

POLETNA ŠOLA
SPLETNE SLOVENŠČINE JANES
ZA ŠTUDENTE

Oddelek za prevajalstvo
Filozofska fakulteta, Ljubljana

4. – 8. julij 2016

UČNA GRADIVA

JANES (••)

ISBN 978-961-237-853-0 (pdf)

KAZALO

PREDGOVOR	4
1 PREDSTAVITEV PROJEKTA IN KORPUSA JANES	5
1.1 Projekt in korpus JANES – izročki.....	5
2 RAZISKOVANJE RAČUNALNIŠKO POSREDOVANE KOMUNIKACIJE	22
2.1 Raziskovanje računalniško posredovane komunikacije – izročki.....	22
2.2 Raziskovanje spletne slovenščine – delavnica	37
2.3 Raziskovanje spletne slovenščine – vaje.....	45
3 SPREMINJANJE JEZIKA	48
3.1 Spreminjanje jezika – izročki	48
4 INTERPRETACIJA KORPUSNIH PODATKOV	57
4.1 Interpretacija korpusnih podatkov – izročki	57
4.2 Interpretacija korpusnih podatkov – delavnica	79
5 JE JANES ŠOLAR?	82
5.1 Je Janes Šolar? – izročki.....	82
6 KNJIŽNOJEZIKOVNI STANDARD IN SPLET	100
6.1 Knjižnojezikovni standard in splet – izročki	100
6.2 Knjižno in spletno – dva jezika, dva svetova? – delavnica	111
6.2.1 Jezikovna neustaljenost	111
7 TIPOLOŠKA ANALIZA SPLETNIH BESEDIL	118
7.1 Tipološka analiza spletnih besedil – izročki	118
8 JEZIK IN SPOL	130
8.1 Jezik in spol – izročki	130
9 SKLADENJSKA ANALIZA (NE)STANDARDNE SLOVENŠČINE.....	165
9.1 Skladenjska analiza (ne)standardne slovenščine – izročki	165
10 KORPUSNA FRAZELOGIJA.....	180
10.1 Korpusna frazeologija – izročki	180

PREDGOVOR

Od 4. do 8. julija 2016 je na Oddelku za prevajalstvo Filozofske fakultete Univerze v Ljubljani potekala 1. poletna šola spletne slovenščine za študente JANES. Poletna šola je bila organizirana v okviru temeljnega nacionalnega projekta 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, sofinancirala pa jo je Slovenska raziskovalna infrastruktura za jezikovne vire in tehnologije CLARIN.SI.

S poletno šolo smo želeli nasloviti ključne teoretično-metodološke vsebine s področja korpusnega jezikoslovja, sociolingvistike in analize diskurza, ki so v trenutnih dodiplomskih in podiplomskih jezikoslovnih programih pri nas zaenkrat še vse premalo zastopane. Poleg tega je bil naš cilj tudi, da študente seznanimo z najsodobnejšimi jezikovnimi viri in tehnologijami za slovenščino, ki so bile razvite v okviru projekta JANES, in jih opremimo s kompetencami, potrebnimi za samostojen študij in raziskovalno delo z nekanoničnimi jezikovnimi različicami. Ker menimo, da so gradiva koristna tudi za druge študente, predavatelje in jezikoslovce, ki se želijo seznaniti s korpusnimi pristopi k proučevanju računalniško posredovane komunikacije in značilnostmi spletne slovenščine, smo jih zbrali v skupni publikaciji, ki je pred vami.

Poglavja so urejena v enakem vrstnem redu, kot so se vsebine zvrstile na poletni šoli, ki je trajala pet dni. Vsak je vseboval po dve pregledni predavanji in eno tematsko obarvano delavnico. Poletno šolo je otvorila *doc. dr. Darja Fišer* s predstavitvijo zajema, označevanja in analize besedil v korpusu Janes (→) ter pregledom metod za proučevanje računalniško posredovane komunikacije (→). Drugi dan je *prof. dr. Marko Stabej* predaval o spreminjanju jezika (→), *doc. dr. Nataša Logar* pa predstavila ključna načela in pasti za kvalitetno interpretacijo korpusnih podatkov (→). Tretji dan je *dr. Tadeja Rozman* primerjala vsebino korpusa Šolar s korpusom Janes (→), *doc. dr. Helena Dobrovoljc* pa razpravljala o razmerju med knjižnojezikovnim standardom in spletno slovenščino (→). Četrty dan je *dr. Mija Michelizza* naredila tipološki pregled besedil na spletu in njihovih značilnosti (→), medtem ko se je *asist. dr. Damjan Popič* posvetil jezikovnim značilnostim uporabnikov glede na sociodemografske podatke (→). Poletno šolo sta sklenili *dr. Špela Arhar Holdt* s predavanjem o skladišnji analizi (ne)standardne slovenščine (→) in *dr. Polona Gantar* s pregledom najsodobnejših metodoloških pristopov h korpusni frazeologiji (→).

doc. dr. Darja Fišer

Ljubljana, avgust 2016

Darja Fišer

1 PREDSTAVITEV PROJEKTA IN KORPUSA JANES

1.1 Projekt in korpus JANES – izročki



Projekt in korpus JANES

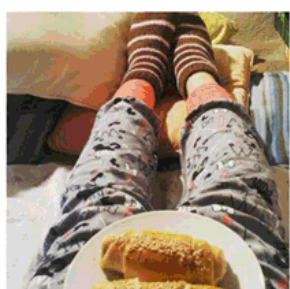
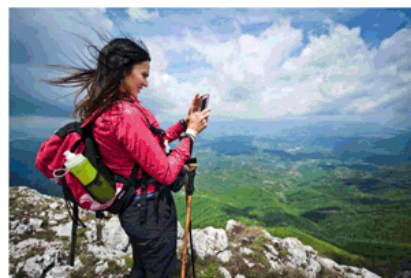
Darja Fišer
Oddelek za prevajalstvo, Filozofska fakulteta Univerze v Ljubljani
Odsek za tehnologije znanja, Inštitut Jožef Stefan

Ljubljana, 4. julij 2016



Predstavitev udeležencev

JAZ JANES, TI ... ?

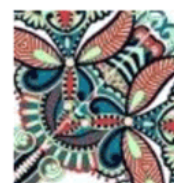


- Viri, orodja in metode za analizo nestandardne spletne slovenščine
 - nacionalni temeljni projekt ARRS
 - 2014-2017
 - <http://nl.ijs.si/janes/>
- 2 instituciji, 9 raziskovalcev
 - Filozofska fakulteta
 - Darja Fišer
 - Jaka Čibej
 - Špela Arhar Holdt
 - Ana Zwitter Vitez
 - Damjan Popič
 - Polona Gantar
 - Inštitut Jožef Stefan
 - Tomaž Erjavec
 - Nikola Ljubešić
 - Senja Pollak

- DS1: Izdelava korpusa
 - N1: Zajem besedil
 - N2: Obdelava besedil
 - N3: Objava korpusa
- DS2: Jezikoslovna analiza
 - N1: Primerjava s standardno slovenščino
 - N2: Primerjava z govorno slovenščino
 - N3: Analiza nestandardne leksike
- DS3: Razvoj orodij za procesiranje
 - N1: Avtomatska standardizacija besedil
 - N2: Prilagajanje tegetja in lematizatorja
 - N3: Izdelava spremljevalnega korpusa

Korpus JANES

IZBOR IN ZAJEM BESEDIL



- osnovna načela
 - spletne uporabniške vsebine
 - javna komunikacija
- 5 zvrsti, 10 virov
 - tviti
 - forumi
 - Medovernet
 - Avtomobilizem
 - Kvarkadabra
 - komentarji na novice
 - RTV Slo
 - Mladina
 - Reporter
 - blogi
 - RTV Slo
 - Publishwall
 - pogovorne in uporabniške strani na Wikipediji

- TweetCat (Ljubešić et al. 2014)
- slovenske semenske besede -> slovenski uporabniki -> njihova mreža
- filtriranje uporabnikov, ki tvitajo pretežno v slovenščini
- metapodatki: uporabniško ime, čas objave, št. retweetov & všečkov

*še, kaj, že, če, ampak, mogoče, jutri, zdaj,
vendar, kje, oziroma, tudi, sploh, spet, všeč,
ravno, kar, končno, kdaj, preveč, očitno*



Katarina Jenko @KatarinaJenko · 11h

Sam mal pa paše, da imajo tudi višje razviti narodi clusterfuck od politike, ne samo mi. Da vidimo, kako bodo oni prišli iz te jebe...

11

- izbor:
 - analiza 96 forumov (Lebar et al. 2012)
 - kriteriji: št. registriranih uporabnikov, št. in dinamika objavljenih sporočil, št. aktivnih tem
- namenski ekstraktorji za vsak forum posebej
- metapodatki: tema, ID objave, URL objave, čas objave, upor. ime



im nobody
Mojster foruma

Prispevkov: 3321

Pridružen: Pe feb 03, 2006 11:18
pm

Kraj: laško
Spletna stran

blindiranje egr-ja material?

So feb 21, 2009 9:27 pm

lep pozdrav....s kakšnim materialom je najbolše blindirat egr ventila?1.5 mm
aluminija,medenina,karkoli?
zanimajo me še opazne spremembe?moč motorja itd

kokr jz vem motor mirneje laufa,ne kadi,bolj odziven...gre se za passata 1.9 tdi 110 konjev...b5
motor...avto gre tudi na čipiranje v bližnji prihodnosti

passat 1.9 tdi 110,235 nm
Na levi pas...do konca gas
viewtopic.php?t=72616&postdays=0&postorder=asc&start=60

JANES Novice & komentarji nanje



- namenski ekstraktorji za vsak novičarski portal posebej
- izbor: politika portala & tehnične rešitve
- metapodatki: URL članka, ID članka, ID komentarja, čas objave, upor. ime

Kadrovski cunami na DUTB: ogorčeni zaposleni pišejo ministru Mramorju

Od vodstva DUTB zahtevajte, da pri izboru vodij zasleduje cilje in zahteve strokovnega in učinkovitega bodočega upravljanja slabih naložb, zato srednji management ne morejo sestavljati zaposleni, ki so povezani z dolžniki DUTB, ki so bili vodje v državnih bankah, ali so bili hkrati sopredlagatelji ter sopotrjevalci predlogov za slabe naložbe, ki so zdaj v skrbništvu DUTB, so v pismo ministru za finance zapisali večinoma tisti zaposleni na Družbi za upravljanje terjatev bank (DUTB), ki so bili izbrani pod Švedi.

