

Viri, orodja za procesiranje in jezikoslovna analiza slovenskih spletnih uporabniških vsebin

Tomaž Erjavec

Odsek za tehnologije znanja
Inštitut “Jožef Stefan”

FDV

20. januar 2016

- 1. Ozadje: korpusno jezikoslovje**
2. Ozadje: jezikovne tehnologije
3. JANES: projekt in korpus
4. Konkordančniki
5. Par raziskav
6. Prihodnje delo

Želimo raziskovati slovenščino!



časopisi, revije

pošta, e-pošta



knjige



besedilni drobiž

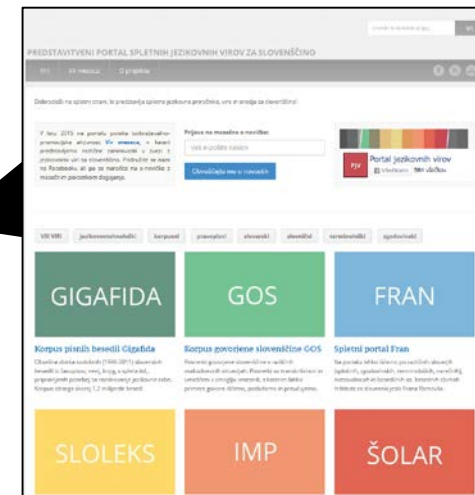
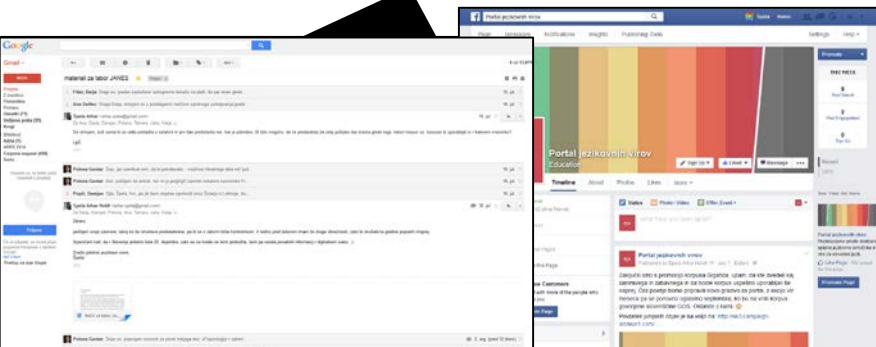


govorjeni jezik

itd.

spletne strani

družbena omrežja



Kako?

Gigafida | Iskanje ▼ | Okolica | Seznam

Pomoč | O korpusu | Slovensko ▼

čofotalnik

Najdi

Uporabljaš [enostavno iskanje](#)[Napredno iskanje](#)

1 2 3 4

[naslednja stran](#) ▶

Prikazujem 1-20 od 80 konkordanc (0.343 sekund).



Osnovne oblike

▶ [čofotalnik \(80\)](#)

Vrsta besedila

- ▶ [Časopisi \(63\)](#)
- ▶ [Revije \(12\)](#)
- ▶ [Internet \(4\)](#)
- ▶ [Stvarna besedila \(1\)](#)
- ▶ [Več](#)

Vir

- ▶ [Dnevnik \(28\)](#)
- ▶ [drugo \(22\)](#)
- ▶ [Delo \(11\)](#)
- ▶ [Večer \(6\)](#)
- ▶ [Gorenjski glas \(5\)](#)
- ▶ [Več](#)

Leto

- ▶ [2003 \(16\)](#)
- ▶ [2004 \(15\)](#)
- ▶ [2001 \(11\)](#)

sta bila od skupno 34 vzorcev dva mikrobiološko neustreznaz bazena **čofotalnika** na kopališču RC Cizej, Orla vas pri Braslovčah,

. To bo kompleks zunanjih bazenov, ki bo v **čofotalniku** ponujal vodno veselje najmlajšim, na grčah in toboganih pa

se lahko zabavajo tudi v zunanjem gusarskem bazenu in notranjem **čofotalniku** .

se lahko zabavajo tudi v zunanjem gusarskem bazenu in notranjem **čofotalniku** .

dodatno razveseljuje obnoven bazen s številnimi vodnimi doživetji, otroški **čofotalnik** in masažni bazen whirlpool. Celovita ponudba Dežele dobrega počutja

park z dvema drčama, toboganom in otroškim bazenom - **čofotalnikom** . Za krepitev zdravja vabi Sawaddee, center tradicionalne tajske

, posebne kopeli, za najmlajše pa je tu otroški **čofotalnik** . Gozdni vodni park ponuja vodne in biserne vrelece,

ki se bodo lahko namakali v dveh bazenih ter otroškem **čofotalniku** , se sprostili v velikem savna centru in telovadili v

park z dvema drčama, toboganom in otroškim bazenom - **čofotalnikom** . Za krepitev zdravja bodo poskrbeli v centru tradicionalne tajske

za dezinfekcijo vode. Kopališču manjkata tudi konpenzacijski bazen in **čofotalnik** za dojenčke ter manjše otroke, ki zdaj pogosto čofotajo

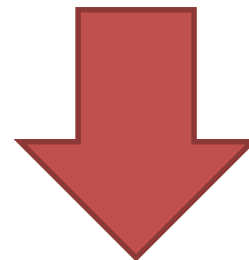
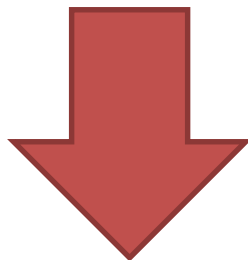
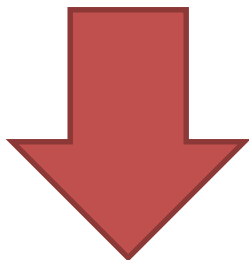
tiste, ki želijo toplejšo vodo. Predlagal bi » **čofotalnik** «.

v treh bazenih: zunanjem, v velikem bazenu in **čofotalniku** . V tamkajšnjem vročem bazenu oziroma savni pa je neustrezen

o., Zimsko kopališče v velikem in malem bazenu ter **čofotalniku** . V Zdravilišču Laško sta imela kemično neustrezno vodo zunanji

. Gre za kompleks zunanjih bazenov, ki bo v **čofotalniku** ponujal vodno veselje najmlajšim obiskovalcem, na grčah in toboganih

Besedilni korpus



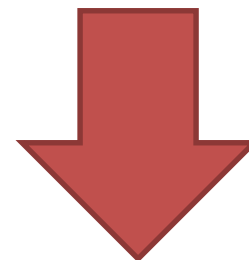
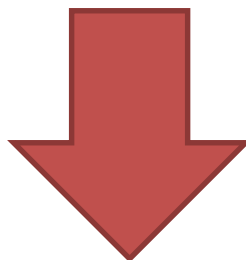
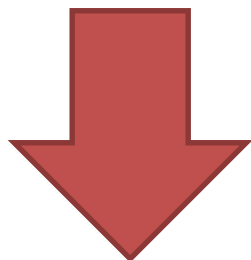
Zberemo veliko količino besedil v elektronski obliki.

Besedila izberemo tako, da predstavljajo vzorec jezika, ki nas zanima.

Besedila računalniško pripravimo, da jih je lažje raziskovati.

Besedila vstavimo v specializiran program: konkordančnik.

Besedilni korpus

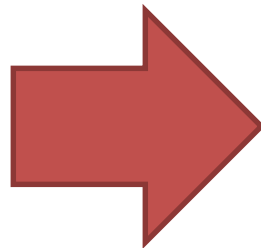


Zberemo veliko količino besedil v elektronski obliki.
Besedila izberemo tako, da predstavljajo **vzorec jezika**.
Besedila pripravimo, da jih je lažje računalniško raziskovati.
Besedila vstavimo v specializiran program: konkordančnik.

Vzorec - primer

Recimo, da nas zanima, kakšno glasbo poslušajo slovenski srednješolci. Kako bomo to ugotovili?

Je dovolj, če ocenimo stanje glede na svoje izkušnje? Vprašamo svoje sošolce? Ali vse dijake na svoji šoli? Dijake z različnih srednjih šol? Morda različnih starosti, spolov ali iz različnih regij?



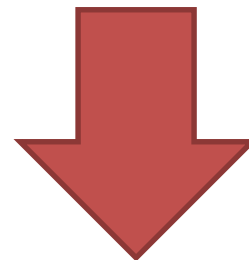
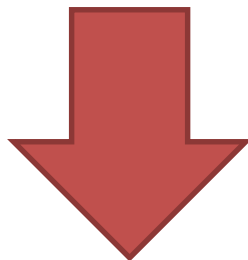
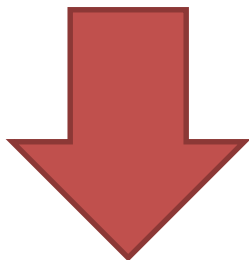
Kakšen **vzorec** moramo zajeti v raziskavi, da lahko na koncu **generaliziramo** rezultate?

Vzorec - korpus

Besedilni korpusi so zgrajeni tako, da predstavljajo **vzorec jezika**. Da lahko pravilno **interpretiramo** in **generaliziramo** ugotovitve, moramo dobro razumeti, kakšna besedila določen korpus vsebuje, kako je bil zgrajen in kakšen je njegov namen.

Korpus Kres	Splošna pisna slovenščina	časopisi, revije, leposlovje, strokovna literatura, spletna besedila, besedilni drobiž
Korpus Gos	Govorjena slovenščina	televizijske in radijske oddaje, javni nastopi, sestanki, zasebna komunikacija
Korpus IMP	Starejša slovenščina	knjige, časopisi, revije 1800-1918
Korpus Šolar	Jezik šolarjev	šolski eseji in testi + učiteljski popravki

Besedilni korpus



Zberemo veliko količino besedil v elektronski obliki.
Besedila izberemo tako, da predstavljajo vzorec jezika.
Besedila **pripravimo**, da jih je lažje računalniško raziskovati.
Besedila vstavimo v specializiran program: konkordančnik.

Priprava korpusnih besedil

1. Besedila **enotno zakodiramo**:

- zapis znakov: Unikod
- struktura: XML

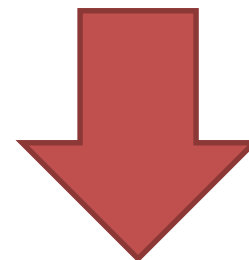
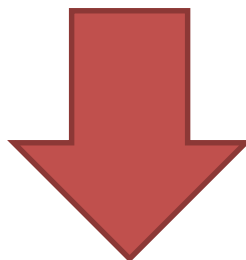
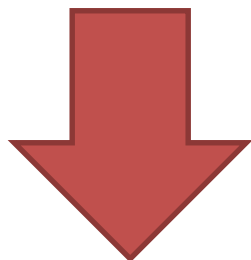
2. Besedilom dodamo **metapodatke**:

- izvor besedila
- kdaj je nastalo oz. izšlo
- avtor
- naslov
- ...

3. Besedila **jezikoslovno označimo** s posebnimi programi:

- vsaki besedi pripišejo osnovno obliko,
- besedno vrsto
- in mogoče še kaj...

Besedilni korpus



Zberemo veliko količino besedil v elektronski obliki.
Besedila izberemo tako, da predstavljajo vzorec jezika.
Besedila pripravimo, da jih je lažje računalniško raziskovati.
Besedila vstavimo v specializiran program: **konkordančnik**.

