Špela Arhar Holdt in Jaka Čibej

1 KORPUSI IN KORPUSNO JEZIKOSLOVJE

1.1 Korpusi in korpusno jezikoslovje – izročki

JANES IJS FILOZOFSKA FAKULTETA

Korpusi in korpusno jezikoslovje

Špela Arhar Holdt
Center za uporabno jezikoslovje ZUS Trojina
Filozofska fakulteta UL

Ljubljana, 4. 7. 2016

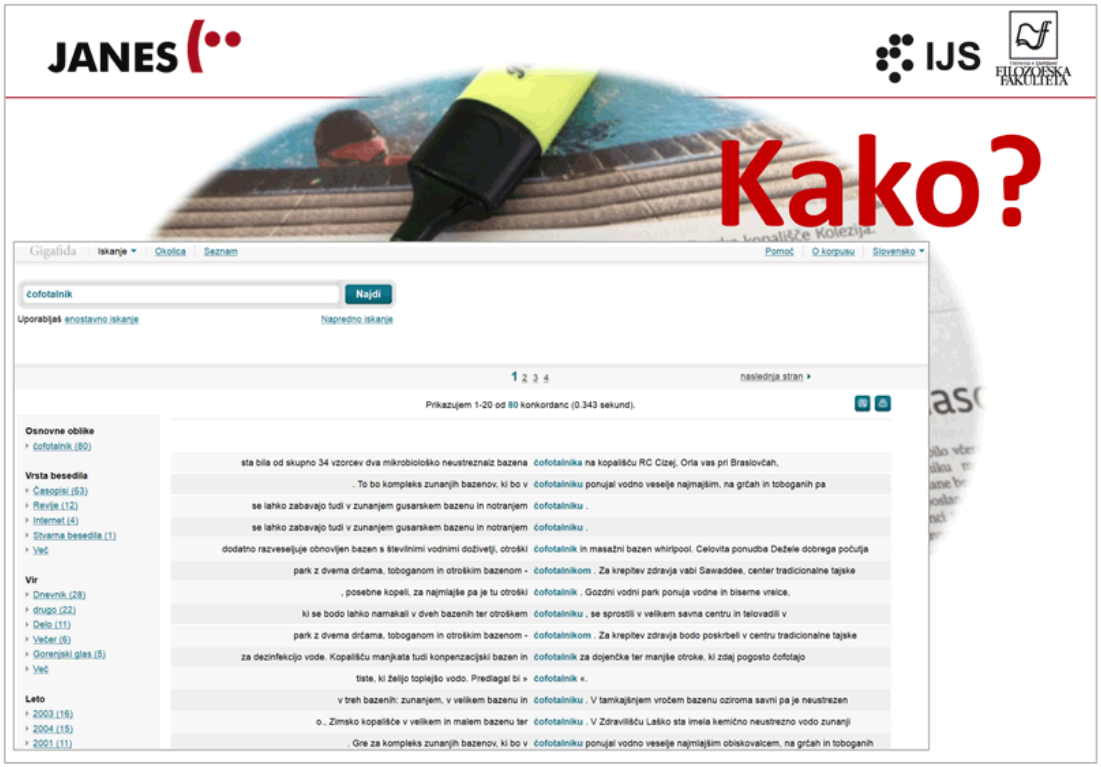
JANES IJS FILOZOFSKA FAKULTETA

Naloga za prvi dan

Spoznati pojme in orodja, ki jih bomo uporabljali ta teden.

- Kaj je korpus?
- Vzorec in generalizacija
- Gradnja korpusov
- Raba korpusov
- Korpus in priročniki
- Slovenski korpusi


JANES
IJS
FILOZOFSKA FAKULTETA
Želimo raziskovati slovenščino!
časopisi, revije
knjige
besedilni drobiž
govorjeni jezik
pošta, e-pošta
itd.
spletne strani
družbena omrežja
GIGAFIDA
GOS
FRAN
SLOLEKS
IMP
ŠOLAR


JANES
IJS
FILOZOFSKA FAKULTETA
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 neustreznaiz bazena čofotalnika na kopališču RC Cizej, Orla vas pri Brasovčah,
. 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 obnovljen bazen s številnimi vodnimi doživetji, otroški čofotalnik in masažni bazen whirlpool. Celovita ponudba Dežele dobrega počutja
park z dvema dričama, 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 vrelce,
ki se bodo lahko namakali v dveh bazenih ter otroškem čofotalniku , se sprostili v velikem savna centru in telovadili v
park z dvema dričama, 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.
Besedila pripravimo, da jih je lažje računalniško 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 - 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 Janes	**Spletna slovenščina**	**tviti, blogi, uporabniški komentarji, forumi**
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.

