

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 korpori
- 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 JAMES

Metapodatki o besedilu

```
<text type="blog" author="-" date="2011" time="-" title="-"  
urldomain="www.blog.nevestica.si" url="http://www.blog.nevestica.si"  
id="janes.blog.062276">
```

Struktura besedila

Besedilo

```
Danes      danes      danes      Rgp   Rsn  
bom        bom        biti Va-fls-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  Vmprls  Ggnsp  
rožice    rožice    rožica   Ncfsg  Sozer
```

```
<g/>  
. . . . Z U  
</s>  
</p>
```

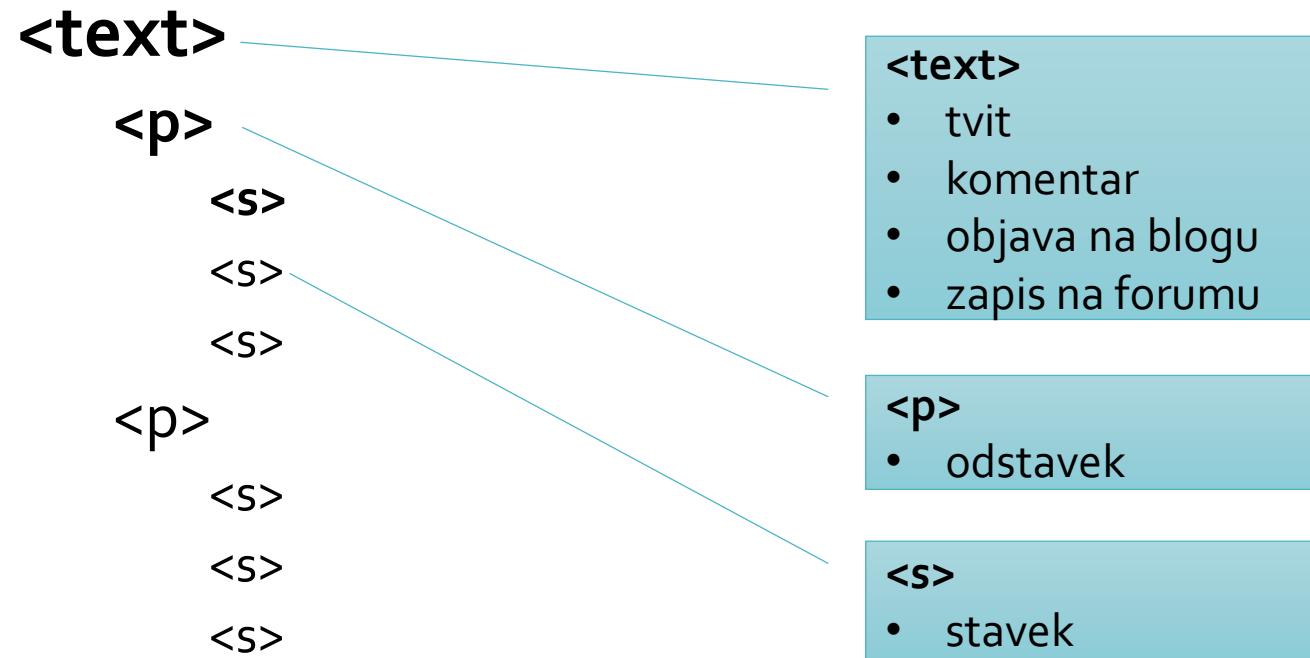
Besedilo: atributi

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: gnezstandart6
- v primeru težav:
http://nl.ijs.si/noske/janes.cgi/first_form?corpname=janes
 - up. ime: janes
 - geslo: gnezstandart6

ključno

dodatno

dobro je vedeti

ISKANJE

Enostavno iskanje

The screenshot shows the Sketch Engine interface with the following elements:

- Sketch Engine logo** at the top left.
- Navigation bar** with "uporabnik: kaja" and "korpus: JANES v0.2".
- Search bar** with "Išči" button.
- Left sidebar menu** (highlighted in blue):
 - Iskanje
 - Seznamni
 - Besedne skice
 - Tezaver
 - Najdi X
 - Primerjalne skice
 - O korpusu
- Search panel**:
 - Korpus: **JANES v0.2**
 - Enostavno iskanje:
 - Izdelaj konkordančni niz
 - Vrste iskanj | Kontekst | Lastnosti besedil
- Text callout boxes** on the right side:
 - A green box explaining the "Enostavno iskanje" input field: "Enostavno iskanje išče po vseh oblikah leme, če tako lema obstaja, in po vseh takih normaliziranih oblikah (lc)."
 - A green box below it: "Možnost vklapljanja in izklapljanja naprednega iskanja."

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

Napredno iskanje: Vrste iskanj

The screenshot shows the Sketch Engine search interface with the following details:

- Top Bar:** Sketch Engine logo, user: kaja, corpus: JANES v0.2, search input field, and results count.
- Left Sidebar:** Navigation links: Iskanje, Seznamni, Besedne skice, Tezaver, Najdi X, Primerjalne skice, O korpusu, and a help icon.
- Search Form:** Korpus dropdown set to JANES v0.2. Enostavno iskanje input field. Buttons: Izdelaj konkordančni niz, Vrste iskanj, Kontekst, Lastnosti besedil. Vrsta iskanja radio buttons: enostavno (selected), lema, zveza, bes. oblika, znak, CQL. Fields for Lema, Zveza, Besedna oblika, Znak, and CQL. Buttons: Izdelaj konkordančni niz, Počisti vse.

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), če odključamo začetnico pa take zapisa besedne oblike. (npr. *zanimivo* → *zanimivo ipd.*)

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

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

Napredno iskanje: Kontekst

Kontekst lahko določamo tako pri enostavnem kot naprednih iskanjih.

The screenshot shows the Sketch Engine search interface. On the left, a sidebar lists navigation options: Iskanje, Seznamni, Besedne skice, Tezaver, Najdi X, Primerjalne skice, O korpusu, and a help icon. The main search area has a 'Korpus:' dropdown set to 'JANES v0.2' and an 'Enostavno iskanje:' input field. Below these are tabs for 'Vrste iskanj', 'Kontekst', and 'Lastnosti besedil'. The 'Kontekst' section contains two filter sections: 'Filter za leme' and 'Filter besednih vrst'. Both filters have dropdowns for 'Razpon:' (set to 'levo') and 'pojavnic.' (set to '1'). In the 'Leme:' field, the word 'vse' is selected. In the 'Besedna vrsta:' field, several categories are listed: samostalnik, glagol, pridevnik, prislov, and zaimek. At the bottom of the search area are buttons for 'iz' and 'Počisti vse'.

