.txt)

```
S_p-koga-česa 8   0.5 10  3.0 S
S_p-komu-čemu 5   0.5 10  1.0 S
S_koga-kaj 8   0.5 100 0.5 S
O_z_lastnim_imenom   8   0.5 8   2.5 O
```
itd.

Parametri po stolpcih:

1. min. frekvenca kolokatorja
2. min .izpostavljenost kolokatorja
3. min. frekvenca relacije
4. min. izpostavljenost relacije
5. Vrsta relacije: skladenjska struktura (S), oznaka (O), vzorci (V), skladenjska zveza (Z)

Potrebuješ:

- python v2.7
- knjižnico httplib
- knjižnico simplejson

# ALLP API: parametri in privzete vrednosti

```
parser.add_option("-l", "--lemmalist", dest="lemmalist", default="",
parser.add_option("-r", "--relations", dest="gramrellist", default="",
parser.add_option("-f", "--frequency", dest="minfreq", default="0",
parser.add_option("-s", "--salience", dest="minsal", default="0.0",
parser.add_option("-F", "--Freqrel", dest="minfreqrel", default="25",
parser.add_option("-S", "--Salrel", dest="minsalrel", default="0.0",
parser.add_option("-n", "--number", dest="number", default="6",
parser.add_option("-m", "--maxCollocs", dest="maxitems", default="10",
parser.add_option("-g", "--gdexconf", dest="gdexconf", default="",
parser.add_option("-H", "--Hierarchical", dest="hierarchical", default=False,
parser.add_option("-U", "--URL", dest="url", default=SKE_URL,
parser.add_option("-u", "--username", dest="username", default=None)
parser.add_option("-p", "--password", dest="password", default=None)
parser.add_option("-d", "--debug", action="store_true",
parser.add_option("-t", "--template", dest="urltmpl", default=
parser.add_option("-C", "--CA_auth", dest="loginpage", default='',
parser.add_option("-e", "--refs", dest="refs", default="",
parser.add_option("-z", "--gdex-size", dest="gdex_size", default="300",
```

# ALLP API: izvoz

Več o strukturi LBS in LBS XML DTD: http://www.slovenscina.eu/spletni-slovar/prenos.

umestitev v ustrezni del baze glede na vrsto relacije [SOVZ]

izpis kolokatorjev za relacijo

izpis zgledov za kolokator

```
<?xml version="1.0" encoding="UTF-8"?>
<clanek>
<glava>
<oblika>
<zapis>dan</zapis>
<iztocnica frek_lema="1471">dan</iztocnica>
</oblika>
<zaglavje>
<besvrs>samostalnik</besvrs>
<oznaka frek_gramrel="386" salience_gramrel="11.9">O_s_števili</oznaka>
</zaglavje>
</glava>
<geslo>
<pomen>
<indikator></indikator>
<pomenska_shema></pomenska_shema>
<skladenjske_skupine>
<skladenjska_struktura>
<struktura frek_gramrel="563" salience_gramrel="5.6">S_kakšen?</struktura>
<kolokacije>
<kolokacija kid="99747" frek_kol="82" salience_kol="12.0"><k>rojsten</k></kolokacija>
<kolokacija kid="99719" frek_kol="55" salience_kol="10.66"><k>cel</k></kolokacija>
<kolokacija kid="99708" frek_kol="112" salience_kol="10.5"><k>dober</k></kolokacija>
<kolokacija kid="99699" frek_kol="29" salience_kol="10.2"><k>današnji</k></kolokacija>
<kolokacija kid="99697" frek_kol="53" salience_kol="10.3"><k>lep</k></kolokacija>
<kolokacija kid="99694" frek_kol="33" salience_kol="9.98"><k>zadnji</k></kolokacija>
<kolokacija kid="99703" frek_kol="23" salience_kol="9.7"><k>naslednji</k></kolokacija>
<kolokacija kid="99723" frek_kol="8" salience_kol="8.7"><k>srečen</k></kolokacija>
<kolokacija kid="99733" frek_kol="8" salience_kol="8.59"><k>bel</k></kolokacija>
</kolokacije>

<zgledi>
<zgled kol="rojsten" kid="99747" pozicija="1" GDEX_score="9.2" >jah [n prašal ti eee kdaj si kdaj maš pa ti <k>rojstn</k> <i id="580445">d
<zgled kol="rojsten" kid="99747" pozicija="2" GDEX_score="9.19" >ker na tork je reku n [gap] eee eee mi je povedov kdaj ma <k>rojstn</k> <i id="1071765">dan</i></zgl
<zgled kol="rojsten" kid="99747" pozicija="3" GDEX_score="9.19" >ne ne pa na [name:personal] <k>rojstn</k> <i id="890633">dan</i> sem bil ta prvič</zgled>
<zgled kol="rojsten" kid="99747" pozicija="4" GDEX_score="8.84" >in potem sta mimo prišla tudi eem [name:personal] in [name:personal] eem ki sta šla v trgovino po da
<zgled kol="rojsten" kid="99747" pozicija="5" GDEX_score="8.8" >pa [name:personal] [name:surname] ima prav tako <k>rojstni</k> <i id="471918">dan</i></zgled>
<zgled kol="rojsten" kid="99747" pozicija="6" GDEX_score="8.73" >[name:personal] pa [name:surname] no pa [name:personal] [name:surname] sta mela <k>rojsn</k> <i id="
<zgled kol="cel" kid="99719" pozicija="1" GDEX_score="9.0" >ja men tud pa sem bil <k>celi</k> <i id="967248">dan</i> stari</zgled>
<zgled kol="cel" kid="99719" pozicija="2" GDEX_score="8.58" >aja sta <k>cel</k> <i id="1070608">dan</i> tle dol?</zgled>
<zgled kol="cel" kid="99719" pozicija="3" GDEX_score="8.5" >in bomo po tisti ulici se v bistvu vozil <k>cel</k> <i id="683521">dan</i> celo dopoldne</zgled>
<zgled kol="cel" kid="99719" pozicija="4" GDEX_score="8.5" >če ne no čo smo hodli <k>cel</k> <i id="398683">dan</i> smo hodli ne</zgled>
```