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. K besedilom dodamo vse informacije, ki so na voljo: od kod besedilo izvira, kdaj je nastalo oz. izšlo, kdo je avtor …
2. Besedila **jezikoslovno označimo:** s posebnim programom pripišemo besedi osnovno obliko, besedno vrsto in druge lastnosti.

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

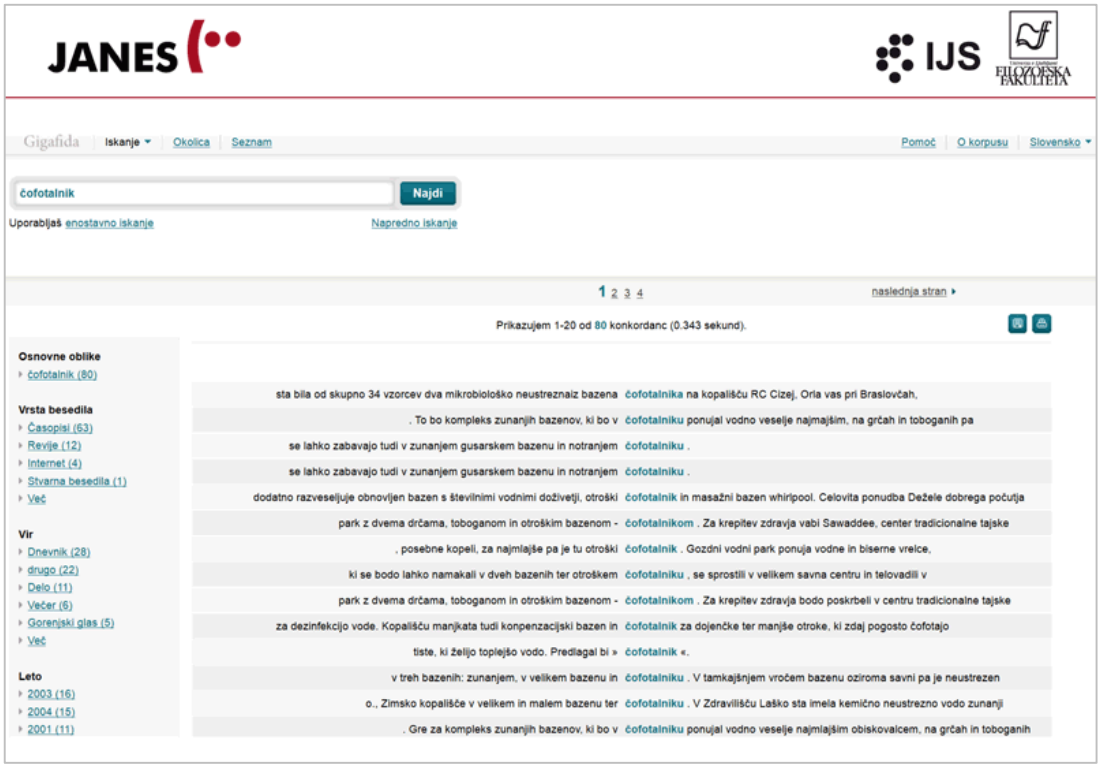
Vaši	**vaš**	svojilni zaimek; 2. oseba, moški spol, množina, imenovalnik, množina svojine
malčki	**malček**	samostalnik, občno ime; moški spol, množina, imenovalnik
lahko	**lahko**	splošni prislov; nedoločena stopnja
varno	**varno**	splošni prislov; nedoločena stopnja
uživajo	**uživati**	glavni glagol; 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

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.