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

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

Napredno iskanje: Lastnosti besedil

The screenshot shows the Sketch Engine interface with the 'korpus: JANES v0.2' selected. The 'Lastnosti besedil' panel is open, displaying various filters. Under 'GROUP.ID', 'comment' and 'tweet' are checked. Under 'GROUP.TYPE', 'comment' and 'forum' are checked. A green box at the bottom left states: 'Lastnosti besedil lahko določamo tako za enostavna kot za napredna iskanja (s filtri ali brez njih).'

TEXT.TYPE

- blog
- comment
- forum
- tweet

Izberi vse

označimo tip besedila

TEXT.URLDOMAIN

- ..
- tweeter.com
- www.rtvslo.si
- med.over.net
- kitchenstories.blog.siol.net
- mobilnipotepi.blog.siol.net
- darjas.blog.siol.net
- napoti.blog.siol.net
- janko13.blog.siol.net
- pikainmamlaz.blog.siol.net
- jonas.blog.siol.net
- sosed.blog.siol.net
- savo.blog.siol.net
- borut.blog.siol.net
- piroman.blog.siol.net
- stasa.blog.siol.net
- www.blog.uporabnastran.si

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 ...*

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 ustrezni atribut (word), na kar pogosto pozabimo.

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

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

CQL: (nekateri) posebni operatorji

operator	pomen	primer iskanja	opis iskanja
.	katerikoli znak	[tag="So..d"]	poiči vse občne samostalnike v dajalniku
[] (znotraj besede)	katerakoli znak znotraj oglatih oklepajev	[lemma="miš[eo]lovka"]	poiče vse oblike lem <i>mišelovka</i> in <i>mišolovka</i>
*	o ali več pojavitev (znaka, črke, besede ...)	[word="tvit.*"]	poiči vse besede, ki se začnejo na tvit- (tudi <i>tvit</i>)
+	1 ali več pojavitev	[word=".+ost"]	poiči vse besede, ki se končajo na –ost (brez <i>ost</i>)
?	o ali 1 pojavitev	[word="Kleme?n.*"]	poiči besede, ki se začnejo na Klemen- ali Klemn-
	disjunkcija ("ali")	[lemma=" (bloger blogar) "]	poiči vse oblike lem <i>bloger</i> in <i>blogar</i>
!	negacija ("ne")	[tag="S.*" & word!="@.*"]	poiči vse samostalnike, ki se ne začnejo z znakom @
(?i)	poljubna velikost črk	[word=" (?i) ozn"]	poiči 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 <i>tistega</i> in <i>dne</i> , med katerimi je vrinjena še ena beseda
{k,n}	interval (med k in n pojavitev znaka, črke, besede)	[tag="P.*"] {3,5} [tag="S.*"] {0,3} – med nič in tri {3} – točno tri {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 <i>živjo</i>
within < s />	išči znotraj stavka	[word="tistega"] [*] [word="dne"] within < s />	v stavku poišči vse zveze besed <i>tistega</i> in <i>dne</i> , 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. *dooooolgčas*). 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!="@.*"] | [code="(?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

podatki o besedilu
(klik: prikaz vseh podatkov)

skice
O korpusu
?

Shrani
Možnosti
prikaza
Usredinjeno
Stavek
Razvrščanje
po levi
po desni
iskani niz
Podatki
Premešaj
Vzorec
Filter

Meni konkordančnega niza
ponuja različne možnosti urejanja.

< previous to bodo objavili , ko bo teh 15 narastlo na ..%%% #vslužbitranke ## @Svarun_K kaj si tale bolna stranka lahko izmišljuje , si ne more nobena na svetu . In to ravno oni.Aja .. v Slo smo #wtf ## RT @MeFollowYouAll : Iščem kakršnokoli delo po podjemni pogodbi , neto 3€ ura , čas , trajanje dela ni bistven-info DM . Dobrim dušam hvala za RT h... ## @vitaminC_si prelepe sanje .. utopistične :))) ## In SDS je ponovno prišel do " vesoljne " ugotovitve next >

Sketch Engine

uporabnik: kaja korpus: JAMES v0.2

štевilo zadetkov

Iskalni niz **iskati** 41,666 (322.6 na milijon)

Stran 1 od 2,084 Pojd Narednja | Zadnja

premikanje po straneh

janes.tweet.430449709724225536 poskušajo pomagati vsem pomoči potrebnim . Ne pa **iskati** frko . Re : cehovska kolegica na #odmevi
janes.tweet.451389652004786176 @petrasovdat Itak , da ne . Ne ljubi se mi **iskati** linke . :P Pa glavne a
janes.tweet.466968317887266816 nove Europe . http://t.co/f0nLnIKBzw ## EU **išče** slovenske odvetnike /
janes.tweet.478241038889648128 v bistvu je underdog . :) @DC43 ##g @DC43 Iščem besedo , ne pol stavki
janes.tweet.48207667610042368 tri četrtnine zagonske denarja in zdaj **iščejo** dva redna sodelavca
janes.tweet.361803902624079872 pravkar končal Fakulteto za strojništvo in **išče** zaposlitev , potem se
janes.tweet.474618184113586176 #NBAFinals2014 ##g " Spurci bodo zmage moralni nujno **iskati** na domačem parketu
janes.tweet.377108132008566784 to je beseda . Danes pa te bom vztrajno **iskal** s tvojim klobukom po tv ekranu , s flašo
janes.tweet.374519789097857024 g RT @kriminolog : Za kolegovo gospo mamo **iščemo** sedež v avtu za naslednjo sredo (ali kak
janes.tweet.381486500816621568 Vsem prostovolj ... ##g RT @jagoda879 : . Iščem dve karti za slov.tekmo ob 21h . #optimist
janes.tweet.384362845917683712 http://t.co/KEYNKwpzXO naj ... ##g RT @MrVates : Iščem urednika , da mi pomaga končati (kratki
janes.tweet.385426028241649666 račune ? junaki ##g RT @ZanKeglic : " GIG " Iščemo fotografa ki ima izkušnje z fotografiranjem
janes.tweet.385110545785159680 pogajamo . ##g RT @Donfarfezi : Please RT : Iščem uporabnika apple notebooka (in iPhoto
janes.tweet.373552898116190208 do wc :)) #dobrojutro ##g RT @PeterFilec : Išče se nekdo , ki bi mi popravil nek vtikač
janes.tweet.375947368866521090 , ljubitelji pierogov ! Poljsko velep . Išče koga , ki bi v torek v CE na tekmi #Eurobasket
janes.tweet.345984403819868160 .. v Slo smo #wtf ##g RT @MeFollowYouAll : Iščem kakršnokoli delo po podjemni pogodbi ,
janes.tweet.343619492150132737 .. posteljno) #kimono ##g RT @vanfranco : Išče se Albin Florjančič iz Trebnjega . Vozil
janes.tweet.342973864629919744 parkrat fajn :::) ##g Nadal raztura , Nole pa Išče fizioterapevta . Nekaj tu ni ok :D #tenis
janes.tweet.342964244079312896 če jih nimaš na nosu in nič ne vidis ::):) Išči jih , ko jih imaš nataknjene :) #travma
janes.tweet.396999500893814784 na psihiatrijo ::))) ##g RT @JsSmRental

