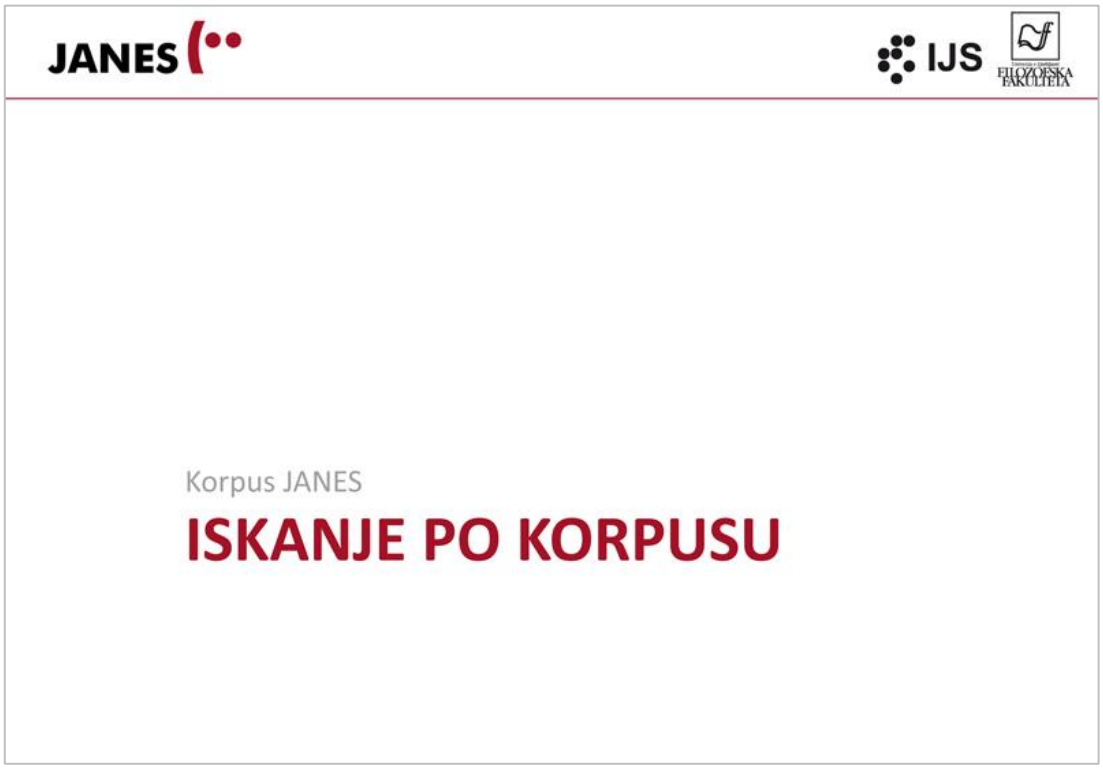
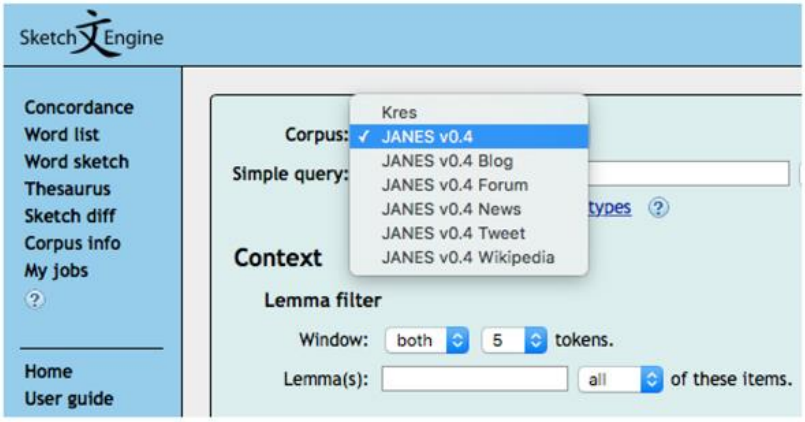
2.2 Raziskovanje spletne slovenščine – delavnica