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.


JANES
IJS
FILOZOFSKA FAKULTETA
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 neustreznaiz bazena čofotalnika na kopališču RC Cizej. Oria vas pri Brasiovčah,
. To bo kompleks zunanjih bazenov, ki bo v čofotalniku ponujal vodno veselje najmajšim, 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 obnovljen 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 vrelce,
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


JANES
IJS
FILOZOFSKA FAKULTETA
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 **oznake,** npr. osnovno obliko besede, besedno vrsto in druge lastnosti besede. Za raziskovanje besedilnih korpusov so besedila umeščena v **konkordančnike** specializirane programe, ki omogočajo napredno iskanje po besedilih, razvrščanje, filtriranje, izvažanje podatkov in podobno.

Zakaj uporabljati korpuse?

- 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 …

Od korpusa do priročnika

korpuse uporabimo kot vir jezikovnih podatkov

uporabniki lahko preberejo opis jezika v priročnikih

Za slovenščino še ne obstaja veliko priročnikov, ki so narejeni na osnovi korpusnih podatkov (v prihodnosti jih bo več). Trenutno je zato najti razlike med korpusnimi podatki in podatki nekaterih priročnikov.

Korpusi in jezikovni priročniki

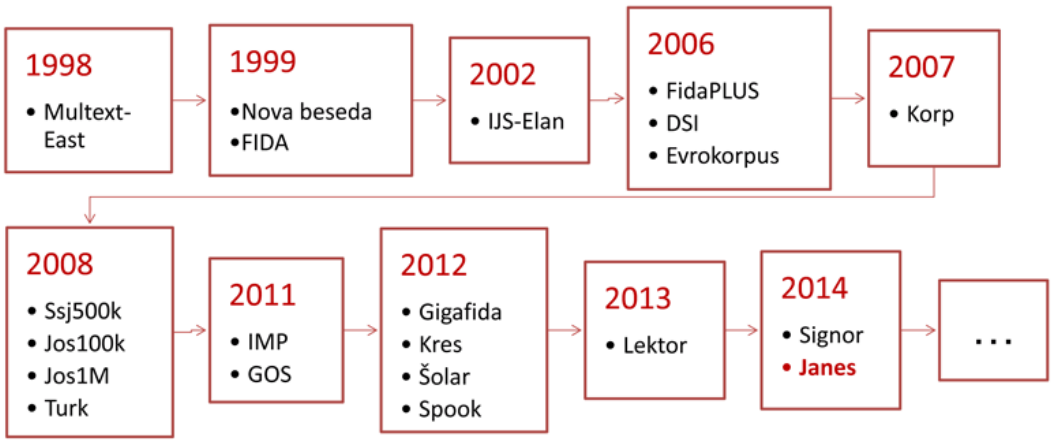
KORPUSI

- Vidimo sobesedilo in podatke o besedilu
- Potrebna je sinteza in interpretacija
- Možnost hitrega posodabljanja
- Razen pripisanih oznak so besedila v izvorni obliki

PRIROČNIKI

- Vsak priročnik služi določenemu namenu.
- Jezikovno gradivo je izbrano in urejeno glede na ta namen.
- Priročniki imajo pogosto (tudi) normativno vrednost.

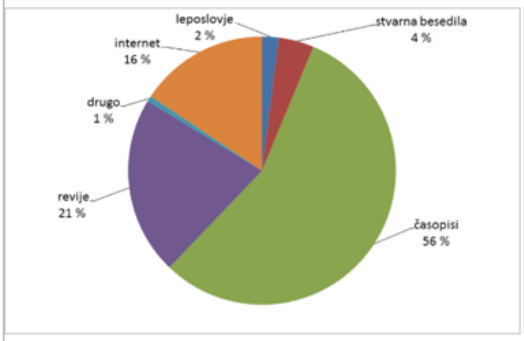
Slovenski korpusi (izbor)