Stran 1 od 2,084 Pojd Narednja | Zadnja

konkordanca

(klik: prikaz širšega konteksta)

group.id janes.tweet
group.type tweet
text.id janes.tweet.373552898116190208
text.type tweet
text.url -
text.urldomain tweeter.com
text.date 2013-08-30
text.time 21:08:45
text.author 123koriz
text.title -
text.wordcount 26

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.

Sketch Engine

uporabnik: kaja korpus: JANES v0.2

Možnosti prikaza

Atributi

Strukturni elementi

Podatki

Število zadetkov na stran: 20

Število znakov v kontekstu: 40

Ikona za kopiranje z enim klikom

Dovoli hkratno izbiro več zadetkov

XML predloga za kopiranje z enim klikom:

Shrani in spremeni možnosti

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

Enostavno razvrščanje

Atribut: word

Razvrsti: Levi kontekst iskani niz Desni kontekst

Število pojavnic za razvrščanje: 3

Ne ločuj m/v črk Ž-A

Razvrsti

Večnivojsko razvrščanje

prvi nivo (Razvrsti po ...)

Atribut: word

Ne ločuj m/v črk Ž-A

Položaj: 3L
2L
1L
iskani niz
1D

drugi nivo (... nato razvrsti po ...)

Atribut: word

Ne ločuj m/v črk Ž-A

Položaj: 3L
2L
1L
iskani niz
1D

tretji nivo (... nazadnje razvrsti po)

Atribut: word

Ne ločuj m/v črk Ž-A

Položaj: 3L
2L
1L
iskani niz
1D

Razvrsti

Konkordance: Vzorec

Sketch Engine

uporabnik: kaja korpus: JAMES v0.2

Iskanje
Seznam
Besedne skice
Tezaver
Najdi X
Primerjalne
skice
O korpusu
?

Naključni vzorec ?

Izdelaj naključni vzorec iz konkordančnega niza.
Število zadetkov v vzorcu: 250

Izdelaj vzorec

lati

Določimo število konkordanc v naključnem vzorcu zadetkov iz konkordančnega niza.

Naključni vzorec danega konkordančnega niza je vedno enak.

Konkordance: Filter

Filter

Prekrivanja

1. zadelek v dokumentu

Privzeti funkciji filtrirata:

- ujemajoče se vrstice
- prvi zadelek v vsakem dokumentu

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

Filtriranje konkordančnega niza [?](#)

Filter: pozitivni negativni

Izbrana pojavnica: prva zadnja

Razpon iskanja: od do vključi iskan niž

Enostavno iskanje: [Vrste iskanj](#) [Lastnosti besedil](#)

Filtriraj

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.

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

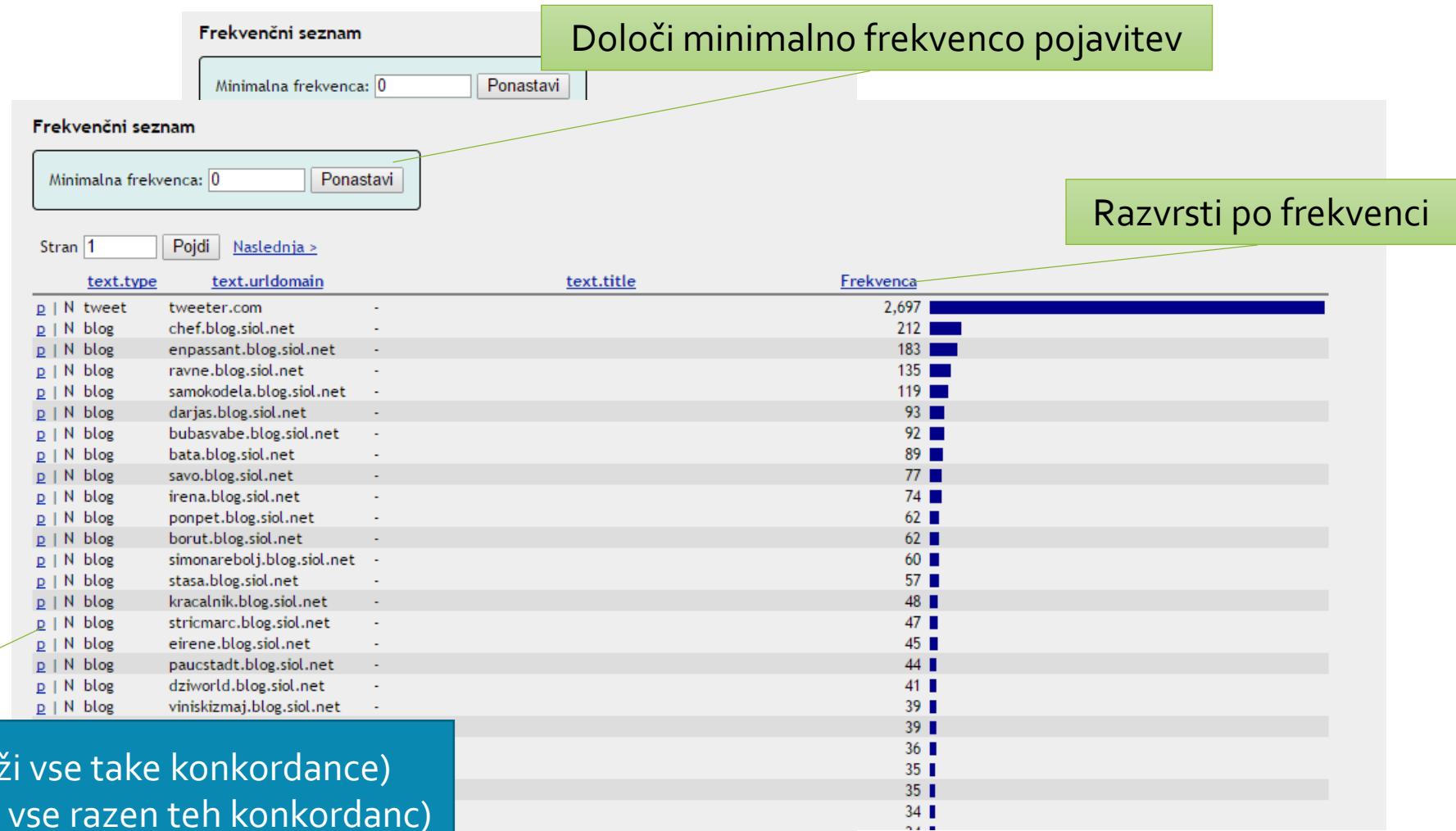
Običajno iskalnega niza ne vključujemo (filtriramo zgolj glede na okolico).