27. junij 2016, ob 14:11 (posodobljeno: 27. junij 2016, ob 14:29)

Tekst: Jože Biščak | Foto: Bobo



Tiskalniku

zeugma 27. junij 2016, ob 19:39

+1

0

-1

cmeravec je izgnal svede, resil savine hotele bandi 21 in pripojil nesnago od dragonjske faktor banke dutb. drugi dragonjski brat pa metodolosko laze poslancem in vseskozi zavaja davkoplacevalce in volilce ki si to seveda povsem zaslužijo. a nje.

JANES Blogi & komentarji nanje



- namenski ekstraktorji za vsak novičarski portal posebej
- izbor: enotna predloga blogov
- metapodatki: URL objave, ID objave, čas objave, upor. ime

Publishall



Blog



Mesta



Oglasnik



Išči

Pametnjakoviči

Objavil/a Samosvoja b.p., dne 2016-06-23 ob 15:25:21



Ne vem od kje in kdaj so vzniknili strašno "pametni" osebk. Je to neka nova okupacija izven zemeljskih energij, ki okupirajo ubogi um navadnega osebk? Imajo se za strašno vsevedega. Najhuje je, da je tak pametnjakoviči sveto prepričan, da ima edini prav.

Taki osebk se nekaj časa potuhnjeno



abram1b2 Cej pred 3 dnevi



pametnjakoviči

Če te vabi Polje, potem se tam javi čimprej. Ne odlašaj predolgo, če ti tvoje mentalno zdravje nagaja.

#38

JANES Uporabniške & pogovorne strani na Wikipediji IJS



- WikitalkExtractor (Ljubešić 2016)
- slovenska koda za uporabnika ("uporabnik") & jezik ("sl")
- minimalna segmentacija objav

Fran Krivic

Pogovor:Slovenščina

Iz Wikipedije, proste enciklopedije

Vsebina [prikaži]

Glasoslovje [uredi kodo]

V razdelku o značilnostih slovenščine piše, da ima 25 črk s katerimi zapisuje 28 glasov. Odkod ta referenca? Povsod kjer gledam piše 29. --Uporabnik:Zevnik 13:12, 25 nov 2011 (CEST)

Če je narobe, popravi in navedi vir. --IP 213 15:54, 25. november 2011 (CET)

Digramma (it.) [uredi kodo]

Na it:Discusione:Lingua slovena se z Boraczkem sprašuje ali so dž, lj in nj v slovenščini nekaj, čemur Italijani rečejo digramma. Verjetno obstaja tudi kak slovenski izraz, ampak gre za to, da dve črki izgovorili kot en glas. Prosim, da po svojih močeh prispevate k debati. --romanm 09:09, 5 apr 2004 (CEST)

Lahko bi kaj rekli tudi o slovenskih narečjih. Veliko jih imamo na tako majhnem prostoru. --Igor 23:19, 12 jul 2004 (CEST)

Vsekakor. Na angleški strani sem pred časom nekaj malega o tem prispeval. Potrebno je samo prenesti sem. --XJam 03:27, 13 jul 2004 (CEST)

Heh, nikoli si nebi mislil, da bom o slovenščini prevajal iz angleščine? Še vedno me boga uporaba v in na. Na Planini in v Zasavju npr., ali lahko prosim kdo preveri, če to pri narečjih ok? --Igor 11:30, 13 jul 2004 (CEST)

Ja, hecno. Pri predlogih v in na velja tista: grem y šolo in potem iz šole, ter pa morje in pišem z morja. Bom preveril. --XJam 12:22, 13 jul 2004 (CEST)

Knjižna slovenščina [uredi kodo]

Slovenska narečja so tako različna, da se različne narečne skupine med seboj le težko razumejo. Zato se uporablja takimenovana knjižna slovenščina, ki je nekakšna slovenska različica esperanta in se jo večinoma srečuje le v knjižni obliki ter javnih obdih. Najbližja knjižni slovenščini je bojda novomeška dolanščina.

Primerjava z esperantom ni po mojem mnenju nikakor prava. Esperanto je umeten jezik, knjižna slovenščina ni umeten jezik. Večina evropskih knjižnih jezikov je nastala kot knjižna slovenščina in na začetku nihče ni govoril v knjižnem jeziku, vsi so govorili v dialektih. Na primer ko se je Italija združila leta 1861, samo 2-3 odstotki prebivalcev so znali italijansko. 97-98 odstotkov je govorilo le v dialektih (ali drugih jezikih). Knjižna slovenščina je normalen knjižni jezik. Esperanto je pa umeten jezik, ki ima svoj umeten pravopis in besede iz različnih evropskih jezikov. Lep pozdrav v Slovenijo in oprostite, če sem

JANES



Korpus JANES

OBDELAVA BESEDIL

1. Stavčna segmentacija & tokenizacija
2. Rediakritizacija
3. Normalizacija
4. Tegiranje & lematizacija
5. Zapis korpusa

lep pozdrav....s kakšnim materialom je najbolše blindirat egr ventila?1.5 mm
aluminija,medenina,karkoli?
zanimajo me še opazne spremembe?moč motorja itd

- temeljita na pravilih v obliki regularnih izrazov
- standardni modul + opcijski nestandardni modul (Ljubešić in Erjavec 2016)
 - pika lahko konča poved, čeprav se naslednja beseda ne začneja z veliko začetnico ali ji celo ne sledi presledek
 - pojavnice, ki se končajo s piko in so na seznamu okrajšav, ki ne končajo povedi, kot npr. *prof.*, ne končujejo povedi
 - emotikoni so ena pojavnica, kot npr. :-], :-PPPP, ^ _ ^
- evalvacija
 - ročno popravljanje stavčne segmentacije in tokenizacije za 4.000 tvitov / 100.000 pojavnic (Čibej et al. 2016)
 - stavčno segmentacijo bi bilo tvitov mogoče še precej izboljšati (86,3 % natančnost)
 - tokenizacija je zadovoljiva (99,2 % natančnost)

cmeravec je izgnal svede, resil savine hotele bandi 21 in pripojil nesnago od
dragonjske faktor banke dutb. drugi dragonjski brat pa metodolosko laze poslancem in
vseskozi zavaža davkoplacevalce in volilce ki si to seveda povsem zaslužijo. a nje.

- temelji na strojnem učenju (Ljubešić in dr. 2016)
 - učenje modela: običajna besedila s šumniki & besedila z odstranjenimi šumniki
 - strategija 1: verjetnost prevoda besede brez šumniki v besedo s šumniki
 - strategija 2: verjetnost besede s šumniki glede na kontekst
- evalvacija
 - najboljši rezultati za model, naučen na standardnih & nestandardnih besedilih (Wikipedija, slWaC, tviti)
 - Wikipedija: 99,62 %
 - tviti: 99,12 %
 - problem: *se/še*

kokr jz vem motor mirneje laufa,ne kadi,bolj odziven...gre se za passata 1.9 tdi 110 konjev...b5
motor...avto gre tudi na čipiranje v bližnji prihodnosti

- *jest, jst, jas, js, jz -> jaz*
- temelji na strojnem učenju
 - učenje prevodnega modela: ročno normaliziran vzorec 4.000 tvitov / 100.000 pojavnic
 - učenje modela ciljnega jezika: korpus Kres & standardni tviti
 - normalizacija poteka na nivoju besede (na nivoju povedi rezultati malo boljši, a je procesiranje veliko počasnejše)

- Nikolex (Ljubešič in Erjavec 2016)
 - 1. korak: tegiranje
 - temelji na strojnem učenju
 - učenje modela: ročno označen korpus ssj500k 1.3 (Krek et al. 2013) & oblikoskladenjski leksikon Sloleks 1.2 (Dobrovoljc et al. 2015)
 - za razliko od klasičnih označevalnikov leksikon uporabljen samo posredno, v obliki značilk
 - nove oznake za specifične elemente RPK:
 - Nw: e-mail naslovi, URL-ji
 - Ne: emotikoni, emojiji :-), ☺
 - Nh: heštagi #kvadogaja
 - Na: @dfiser3
 - 2. korak: lematizacija
 - upošteva oblikoskladenjsko oznako iz 1. koraka & oblikoskladenjski leksikon
 - strojno naučen model se uporabi samo v primerih, ko para *oblikoskladenjska oznaka* : *besedna oblika* ni v leksikonu
- evalvacija
 - natančnost: 94,3 %
 - zmanjšanje relativne napake: 25 %