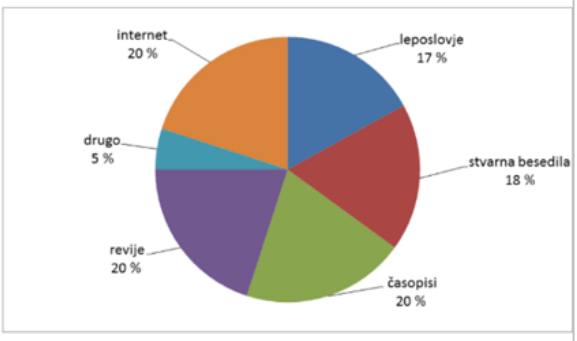


Gigafida in Kres

- **Gigafida** je obsežna zbirka sodobnih (1990-2011) slovenskih besedil iz časopisov, revij, knjig, s spleta itd. Korpus obsega skoraj 1,2 milijarde besed.
- **Kres** je manjša različica tega korpusa, prinaša cca. 100 milijonov besed.
- Korpuse, ki prinašajo splošni, nespecializirani jezik, imenujemo **referenčni korpusi.** Ti se uporabljajo za izdelavo referenčnih priročnikov, v raziskavah pa jih 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





GOS

- **GOS** je prvi korpus govorjene slovenščine. Prinaša posnetke govora v različnih vsakodnevnih situacijah.
- Posnetki so **transkribirani** in umeščeni v zmogljiv konkordančnik, s katerim lahko primere govora iščemo, poslušamo in preučujemo.
- Korpus obsega okrog **milijon besed.**
- Namenjen je **raziskovanju govora.**

Transkribcija 1: *ja ne vem po **kermu kluču** se bomo odločali eee koga bomo poslali v samo trgovanje z volno eem ne vem a bo to žrebanje al al glasvanje al kekrkoli*

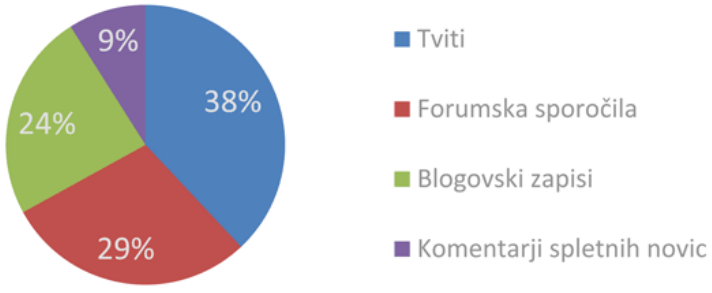
Transkribcija 2: *ja ne vem po **kateremu ključu** se bomo odločali eee koga bomo poslali v samo trgovanje z volno eem ne vem a bo to žrebanje ali ali glasovanje ali kakor*

Šolar

- **Šolar** vsebuje pisna besedila, ki so jih učenci in dijaki slovenskih šol tvorili pri pouku.
- V precejšnjem delu besedil so posebej označene tudi jezikovne napake, ki so jih v spisih **popravili učitelji.**
- Po jezikovnih napakah oz. učiteljskih popravkih lahko s pomočjo specializiranega konkordančnika tudi iščemo.
- Korpus vsebuje približno **milijon besed.**
- Korpus je primarno namenjen raziskavam šolske pisne produkcije oz. jezikovne zmožnosti šolarjev in pripravi učnih gradiv.

Janes

- **Janes** je korpus spletne slovenščine. Vsebuje besedila, ki so jih na spletu tvorili uporabniki. Korpus obsega okrog **134 milijonov besed.**



JANES IJS

Janes

Janes je tudi razlog, da smo tu:

- *JANES – Jezikoslovna analiza nestandardne slovenščine* je nacionalni raziskovalni projekt (J6—6842), ki ga od 1. 7. 2014 do 30. 6. 2017 financira Javna agencija za raziskovalno dejavnost Republike Slovenije.
- Cilj projekta je zgraditi obsežen korpus spletne slovenščine, s pomočjo katerega bomo omogočili empirično podprto jezikoslovno analizo nestandardne spletne slovenščine, izboljšali jezikovnotehnološka orodja za obdelavo besedil, napisanih v nestandardnem jeziku, in izdelali slovarček spletne slovenščine.