Konkordance: Frekvence

Frekvence
Oznake niza
Oblike niza
Dokumenti

klik do konkordanc

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



Konkordance: Frekvence

Večnivojska razporeditev po frekvenci (?)

Minimalna frekvencna: 5

prvi nivo Atribut: word

drugi nivo Atribut: lc

tretji nivo Atribut: lemma

četrти nivo Atribut: tag

Ne ločuj m/v črk

Ne ločuj m/v črk

Ne ločuj m/v črk

Ne ločuj m/v črk

6L
5L
4L
3L
2L
1L
iskani niz
1D
2D

6L
5L
4L
3L
2L
1L
iskani niz
1D
2D

6L
5L
4L
3L
2L
1L
iskani niz
1D
2D

6L
5L
4L
3L
2L
1L
iskani niz
1D
2D

Položaj: 2D

Položaj: 2D

Položaj: 2D

Položaj: 2D

Izdelaj frekvenčni seznam

Frekvenčna razporeditev po lastnostih besedil

Minimalna frekvencna: 0

Prikaži kategorije brez zadetkov:

group.id
group.type
text.id
text.type
text.url
text.urldomain
text.date
text.time

Izdelaj frekvenčni seznam

Določimo minimalno frekvenco.

Določimo do štiri **nivoje izpisa seznama** in nastavitev 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

Sketch Engine

uporabnik: kaja korpus: JANES v0.2

Išči drevlo JAMES v0.2

Iskanje
Sezname
Besedne skice
Tezaver
Najdi X
Primerjalne skice
O korpusu ?

Shrani < Konkordance Vzorec Zadnja (250) Filter Prekrivanja 1. zadetek v dokumentu Frekvence Oznake niza Oblike niza Dokumenti Kolokacije Opis niza Vizualiziraj ?

Frekvenčni seznam

Minimalna frekvanca: 5 Ponastavi

Stran 1 Pojdi Naslednja >

text.author	Frekvanca	Rel. frekvanca (%)
D N druga	9	89,960.5
D N Ally*	5	48,254.6
D N ObnovimoGozdove	10	20,410.4
D N Zima1	7	18,374.1
D N Manteja	7	8,755.9
D N cicmen	9	7,125.6
D N Mazni	7	4,637.0
D N bdbstone	6	4,142.5
D N ZaMestoPoDveh	6	4,006.6
D N matjasec	75	3,626.2
D N Radio_Si	5	3,617.1
D N Genki_Dashite	13	3,420.4
D N mutawa	11	3,066.4
D N MajaSimoneti	38	2,898.9
D N AlesCerin	6	2,865.6
D N belapotonka	5	2,524.8
D N Loccitane_Slo	5	2,350.4
D N RislzpodRoznika	5	2,269.9
D N FriLLox	5	1,946.6
D N pesemsi	9	1,838.9
D N SuzanaRengeo	7	1,394.5
D N hepimen	5	1,260.3
D N Pitka	16	1,254.0
D N karolina83	8	1,226.4

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_korp) / 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 okolini iskanega niza je statistično pomembna.

The screenshot shows the Sketch Engine interface with the title 'Sketch Engine' at the top. Below it, the user information 'uporabnik: kaja korpus: JANES v0.2' and search buttons 'Išči' and 'scena'. On the left, there's a sidebar with links like 'Analize', 'Vzorci', 'Besedne skice', 'Rezaver', 'Najdi X', 'Imperialne', 'korpusu', 'Konkordance', 'zorec', 'Zadnja (250)', 'Filter', and 'Prekrivanja'. The main panel is titled 'Kolokacijski kandidati'. It contains a search form with the following fields:

- 'Atribut:' dropdown set to 'word' with input fields 'V razponu od: -5 do: 5'.
- 'Minimalna frekvenca v korpusu:' input field set to '5'.
- 'Minimalna frekvenca v danem razponu:' input field set to '3'.
- Two dropdown menus for 'T-score' and 'logDice'.
- 'Pokaži vrednosti za:' dropdown set to 'logDice'.
- 'Razvrsti glede na:' dropdown set to 'logDice'.
- 'Izdelaj seznam' and 'Shrani nastavitev' buttons.

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

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

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).

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>

The screenshot shows the Sketch Engine interface with a green sidebar on the left and a main search form on the right.

