

## 2.2 Raziskovanje spletne slovenščine – delavnica

**JANES**

IJS  
FILOZOFSKA  
FAKULTETA

**Delavnica 1:  
Raziskovanje spletne slovenščine**

Darja Fišer  
Oddelek za prevajalstvo, Filozofska fakulteta Univerze v Ljubljani  
Odsek za tehnologije znanja, Inštitut Jožef Stefan

Ljubljana, 4. julij 2016

**JANES**

IJS  
FILOZOFSKA  
FAKULTETA

Korpus JANES

**ISKANJE PO KORPUSU**

- konkordančnik
  - [https://sketch.cjvt.si/bonito/janes04.cgi/first\\_form](https://sketch.cjvt.si/bonito/janes04.cgi/first_form)
- navodila za uporabo

The screenshot shows the Sketch Engine interface. On the left is a navigation menu with options: Concordance, Word list, Word sketch, Thesaurus, Sketch diff, Corpus info, My jobs, Home, and User guide. The main area is titled 'Sketch Engine' and contains a 'Corpus:' dropdown menu with a list of options: Kres, JANES v0.4 (selected), JANES v0.4 Blog, JANES v0.4 Forum, JANES v0.4 News, JANES v0.4 Tweet, and JANES v0.4 Wikipedia. Below the dropdown is a 'Simple query:' input field. Further down is a 'Context' section with a 'Lemma filter' section containing a 'Window:' dropdown set to 'both', a '5' token count, and a 'Lemma(s):' input field with a dropdown set to 'all'.

The screenshot shows the Sketch Engine search interface. At the top, the 'Corpus:' dropdown is set to 'JANES v0.4'. Below it is a 'Simple query:' input field with a 'Make Concordance' button. There are links for 'Query types', 'Context', and 'Text types'. The 'Query type' section has radio buttons for 'simple', 'lemma' (selected), 'phrase', 'word', 'character', and 'CQL'. Below this are input fields for 'Lemma:', 'Phrase:', 'Word form:', and 'Character:'. The 'PoS:' dropdown is set to 'unspecified'. There is a 'match case' checkbox. At the bottom, there is a 'CQL:' input field and a 'Default attribute:' dropdown set to 'word'. A 'Tagset summary' link is at the bottom left.

## JANES SketchEngine: filtriranje po metapodatkih

**Text types**

Subcorpus: None (whole corpus) [Info create new](#)

**TEXT.USER**

**TEXT.SOURCE**

 corporate  
 private

**TEXT.SEX**

 female  
 male  
 neutral

**TEXT.LANG**

 eng  
 hbs  
 slv  
 und

**TEXT.SENTI**

 negative  
 neutral  
 positive

**TEXT.STD\_TECH**

 T1  
 T2  
 T3

**TEXT.STD\_LING**

 L1  
 L2  
 L3

**TEXT.YEAR**

 2013  
 2014  
 2015  
 2016

## JANES SketchEngine: konkordance

Iskalni niz **boljše** 27,257 > Premešaj 27,257 (169.0 na milijon)

[Prva](#) | [Prejšnja](#) | Stran 2 od 1,363 | [Pojdi](#) | [Naslednja](#) | [Zadnja](#)