- metapodatki: lastni XML
- anotacije: TEI P5

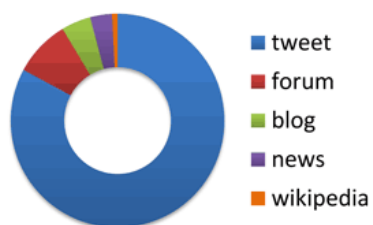
```
<s>
  <w lemma="pa" ana="#Cc">pa</w>
  <c> </c>
  <w lemma="še" ana="#Q">še</w>
  <c> </c>
  <choice>
    <orig>
      <w>tamali</w>
    </orig>
    <reg>
      <w lemma="ta" ana="#Pd-fsn">ta</w>
      <c> </c>
      <w lemma="mali" ana="#Agmpn">mali</w>
    </reg>
  </choice>
  <pc ana="#Z">.</pc>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

Korpus JANES

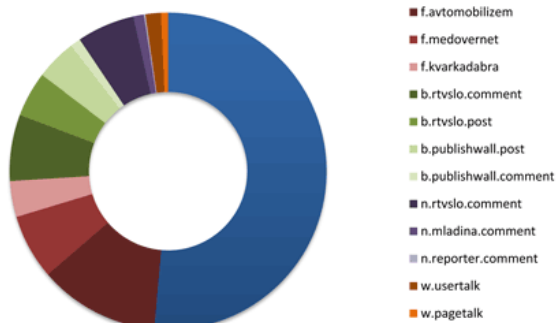
ANALIZA KORPUSA

Janes v0.4	
Št. besedil	9.055.351
Št. besed	175.134.545
Št. pojavnic	208.261.725
Št. besed/besedilo	19,3
Št. avtorjev	96.648

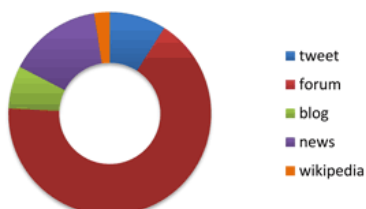
Št. besedil



Št. pojavnic



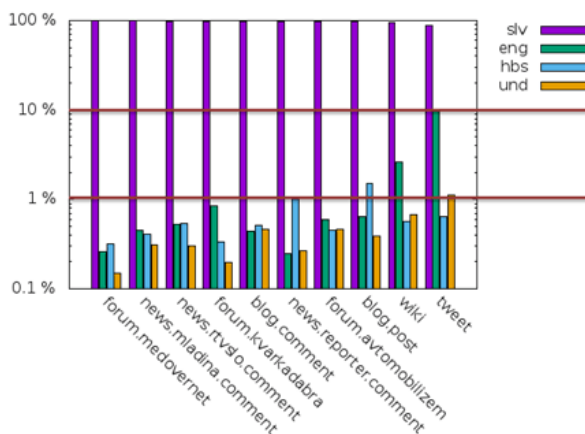
Št. avtorjev



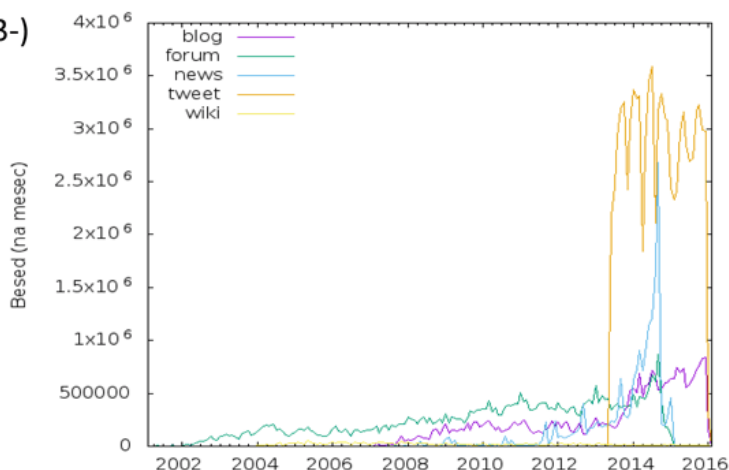
Podkorpus	Pov. št. besed / besedilo	Št. besed/uporabnika	Št. besedil/uporabnika
tweet	12	10.307,5	857,6
forum	51,4	616,5	12,0
avtomobilizem	38,5	1.713,5	44,5
medovernet	94,7	234,7	2,5
kvarkadabra	77,1	2.814,1	36,5
blog	71,3	4.373,8	61,3
rtvslo.comment	35,8	3.705,5	103,4
rtvslo.post	343,7	33.261,7	96,8
publishwall.post	394	11.860,8	30,1
publishwall.comment	48,4	599,6	12,4
news	41,8	867,7	20,7
rtvslo	38,6	800,5	20,7
mladina	72,7	1.484,8	20,4
reporter	54,4	1.221,7	22,5
wikipedia	50,8	1.609,3	31,7
usertalk	52,2	1.765,5	33,8
pagetalk	48	1.349,1	28,1

- Glede na način označevanja
 - avtomatsko
 - jezik
 - spol
 - stopnja standardnosti
 - sentiment
 - regija (samo za tvite)
 - ročno
 - tip
 - spol (samo za tvite)
- Glede na nivo označevanja
 - na nivoju uporabnika
 - spol
 - tip
 - regija (samo za tvite)
 - na nivoju besedila
 - jezik
 - sentiment
 - stopnja standardnosti

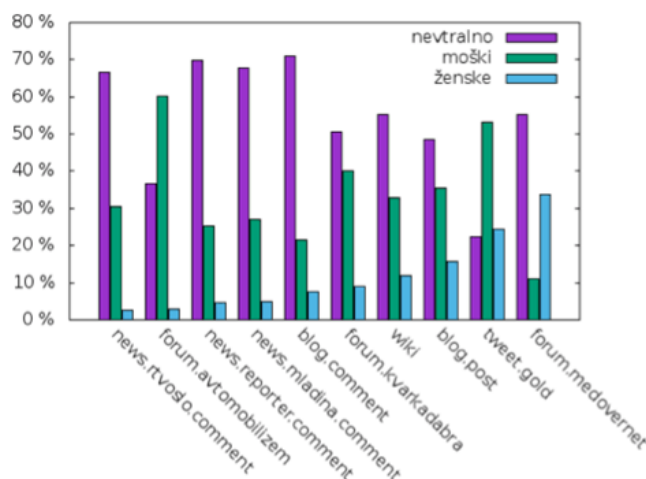
- detekcija:
 - *langpy*
 - slv, eng, hbs, und
- > 1 % tujejezičnih besedil samo wiki & tviti
- wiki: 2,6 % ang besedil
- tviti:
 - 9,6 % ang besedil
 - 1,1 % drugo



- zajeto obdobje:
 - 2001–2015
- najstarejši viri:
 - forumi (2001-)
 - Wikipedija (2003-)
 - blogi (2006-)
- najmlajši viri:
 - komentarji na novice (2014, politika portala)
 - tviti (2014, začetek zajema)

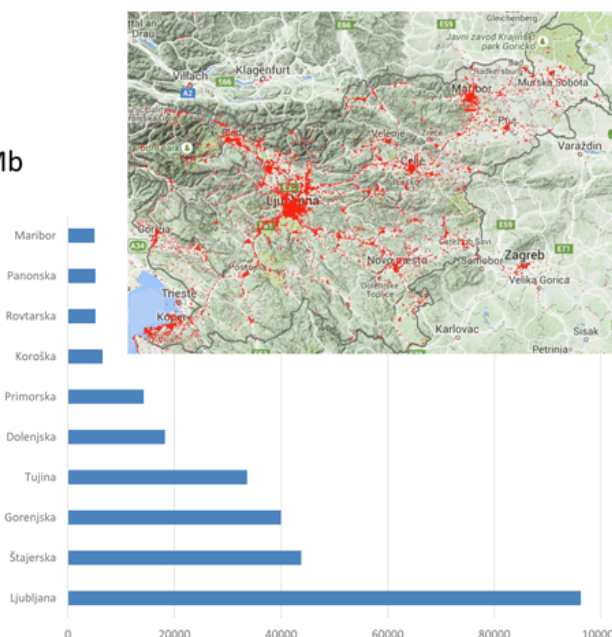


- detekcija:
 - 1. os. ed. pom. gl. + deležnik na -l (*sem/nisem/bom mislil/a*)
 - > 0.7 odkritih ž/m > 1 % besedil
- evalvacija (tviti):
 - 76 % natančnost
 - 5 % napačni spol
 - 19 % nevtralni spol
- komentarji prevladuje N
- tviti & avtomobilizem prevladuje M
- medovernet prevladuje Ž

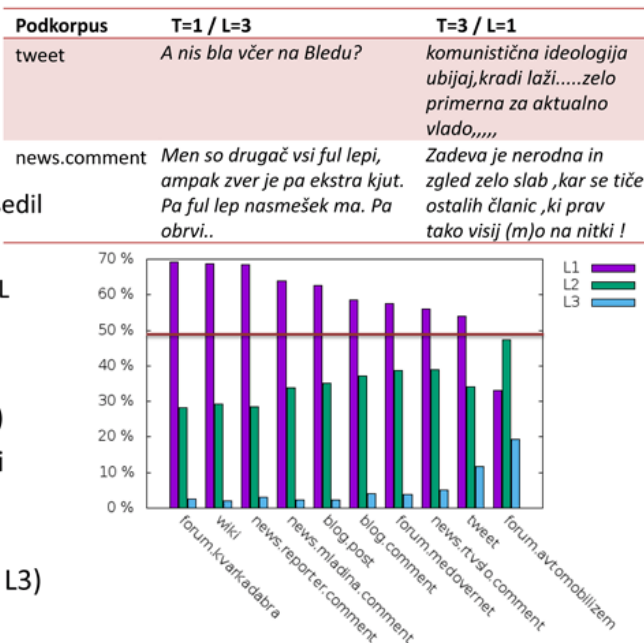


- tip avtorja
 - osebni računi posameznikov (prosti čas)
 - uradni računi medijskih hiš, institucij, podjetij (profesionalna raba)
- označevanje
 - ročno (analiza profila uporabniškega računa & zgodovino objav)
 - tviti (blogi kmalu)
- rezultati
 - 76 % zasebnih uporabnikov
 - 24 % korporativnih uporabnikov
 - 84 % N
 - 13 % M
 - 3 % Ž

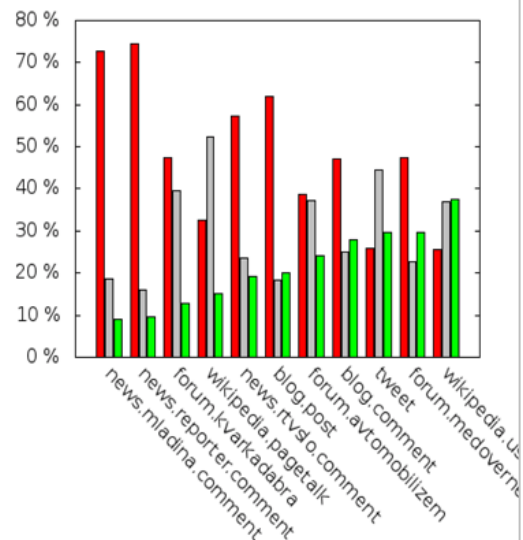
- detekcija
(Čibej in Ljubešić 2015):
 - geolokacija tvitov
 - 7 narečnih skupin + Lj + Mb
 - > 90 % tvitov iz 1 regije
 - > 2 tvita
- rezultati:
 - 22 % uporabnikov
 - 2 % podkorpusa
 - redno osveževanje



- stopnje standardnosti
(Ljubešić et al. 2015):
 - tehnična & lingvistična 1-3
- učenje modela:
 - ročno označenih 1.200 besedil
- evalvacija:
 - povp. abs. n. 0,38 T / 0,42 L
- rezultati:
 - najbolj nestandarden avtomobilizem (20 % L3)
 - precej nestandardni tudi tviti (12 % L3)
 - najbolj standardni kvarkadabra & wiki (2 % L3)



- sentiment (Smailović 2014):
 - + / - / 0
- učenje modela:
 - 5000 ročno označenih tвитov
- evalvacija:
 - 600 besedil, 3 anotatorji
 - ujemanje med A. 0,563, sistema 0,432
 - najboljše blogi, najslabše forumi
 - razmeroma nenatančno na nivoju posameznih besedil, a zelo natančno na nivoju podkorpusov



Korpus JANES

NERAZREŠENI PROBLEMI

- Sestava korpusa
 - reprezentativnost
 - celovitost
 - uravnoteženost
- Metapodatki
 - dodajanje starosti uporabnikov
 - podkorpus politikov, mikrovezdnikov, etc.
- Zapis
 - struktura dokumenta (specifični elementi RPK)
- Spremljevalni korpus
 - Twitter & Wikipedia
- Uporabnost korpusa izven konkordančnika
 - sociolingvistika, analiza diskurza, žanrska analiza
 - discussion threads, layout, nebesedilni elementi

- Problemi
 - pogoji uporabe (Twitter)
 - avtorske pravice (forumi, blogi, novičarski portali)
 - pravica do zasebnosti (informacijska pooblaščenka)
- Rešitve
 - anonimizacija
 - premešanje
 - vzorčenje

Darja Fišer

2 RAZISKOVANJE RAČUNALNIŠKO POSREDOVANE KOMUNIKACIJE

2.1 Raziskovanje računalniško posredovane komunikacije – izročki



Raziskovanje računalniško posredovane komunikacije

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