Left Sidebar (Tezaver):

- Najdi X
- Primerjalne skice
- O korpusu
- < Konkordance
- Vzorec
- Zadnja (250)
- Filter
- Prekrivanja

Search Form:

- Išči (Search) button
- scena (Scene) button
- Minimalna frekvanca v korpusu:
- Minimalna frekvanca v danem razponu:
- T-score dropdown menu:
 - MI
 - MI3
 - log likelihood
 - min. sensitivity
 - logDice (selected)
- Razvrsti glede na: dropdown menu:
 - MI
 - MI3
 - log likelihood
 - min. sensitivity
 - logDice (selected)
- Pokaži vrednosti za: dropdown menu:
 - Izdelaj seznam (Create list)
 - Shrani nastavitev (Save settings)

Konkordance: Kolokacije

Sketch Engine

uporabnik: kaja korpus: JAMES v0.2 Išči scena

Iskanje
Seznam
Besedne skice
Tezaver
Najdi X
Primerjalne
skice
O korpusu
?

Shrani
< Konkordance
Vzorec
Zadnja (250)
Filter
Prekrivanja
1. zadetek v
dokumentu
Frekvenca

Kolokacijski kandidati

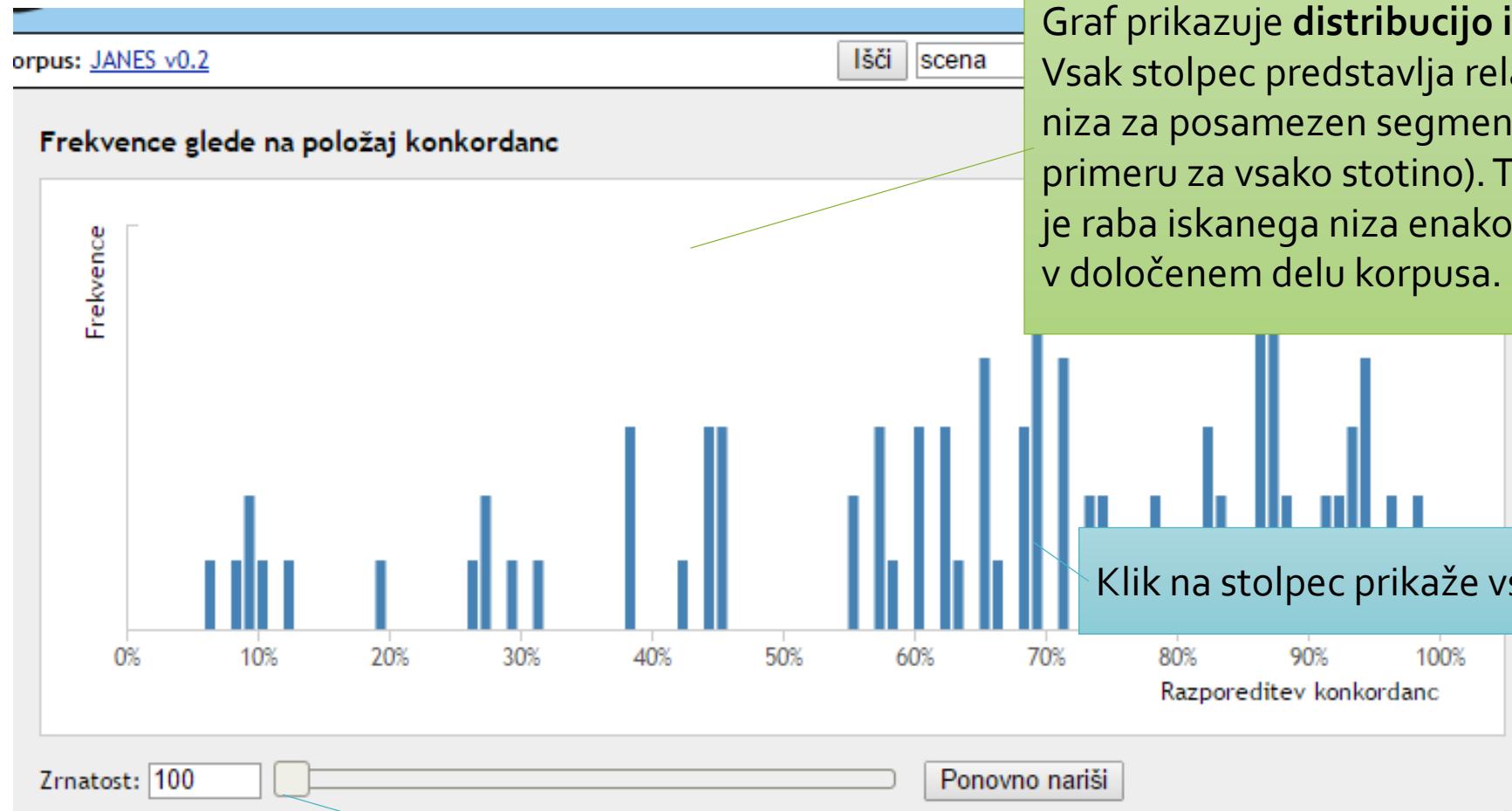
Stran 1 Pojdi Naslednja >

	Freq	T-score	MI	logDice
p N glasbeni	160	12.639	10.225	8.994
p N politični	200	14.114	8.983	8.828
p N glasbene	134	11.567	10.309	8.801
p N glasbeno	105	10.235	9.797	8.421
p N slovenski	232	15.154	7.614	8.143
p N slovenske	173	13.087	7.640	8.017
p N pop	78	8.811	8.747	7.855
p N slovensko	115	10.661	7.405	7.639
p N stand-up	51	7.127	8.909	7.407
p N medijski	50	7.054	8.712	7.348
p N slovenska	76	8.663	7.326	7.313
p N domači	58	7.582	7.827	7.282
p N politično	78	8.773	7.222	7.282
p N alter	40	6.321	10.633	7.257
p N medijske	44	6.620	8.937	7.232
p N @FNSlo	45	6.692	8.692	7.220
p N glasbena	41	6.394	9.438	7.205

S klikom na p si ogledamo
primere rabe kolokacij.

S klikom na naslov stolpca
razvrščamo po poljubni statistični
vrednosti.

Konkordance: Vizualiziraj



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 liste, kolokacije ipd.

Vsaki konkordanci pripisi zaporedno številko.

Usredinjen ali stavčni izpis.

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

Shrani konkordance ?