<b>blog</b>	distančirali tudi terminološko, če se jih ima večina itak za	<b>boljše</b> /boljše/dober/Agcempa	od novinarjev in njih cenzorskih, politično nastavljenih
<b>tweet</b>	nič spornega http://t.co/hSLID7A6Dv ##g @BozoPredalic	<b>Boljšje</b> /boljšje/dobra/Rgc	za marsikoga, da je ne. Lahko katero prime, da ga
<b>blog</b>	Sicer je super, sedaj me zanima, če je lahko še	<b>boljšje</b> /boljšje/dober/Agcfn	. Grrrr g heh, skrajni čas, bejbi. Jst to že nekej časa
<b>tweet</b>	Affleck bo naslednji Batman. Buuu ##g A ne bi bilo ful	<b>boljšje</b> /boljšje/dober/Agcnsn	, če bi tvite z deli, zaposlitvami opremili z enim
<b>forum</b>	rdečic po telesu, včasih tudi po obrazu. Sedaj so vse	<b>boljšje</b> /boljšje/dober/Agcfn	, niso več rdeče le srbi jo še večkrat (včasih se
<b>tweet</b>	@maticslapsak Vsakemu svoje veselje. Kaj č'mo. Vseeno	<b>boljšje</b> /boljšje/dobra/Rgc	kot nazi ikonografija. #alwayslookonthebrightsideoflife
<b>tweet</b>	slovenske oblasti in sodišča ji stojijo na poti v	<b>boljšje</b> /boljšje/dober/Agcnsa	živetjenje http://t.co/6LE4s7GEzb ##g Vsaka tretja ženska
<b>forum</b>	malo neprijetno. Savine so sicer v snegu bile veliko	<b>boljšje</b> /boljšje/dober/Agcfn	, na mokri in mastni podlagi pa prava katastrofa v
<b>blog</b>	13.09.2012 ob 16:57 g ne vem, meni se zdi, da bi veliko	<b>boljšje</b> /boljšje/dobra/Rgc	(in bolj seksi) izpadlo, če bi imela zgornji del
<b>forum</b>	kaksne narezane diske (npr. ATE, breombo... ) in pa	<b>boljšje</b> /boljšje/dober/Agcempa	zavorne plosvice. ##g Pri nekaterih avtjih je res potrebno
<b>tweet</b>	koga ##g @Razdelilec tista Gradišnikova je huda, ja,	<b>boljšj</b> /boljšje/dobra/Rgc	da nima prav ##g hm, a ni Al Gore 07 pokasiral Nobelove
<b>blog</b>	dalje), ker morda v takem moodu kaj lepega zamujaš ... g	<b>boljšje</b> /boljšje/dobra/Rgc	da neham besedičit .... lačna sem pa se hočem zamotit
<b>forum</b>	cevi, če bi bilo morda res kej v dovodu nafte, pa nič	<b>boljš</b> /boljšje/dobra/Rgc	g - ni 4motion g - kompresija bi avto skos zajebavala
<b>tweet</b>	##g @TamaraSvetina Sem zelo iz vaje, a če ne najdeš	<b>boljšje</b> /boljšje/dober/Agcempa	(ga) se lahko potrudim. @ales_gantar ##g @_lnja ... Kaj
<b>forum</b>	Tudi meni ni všeč Astra enjoy, sport mi pa je. Mnogo	<b>boljšje</b> /boljšje/dober/Agcempa	sedeže ima, pa el. ročno in tudi ni veliko dražji
<b>tweet</b>	pol sm si pa kupu frušt in sm še zmer u minusu...	<b>Boljšj</b> /boljšje/dobra/Rgc	da bi šou u ošterijo: D http://t.co/SHzwwSnpKF ##g
<b>comment</b>	koncno		
<b>forum</b>	njegovih		
<b>tweet</b>	za st		
<b>forum</b>	original		

**Prva** | **Prejšnja** | **Stran**

text.type forum  
text.author Goggy  
text.title VW Passat BKP trese - NUJNO POMOČ  
text.date 2011-03-15  
text.url http://www.avtomobilizem.com/forum/viewtopic.php?f=6&amp;t=84268&amp;start=20#p1423667  
text.id janes.forum.avtomobilizem.6.84268.1423667

## JANES SketchEngine: obdelava konkordanc

IJS FILOZOFSKA  
FAKULTETA

- Save
- as subcorpus
- View options
  - KWIC
  - Sentence
- Sort
  - Left
  - Right
  - Node
  - References
  - Shuffle
- Sample
  - Last (1000)
- Filter
  - Overlaps
  - 1st hit in doc
- Frequency
  - Node tags
  - Node forms
  - Doc IDs
- Collocations
- ConcDesc
- Visualize