Ljubljana, 4. julij 2016



Osnove RKP

OSNOVNI POJMI

- Kaj je RPK:
 - email, chat, pogovorne strani, družbena omrežja (Herring 2001)
- Koga je RPK najprej zanimal:
 - pragmatika
 - analiza diskurza
 - sociolingvistika
- Koga RPK zanima danes:
 - Podatkovna analitika
 - oglaševanje
 - podatkovno novinarstvo
 - Procesiranje šumnih besedil
 - strojno prevajanje
 - analiza sentimenta
 - Analiza vedenja uporabnikov / skupnosti
- Variantnost
 - dialektologija
- Analiza leksikalnih sprememb
 - ortografske spremembe
 - neologizmi
 - pomenski premiki
- Večjezičnost
 - mešanje jezikov
 - preklapljanje med jeziki / pisavami
- Pismenost
 - usvajanje 1. jezika
 - usvajanje tujega jezika
- Glavni problemi RPK
 - metodologija zbiranja in označevanja gradiva
 - metodologija analize gradiva

- Zbiranje gradiva z interneta vse prej kot trivialno
 - velikost & reprezentativnost korpusa
 - procesiranje korpusa
 - ločevanje med žanri
 - vrsta & količina potrebnih kontekstualnih informacij
 - etična vprašanja (anonimnost & varovanje zasebnosti)
- Slaba praksa
 - majhni, ad-hoc vzorci besedil
- Problemi
 - standardizirane smernice za načrtovanje korpusov RPK
 - pomanjkanje javno dostopnih korpusov RPK (Beißwenger & Storrer 2008)

- Popularne številne metode in pristopi tradicionalne lingvistike:
 - pragmatika
 - konverzacijska analiza
 - sociolingvistika
 - žanrska analiza
 - etnografska analiza
- Predmet raziskav:
 - raba jezikovnih sredstev za vzpostavljanje stika, vzdrževanje interakcije in gradnjo identitete v računalniških omrežjih
- Slaba praksa:
 - kritična refleksija problemov in izzivov ob uporabi tradicionalnih pristopov na nova okolja in okoliščine RPK
- Ključni elementi:
 - interakcijska koherenca, participatorni okvir, intertekstualnost, jezikovna identiteta, skupnost (Herring 2004)

Osnove RKP

SORODNI PROJEKTI

Projekt: Deutsches Referenzkorpus zur internetbasierten Kommunikation (DeRiK)

Kurzbeschreibung

Das Vorhaben, das in Kooperation mit dem Berliner Akademieprojekt "Digitales Wörterbuch der deutschen Sprache" durchgeführt wird, zielt auf den Aufbau eines ausgewogenen Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation (IBK). Ausgewogenheit wird in zweierlei Hinsicht angestrebt: Einerseits sollen für jedes Kalenderjahr gleich große Teilkorpora erhoben werden, andererseits sollen die einzelnen Teilkorpora einen ausgewogenen Querschnitt der jeweils populärsten IBK-Genres repräsentieren.

Zu Projektbeginn (Ende 2010) wurde an der TU Dortmund ein **Testdatenset** (112.000 Tokens) erhoben, das Sprachdaten aus diversen korpusrelevanten Genres (E-Mails, Mailinglisten, Forendiskussionen, Chats, ICQ, Kommunikation in sozialen Netzwerken, Videoplattformen, Online-Tauschbörsen, Online-Games, Wikipedia-Diskussionen etc.) und Nutzungskontexten umfasst. Dieses Datenset bildet das Testbett für die Überprüfung von Methoden zur automatischen Aufbereitung der Korpusdaten.

Im ersten Projektabschnitt (Dezember 2010 bis September 2011) wurde für die Repräsentation der korpusrelevanten IBK-Genres ein **Repräsentationsschema auf Basis der Formate der Text Encoding Initiative (TEI)** (Version TEI-P5) erarbeitet. Das Schema wurde im Oktober 2011 im Rahmen der internationalen Konferenz "Philology in the Digital Age" (2011 Annual Conference and Members' Meeting of the TEI Consortium) vorgestellt und ist auf der Projektseite **XML Schema for the Representation of CMC Genres in TEI** dokumentiert. Eine ausführliche Begründung und Darstellung des Schemas wurde 2012 in einem *Forschungsartikel für das Journal of the Text Encoding Initiative (JTEI)* publiziert.

Gegenwärtig befindet sich das Projekt in der Phase der Datenerhebung. Parallel dazu werden Verfahren für die linguistische Aufbereitung der Daten und für ihre Integration in das Korpus entwickelt.



In the project *DiDi* we have analysed the linguistic strategies employed by users of social network sites (SNS). The data analysis focused on South Tyrolean users and we investigated how they communicate with each other. In regions of the German speaking area where dialect is frequently used in different communicative contexts, regional and social codes are often also used in written computer mediated communication. Another interesting but more general aspect of the new media is connected to the emerging linguistic and social practices (*new literacy*). One of the main research questions in *DiDi* was whether people of different age use language on SNS in a similar way or in an age-specific manner.

The purpose of the study was:

1. to record the contemporary language use of South Tyrolean German in the new media (cf. the **DiDi Corpus**)
2. to describe the everyday usage of language of South Tyrolean SNS users with L1 German with respect to their choice of languages and varieties as well as with respect to their usage of specific cmc phenomena.

JANES

CoMeRe

IJS

FILIZOVSKA
FAKULTETA



A propos de CoMeRe

Site du projet CoMeRe, corpus de communication médiée par les réseaux.

CoMeRe a pour objectif, à l'horizon 2014, de créer un noyau de corpus de communication médiée par les réseaux (*Computer Mediated Communication – CMC*) en français. Chaque corpus rassemblera un ensemble de conversations intervenant sur la Toile et les réseaux. Nous nous intéressons à une variété de systèmes de communication synchrone ou asynchrone, mono ou multimodaux (éventuellement) : blogues, tweets, SMS / textos, courriels, clavardage, forums, etc.

Les corpus et leurs métadonnées seront structurés suivant des formats standard : TEI (*Text Encoding Initiative*), CLARIN, OLAC. La banque de corpus sera diffusée en accès libre en 2014 sur le site Ortolang. L'assemblage des corpus se fera sur les serveurs de la MSH (Maison des Sciences de l'Homme) de Clermont-Ferrand et du Laboratoire de Recherche sur le Langage (LRL). Le travail s'effectue avec partenariat européen sur la TEI (groupe d'annotation TEI-CMC) avec relation avec l'infrastructure DARIAH. Ce noyau de corpus sera intégré au futur « Corpus de référence du français »

Les membres du projet CoMeRe appartiennent au groupe de travail « Nouvelles formes de communication » du consortium Corpus-écrits. Le projet a reçu l'appui de Corpus-écrits et de Ortolang.

JANES

What's up, Switzerland?

IJS

FILIZOVSKA
FAKULTETA

Language, Individuals and Ideologies in mobile messaging

What's up, Switzerland?

- Project duration: 36 months (1/1/2016 – 31/12/2018)
- Overall Research Questions:
 1. What do Swiss WhatsApp messages look like? What has changed overall between Swiss SMS and Swiss WhatsApp messages, and why (as regards linguistic structures, use of images in a broad sense, spelling, register-specific style, individualization vs. accommodation)
 2. What is said / done by the individual users and the media in/on WhatsApp messages and chats, in relation to the findings for question 1?
- Subprojects: The project consists of four subprojects:
 - Language(s) of WhatsApp: Verbal Periphrases and Argument Drop
 - Language Design in WhatsApp: Icono/Graphy
 - Individuals in WhatsApp
 - The Cultural Discourses and Social Meanings of Mobile Communication

Osnove RPK

RAZISKAVE NA KORPUSU JANES

- V čem raba vejice v USV odstopa od norme, ali se ta odstopanja bistveno razlikujejo od tradicionalnih besedil, ali na odstopanja vplivata stopnja formalnosti komuniciranja in interaktivnost medija? (Popič et al., v tisku)
 - označevanje:
 - 500 naključnih tvtov (250 T1L3 + 250 T3L3), 2 anotatorja + kuriranje
 - razvoj & testiranje tipologije
 - rezultati:
 - nestandardna raba vejice na Twitterju je vezana predvsem na skladenjsko rabo
 - kot najbolj problematični sta kategoriji manjkajoče vejice v pristavčnih strukturah (novo) in odvisnikih (univerzalno)
 - uporabniki Twitterja z rabo odvečne vejice nimajo težav (presenetljivo)

- Ali je spletna slovenščina res zapisana govorjena slovenščine? (Zwitter & Fišer 2015)
 - analiza:
 - triangulacija ključnih prvin v pisni, govorjeni in spletni slovenščini
 - rezultati:
 - Janes sicer je med govorom in standardnim pisnim jezikom, a veliko bliže pisni slovenščini
 - podobnosti z govorno slovenščino:
 - os. zaim. v im. (*jaz*), gl. 2. os. v sed. (*veš*) v vlogi diskurzivnih označevalcev, kaz. zaim. v vlogi besed. aktualizatorjev (*un ta rdeč*)
 - izrazi interakcije (*lej*), deiktični izrazi (*tale*), izrazi nestandardne izreke/zapisa (*vidla*)
 - razlike z govorno slovenščino:
 - potek načrtovanja in tvorjenja besedil (ekspresivni izrazi *eee* v govoru, *lol* v spletnih žanrih)
 - prostorska oddaljenost udeležencev (specifična raba svojilnih in kazalnih zaimkov v spletnih žanrih *tale*, *tvoj*)

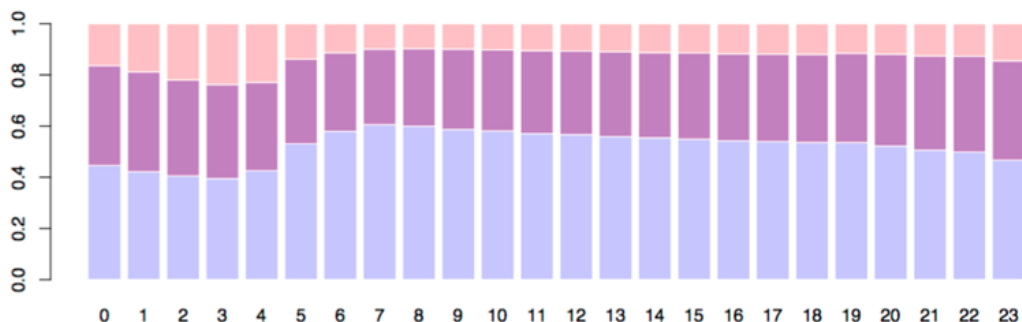
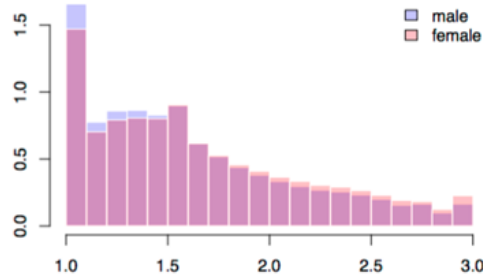
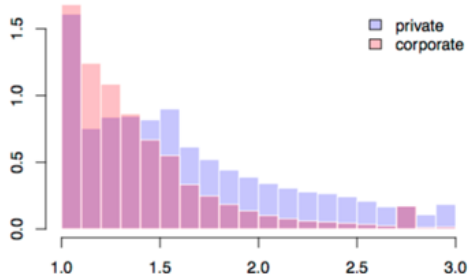
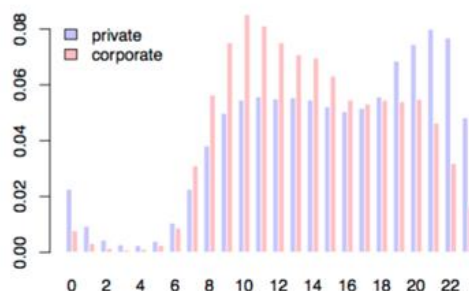
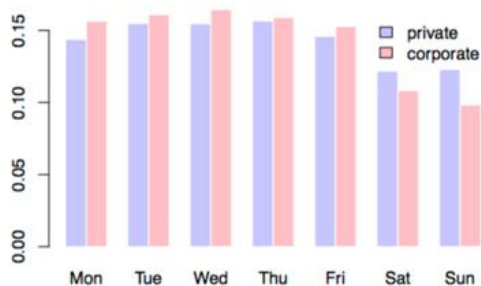
- Pogostost & načini krajšanja slovenskih tvitov (Goličič et al., v tisku)
- analiza:
 - 800 naključnih tvitov (400 T1L1 + 400 T3L3), 2 anotatorja + kuriranje
 - razvoj & testiranje tipologije
- rezultati:
 - trend krajšanja je zelo pogost (90 % vseh tvitov)
 - v nestandardnih tvitih bistveno več krajšanja kot v standardnih