čofotalnik

Najdi

Uporabljaš [enostavno iskanje](#)[Napredno iskanje](#)

1 2 3 4

[naslednja stran](#) ▶

Prikazujem 1-20 od 80 konkordanc (0.343 sekund).



Osnovne oblike

▶ [čofotalnik \(80\)](#)

Vrsta besedila

- ▶ [Časopisi \(63\)](#)
- ▶ [Revije \(12\)](#)
- ▶ [Internet \(4\)](#)
- ▶ [Stvarna besedila \(1\)](#)
- ▶ [Več](#)

Vir

- ▶ [Dnevnik \(28\)](#)
- ▶ [drugo \(22\)](#)
- ▶ [Delo \(11\)](#)
- ▶ [Večer \(6\)](#)
- ▶ [Gorenjski glas \(5\)](#)
- ▶ [Več](#)

Leto

- ▶ [2003 \(16\)](#)
- ▶ [2004 \(15\)](#)
- ▶ [2001 \(11\)](#)

sta bila od skupno 34 vzorcev dva mikrobiološko neustreznaz bazena **čofotalnika** na kopališču RC Cizej, Orla vas pri Braslovčah,. To bo kompleks zunanjih bazenov, ki bo v **čofotalniku** ponujal vodno veselje najmlajšim, na grčah in toboganih pase lahko zabavajo tudi v zunanjem gusarskem bazenu in notranjem **čofotalniku** .se lahko zabavajo tudi v zunanjem gusarskem bazenu in notranjem **čofotalniku** .dodatno razveseljuje obnovljen bazen s številnimi vodnimi doživetji, otroški **čofotalnik** in masažni bazen whirlpool. Celovita ponudba Dežele dobrega počutjapark z dvema drčama, toboganom in otroškim bazenom - **čofotalnikom** . Za krepitev zdravja vabi Sawaddee, center tradicionalne tajske, posebne kopeli, za najmlajše pa je tu otroški **čofotalnik** . Gozdni vodni park ponuja vodne in biserne vrele,ki se bodo lahko namakali v dveh bazenih ter otroškem **čofotalniku** , se sprostili v velikem savna centru in telovadili vpark z dvema drčama, toboganom in otroškim bazenom - **čofotalnikom** . Za krepitev zdravja bodo poskrbeli v centru tradicionalne tajskeza dezinfekcijo vode. Kopališču manjkata tudi konpenzacijski bazen in **čofotalnik** za dojenčke ter manjše otroke, ki zdaj pogosto čofotajotiste, ki želijo toplejšo vodo. Predlagal bi » **čofotalnik** «.v treh bazenih: zunanjem, v velikem bazenu in **čofotalniku** . V tamkajšnjem vročem bazenu oziroma savni pa je neustrezeno., Zimsko kopališče v velikem in malem bazenu ter **čofotalniku** . V Zdravilišču Laško sta imela kemično neustrezno vodo zunanji. Gre za kompleks zunanjih bazenov, ki bo v **čofotalniku** ponujal vodno veselje najmlajšim obiskovalcem, na grčah in toboganih

Besedilni korpus - definicija

Besedilni korpusi so **obsežne zbirke realnih besedil** v elektronski obliki. Besedila so zajeta iz različnih virov na način, da predstavljajo **vzorec jezikovne rabe določene vrste**. Korpusna besedila tipično vsebujejo programsko ali ročno pripisane jezikoslovne **oznake**. Za raziskovanje besedilnih korpusov so besedila umeščena v **konkordančnike**, specializirane programe, ki omogočajo napredno iskanje po besedilih, razvrščanje, filtriranje, izvažanje podatkov in podobno.

Dober korpus je:

- sestavljen po vnaprej določenih kriterijih
- velik
- enotno kodiran
- dokumentiran

Vrste korpusov

- Referenčni oz. korpusi podjezikov
- Pisni oz. govorni korpusi
- Celoviti oz. vzorčni korpusi
- Enojezični oz. večjezični korpusi
- Statični oz. spremljevalni korpusi

Zakaj uporabljati korpus?

- Ker naša **jezikovna intuicija** ne more natančno predvideti, kako se jezik v širši rabi obnaša (kot ne moremo na pamet predvideti, kakšno glasbo imajo radi slovenski srednješolci).
- Ker s pomočjo računalnika lahko obdelamo večje količine podatkov na naprednejše načine in tako lažje poiščemo relevantne **jezikovne vzorce in trende** (to je posebej pomembno za večje projekte, npr. gradnjo slovarja),
- Ker so (v primerjavi z Googlom, na primer) zgrajeni na transparenten in dokumentiran način, da lahko podatke ustrezno **interpretiramo in generaliziramo**.

Raba besedilnih korpusov

Uporabno jezikoslovje

- slovarji, slovnice, pravopis, šolska gradiva ...

Teoretično jezikoslovje

- nova dognanja o jezikovni rabi in sistemu

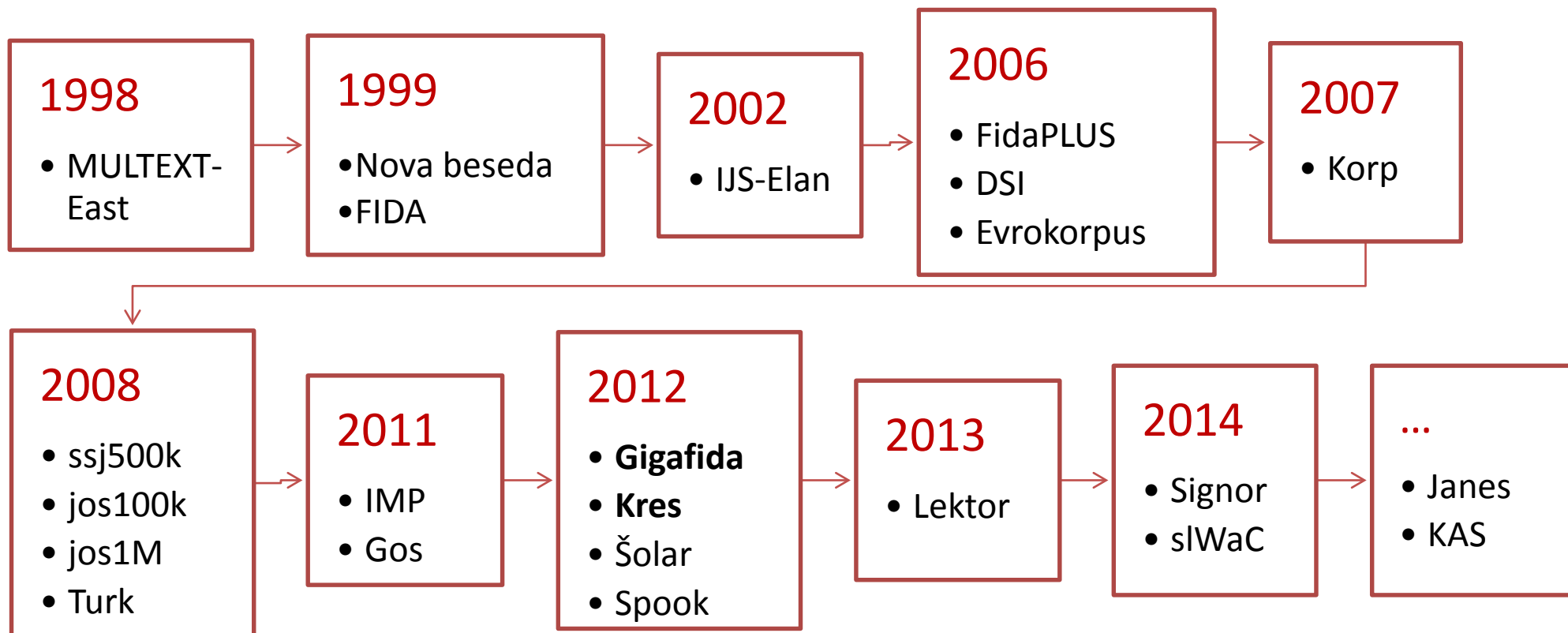
Jezikovna produkcija

- pisanje, prevajanje, lektoriranje ...

Ljubiteljsko raziskovanje

- preverjanje jezikovne rabe, zanimivosti ...

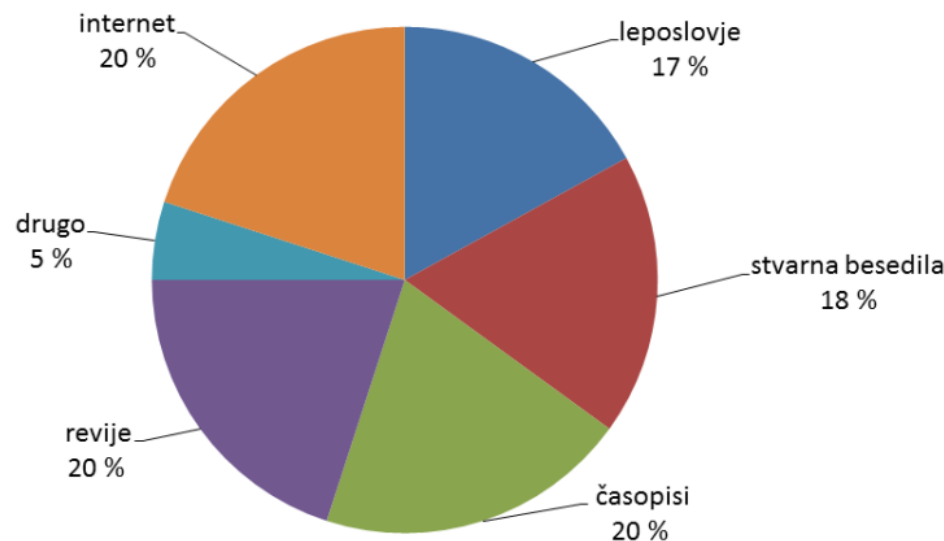
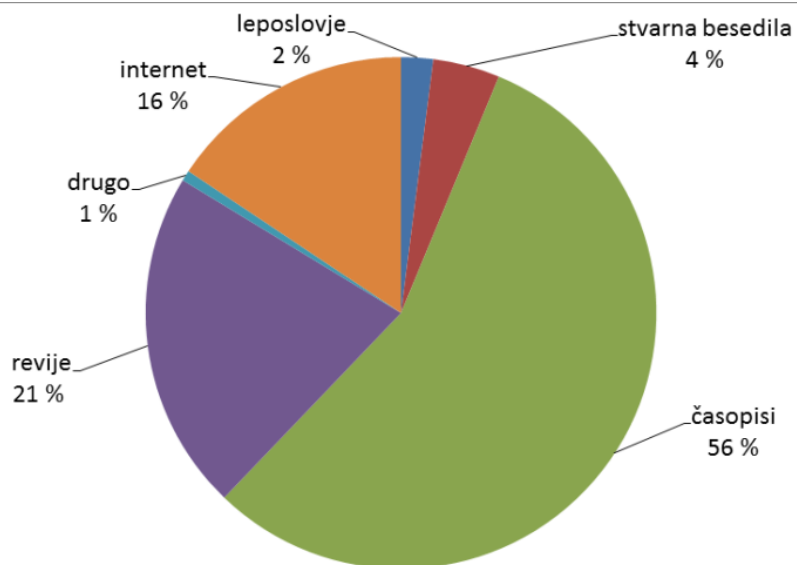
Slovenski korpusi (izbor)



Gigafida in Kres

- **Gigafida** je referenčni korpus: sodobna „splošna“ slovenščina
- Vsebuje sodobna (1990-2011) slovenska besedila iz časopisov, revij, knjig, s spleta itd.
- Korpus obsega skoraj 1,2 milijarde besed.
- **Kres** je manjša različica tega korpusa, cca. 100 milijonov besed.
- ccGigafida, ccKRES dostopna tudi za prenos (10 % velikosti).
- V raziskavah referenčne korpuse pogosto uporabljamo tako, da z njimi primerjamo rezultate iz drugih korpusov (npr. rezultate raziskovanja po korpusu Janes primerjamo z referenčnim korpusom, da vidimo, kaj je specifično za spletni jezik, kaj pa se pojavlja tudi v splošni jezikovni rabi).