## JANES SketchEngine: izdelava frekvenčnih seznamov

IJS FILOZOFSKA  
FAKULTETA

### Multilevel frequency distribution

Frequency limit: 0

first level	second level	third level	fourth level
Attribute: word	Attribute: word	Attribute: word	Attribute: word
Ignore case <input type="checkbox"/>	Ignore case <input type="checkbox"/>	Ignore case <input type="checkbox"/>	Ignore case <input type="checkbox"/>
6L	6L	6L	6L
5L	5L	5L	5L
4L	4L	4L	4L
3L	3L	3L	3L
2L	2L	2L	2L
1L	1L	1L	1L
Node	Node	Node	Node
1R	1R	1R	1R
2R	2R	2R	2R

Make frequency list

### Text type frequency distribution

Frequency limit: 0

Include categories with no hits:

- group.type
- group.title
- group.urldomain
- group.uri
- group.year
- group.month
- group.date
- group.time

Make frequency list

**JANES** **SketchEngine: izdelava frekvenčnih seznamov** IJS FILOZOFSKA FAKULTETA

word	Frequency	tag	Frequency
mi	28,577	Zop-ed--k	26,758
nas	20,637	Zop-el	18,216
nam	13,254	Zop-md	13,551
me	12,546	Zop-et--k	13,053
jaz	10,255	Zop-mt	11,685
Jaz	6,089	Zop-mm	7,341
mene	2,810	Zopmmi	4,773
MI	2,597	Zop-ed	3,799
meni	2,396	Zop-er	2,446
Meni	1,894	Zop-mr	1,824
Me	1,506	Zop-et	1,535
nami	1,313	Zop-mo	1,323
Mene	1,110		
jst	634		
JS	524		
mano	485		
nama	341		
MI	330		
js	303		
naju	270		
Nam	219		
Jst	209		
menoj	202		
midva	183		
Nas	172		
Js	126		

group.type	Frequency	Rel [%]
news.rtvsllo	86,647	277,660.70
news.mladina	20,788	115,597.10
news.reporter	2,367	42,154.00

text.std_ling	Frequency	Rel [%]
L2	56,163	104.40
L1	48,581	92.10
L3	5,058	154.40

text.senti	Frequency	Rel [%]
negative	80,128	106.00
neutral	15,784	73.50
positive	13,890	109.30

**JANES** **SketchEngine: besedni sezname** IJS FILOZOFSKA FAKULTETA

Corpus: JANES v0.4 News

Subcorpus: None (whole corpus) Info create new

Search attribute: word

use n-grams. Value of n: 2

**Filter options:**

Filter word list by: Regular expression:

Minimum frequency: 5

Maximum frequency: 0 (0 = no maximum frequency)

Whitelist: Choose File no file selected Clear

Blacklist: Choose File no file selected Clear format

Include non-words

**Output options:**

Frequency figures:  Hit counts  Document counts  ARF

Output type:  Simple  Keywords

Reference (sub)corpus: JANES v0.4 News (whole corpus)

Prefer: rare words  common words 1

Change output attribute(s)

## JANES SketchEngine: besedni sezname IJS



word	lc	lemma	Frekvenca
p   N jaz	jaz	jaz	10,255
p   N Jaz	jaz	jaz	6,089
p   N jst	jaz	jaz	634
p   N js	jaz	jaz	303
p   N JAZ	jaz	jaz	70
p   N jes	jaz	jaz	33
p   N jast	jaz	jaz	24
p   N JAz	jaz	jaz	7
p   N jales	jaz	jaz	5