 - močno prevladujejo strategije krajšanja na ortografskem nivoju (87 %), na leksikalnem nivoju smo zabeležili približno 11,5 %, na skladenjskem pa nekaj manj kot 1,5 % odstotka vseh krajšanj
 - najpogostejše rabljena strategija krajšanja je opuščanje presledkov pri ločilih, sledi izpuščanje samoglasnikov na koncu besede, najredkeje pa sta rabljena izpusta predloga in zaimka

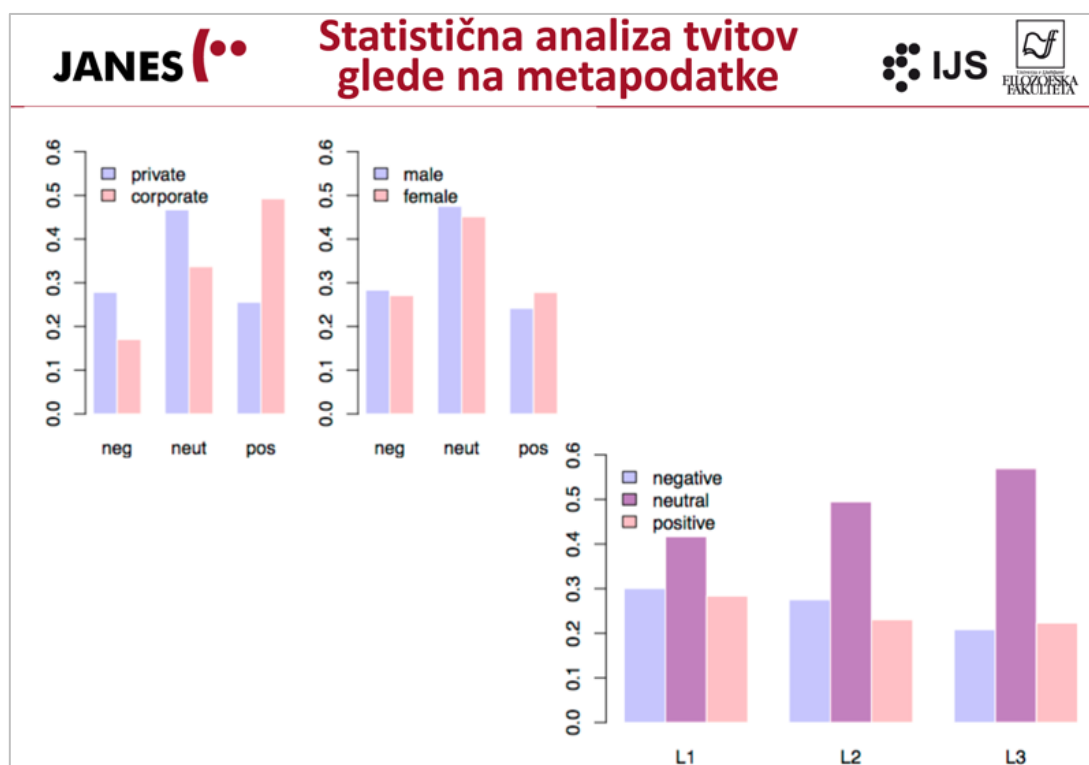
- Ali se raba splošnega besedišča na družbenih omrežjih v čem razlikuje od standardne slovenščine?
 - analiza:
 - nove kolokacije splošnega besedišča
 - kolokacije spletnega besedišča
 - rezultati:
 - neformalni kolokatorji (*nategovati ljudi, frej dan*)
 - aktualne tematike (*feminizacija moških, transspolna oseba, privatizacija vode*)
 - pomenski premiki (*brisanje zgodovine na računalniku*)
 - tujejezične prvine (*rimejk filma, startup podjetje*)
 - terminologija (*evklidski prostor, prva/glavna/spletna/desna stran bloga*)
 - frazeologija (*muca jezik papala*)

- Kakšna je raven specializiranosti, oblika terminov in stopnja nestandardnosti terminov na forumih? (Vintar 2015)
 - analiza:
 - primerjava moderiranih & nemoderiranih forumov
 - objave strokovnjakov & laikov
 - rezultati:
 - terminološka bogatost foruma ni povezana z stopnjo (ne)standardnosti izrazja (avtomobilizem izstopa tako po št. terminov kot po deležu pogovornih, žargonskih in nestandardno zapisanih izrazov, medovernet pa prav tako uporablja veliko terminologije, a z majhnimi odstopanji od standarda)
 - razpravljalci na forumu se izražajo bolj standardno, če je forum moderiran
 - strokovna področja se močno razlikujejo med seboj po terminoloških posebnostih spletne komunikacije (kratice, ki so se ustalile znotraj tega spletnega žanra: *ZM – zadnja menstruacija, G – ginekolog, KT – kontracepcijske tabletki*)

- Ali se vedenje različnih tipov uporabnikov družbenega omrežja Twitter razlikuje glede na dinamiko objavljanja, stopnjo standardnosti & sentiment sporočil? (Ljubešić & Fišer, v tisku)

	retweeted	favorited
private	8.5%	30.2%
corporate	16.3%	18.0%
male	9.4%	29.2%
female	6.8%	32.9%





JANES **IJS** **FILOZOFSKA FAKULTETA**

Osnove RPK

SORODNE RAZISKAVE

Cristian Danescu-Niculescu-Mizil
Stanford University
Max Planck Institute SWS
cristiand@cs.stanford.edu

Robert West
Stanford University
west@cs.stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

Christopher Potts
Stanford University
cgpotts@stanford.edu

- spletne skupnosti so dinamične, interakcijske norme se spreminjajo
- jezikovne spremembe – inovacije, ki postanejo norma – so bistveni del tega dinamičnega procesa
- jezikovne spremembe omogočajo tako možnost individualnega izražanja kot tudi vzpostavljanje kolektivne identitete
- dovzetnost za jezikovne spremembe v spletnih skupnosti ima dve fazi:
 - v fazi učenja uporabniki na jezikovno inovativen način prevzemajo jezik skupnosti
 - v konzervativni fazi uporabniki svoj jezik ne hajo spreminjati in razvijajoče se norme v skupnosti jih prehitijo

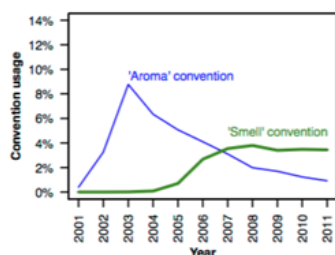
Cristian Danescu-Niculescu-Mizil
Stanford University
Max Planck Institute SWS
cristiand@cs.stanford.edu

Robert West
Stanford University
west@cs.stanford.edu

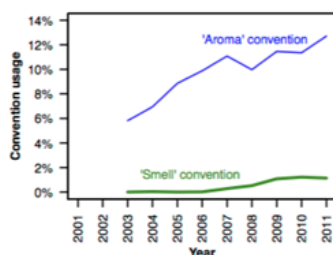
Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

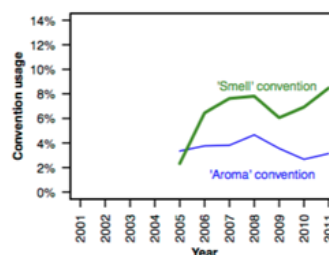
Christopher Potts
Stanford University
cgpotts@stanford.edu



(a) 'Aroma' was the dominant convention by 2003, but it was supplanted by 'S' (for 'Smell') around 2007.

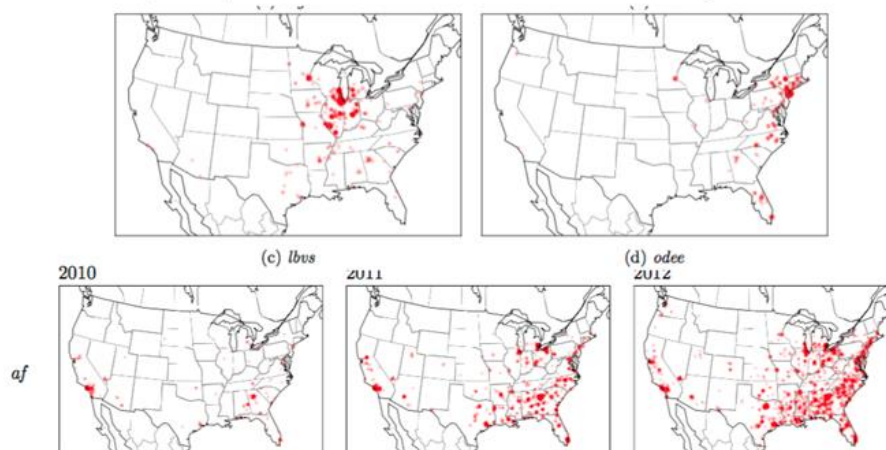


(b) Users who joined in 2003 hung on to the 'Aroma' convention of their youth.



(c) Users who joined in 2005 were more receptive to the emerging 'S' norm.

- avtomatsko odkrivanje značilnih leksikalnih variabel pri uporabnikih družbenega omrežja Twitter iz istih geografskih območij
- razširitev pristopa na diahrono analizo leksikalnih sprememb



Zeeraq Waseem
University of Copenhagen
Copenhagen, Denmark
csp265@alumni.ku.dk

Dirk Hovy
University of Copenhagen
Copenhagen, Denmark
dirk.hovy@hum.ku.dk

- ročna anotacija korpusa 17.000 rasističnih, seksističnih ali nevtralnih tvitov
- analiza korpusa in iskanje značilke za avtomatsko prepoznavanje sovražnega govora
- najboljše delujejo značilke na nivoju znakov
- sociodemografski podatki (regija uporabnika) in kompleksnejše mere (dolžina besede), ne izboljšajo prepoznavanja, edina izjema je spol

	All	Racism	Sexism	Neither
Men	50.08%	33.33%	50.24%	50.92%
Women	02.26%	0.00 %	02.28%	01.74%
Unidentified	47.64%	66.66%	47.47%	47.32%

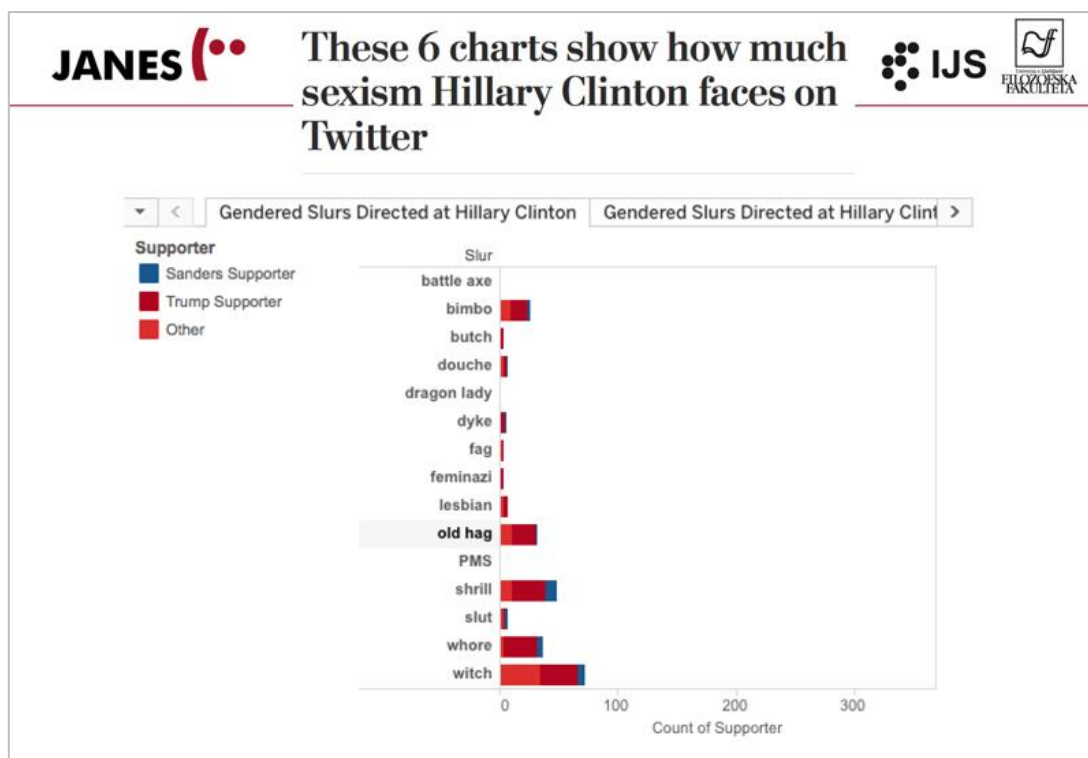
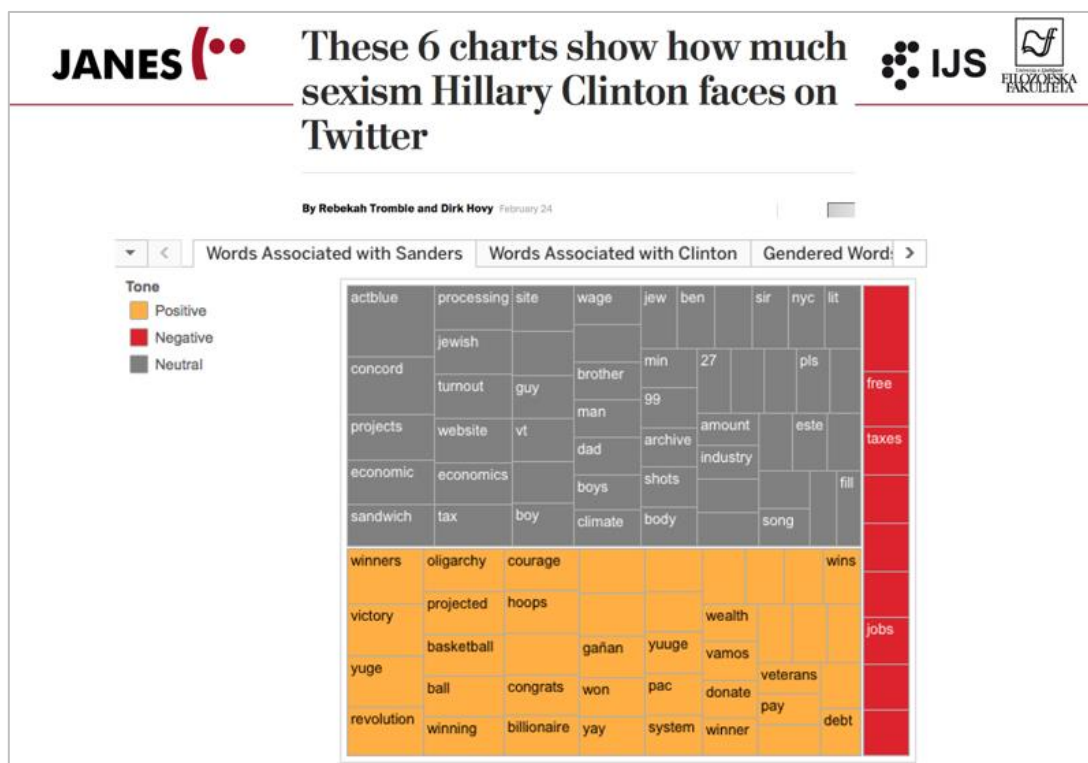
1