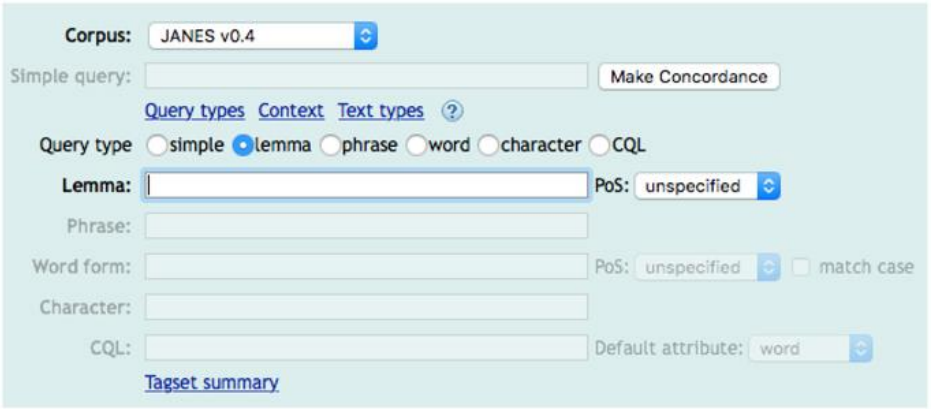


Dostop do korpusa

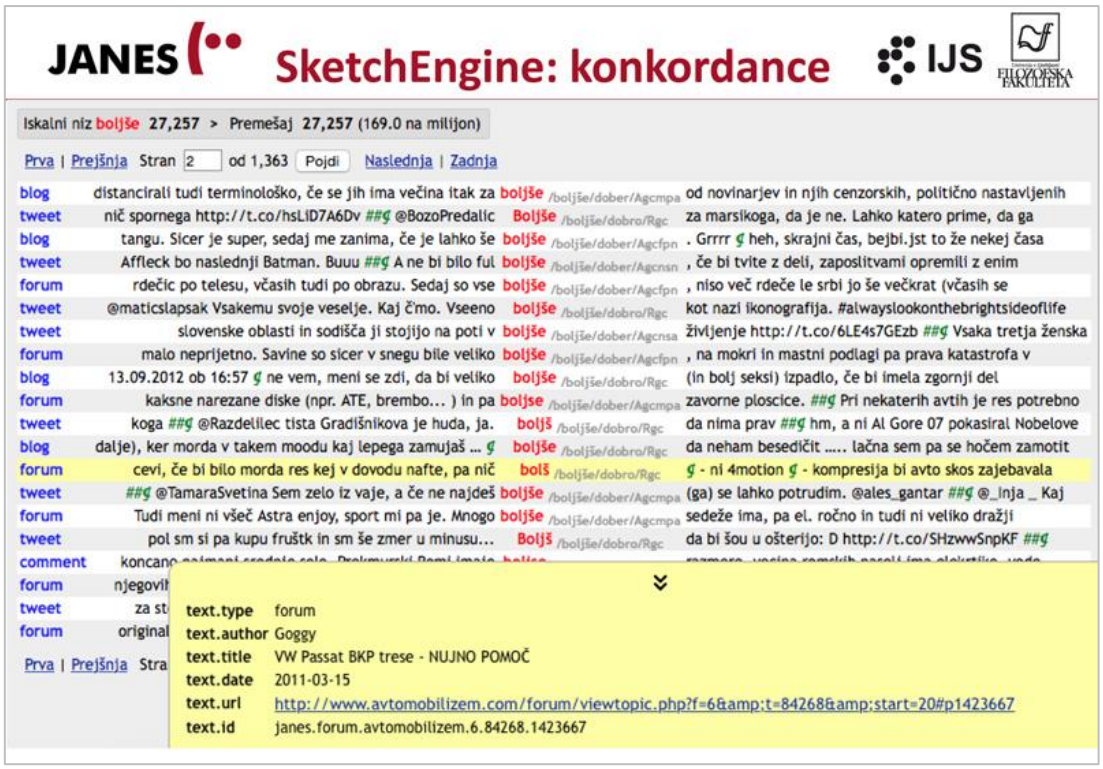
- konkordančnik
  - https://sketch.cjvt.si/bonito/janes04.cgi/first_form
- navodila za uporabo