Gigafida in Kres



1. Ozadje: korpusno jezikoslovje
- 2. Ozadje: jezikovne tehnologije**
3. JANES: projekt in korpus
4. Konkordančniki
5. Par raziskav
6. Prihodnje delo

- Računalniške metode za obdelavo jezikov
- Druga imena: računalniško jezikoslovje, procesiranje naravnega jezika
- Strojno prevajanje, iskanje po besedilih, povzemanje besedil, generiranje besedil; generiranje in razpoznavanje govora, identifikacija govorca, ...
- V kontekstu korpusnega jezikoslovja:
 - Jezikoslovno označevanje besedil
 - Tokenizacija, oblikoskladenjsko označevanje, lematizacija
 - Skladenjsko razčlenjevanje
 - Identifikacija in razvrščanje imen (osebno, geografsko, ustanova...)
 - Luščenje terminov itd.
 - Klasifikacija besedil

1. Tokenizacija (kaj je beseda, kaj ločilo)
2. Stavčna segmentacija (kaj je poved)
3. Oblikoskladenjsko označevanje (gramatične lastnosti)
4. Lematizacija (osnovna oblika besede)

Vaši malčki lahko varno uživajo v otroškem čofotalniku. (časopis Celjan, 2009)

Vaši	vaš	svojilni zaimек; 2. oseba, moški spol, množina, imenovalnik, množina svojine
malčki	malček	samostalnik, občno ime; moški spol, množina, imenovalnik
lahko	lahko	splošni prislov; nedoločena stopnja
varno	varno	splošni prislov; nedoločena stopnja
uživajo	uživati	glavni glagol; nedovršni, sedanjik, 3. oseba, množina
v	v	predlog; mestnik
otroškem	otroški	splošni pridevnik; nedoločena stopnja, moški spol, ednina, mestnik
čofotalniku	čofotalnik	samostalnik, občno ime; moški spol, ednina, mestnik
.	.	ločilo

- Naloga: pripisati vsaki besedi v besedilu njene oblikoskladenjske lastnosti (besedna vrsta, spol, sklon, ...)
- Primera:
 - „Hotel je drag“: **Somei** $\equiv$ samostalnik vrsta=občno_ime spol=moški število=ednina sklon=imenovalnik
 - „Hotel je domov“: **Ggnd-em** $\equiv$ glagol vrsta=glavni vid=nedovršni oblika=deležnik število=ednina spol=moški
- Potrebno (tudi) za lematizacijo:
 - *hotel / Somei* $\rightarrow$ *hotel*
 - *hotel / Ggnd-em* $\rightarrow$ *hoteti*
- Kaj so možne oznake?
 - MULTEXT-East V4, 1902 oznak za slovenščino
 - V slovenščini in angleščini: Somei $\equiv$ Ncmsn

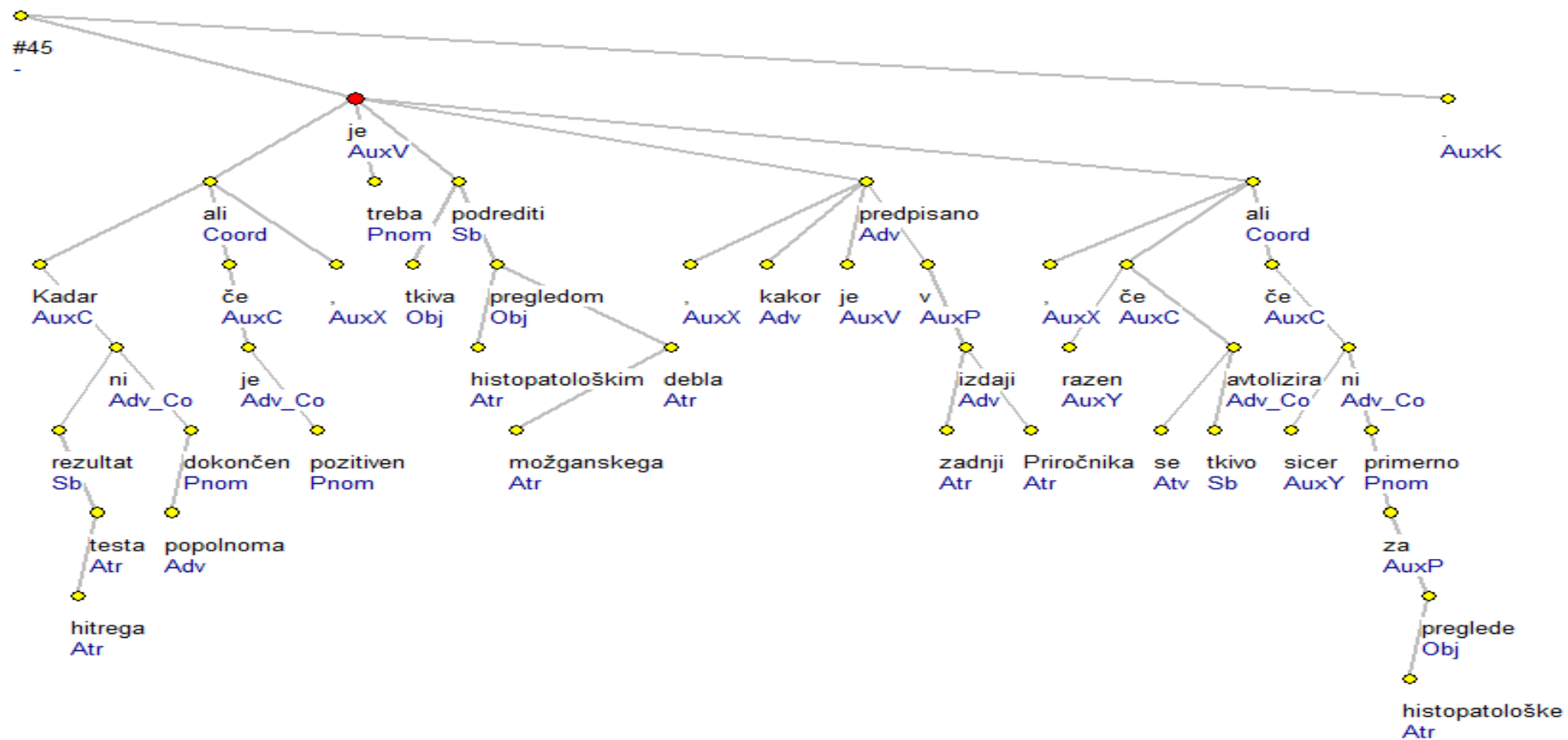
TTree Editor Default(2/1): C:\home\tomaz\My Projects\Syn\SVEZ\SVEZ_V0.1\svez-10-nina.fs

TredMacro

<From File>

45/50

#45 Kadar rezultat hitrega testa ni popolnoma dokončen ali če je pozitiven , je treba tkiva podrediti histopatološkim pregledom možganskega debela , kakor je predpisano v zadnji izdaji Priročnika , razen če se tkivo avtolizira ali če sicer ni primerno za histopatološke preglede .



- Za označevanje se tipično uporablja **nadzorovano strojno učenje**:
 - Ročno označena učna množica (drago in zamudno!)
 - Izbira primerne programa:
 - Metoda podpornih vektorjev (SVM)
 - Statistično strojno prevajanje (STM)
 - Skriti model Markova (HMM)
 - Pogojna naključna polja (CRF)
 - Nevronske mreže (NN)
 - Izbor značilk
 - Šolanje in testiranje programa
 - Uporaba
- Metode so splošne, učne množice jezikovno specifične (Vsak narod si bo učne množice označeval sam!)
- Za slovenski jezik: jos100k, jos1M, ssj500k, goo300k

1. Ozadje: korpusno jezikoslovje
2. Ozadje: jezikovne tehnologije
- 3. JANES: projekt in korpus**
4. Konkordančniki
5. Par raziskav
6. Prihodnje delo

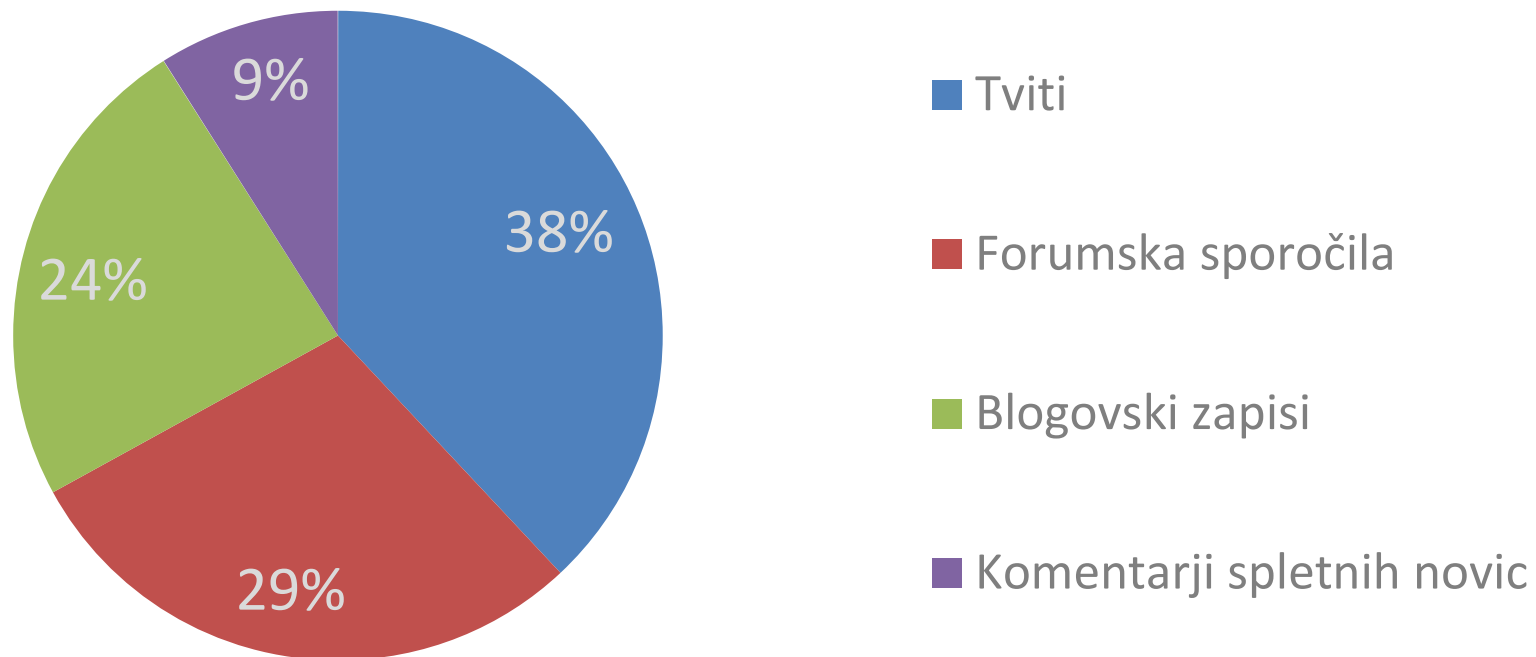
Uporabniško generirane spletne vsebine:

- Facebook, Twitter, blogi, forumi, komentarji na novice, ...
- Z razmahom družbenih omrežij strm porast takih vsebin
- Zaenkrat malo raziskav o „nestandardni“ slovenščini:
 - Nelektorirana, velikokrat na hitro napisna besedila
 - Tehnične bližnjice (presledki, ločila, csz)
 - Komunikacija nekje med pisno in govorno
 - Jezikovno inventivno (fonetični zapis, sleng, dialekti)
- Računalniška orodja za standardno slovenščino delajo slabo
- Zanimivo tudi za zajem informacij: tviti

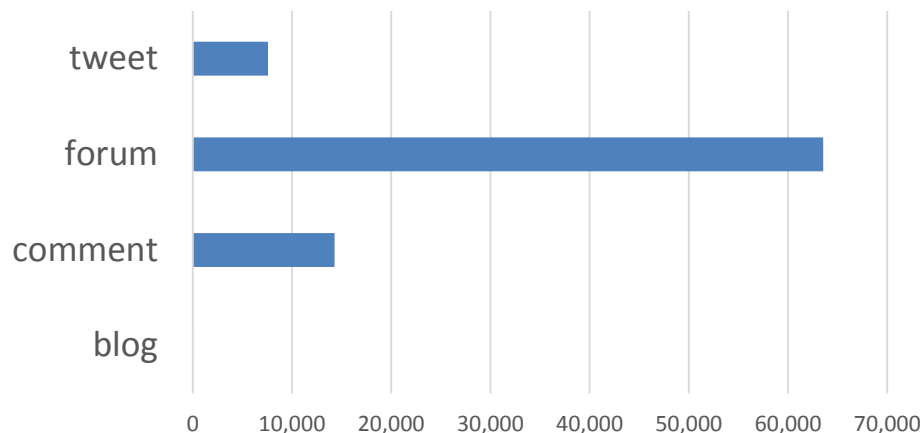
- Projekt JANES: <http://nl.ijs.si/janes>
 - Temeljni projekt ARRS, 3 leta: 2014-2017
 - Vodja: Darja Fišer, FF
 - Sodeluje: IJS
 - Cilj: razvoj virov, orodij in metod za analizo uporabniških spletnih vsebin
- Delovni sklopi:
 1. Izdelava korpusa uporabniško ustvarjene slovenščine
 2. Jezikoslovne študije
 3. Orodja za procesiranje nestandardne slovenščine

- Delovna različica korpusa Janes: v0.3
- Za vsako enoto:
 - Metapodatki
 - Besedilo
 - Jezikoslovne oznake besedila
- Trenutno dostopen samo za sodelavce projekta
 - Varovanje osebnih podatkov
 - Pogoji uporabe
 - Avtorske pravice

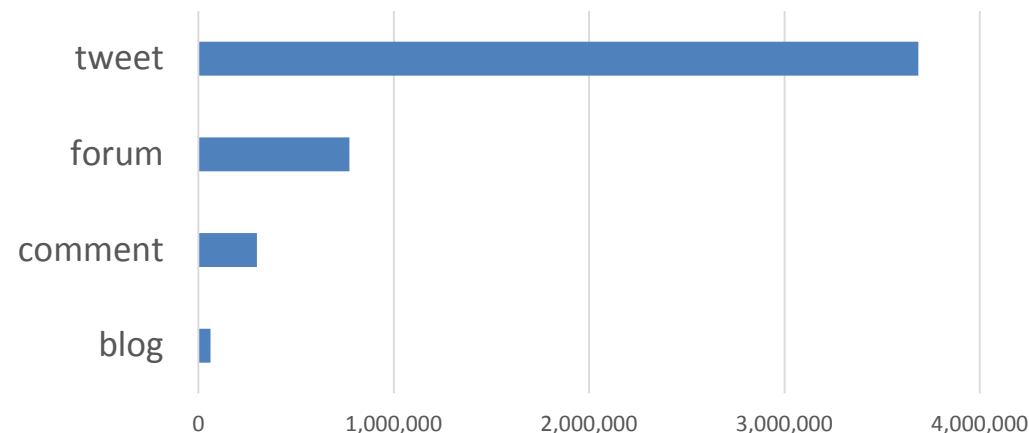
- Korpus trenutno obsega okrog **134 milijonov besed**
 - Blogi so zasilni (vzeti iz slWaC)
 - Manjkajo še komentarji s slovenske Wikipedije
 - Dodati nove tvite od sredine 2015 naprej