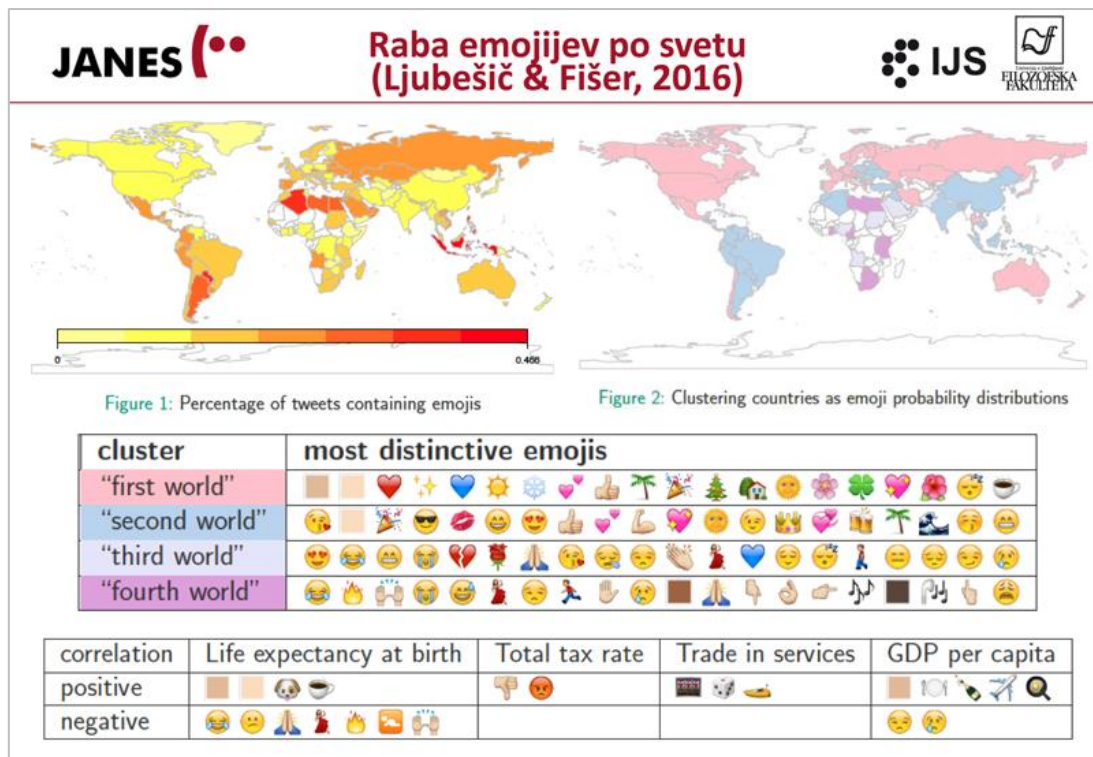
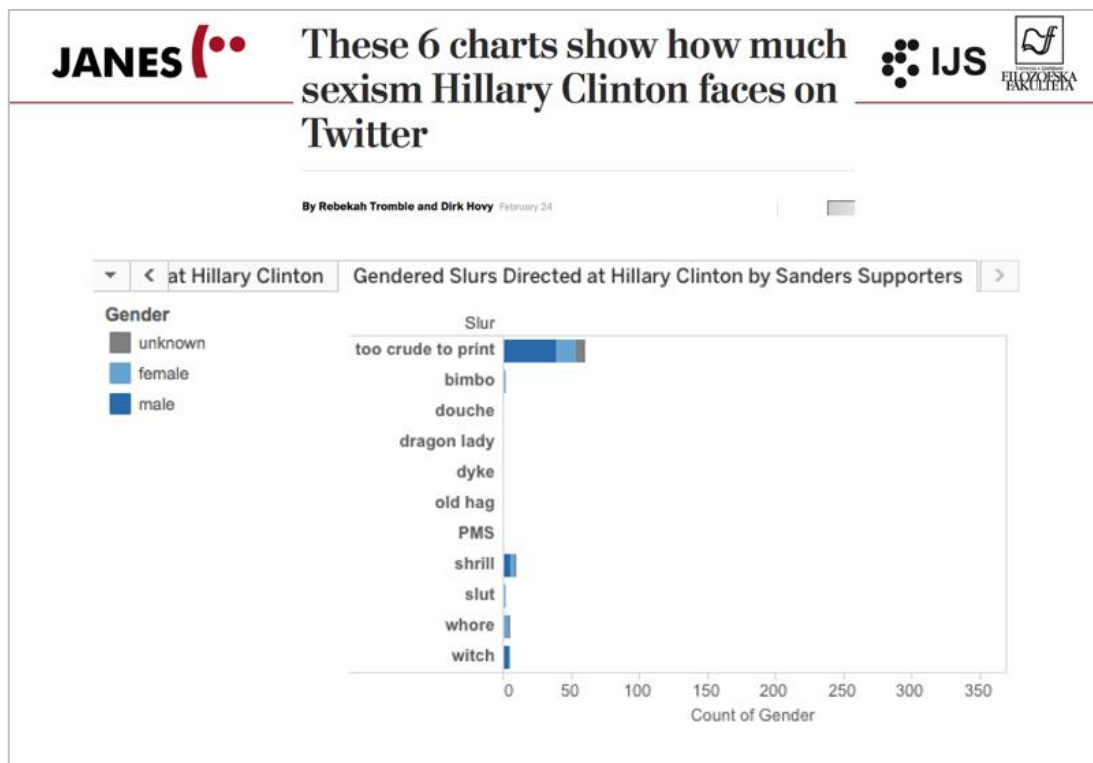
- 

Positive
 Negative
 Neutral

lgb

tea





2.2 Raziskovanje spletne slovenščine – delavnica



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



Korpus JANES

ISKANJE PO KORPUSU

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

The screenshot shows the Sketch Engine web interface. On the left is a navigation menu with links: 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 that is open, showing a list of corpora: 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' and a 'Window:' dropdown set to 'both' and a 'tokens.' dropdown set to '5'. At the bottom, there is a 'Lemma(s):' input field and a dropdown set to 'all' of these items.

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 and 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 is a 'Lemma:' input field and a 'PoS:' dropdown set to 'unspecified'. There is also a 'Phrase:' input field. The 'Word form:' section has an input field and a 'PoS:' dropdown set to 'unspecified', with a 'match case' checkbox. The 'Character:' section has an input field. The 'CQL:' section has an input field and a 'Default attribute:' dropdown set to 'word'. At the bottom, there is a link for 'Tagset summary'.

JANES 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](#)

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

blog	distančirali tudi terminološko, če se jih ima večina itak za	boljše /boljše/dober/Agcmpa	od novinarjev in njih cenzorskih, politično nastavljenih
tweet	nič spornega http://t.co/hsLID7A6Dv ##g @BozoPredalic	boljše /boljše/dobra/Rgc	za marsikoga, da je ne. Lahko katero prime, da ga
blog	Sicer je super, sedaj me zanima, če je lahko še	boljše /boljše/dober/Agcfn	. Grrrr g heh, skrajni čas, bejbi. Jst to že nekeč časa
tweet	Affleck bo naslednji Batman. Buuu ##g A ne bi bilo ful	boljše /boljše/dober/Agcnsn	, če bi tvite z deli, zaposlitvami opremili z enim
forum	rdečic po telesu, včasih tudi po obrazu. Sedaj so vse	boljše /boljše/dober/Agcfn	, niso več rdeče le srbi jo še večkrat (včasih se
tweet	@maticslapsak Vsakemu svoje veselje. Kaj čmo. Vseeno	boljše /boljše/dobra/Rgc	kot nazi ikonografija. #alwayslookonthebrightsideoflife
tweet	slovenske oblasti in sodišča ji stojijo na poti v	boljše /boljše/dober/Agcnsa	življenje http://t.co/6LE4s7GEzb ##g Vsaka tretja ženska
forum	malo neprijetno. Savine so sicer v snegu bile veliko	boljše /boljše/dober/Agcfn	, na mokri in mastni podlagi pa prava katastrofa v
blog	13.09.2012 ob 16:57 g ne vem, meni se zdi, da bi veliko	boljše /boljše/dobra/Rgc	(In bolj seksi) izpadlo, če bi imela zgornji del
forum	kaksne narezane diske (npr. ATE, breombo...) in pa	boljše /boljše/dober/Agcmpa	zavorne plosvice. ##g Pri nekaterih avtih je res potrebno
tweet	koga ##g @Razdelilec tista Gradišnikova je huda, ja.	boljš /boljše/dobra/Rgc	da nima prav ##g hm, a ni Al Gore 07 pokasiral Nobelove
blog	dalje), ker morda v takem moodu kaj lepega zamujaš ... g	boljš /boljše/dobra/Rgc	da neham besedičit lačna sem pa se hočem zamotit
forum	cevi, če bi bilo morda res kej v dovodu nafte, pa nič	boljš /boljše/dobra/Rgc	g - ni 4motion g - kompresija bi avto skos zajebavala
tweet	##g @TamaraSvetina Sem zelo iz vaje, a če ne najdeš	boljš /boljše/dober/Agcmpa	(ga) se lahko potrudim. @ales_gantar ##g @_lnja _ Kaj
forum	Tudi meni ni všeč Astra enjoy, sport mi pa je. Mnogo	boljš /boljše/dober/Agcmpa	sedeže ima, pa el. ročno in tudi ni veliko dražji
tweet	pol sm si pa kupu frustk in sm še zmer u minusu...	boljš /boljše/dobra/Rgc	da bi šou u ošterijo: D http://t.co/SHzwwSnpKF ##g
comment	koncno		
forum	njegovi		
tweet	za st		
forum	original		

Prva | Prejšnja | Stran 2 od 1,363 | Pojdi | Naslednja | Zadnja

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

JANES

SketchEngine: obdelava konkordanc

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

Multilevel frequency distribution

Frequency limit: 0

first level

Attribute: word

Ignore case

6L

5L

4L

3L

2L

1L

Node

1R

2R

Position: 1R

Position: 2R

second level

Attribute: word

Ignore case

6L

5L

4L

3L

2L

1L

Node

1R

2R

Position: 1R

Position: 2R

third level

Attribute: word

Ignore case

6L

5L

4L

3L

2L

1L

Node

1R

2R

Position: 1R

Position: 2R

fourth level

Attribute: word

Ignore case

6L

5L

4L

3L

2L

1L

Node

1R

2R

Position: 1R

Position: 2R

Make frequency list

Text type frequency distribution

Frequency limit: 0

Include categories with no hits:

group.type

group.title

group.urldomain

group.uri

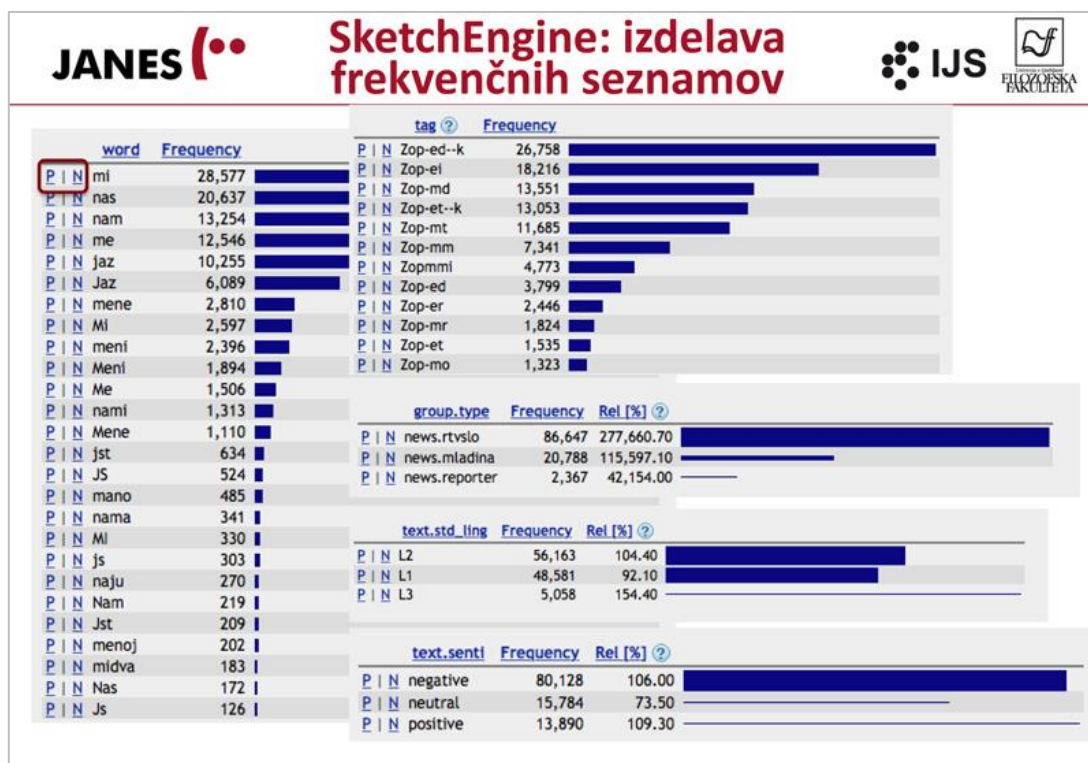
group.year

group.month