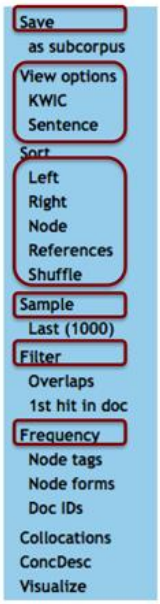
SketchEngine: načini iskanja



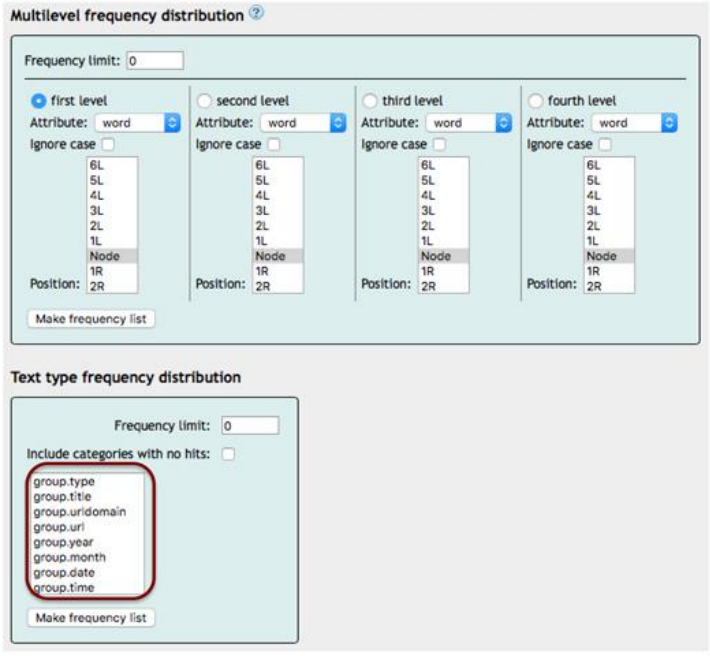

SketchEngine: filtriranje po metapodatkih
Text types
Subcorpus: None (whole corpus) Info create new
TEXT.USER
TEXT.SOURCE
corporate
private
Select All
TEXT.SEX
female
male
neutral
Select All
TEXT.LANG
eng
hbs
slv
und
Select All
TEXT.SENTI
negative
neutral
positive
Select All
TEXT.STD_TECH
T1
T2
T3
Select All
TEXT.STD_LING
L1
L2
L3
Select All
TEXT.YEAR
2013
2014
2015
2016
Select All