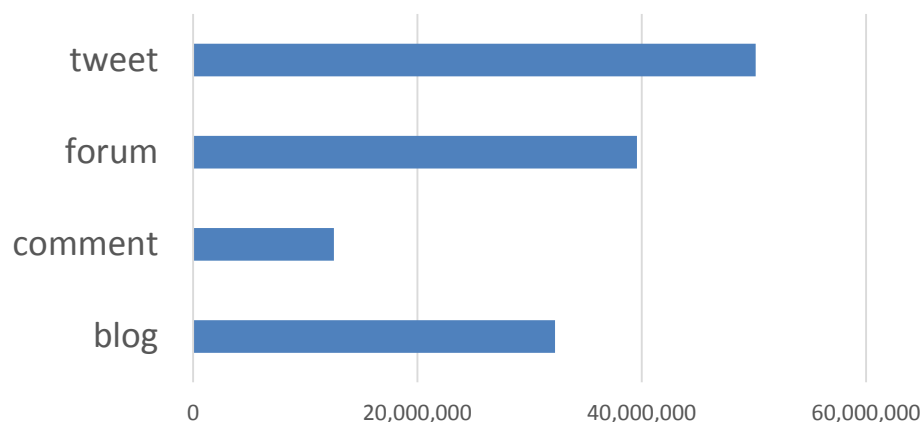
85.429 avtorjev



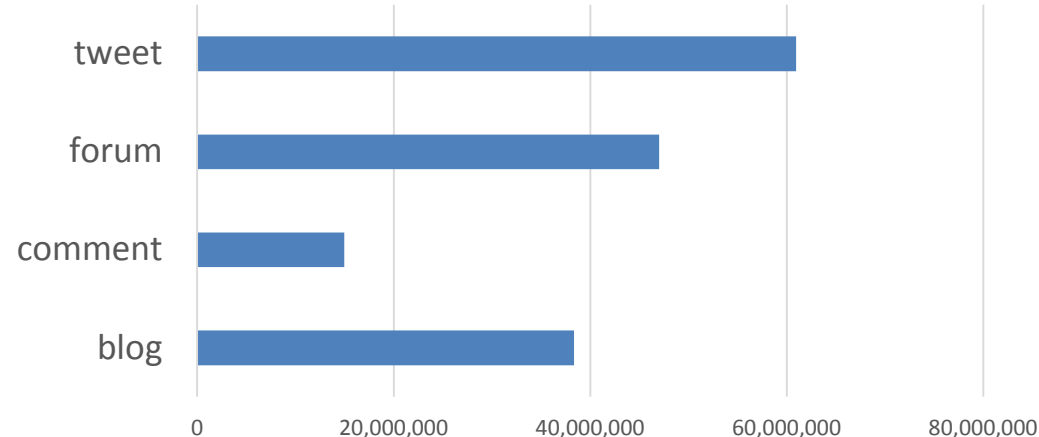
4.819.558 besedil



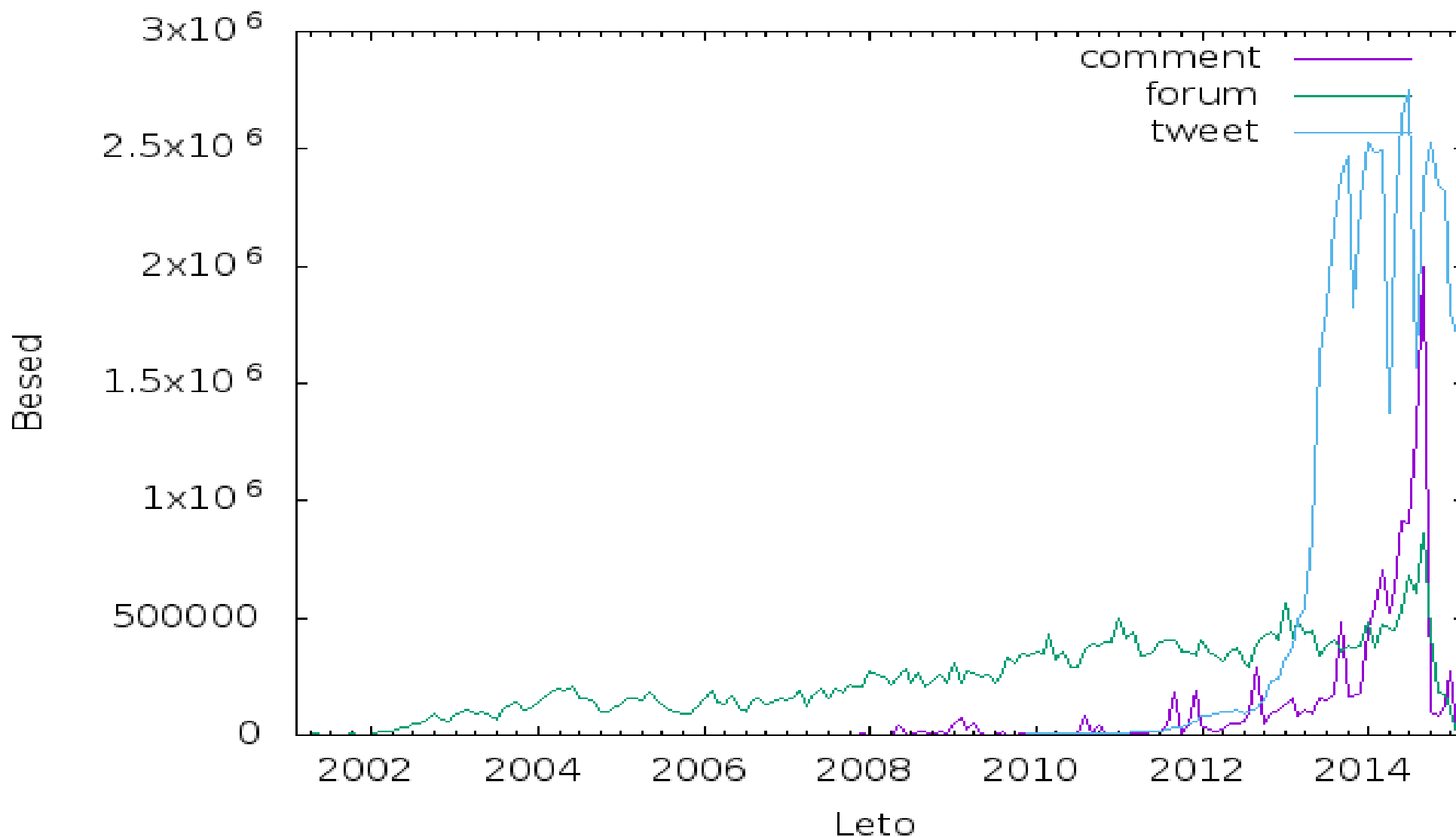
134.543.613 besed



161.289.153 pojavnic



Starost podkorpusov JANES v0.3

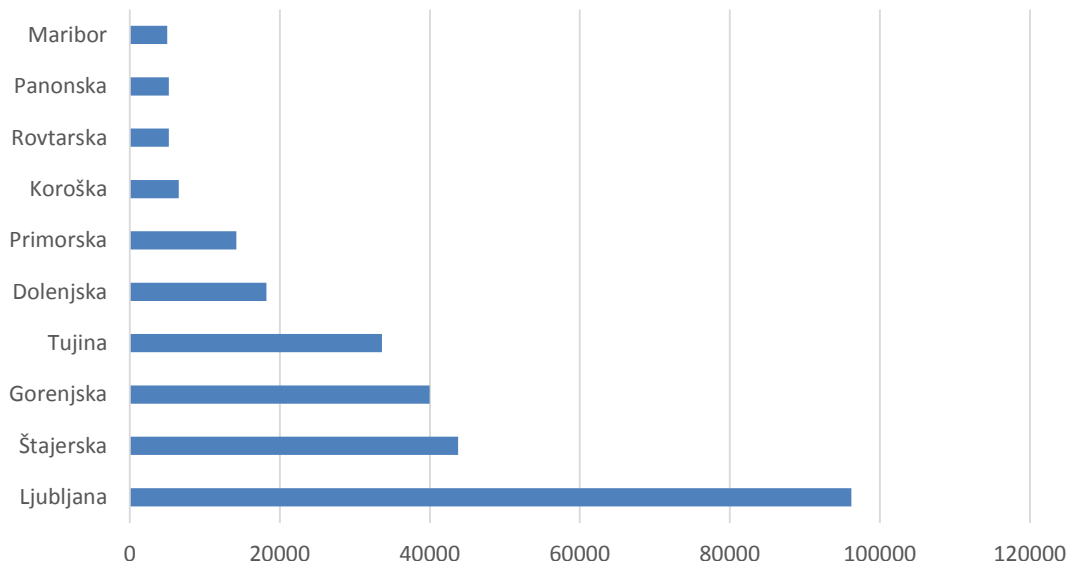


- TweetCat (Ljubešić et al. 2014):
 1. Twitter API:
vrne tvite uporabnika ali tvite, ki vsebujejo podane besede
 2. Začnemo s majhnim naborom tipično slovenskih besed
 3. Dobimo nekaj slovenskih tviterašev & zajamemo njihove tvite
 4. Pogledamo, komu sledijo in kdo sledi njim
 5. Zajamemo tudi te uporabnike in njihove tvite
 6. Pri vsakem testiramo, ali so slovenski
 7. Nazaj na 4
- Sprotni zajem
- Velik del slovenske tvitosfere:
4,3M tvitov, 53M besed (~Enciklopedija Britannica)

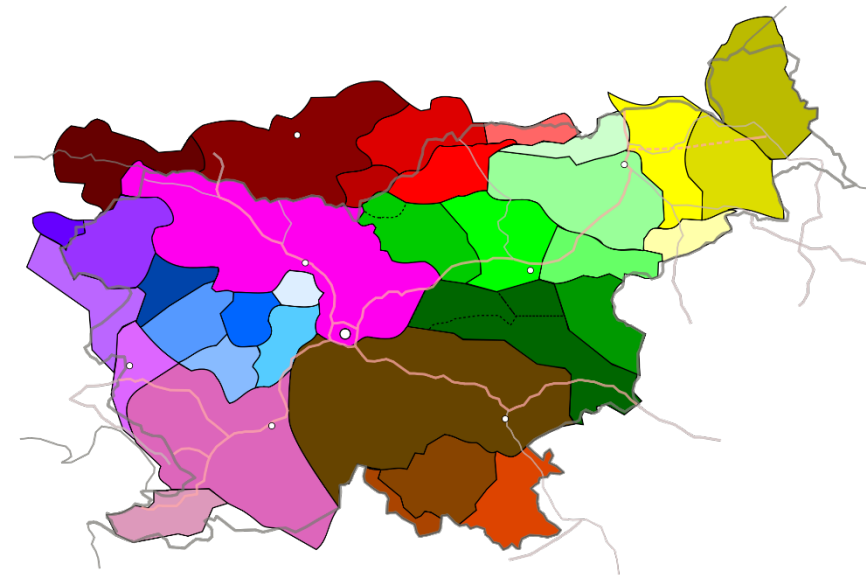
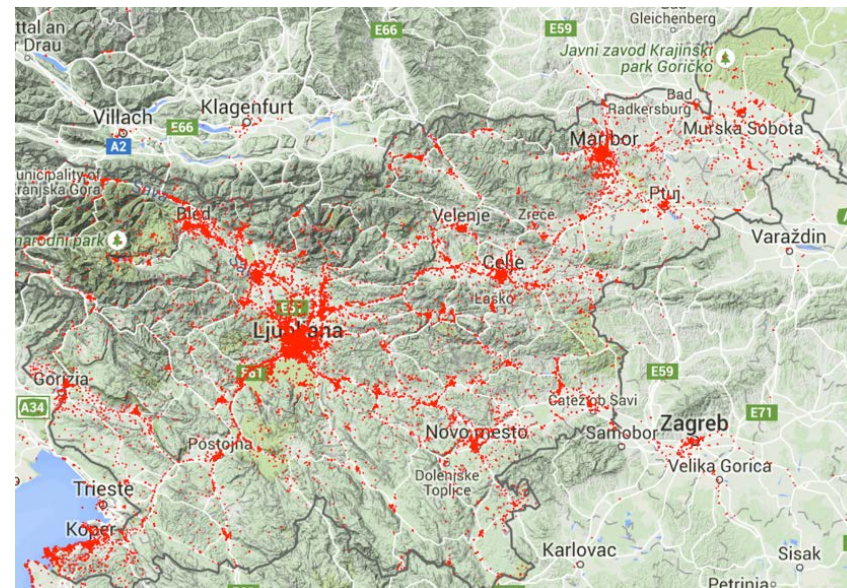
- Neposredni zajem:
 - Uporabniško ime, čas objave, št. všečkov in posredovanj
- Dodano za vsakega avtorja:
 - Vir: zasebni, organizacija (ročno)
 - Spol: moški, ženski, nevtrarno (ročno)
 - **Regija (avtomatsko)**
- Dodano vsakemu tvitu:
 - **Sentiment besedila (avtomatsko)**
 - **Stopnja standardnosti (avtomatsko)**

(Čibej & Ljubešić 2015)

- Zajem podatkov
 - 130K geolociranih tvitov
 - 1,7K različnih uporabnikov
- Določanje regije: 7 regij + Lj & MB

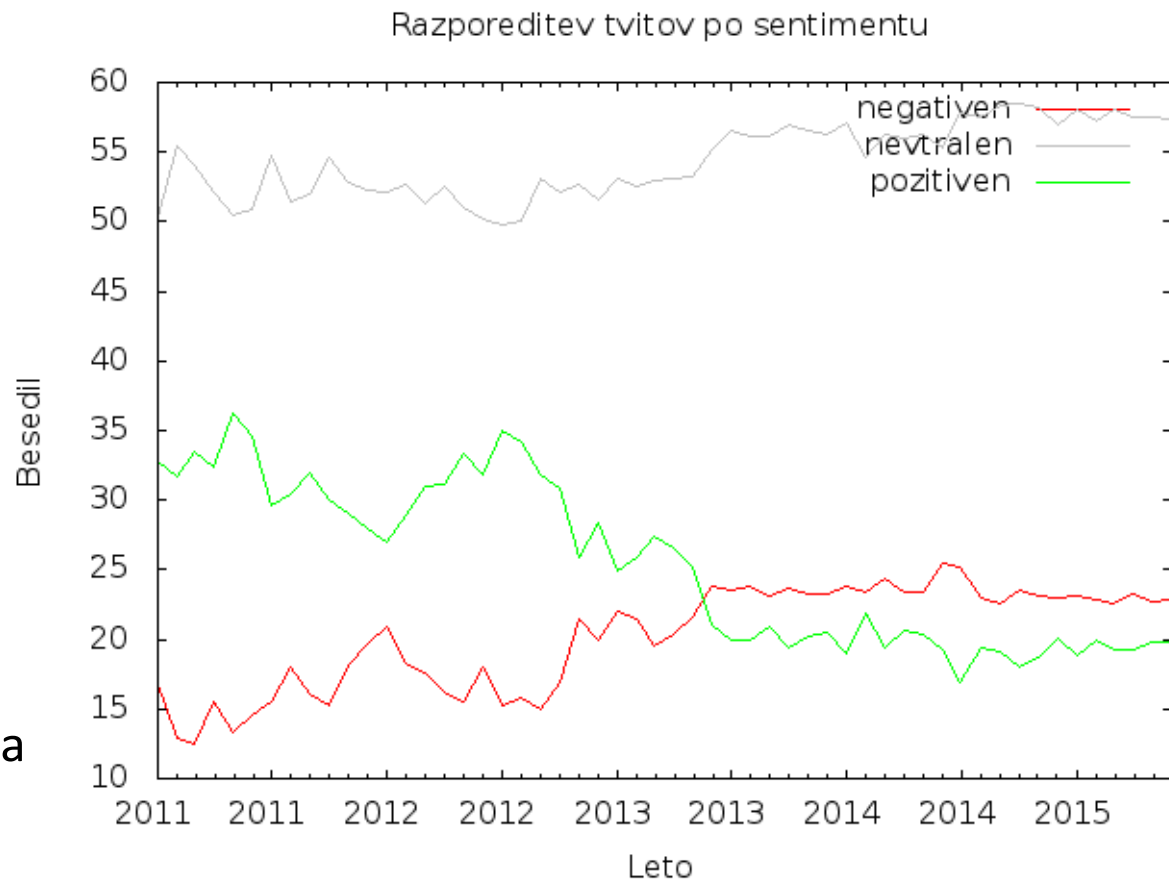


Osnova za dialektološke študije
„saj te pa lejko s trolo ideva no. :)“



Jasmina Smailović, IJS

- Avtomatsko označevanje
 - velika ročno označena podatkovna množica
 - SVM
- Evalvacija na 1,000 tvitih
 - Osnova = 37.7%
 - 1 anotator ~ 57.3%
 - 2 anotatorja = 62.1%
 - ujemanje med anotatorjema = 76.5%



„A veš tist, k si že itak tečen kt svina, potem je pa še twitter poln sranja o #znanobraz“
„0,5 m snegeca: D prav lepo je zunej:) pes srečen, micka se ne obremenjuje“

- Ljubešić et al., RANLP 2015
- Tehnična in jezikovna: T1 – T3, L1 – L3

T=1 / L=3

Original: *Ma men se zdi tole s poimenovanji oz s poslovenjenjem imen mest čist mem.*

Standardised: *Meni se zdi to s poimenovanji oz. s poslovenjenjem imen mest čisto mimo.*