Shrani konkordance kot: txt XML

Shrani strani: Vse
 Samo stran:

Dodaj glavo:

Oštrevilči zadetke:

Poravnaj po iskanem nizu:

Maksimalno število zadetkov: (max. 100,000)

Shrani konkordance

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

Shrani vse konkordance ali samo določeno stran

Shrani tudi **dodate 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 najpogostejsa 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 način izpisa:

- enostavno: iskani atribut s frekvenco
- ključne besede: (podrobneje v nadaljevanju)
- večnivojsko: izpis z dodatnimi informacijami o iskanem atributu (npr. lema in msd)

The screenshot shows the Sketch Engine interface with the following search parameters:

- korpus: JANES v0.2
- Podkorpus: izdelaj novega
- Išči atribut: word
- Minimalna frekvenca: 5
- Maksimalna frekvenca: 0 (0 = brez minimalne frekvence)
- Način izpisa: Enostavno
- Referenčni (pod)korpus: JANES v0.2 (celotni korpus)
- Preferiraj: redke besede pogoste besede 100

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:

- števec zadetkov: kolikokrat se besede pojavijo v korpusu
- število besedil: v koliko različnih besedilih se besede pojavijo
- PRF: varianca frekvenčnega seznama, ki ne šteje pojavitev iste besede, ki se pojavlja skupaj, npr. v istem dokumentu

Seznami: besedni nizi

Izberemo atribut.

Določimo dolžino niza (število besed).

Določimo frekvenčni prag.

Sketch Engine

uporabnik: kaja korpus: JANES v0.2

Išči | v JANES v0.2

Iskanje
Seznam
Besedne skice
Tezaver
Najdi X
Primerjalne
skice
O korpusu
?

Vse besede
Vse teme
?

Položaj menija

Možnosti Seznamov besed ②

Korpus: JANES v0.2

Podkorpus: [izdelaj novega](#)

Išči atribut: word

uporabi n-gramme. Vrednost n: 2

Možnosti filtriranja:

Filtriraj seznam besed po: Regular expression:

Minimalna frekvenca: 5

Maksimalna frekvenca: 0 (0 = brez minimalne frekvence)

Poisci besede: Izberi datoteko Nobena datoteka ni izbrana Počisti

Izloči besede: Izberi datoteko Nobena datoteka ni izbrana Počisti format

Upoštevaj nebesede

Možnosti izpisa:

Frekvenčne vrednosti: Števec zadetkov Število besedil Povprečna relativna frekvenca (PRF)

Način izpisa: Enostavno Ključne besede

Referenčni (pod)korpus: JANES v0.2 (celotni korpus)

Preferiraj: redke besede pogoste besede: 100

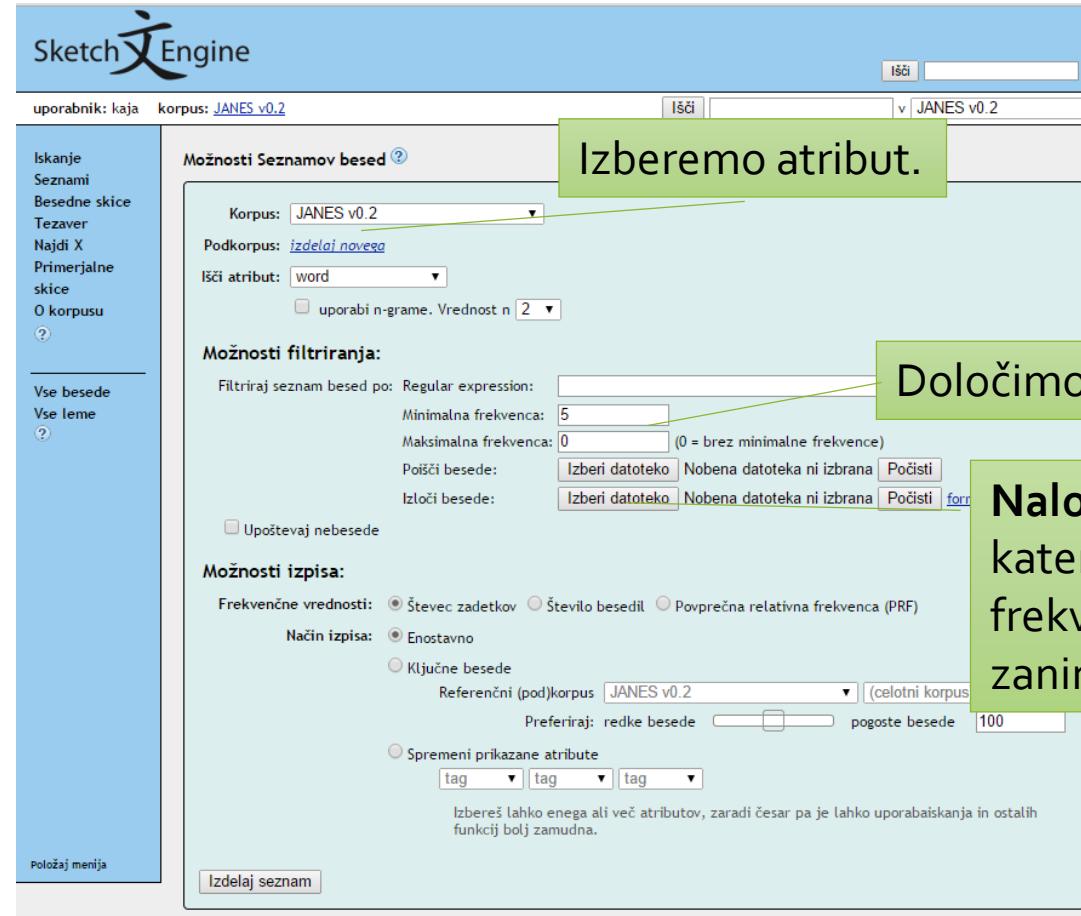
Spremeni prikazane attribute: tag tag tag

Izberete lahko enega ali več atributov, zaradi česar pa je lahko uporaba istka in ostalih funkcij bolj zamudna.

Izdelaj seznam

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.