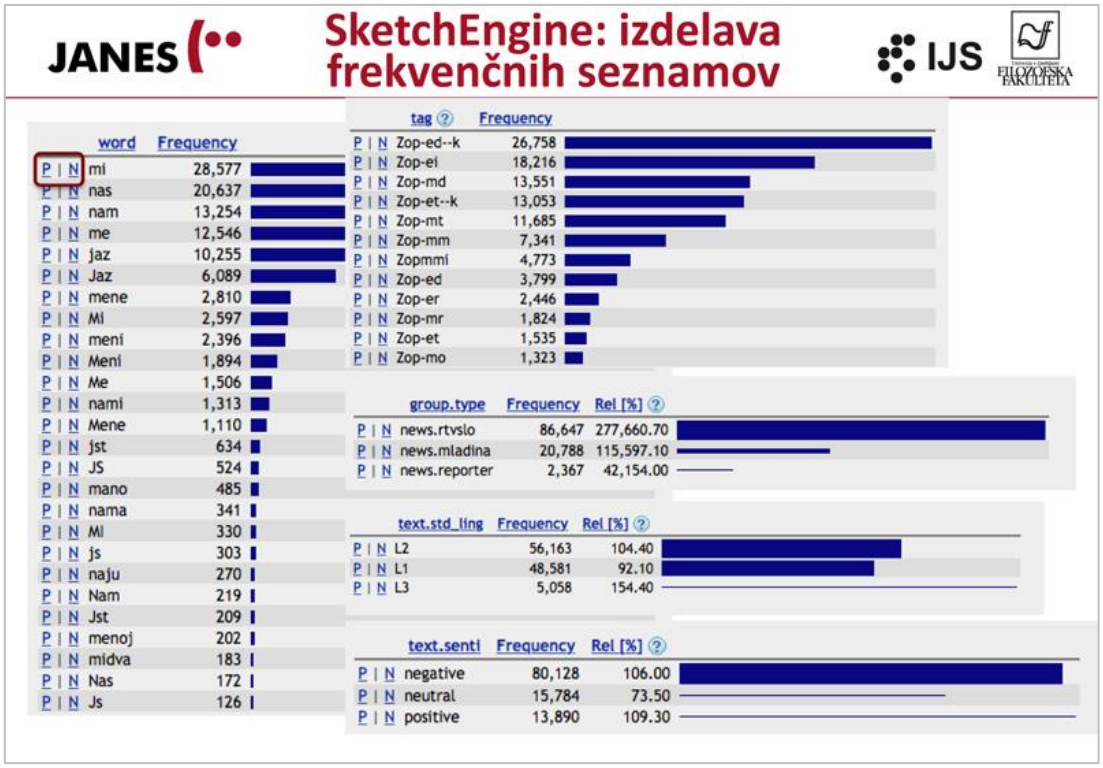

SketchEngine: konkordance
Iskalni niz boljše 27,257 > Premešaj 27,257 (169.0 na milijon)
Prva | Prejšnja Stran 2 od 1,363 Pojdi Naslednja | Zadnja
blog distancirali tudi terminološko, če se jih ima večina itak za boljše od novinarjev in njih cenzorskih, politično nastavljenih
tweet nič spornega http://t.co/hsLiD7A6Dv ##ǥ @BozoPredalic Boljše za marsikoga, da je ne. Lahko katero prime, da ga
blog tangu. Sicer je super, sedaj me zanima, če je lahko še boljše . Grrrr ǥ heh, skrajni čas, bejbi.jst to že nekej časa
tweet Affleck bo naslednji Batman. Buuu ##ǥ A ne bi bilo ful boljše , če bi tvite z deli, zaposlitvami opremili z enim
forum rdečic po telesu, včasih tudi po obrazu. Sedaj so vse boljše , niso več rdeče le srbi jo še večkrat (včasih se
tweet @maticslapsak Vsakemu svoje veselje. Kaj čmo. Vseeno boljše kot nazi ikonografija. #alwayslookonthebrightsideoflife
tweet slovenske oblasti in sodišča ji stojijo na poti v boljše življenje http://t.co/6LE4s7GEzb ##ǥ Vsaka tretja ženska
forum malo neprijetno. Savine so sicer v snegu bile veliko boljše , na mokri in mastni podlagi pa prava katastrofa v
blog 13.09.2012 ob 16:57 ǥ ne vem, meni se zdi, da bi veliko boljše (in bolj seksi) izpadlo, če bi imela zgornji del
forum kaksne narezane diske (npr. ATE, brembo...) in pa boljse zavorne ploscice. ##ǥ Pri nekaterih avtih je res potrebno
tweet koga ##ǥ @Razdelilec tista Gradišnikova je huda, ja. boljš da nima prav ##ǥ hm, a ni Al Gore 07 pokasiral Nobelove
blog dalje), ker morda v takem moodu kaj lepega zamujaš ... ǥ boljše da neham besedičit ..... lačna sem pa se hočem zamotit
forum cevi, če bi bilo morda res kej v dovodu nafte, pa nič bolš ǥ - ni 4motion ǥ - kompresija bi avto skos zajebavala
tweet ##ǥ @TamaraSvetina Sem zelo iz vaje, a če ne najdeš boljše (ga) se lahko potrudim. @ales_gantar ##ǥ @_inja _ Kaj
forum Tudi meni ni všeč Astra enjoy, sport mi pa je. Mnogo boljše sedeže ima, pa el. ročno in tudi ni veliko dražji
tweet pol sm si pa kupu fruštk in sm še zmer u minusu... Boljš da bi šou u ošterijo: D http://t.co/SHzwwSnpKF ##ǥ
comment koncano
forum njegovi
tweet za st
forum original
Prva | Prejšnja Stra
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&t=84268&start=20#p1423667
text.id janes.forum.avtomobilizem.6.84268.1423667