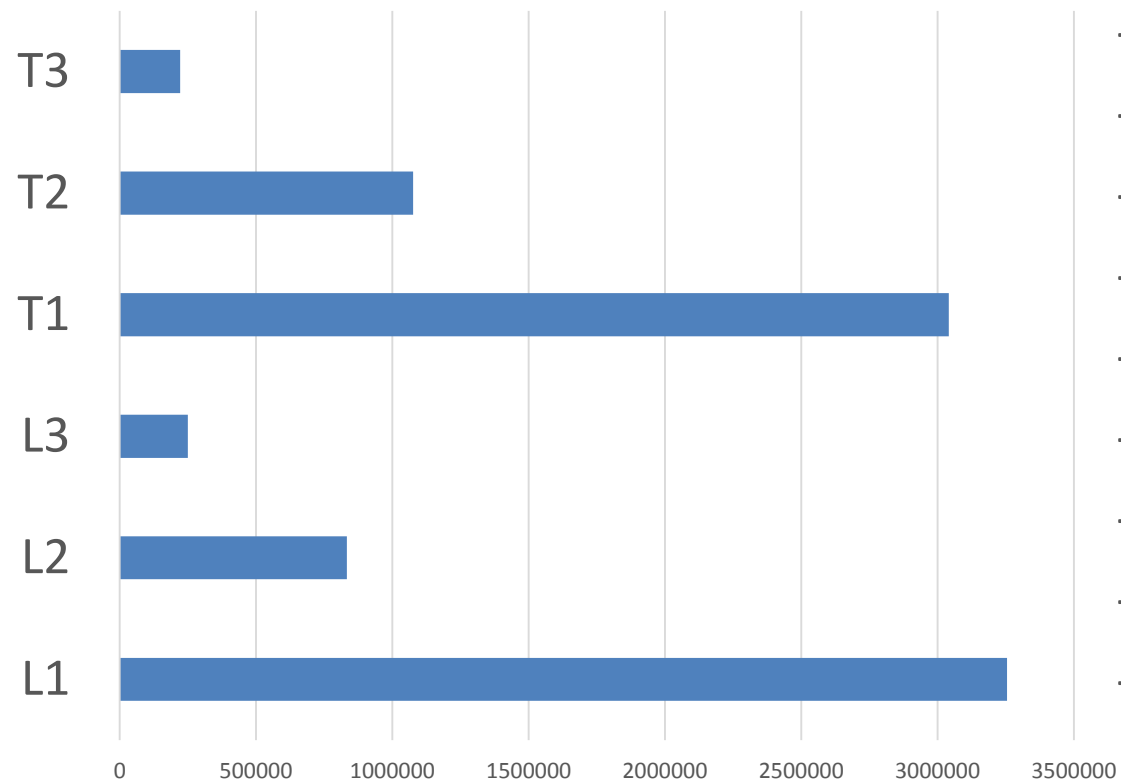
T=3 / L=1

Original: *se pravi, da predvidevaš razveljavitev*

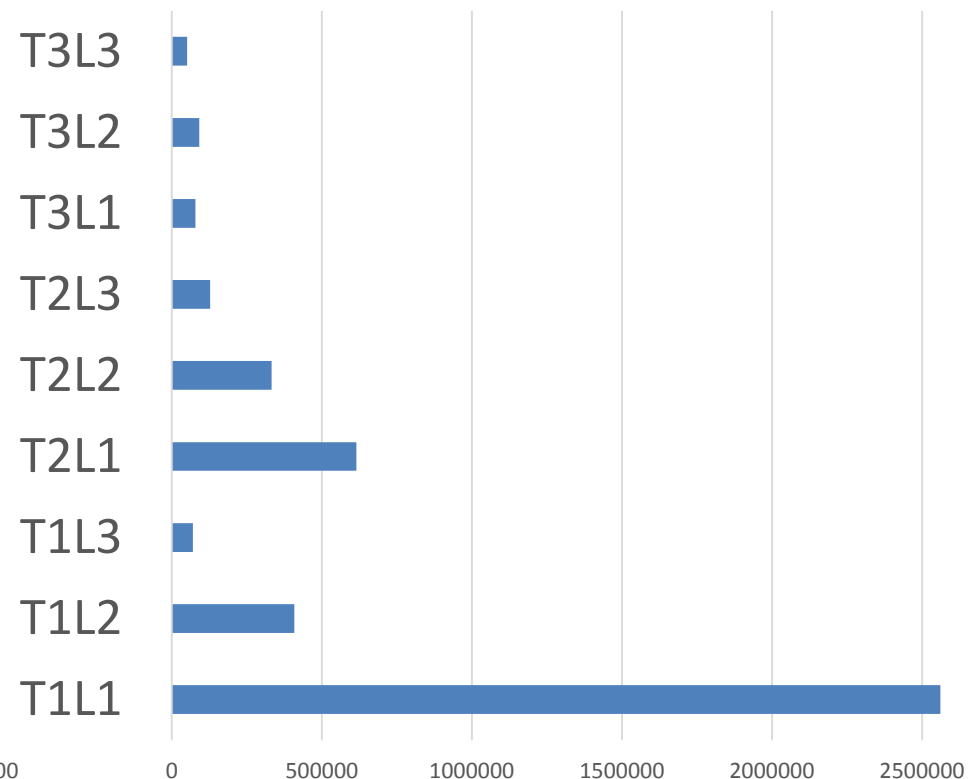
Standardised: *Se pravi, da predvidevaš razveljavitev?*

- Ročno označevanje
 - 50:50 standardna & nestandardna besedila
 - 900 tvitov označil 1 anotator (razvojna množica)
 - 400 tvitov označila 2 anotatorja (testna množica)
- Izbor značilk (29 značilk)
 - Na osnovi znakov:
 - ponavljanje znakov
 - razmerje med črkami in ne-črkami, med samoglasniki in soglasniki, ...
 - Na osnovi pojavnic:
 - delež zelo kratkih besed, besed z veliko začetnico, ...
 - delež besed, ki jih ni v Sloleksu
- Regresor
 - Optimizacija hiperparametrov z 10-kratnim prečnim preverjanjem na SVR regresorju z uporabo jedra RBF
- Evalvacija
 - Povprečna absolutna napaka: 0.377 T, 0.424 L

Standardnost T in L



Skupna standardnost



id="tid.392972411765018626"
name="007_delic"
created="2013-10-23T11:14:58"
retrieved="2013-10-24T05:20:32.036451"
favorited="0"
retweeted="0"
in_reply="tid.392965997352591361"
lang="sl" lang_prob="0.996788622457"
standard_tech="T1" standard_tech_n="1.1"
standard_ling="L1" standard_ling_n="1.2"
sentiment="neutral"
source="private"
sex="female"
geo="-"

- Metoda za nestandardni jezik:
 1. Standardizacija posameznih besed
 2. Uporaba metod za standardni jezik
- Prednosti:
 - Ne potrebujemo nove učne množice
 - Lahko tudi iščemo po standardnih besedah oz. lemah
- Normalizacija besed (Ljubešić et al. 2013)
 - Ročno izdelan leksikon parov
nestandardna beseda : standardna beseda
 - Statistično strojno prevajanje znakov (GIZA++ in Moses)

Source	g	l	e	d	u	
						\
Target	g	l	e	d	a	l

Trajnost, izmenljivost in ponovna uporaba:

1. Unikod (UTF-8)
2. XML
3. TEI: Text Encoding Initiative
4. ISO (kode jezikov, zapis časa)
5. MULTEXT-East

```
<s>
  <w lemma="pa" ana="#Cc">pa</w>
  <c> </c>
  <w lemma="še" ana="#Q">še</w>
  <c> </c>
  <choice>
    <orig>
      <w>tamali</w>
    </orig>
    <reg>
      <w lemma="ta" ana="#Pd-fsn">ta</w>
      <c> </c>
      <w lemma="mali" ana="#Agpmpn">mali</w>
    </reg>
  </choice>
  <pc ana="#Z">.</pc>
</s>
```

“tut drgač se da to napisat”

Poletni raziskovalni tabor
spletne slovenščine JANES
za srednješolce

Oddelek za prevajalstvo, Filozofska fakulteta,
Ljubljana

24.–28. avgust 2015

Poletni raziskovalni tabor spletne slovenščine za srednješolce je potekal na **Filozofski fakulteti** v Ljubljani.

Tabor je bil namenjen dijakom 3. in 2. letnika gimnazije, ki jih zanimata jezikoslovje in slovenščina, vključeval pa je predavanja, vaje in delavnice na temo slovenščine na spletu in v novih medijih. Poskrbljeno je bilo tudi za pester kulturni program.

Raziskovalni tabor je bil sofinanciran s sredstvi razpisa Ministrstva za kulturo za predstavljanje, uveljavljanje in razvoj slovenskega jezika (JPR-UPRS-2015) ter s sredstvi nacionalnega temeljnega projekta “Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine” (J6-6842), ki ga financira ARRS.

Raziskovalnemu taboru je novinarka Špela Šebenik posvetila oddajo **Gymnasium** na 1. programu Radia Slovenija, novinarka Nataša Ličen pa **Dijaško oddajo**, ki jo je pospremila tudi s **pisnim prispevkom**. **Poročilo o taboru** je bilo objavljeno tudi v reviji Slovenščina 2.0.

Program tabora

- Korpusi in korpusno jezikoslovje (**Špela Arhar Holdt** in **Jaka Čibej**)
- Standardna in spletna slovenščina (**Damjan Popič** in **Katja Zupan**)
- Frazeologija v spletni slovenščini (**Polona Gantar** in **Jaka Čibej**)
- Govor in spletna slovenščina (**Ana Zwitter Vitez** in **Špela Arhar Holdt**)
- Slovenščina v stiku z drugimi jeziki (**Tamara Mikolič Južnič** in **Katja Zupan**)

JANES Ekspres

December 2015
Ljubljana – Zagreb – Beograd

Utrinki

JANES Ekspres v okviru razpisa za promocijo slovenske znanosti v tujini sofinancira Javna agencija za raziskovalno dejavnost republike Slovenija (ARRS). Cilj projekta je skozi predavanja in delavnice predstaviti gradnjo, označevanje in analizo korpusa spletne slovenščine JANES raziskovalcem in študentom na Hrvaškem in v Srbiji.

Institut “Jožef Stefan”, Ljubljana, 13.–14. november 2015

Petek, 13. november

15.00–15.45	Predavanje: » Označevanje korpusov z orodjem WebAnno « (Kaja Dobrovoljc)
15.45–16.30	Demo: » Potek dela v WebAnno « (Kaja Dobrovoljc)
16.30–17.00	Odmor za kavo
17.00–18.00	Tutorial: » Označevanje v WebAnno « (Kaja Dobrovoljc)

Sobota, 14. november

1. Ozadje: korpusno jezikoslovje
2. Ozadje: jezikovne tehnologije
3. JANES: projekt in korpus
- 4. Konkordančniki**
5. Par raziskav
6. Prihodnje delo

- Orodje za raziskovanje korpusov
- Indeksirane vsebine za hitro iskanje
- Statistične mere (kolokacije)
- Za slovenščino jih obstaja več:
 - Projekt Sporazumevanje v slovenskem jeziku:
 - <http://www.gigafida.net/>
 - <http://www.korpus-kres.net/>
 - <http://www.korpus-gos.net/>
 - noSketchEngine na IJS:
 - <http://nl.ijs.si/noske/>

Najdi

Uporabljaš [enostavno iskanje](#)[Napredno iskanje](#)

1 2 3 4

[naslednja stran](#) ▶Prikazujem 1-20 od **80** konkordanc (0.0 sekund).