## JANES SketchEngine: ključne besede IJS



word	JANES v0.4 News		JANES v0.4		Score
	Freq	Freq/mill	Freq	Freq/mill	
MIRNČAN	618	28.8	625	2.9	7.6
K_ris	587	27.4	587	2.7	7.6
law1	523	24.4	523	2.4	7.4
zapatist	495	23.1	497	2.3	7.3
Dandet	462	21.5	467	2.2	7.1
Jethros	440	20.5	447	2.1	7.0
ČAN	419	19.5	421	2.0	6.9
vojnasio91	426	19.9	432	2.0	6.9
Binder	445	20.8	460	2.1	6.9
Ramus	353	16.5	358	1.7	6.5
Tunek	320	14.9	330	1.5	6.3
šurda	299	13.9	303	1.4	6.2
Forex	328	15.3	371	1.7	6.0
IJJ	506	23.6	671	3.1	6.0
Mirnčan	263	12.3	268	1.2	5.9
Cmokc	237	11.1	237	1.1	5.7
oliva	292	13.6	335	1.6	5.7
rimos	228	10.6	228	1.1	5.6
olimpija	458	21.4	637	3.0	5.6
silvester	246	11.5	273	1.3	5.5
minuse	491	22.9	726	3.4	5.5
binbon	210	9.8	211	1.0	5.4
lojzek	232	10.8	257	1.2	5.4
gesan	198	9.2	198	0.9	5.3
SDS-a	712	33.2	1,204	5.6	5.2
martinove	189	8.8	197	0.9	5.1
čan	220	10.3	262	1.2	5.1
ti-ne	172	8.0	174	0.8	5.0
generusus	168	7.8	168	0.8	5.0

  

word	JANES v0.4 News		Kres		Score
	Freq	Freq/mill	Freq	Freq/mill	
Bratušek	1,615	75.3	15	0.1	67.9
KPK	1,334	62.2	35	0.3	49.0
Bratuškova	1,024	47.8	0	0.0	48.8
SMC	1,188	55.4	32	0.3	44.6
Cerarja	1,103	51.4	29	0.2	42.3
JJ	4,831	225.3	569	4.7	39.5
Prijavi	1,320	61.6	79	0.7	37.8
MIRNČAN	618	28.8	0	0.0	29.8
ZL	1,048	48.9	89	0.7	28.7
K_ris	587	27.4	0	0.0	28.4
Cerar	3,206	149.5	561	4.7	26.6
DUTB	528	24.6	0	0.0	25.6
law1	523	24.4	0	0.0	25.4
zapatist	495	23.1	0	0.0	24.1
Juncker	580	27.0	20	0.2	24.1
IJJ	506	23.6	7	0.1	23.2
Dandet	462	21.5	1	0.0	22.4
Janši	2,280	106.3	458	3.8	22.3
sds	683	31.9	59	0.5	22.1
Jethros	440	20.5	0	0.0	21.5
Ukrajini	1,583	73.8	299	2.5	21.5
vojnasio91	426	19.9	0	0.0	20.9
ČAN	419	19.5	1	0.0	20.4
Janše	3,033	141.4	735	6.1	20.1
Janšo	2,990	139.4	730	6.1	19.9
Patria	1,346	62.8	280	2.3	19.2
Bravo	3,963	184.8	1,051	8.7	19.1
rust	576	26.9	58	0.5	18.8
Zoki	815	38.0	131	1.1	18.7
Bratuškove	377	17.6	0	0.0	18.6
Bratuškovo	376	17.5	0	0.0	18.5
levica	1,545	72.1	356	3.0	18.5
Ukrajino	855	39.9	147	1.2	18.4

Korpus JANES

## DELAVNICA

- 5 skupin
- vsaka skupina si izbere eno temo
- nalogo skupaj s 5-minutno predstavitevijo je treba pripraviti v 60 minutah
- v zadnjih 30 minutah bo predstavitev dela vsake skupine
- struktura predstavitve:
  1. Ozadnje in motivacija
  2. Zasnova raziskave
  3. Rezultati
  4. Interpretacija in diskusija
  5. Sklepi

- Tema 1:
  - stopnja variantnosti na ortografski ravni med skupinami uporabnikov / pri posameznih uporabnikih
- Tema 2:
  - primerjava rabe izbranih slovničnih besednih vrst v spletni in govornjeni slovenščini
- Tema 3:
  - analiza sentimentnih korpusov
- Tema 4:
  - jezik komentarjev na ženske / moške politike na različnih portalih
- Tema 5:
  - mešanje jezikov / preklapljanje med jeziki