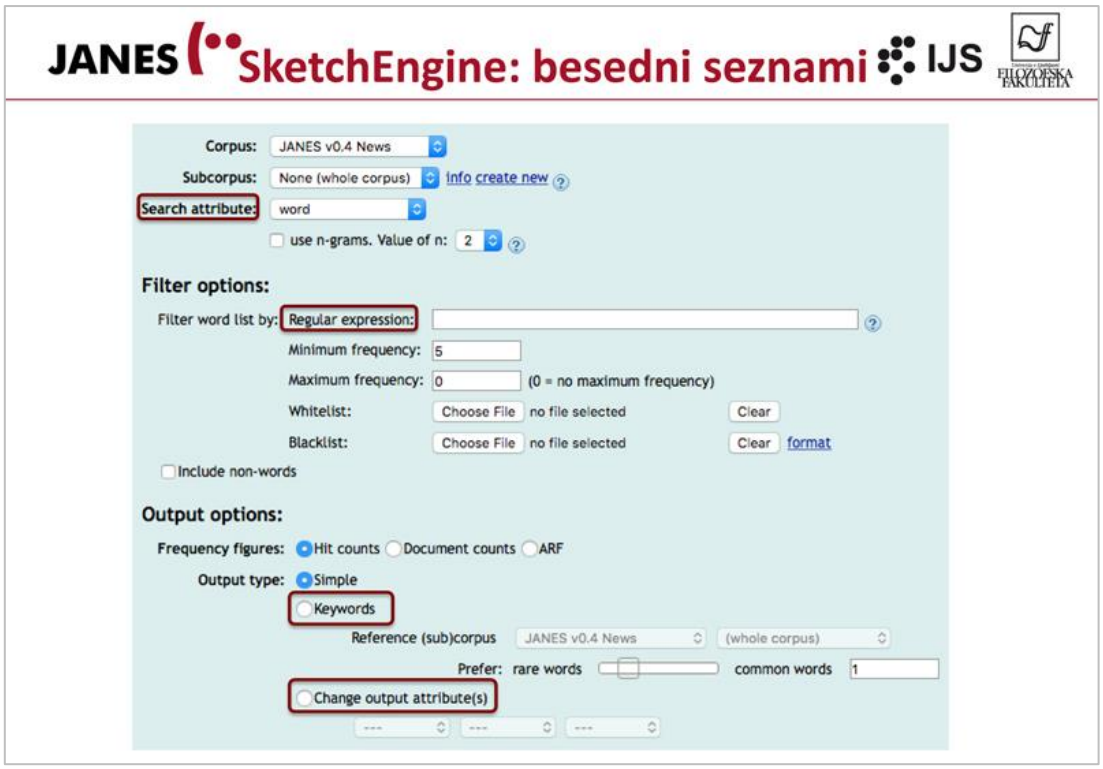
SketchEngine: obdelava konkordanc



SketchEngine: izdelava frekvenčnih seznamov




JANES
SketchEngine: izdelava frekvenčnih seznamov
IJS
word Frequency
P | N mi 28,577
P | N nas 20,637
P | N nam 13,254
P | N me 12,546
P | N jaz 10,255
P | N Jaz 6,089
P | N mene 2,810
P | N Mi 2,597
P | N meni 2,396
P | N Meni 1,894
P | N Me 1,506
P | N nami 1,313
P | N Mene 1,110
P | N jst 634
P | N JS 524
P | N mano 485
P | N nama 341
P | N MI 330
P | N js 303
P | N naju 270
P | N Nam 219
P | N Jst 209
P | N menoj 202
P | N midva 183
P | N Nas 172
P | N Js 126
tag Frequency
P | N Zop-ed--k 26,758
P | N Zop-ei 18,216
P | N Zop-md 13,551
P | N Zop-et--k 13,053
P | N Zop-mt 11,685
P | N Zop-mm 7,341
P | N Zopmmi 4,773
P | N Zop-ed 3,799
P | N Zop-er 2,446
P | N Zop-mr 1,824
P | N Zop-et 1,535
P | N Zop-mo 1,323
group.type Frequency Rel [%]
P | N news.rtvslo 86,647 277,660.70
P | N news.mladina 20,788 115,597.10
P | N news.reporter 2,367 42,154.00
text.std_ling Frequency Rel [%]
P | N L2 56,163 104.40
P | N L1 48,581 92.10
P | N L3 5,058 154.40
text.senti Frequency Rel [%]
P | N negative 80,128 106.00
P | N neutral 15,784 73.50
P | N positive 13,890 109.30


JANES
SketchEngine: besedni seznami
IJS
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 seznami

	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

	JANES v0.4 News		JANES v0.4		
word	Freq	Freq/mill	Freq	Freq/mill	Score
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
vojnaslo91	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

	JANES v0.4 News		Kres		
word	Freq	Freq/mill	Freq	Freq/mill	Score
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
vojnaslo91	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
rusi	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

Potek delavnice

- 5 skupin
- vsaka skupina si izbere eno temo
- nalogo skupaj s 5-minutno predstavitvijo 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

Raziskovalna vprašanja

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