The screenshot shows the 'Seznamy' (Lists) tool in the JANES v0.2 interface. The left sidebar has links for 'Seznamy', 'Besedne skice', 'Tezaver', 'Najdi X', 'Primerjalne skice', 'O korpusu', 'Vse besede', and 'Vse teme'. The main panel has sections for 'Korpus:' (set to JANES v0.2), 'Podkorpus:' (set to 'izdelai novega'), 'Išči atribut:' (set to 'word'), and a checkbox for 'uporabi n-gramme. Vrednost n' (set to 2). Below this is a 'Možnosti filtriranja:' section with fields for 'Minimalna frekvenca' (5) and 'Maksimalna frekvenca' (0). It also includes buttons for 'Izberi datoteko', 'Počisti', and 'format'. A green callout box labeled 'Izberemo atribut.' points to the 'Išči atribut:' dropdown. Another green callout box labeled 'Določimo frekvenčni prag.' points to the 'Minimalna frekvenca' field. A large green callout box labeled 'Ključne besede: Izberemo referenčni korpus (npr. Kres)' points to the 'Referenčni (pod)korpus:' dropdown set to 'JANES v0.2'. The bottom of the panel has sections for 'Možnosti izpisa:' and 'Spremeni prikazane atribute'.

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)

lemma	JANES v0.2		Kres		Rezultat
	Frekvence	Frekvence/mil	Frekvence	Frekvence/mil	
slovenija	129,490	1002.6	1,862	15.5	9.6
ljubljana	46,228	357.9	284	2.4	4.5
janska	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
rabit	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
desjni	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

Korpus: JAMES v0.2

Ime novega podkorpusa: JAMES v0.2 blogi

Število besedil Število pojavnic

GROUP.ID	#
janes.blog.slwac	54689599
janes.comment.rtvslosi	12330236
janes.forum.medovernet	13817612
janes.tweet	48310694

Izberi vse

GROUP.TYPE	#
blog	54689599
comment	12330236
forum	13817612
tweet	48310694

Izberi vse

TEXT.ID

TEXT.TYPE
<input checked="" type="checkbox"/> blog
comment
forum
tweet

Izberi vse

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.

Vnesemo ime podkorpusa.

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

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)

furati (-g)
JANES v0.2 frekvenca = [2.308](#) (17.9 na milijon)

Constructions

O_tretja_oseba	560	1.5
O_povratni_se	319	3.0
O_nedoločnik_cs	101	2.7
V_lahko_G	44	4.6

S_v_tožil

safr	38	10.43
imidž	18	8.14
irharec	5	7.79
bb-je	5	7.75
enakost	5	5.73
biznis	4	5.33
kariera	11	5.06
drzava	6	5.01
finance	5	4.6
stil	9	4.56
scena	9	4.54
komunikacija	4	4.42
kampanja	6	4.31
študij	4	4.23
politika	24	4.2
logika	4	3.68
gospodarstvo	5	3.54
zadeva	12	3.19
država	27	2.9
projekt	4	2.31
avto	5	2.25
življenje	13	2.21
oddaja	5	2.08
trg	4	1.99
odnos	5	1.96

S_kako-kdaj_g?

safr	9	8.77
uspešno	14	5.76
dalje	5	4.87
doslej	4	4.83
naprej	25	4.59
zmeraj	4	3.79
ponavadi	4	3.18
dejansko	7	3.12
trenutno	5	2.98
nekako	6	2.97
dolgo	6	2.91
vedno	33	2.69
dobro	19	2.45
težko	7	2.25
zdaj	12	2.05
potem	13	2.0
spet	10	1.96
rad	12	1.93
včasih	4	1.87
skupaj	5	1.8
kdaj	5	1.55
sedaj	5	1.48
zato	5	1.38
danes	11	1.38
lahko	37	0.9

S_osebek_ie

ustluzbenec	6	8.4
debil	5	6.4
imidž	4	6.2
klovn	5	6.1
bus	7	6.05
župan	4	3.55
tip	6	2.44
k	5	1.8
a	4	0.48
človek	10	0.36

Število vseh kolokacijskih kandidatov v relaciji
(klik do konkordanc)

S_namenilnik	112	24.0
znati	20	2.6
začeti	10	1.02
dati	10	0.82
morati	12	0.26
iti	15	0.06

koga-kaj_g4	103	
S_na_g4	43	3.5
S_v_g4	32	3.1
S_za_g4	10	0.9
S_skozi_g4	8	17.8
S_čež_g4	6	8.4
S_skoz_g4	4	157.4

S_v_rodil	95	5.8
safr	4	8.93
pozitiv	5	8.59
zvezda	4	1.72

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 besednjem 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	36
+ TRINARY	
DUAL	23
UNARY	2
CONSTRUCTION	13
CONSTRUCTION	6
+UNARY	
COLLOC	3
SYMMETRIC	2
brez	18
skupaj	103

Tabela 1: slovnične relacije po direktivah

Besedne skice: posebne direktive

peljati (-g)
JANES v0.2 frekvenca = 15.055 (116.6 na milijon)

Constructions	Count	Frequency
O_povratni_se	4,685	5.7
O_tretja_oseba	4,378	1.5
O_nedoločnik_cs	1,149	4.0
V_lahko_G	343	4.6
O_povratni_si	179	1.2

S_kako-kdaj_g?	Count	Frequency
kam	436	9.33
lahko	329	4.05
danes	102	4.56
mimo	81	7.74
rad	79	4.61
potem	76	4.51
prvič	65	6.55
aj	63	5.14
kamor	60	8.19
a	54	6.82
po	52	4.81
bro	52	3.87
mov	47	6.45
et	45	4.09
utraj	42	6.33
upaj	40	4.72
ikrat	39	4.77
malo	34	2.7
tako	34	1.53
jutri	32	5.0
treba	32	2.84
počasi	31	5.46
vedno	30	2.53
hitro	28	4.6
zdaj	27	3.18

koga-kaj_g4	Count	Frequency
S_na_g4	1,490	15.6
S_v_g4	1,148	14.4
S_cez_g4	136	24.8
S_skozi_g4	123	35.5
S_za_g4	90	1.1
S_z_g4	21	3.2
S_po_g4	16	7.5
S_pred_g4	10	5.9
S_med_g4	7	3.6
S_mimo_g4	7	279.9
S_do_g4	6	8.7
S_nad_g4	5	7.5
S_pod_g4	3	1.5
S_skoz_g4	3	15.3

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

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

Besedne skice: napredne nastavitev

Spremeni
nastavitev
Grupiranje
Razvrsti
Relacije
Več podatkov
Manj podatkov

Prizete možnosti nastavitev
v meniju na levi.