Osnovne oblike

▶ [čofotalnik](#) (80)

Vrsta besedila

▶ [Časopisi](#) (63)▶ [Revije](#) (12)▶ [Internet](#) (4)▶ [Stvarna besedila](#) (1)▶ [Več](#)

Vir

▶ [Dnevnik](#) (28)▶ [drugo](#) (22)▶ [Delo](#) (11)▶ [Večer](#) (6)▶ [Gorenjski glas](#) (5)sta bila od skupno 34 vzorcev dva mikrobiološko neustrezna iz bazena **čofotalnika** na kopališču RC Cizej, Orla vas pri Braslovčah,. To bo kompleks zunanjih bazenov, ki bo v **čofotalniku** ponujal vodno veselje najmlajšim, na grčah in toboganihse lahko zabavajo tudi v zunanjem gusarskem bazenu in notranjem **čofotalniku**.se lahko zabavajo tudi v zunanjem gusarskem bazenu in notranjem **čofotalniku**.dodatno razveseljuje obnovljen bazen s številnimi vodnimi doživetji, otroški **čofotalnik** in masažni bazen whirlpool. Celovita ponudba Dežele dobpark z dvema drčama, toboganom in otroškim bazenom - **čofotalnikom**. Za krepitev zdravja vabi Sawaddee, center tradiciona, posebne kopeli, za najmlajše pa je tu otroški **čofotalnik**. Gozdni vodni park ponuja vodne in biserne vrelece,ki se bodo lahko namakali v dveh bazenih ter otroškem **čofotalniku**, se sprostil v velikem savna centru in telovadili vpark z dvema drčama, toboganom in otroškim bazenom - **čofotalnikom**. Za krepitev zdravja bodo poskrbeli v centru tradicionaza dezinfekcijo vode. Kopališču manjkata tudi konjeniški bazen in **čofotalnik** za dojenčke ter manjša otroka, ki zdaj pogosto čofotajo

5 4 3 2 1 0

0 1 2 3 4 5

Uporabljaš [iskanje v okolici](#)

Najdi

1 2 3 4 5 6 7

Prikazujem 1-20 od **122** besed v okolici (0.0 sekund).

Osnovne oblike

▶ [čofotalnik \(80\)](#)

Besedna vrsta

- ▶ [pridevnik \(100\)](#)
- ▶ [samostalnik \(97\)](#)
- ▶ [veznik \(55\)](#)
- ▶ [predlog \(50\)](#)
- ▶ [glagol \(35\)](#)
- ▶ [Več](#)

Vrsta besedila

- ▶ [Časopisi \(63\)](#)
- ▶ [Revije \(12\)](#)
- ▶ [Internet \(4\)](#)
- ▶ [Stvarna besedila \(1\)](#)

beseda v okolici	pojavitve v okolici	pojavitve v korpusu	▼ LL
1 otroški	43	141.626	144,722
2 bazen	33	70.616	115,900
3 in	36	30.334.415	33,917
4 masažen	6	5.465	22,768
5 voden	7	94.664	18,382
6 urica	5	10.176	17,212
7 otrok	9	946.600	15,690
8 dojenček	5	27.365	15,064
9 za	17	15.515.714	14,411
10 žvrkljajoč	2	10	12,180

čofot*

Najdi

Uporabljaš [iskanje po seznamih](#)[Dodatno](#) ▼

1 2

Prikazujem 1-20 od 32 besed (0.0 sekund).

Besedna vrsta

- ▶ [glagol \(988\)](#)
- ▶ [samostalnik \(792\)](#)
- ▶ [pridevnik \(27\)](#)
- ▶ [prislov \(21\)](#)
- ▶ [Več](#)

Vrsta besedila

- ▶ [Časopisi \(824\)](#)
- ▶ [Revije \(701\)](#)
- ▶ [Internet \(144\)](#)
- ▶ [Leposlovje \(107\)](#)
- ▶ [Stvarna besedila \(49\)](#)
- ▶ [Več](#)

Vir

☒ prikaži seznam besed v osnovnih oblikah ☐ prikaži seznam besed v vseh oblikah

Osnovna oblika	Število pojavitev
čofotati	986
čofotanje	659
čofotalnik	80
čofotajoč	30
čofot	14
čofotalec	12
čofotaje	10
čofotač	5

1. Ozadje: korpusno jezikoslovje
2. Ozadje: jezikovne tehnologije
3. JANES: projekt in korpus
4. Konkordančniki
- 5. Par raziskav**
6. Prihodnje delo

[O projektu](#)[Raziskovalci](#)[Publikacije](#)[Viri](#)[Dogodki](#)[Interni](#)[English](#)

Konferenca

Slovenščina na spletu in v novih medijih

25.–27. november 2015

dvorana Zemljepisnega muzeja GIAM ZRC SAZU

Gosposka ulica 16

Ljubljana

Zbornik

Predkonferenčni tutorial

- **Maja Miličević:** »Beyond example extraction: Quantitative analysis of the JANES corpus«

Materiali:

- [\[izroček\]](#) [\[testdata1\]](#) [\[testdata2\]](#)

Obvestila

Janes v Izvidnici na Valu 202

Božično Izvidnico na Valu 202 so posvetili slovenščini, kot jo uporabljamo na spletu, v goste so povabili tudi sodelavce projekta Janes.

Posnetki predavanj s konference na videolectures

Posnetka vabljenega in plenarnega predavanja s konference Slovenščina na spletu in v novih medijih sta objavljena na portalu videolectures.net, vabljeni k ogledu.

[CMC-corpora]

Raziskovalci, ki se ukvarjamo s proučevanjem računalniško posredovane komunikacije, smo se povezali tudi na mednarodni ravni: <https://sites.google.com/site/cmccorpora/>

- 15 recenziranih prispevkov, 23 avtorjev, 50 udeležencev
- Nagrada za najboljši študentski prispevek
- Vabljen predavatelj (Michael Beißwenger, TU Dortmund)
- Panel
- Tutorial „Statistika in R za jezikoslovce (Maja Miličević, Uni Belgrade)
- VideoLectures

»Slovenščina Janes: pogovorna, nestandardna, spletna ali spretna?«

- **Vodila:** Špela Arhar Holdt
- **Razpravljali:** Marko Stabej, Helena Dobrovoljc, Simon Krek, Polona Gantar, Damjan Popič

Predstavitve prispevkov

- **Špela Arhar Holdt, Kaja Dobrovoljc:** »Zveze samostalnika z nesklonljivim levim prilastkom v korpusih Janes in Kres« [\[pdf\]](#)
- **Jaka Čibej, Nikola Ljubešić:** »S kje pa si? Metapodatki o regionalni pripadnosti uporabnikov družbenega omrežja Twitter« [\[pdf\]](#)
- **Špela Vintar:** »Terminologija v spletnih forumih« [\[pdf\]](#)
- **Senja Pollak:** »Identifikacija spletno specifičnih kolokacij pogostega besedišča« [\[pdf\]](#)
- **Eneja Osrajnik, Darja Fišer, Damjan Popič:** »Ekspresivna raba ločil v tvitih slovenskih uporabnic in uporabnikov« [\[pdf\]](#)
- **Martin Justin, Nejc Hirci, Polona Gantar:** »Rana ura, slovenskih fantov grob: analiza frazeoloških prenovitev v spletni slovenščini« [\[pdf\]](#)
- **Urška Vranjek Ošlak, Ajda Centa:** »Prava frekvenca – analiza žaljivega govora v spletnih komentarjih« [\[pdf\]](#)
- **Teja Rebernik:** »Slovenščina pod palcem interneta: vezajne in dvozačetniške tvorjenke« [\[pdf\]](#)
- **Iza Škrjanec, Darja Fišer, Damjan Popič:** »Arheologija začetnice pri stvarnih lastnih imenih« [\[pdf\]](#)
- **Teja Goli, Darja Fišer, Damjan Popič:** »Velika in mala dilema pri imenih industrijskih izdelkov na družbenem omrežju Twitter« [\[pdf\]](#)
- **Anja Krajnc, Marko Robnik-Šikonja:** »Postavljanje vejic v slovenščini s pomočjo strojnega učenja in izboljšane korpusa Šolar« [\[pdf\]](#)
- **Ana Zwitter Vitez, Darja Fišer:** »Elementi interakcije v govorjenih in spletnih besedilih« [\[pdf\]](#)
- **Kaja Dolar:** »Kaj nam o slovenščini lahko pove kolaborativni spletni slovar?« [\[pdf\]](#)
- **Mija Michelizza:** »Leksika na spletu (in kje jo iskati)« [\[pdf\]](#)

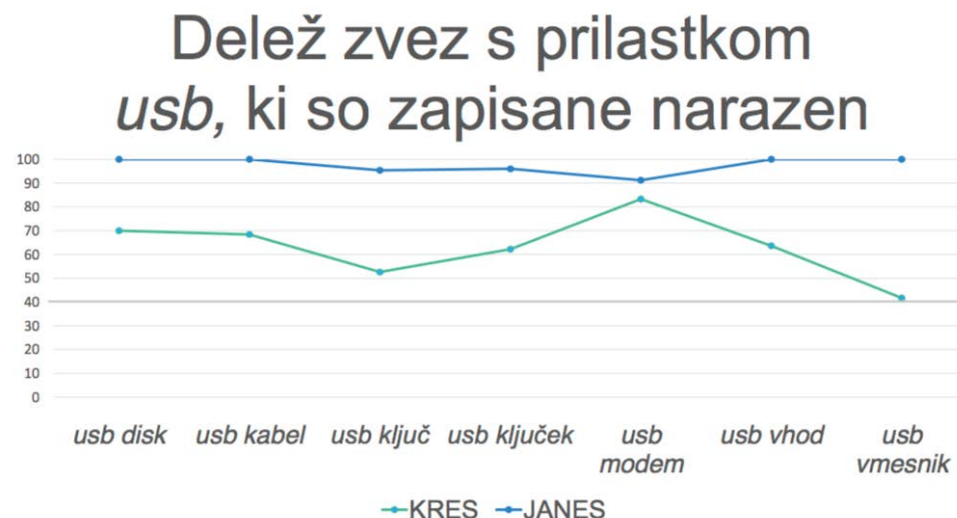
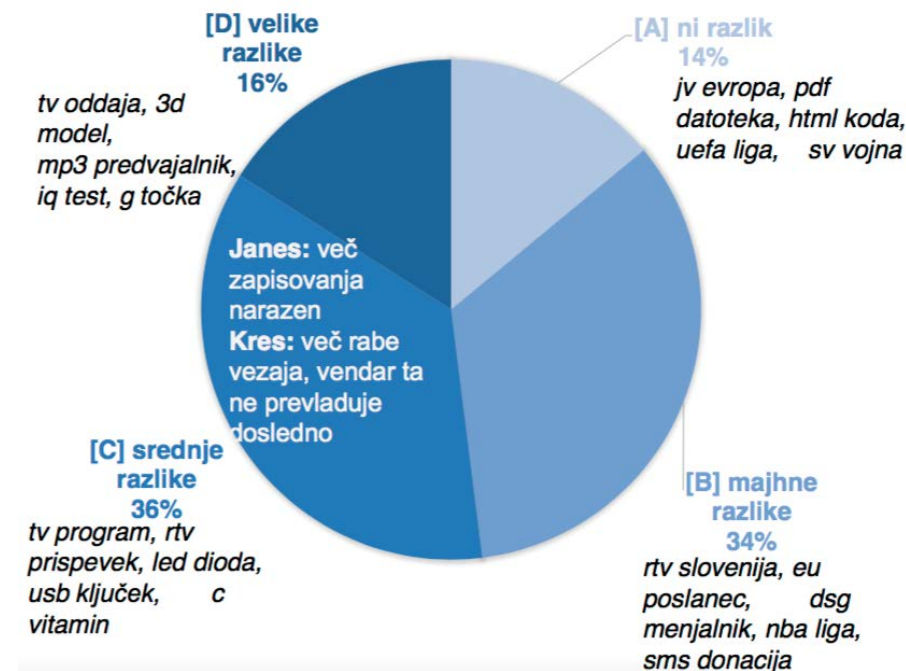
Arhar Holdt, Š., in Kaja Dobrovoljc, 2015.

Zveze samostalnika z nesklonljivim levim prilastkom v korpusih Janes in Kres.

Fišer, D. (ur.). *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete, 4–9.

<http://nl.ijs.si/janes/wp-content/uploads/2015/11/JANES15-01-Zveze-samostalnika.pdf>

- Rezultati v Kres in Janes se pomembno razlikujejo.
- Označevanje korpusov z različnimi označevalniki otežuje primerjavo podatkov.
- Skladenjski vzorec z nesklonljivim levim prilastkom presega primere, na podlagi katerih tipično teče jezikoslovna razprava.
- Obstoječo jezikoslovno tipologizacijo prilastkov je v praksi težko aplicirati.
- Jezikovna regulacija v smer zapisa skupaj/z vezajem krepi variantnost v jezikovni rabi.
- Močan trend k zapisu narazen, ampak variantnost je močno prisotna, tudi na ravni posameznih prilastkov, ki se v različnih zvezah različno zapisujejo.



Pollak, S. 2015.

Identifikacija spletno specifičnih kolokacij pogostega besedišča.

Fišer, D. (ur.). *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete, 57–56.

<http://nl.ijs.si/janes/wp-content/uploads/2015/11/JANES15-09-Primerjava-rabe.pdf>

- Izluščeni kandidati
 - Neformalni kolokatorji
 - Aktualne tematike, novi pojmi
 - Novi pomeni
 - Terminologija
 - Idiomatika
- Napake predprocesiranja
 - Izpuščanje diakritik
 - Napačna normalizacija
 - Napačna lematizacija
 - Neuspešna deduplikacija

Kolokacije (lema: družina)	
Kres	Janes
član družine	enostarševska družina
ribiška družina	član družine
kraljeva družina	primarna družina
lovska družina	kraljeva družina
ustvariti družino	mlada družina
rejniška družina	ogrožena družina
mlada družina	cela družina
kmečka družina	ustvariti družino
njegova družina	(n) članska družina
svoja družina	romska družina

Janes:8
199, 1,23/M

Kres:5,7
63; 0,52/M

Lema	Kolokatorji
država	vitek, zadolžiti, rušiti, pravično, ugrabiti, koruptiven, zavoziti, gnili, pokrasti, ugrabljen, zajebati
jezik	muca, eksotičen, tuja, ang, venetski, šparati
mama	foto, yo, platišče, mreža, odkar, doječ, vreme, kurac, feltna
moški	hetera, napasti, feminizacija, kastracija
zgodovina	servisen, preverljiv, smetisce, spisati, retuširanje, brisanje, odpasti

Justin, M., Hirci, N. in Polona Gantar, 2015.

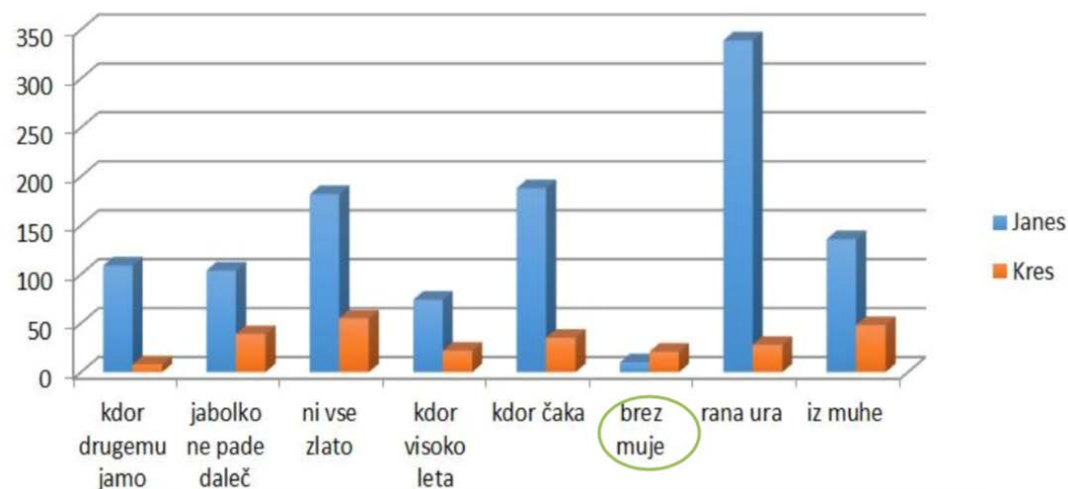
Rana ura, slovenskih fantov grob: analiza frazeoloških prenovitev v spletni slovenščini.

Fišer, D. (ur.). *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete, 33–37.

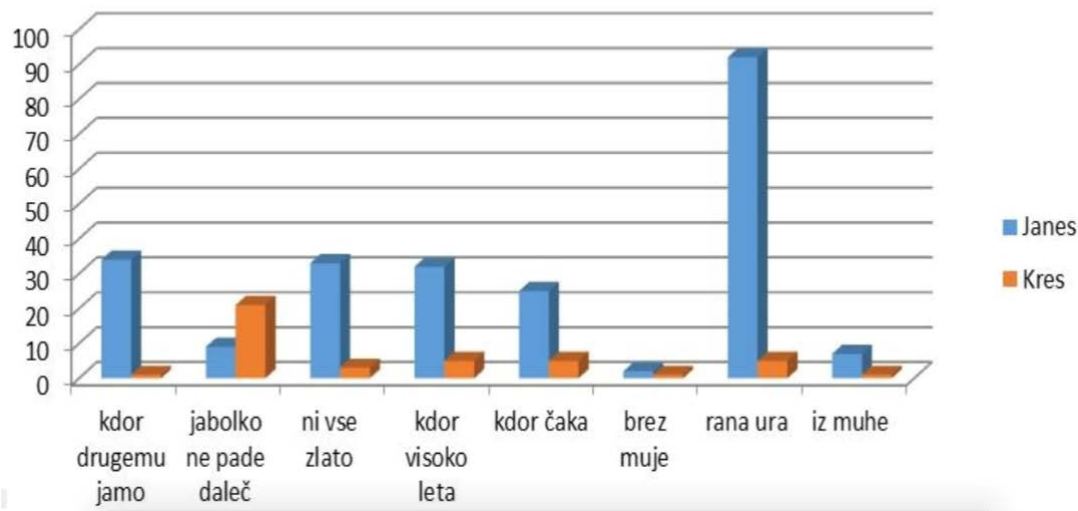
<http://nl.ijs.si/janes/wp-content/uploads/2015/11/JANES15-06-Rana-ura.pdf>

- Raba FE je za spletno specifična besedila tipična in glede na nespletna besedila pogostejša.
- Velik delež frazeoloških prenovitev v spletno specifičnih besedilih je mogoče obravnavati kot priljubljen način ustvarjanja humornega učinka.
- Doseganje humornega učinka poteka z izrabo ključnih pomenskih lastnosti sestavin izhodiščne FE (podobesedenje, nasprotni pomen, nepričakovana beseda itd.)

Frazeološke enote: pogostnost



Frazeološke prenovitve: pogostnost



- *kdor drugemu jamo koplje, je prvega že zasul*
- *kdor drugemu jamo koplje, je grobar*
- *kdor visoko leta, je pilot/ptič/stevardesa*
- *kdor visoko leta, daleč vidi*
- *kdor čaka, ima dobre živce*
- *kdor čaka, postane star*
- *rana ura, zaspana ura*

- *iz muhe delati **Airbus/Dumbota***
- *jabolko ne pade daleč od **hruške***
- *brez muje se še **papuč** ne sezuje*
- *ni vse **zlato**, kar je **nafta** (nafta = črno zlato)*
- ***žena** ne pade daleč od drevesa*

1. Ozadje: korpusno jezikoslovje
2. Ozadje: jezikovne tehnologije
3. JANES: projekt in korpus
4. Konkordančniki
5. Par raziskav
6. **Prihodnje delo**

- JANES v0.4 (marec 2016)
 - Komentarji z Wikipedije
 - Blogi (bolj zares)
 - Posodobljeni tviti
- Glavna naloga zadnjega leta projekta:
izboljšati označevanje na vseh ravneh
 - Tokenizacija, normalizacija, oblikoskladenjsko označevanje, lematizacija
 - Ročno označeni podatki (WebAnno)
 - Boljši programi (RE, CRF)
- = JANES v1.0

- Obljuba:
 - Prosto dostopno na konkordančniku
 - Odprto dostopno (CC) na repozitoriju CLARIN.SI
- Vendar:
 - Avtorske pravice
 - Zaščita zasebnosti
 - Pogoji uporabe
- Rešitve:
 - Anonimizacija, premešanje, vzorčenje
 - Mogoče različni pogoji za različne uporabnike
 - + spremljevalni korpus

Zahvala

- Špeli Arhar Holdt in Jaki Čibeju za prosojnice iz:

Arhar Holdt, Š. in Jaka Čibej: Korpusi in korpusno jezikoslovje, 2015. Fišer, D. (ur.): *Poletni raziskovalni tabor slovenščine Janes za srednješolce. Učna gradiva*. 1. izd. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. 9—22.

<http://nl.ijs.si/janes/wp-content/uploads/2015/10/Gradiva.pdf>

- Nikoli Ljubešiću za delo na večini orodij za označevanje Janesa
- Darji Fišer za dobro vodenje projekta

KAS KORPUS
AKADEMSKE
SLOVENŠČINE

O PROJEKTU

PARTNERJI

INTERNO

ENGLISH

Projekt ARRS J6-7094 "Slovenska znanstvena
besedila: viri in opis"

Razvoj in uporaba slovenskega znanstvenega jezika v visokem šolstvu ter znanosti je zadnja leta eno osrednjih vprašanj slovenske jezikovne politike. Problem je izpostavljen tudi v Resoluciji o Nacionalnem programu za jezikovno politiko 2014–2018, na vpliv, ki ga imata znanje in razvoj znanstvenega jezika na vitalnost posameznega jezika, pa opozarja tudi več evropskih študij. Zaradi tega je nujno izdelati sodobne referenčne jezikovne vire, ki bodo pomagali opolnomočiti slovenski

<http://nl.ijs.si/kas/>

- Projekt se je začel s 1.1.2016
- Partnerj:
 - IJS
 - UniMb
 - FF
 - FDV

- OpenScinece:

44083 Diplomsko delo

6647 Magistrsko delo

766 Doktorska disertacija

8167 Ekonomska fakulteta

7176 Ekonomsko-poslovna fakulteta

7075 Fakulteta za družbene vede