Izdelaj besedno skico [?](#)

Korpus: JANES v0.2 ▾
Lema: peljati
Besedna vrsta: glagol ▾
[Napredne nastavitev](#)

Napredne nastavitev

Podkorpus: Nič (celoten korpus) [info](#)
Minimalna frekvenca: 3
Minimalen rezultat: 0.0
Največje število besed v slovnični relaciji: 25
Razvrsti kolokatorje glede na: Rezultat frekvenco
Predloga za klikskografijsko: Nič ▾ Zgledov na kolokator: 6
Grupiraj kolokacije:
Razporedi besedno skico po relacijah:
Minimalna podobnost med besedami v grupi: 0.15
Minimalen rezultat za unarne relacije: 5.0
Izberi slovnične relacije: Vse

<input checked="" type="checkbox"/> O_količina	<input checked="" type="checkbox"/> O_nedoločnik_cs	<input checked="" type="checkbox"/> O_povratni_se	<input checked="" type="checkbox"/> O_povratni_si
<input checked="" type="checkbox"/> O_s_števili	<input checked="" type="checkbox"/> O_tretja_oseba	<input checked="" type="checkbox"/> O_z_lastnim_imenom	<input checked="" type="checkbox"/> O_zanikanje
<input checked="" type="checkbox"/> S_*_g2	<input checked="" type="checkbox"/> S_*_g3	<input checked="" type="checkbox"/> S_*_g4	<input checked="" type="checkbox"/> S_*_g5
<input checked="" type="checkbox"/> S_*_g6	<input checked="" type="checkbox"/> S_*_p2	<input checked="" type="checkbox"/> S_*_p3	<input checked="" type="checkbox"/> S_*_p4

Bilingual word sketch

Korpus: Nič
Lema:

[Izdelaj besedno skico](#) [Shrani nastavitev](#)

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

Iskanje
Seznamni
Besedne skice
Tezaver
Najdi X
Primerjalne
skice
O korpusu
?

Grupiranje
Shrani

Položaj menija

Lema	Rezultat	Frekvenca
oče	0.287	19,821
starš	0.271	18,834
mati	0.269	16,488
mož	0.256	16,602
žena	0.247	15,718
punca	0.239	17,127
ženska	0.233	47,343
fant	0.228	23,472
prijatelj	0.22	26,256
oseba	0.215	37,280
zdravnik	0.204	22,886
partner	0.2	12,216
mamica	0.2	5,896
moški	0.199	36,851
sin	0.198	13,200
otrok	0.197	108,146
družina	0.192	25,543
sestra	0.191	8,925
sosed	0.189	11,951
politik	0.188	19,031
tip	0.185	24,419
brat	0.183	12,953
dekle	0.181	11,127
pes	0.178	27,351
večina	0.176	42,235
bog	0.174	28,133
kolega	0.173	11,933
novinar	0.171	19,616
prijateljica	0.17	6,870
lastnik	0.168	13,585
gospa	0.168	10,535
folk	0.168	12,073
babica	0.166	5,630
igralec	0.163	23,626



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

Seznam lahko razvrstimo tudi
po skupinah podobnosti.

Primerjalne skice

peljati/furati						
JANES v0.2 frekvence = 15,055 2,308						
	6.0	4.0	2.0	0	-2.0	-4.0
peljati	6.0	4.0	2.0	0	-2.0	-4.0
S_v_rodil	165	95	1.3	5.8		
preklinjevalec	8	0	10.4	--		
pes	18	0	4.3	--		
otrok	14	0	1.9	--		
S_v_dajal	132	8	2.9	1.3		
miroslav	13	0	9.5	--		
veterinar	12	0	8.2	--		
kolegica	27	0	7.7	--		
zdravnik	9	0	3.4	--		
mama	6	0	2.8	--		
S_v_tožil	1,402	728	2.3	9.1		
roba	28	0	8.1	--		
julija	9	0	6.5	--		
pes	63	0	6.0	--		
avto	69	0	6.0	--		
gospodična	6	0	5.6	--		
vlak	15	0	5.6	--		
krog	30	0	5.4	--		
otrok	121	0	5.0	--		
žena	12	0	4.9	--		
sin	13	0	4.7	--		
punca	12	0	4.4	--		
avtobus	6	0	4.2	--		
pomoč	29	0	4.2	--		
dekle	8	0	4.2	--		
mama	14	0	4.0	--		
kolo	9	0	3.9	--		
družina	14	0	3.8	--		
politika	20	24	3.9	4.2		
kampanja	0	6	--	4.3		
scena	0	9	--	4.5		
stil	0	9	--	4.6		
država	0	6	--	5.0		
kariera	0	11	--	5.1		
imidž	0	18	--	8.1		
safir	0	38	--	10.4		

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

Primerjaj besedne skice

Korpus: Gigafida

Lema: upravljati

Besedna vrsta: glagol

Način primerjave skic: lema

Določimo opazovano lemo.

podkorpus

Prvi podkorpus: Gigafida [1990-2000] [info izdelaj novega](#)

Drugi podkorpus: Gigafida [2001-2011] [info izdelaj novega](#)

besedna oblika

Prva besedna oblika:

Druga besedna oblika:

[Napredne nastavitev](#)

[Pokaži primerjavo](#)

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

Določimo korpusa, med katerima bomo primerjali rabe izbrane leme.

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 9neztandard6 -l lemmalist.txt -r gramrellist.txt
```

Seznam lempos-ov (.txt)

dan-s
priden-p
hitro-r
imetи-g
итд.

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

итд.

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

```
<?xml version="1.0" encoding="UTF-8"?>
<clanek>
<glava>
<oblika>
<zapis>dan</zapis>
<iztocnica frek_lema="1471">dan</iztocnica>
</oblika>
<zaglavje>
<hesava>samostalnik</hesava>
<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_structura>
<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="10.57"><k>dober</k></kolokacija>
<kolokacija kid="99699" frek_kol="29" salience_kol="10.57"/><k>današnji</k></kolokacija>
<kolokacija kid="99697" frek_kol="53" salience_kol="10.33"><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="9.77"><k>naslednji</k></kolokacija>
<kolokacija kid="99723" frek_kol="8" salience_kol="8.77"><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></zgled>
<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] 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="471918">dan</i></zgled>
<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>
```

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

izpis kolokatorjev
za relacijo

izpis zgledov za
kolokator

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