group.date

group.time

Make frequency list



JANES **SketchEngine: besedni seznam** **IJS** **FILZOVSKA 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 seznam



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
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
blnbon	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
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 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

- 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

2.3 Raziskovanje spletne slovenščine – vaje

Potek delavnice in vsebina raziskovalnih nalog

Študenti so razdeljeni v 5 skupin. Vsaka skupina si izbere en vsebinski sklop po želji.

1. Variantnost na ortografski ravni med skupinami uporabnikov in pri posameznih uporabnikih

- a) moški/ženske
- b) osebni/korporativni računi
- c) zelo produktivni/malo produktivni uporabniki
- d) različna časovna obdobja
- d) različni viri/tipi besedil

2. Analiza in evalvacija avtomatsko pripisanega sentimenta

- a) glede na ključne besede
- b) glede na oblikoskladenjske oznake

Namen raziskave vsake skupine je priprava **novega opisa izbrane tematike**. Celotno nalogo skupaj s 5-minutno predstavitvijo je treba pripraviti v 60 minutah. V zadnjih 30 minutah bo sledila predstavitev dela vsake skupine.

Skupine v okviru izbrane tematike pripravijo besedilo z naslednjo strukturo:

- I. Ozadnje in motivacija
- II. Zasnova raziskave
- III. Rezultati
- IV. Interpretacija in diskusija
- V. Sklepi

V korpus Janes 0.4 vstopite tu:

- noSkE: http://nl.ijs.si/noske/janes04.cgi/first_form?corpname=janes.04
- friSkE: http://nl.ijs.si/noske/janes04.cgi/first_form?corpname=janes.04

Način iskanja: pomoč je na naslednji strani.

Iskalni koraki:

1. Izberete ISKANJE.

2. Pod vrsto iskanja izberete CQL. V okence vpišite: [lemma=".+ič" &tag="So.*"], privzeti atribut: **lemma**.

POZOR

ič se ustrezno zamenja z drugim
obrazilom

3. Kliknite: IZDELAJ KONKORDANČNI NIZ (čisto spodaj).

uporabnik: janes korpus: JANES v0.4

4. Ko je konkordančni niz izdelan, pojdite na FREKVENCE (levi stolpec).

iskani niz	Podatki	Premešaj	Vzorec	Filter
wikipedia.pagetalk,slv,comment,XJamRastafire,neutral	Hvala za opozorilo. Pri očetovem imenu pa potem r			
wikipedia.pagetalk,slv,comment,XJamRastafire,negative	jezik od nemščine. Prav nasprotno. Ali kakor je v :			
wikipedia.pagetalk,slv,comment,romanm,neutral	stranka, ki se uradno imenuje Komunistična partij			
wikipedia.pagetalk,slv,comment,XJamRastafire,negative	Vau XJam! Hudo! Zlati rez pa to. Kje jih najdeš? Že l			
wikipedia.pagetalk,slv,comment,XJamRastafire,negative	in nemogočimi igrami me pa nadvse zanima. Pa			

Stran 1 od 10,002 Pojdi Naslednja Zadnja

5. V okencu namesto 0 določite 2 (lahko tudi višjo številko, npr. 5), v prvem stolpcu pa izberite atribut: **lemma**. Nato kliknite Izdelaj frekvenčni seznam.

Iskanje
Seznam
O korpusu
< Konkordance
Vzorec
Filter
Prekrivanja
1. zadetek v dokumentu
Frekvence
Oznake niza
Oblike niza
Dokumenti
Kolokacije
Opis niza
Vizualiziraj

Večnivojska razporeditev po frekvenci

Minimalna frekvenca: 2

☒ prvi nivo ☐ drugi nivo ☐ tretji nivo ☐ četrti nivo

Atribut: lemma Atribut: word Atribut: word Atribut: word

Ne ločuj m/v črk Ne ločuj m/v črk Ne ločuj m/v črk Ne ločuj m/v črk

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

iskani niz iskani niz iskani niz iskani niz

Položaj: 6D Položaj: 6D Položaj: 6D Položaj: 6D

Izdelaj frekvenčni seznam

Frekvenčna razporeditev po lastnostih besedil

Minimalna frekvenca: 0

Prikaži kategorije brez zadetkov: ☐

6. V seznamu, ki ga dobite, izberite besede, ki so tvorjenke in ustrezajo vaši nalogi (pomanjševalnice/feminativi), interpretirajte podatke ter zaključite raziskavo.

Marko Stabej

3 SPREMINJANJE JEZIKA

3.1 Spreminjanje jezika – izročki

JANES

IJS

FILOZOFSKA
FAKULTETA

Spreminjanje jezika

Marko Stabej
Oddelek za slovenistiko
FF UL
5. 7. 2016

JANES

IJS

FILOZOFSKA
FAKULTETA

Variatio delectat.
(latinski pregovor)

Plus ça change, plus c'est la même chose.
(Jean-Baptiste Alphonse Karr, 1849)

Some older citiziens welcome the new music and dances,
the new electronic devices and computers. But no one has
ever been heard to say „It's wonderful the way young
people talk today. It's so much better than the way we
talked when I was a kid.“
(William Labov, 2001)

Kakšni bodo avtomobili prihodnosti

(1972)



Kakšni bodo avtomobili prihodnosti

Govorili smo že o »prednikih«, o avtomobilih preteklosti, pregledali smo množico modelov in znamk sodobnih avtomobilov. Zakaj ne bi spregovorili še o avtomobilih jutrišnjega dne? Če je v petdesetih letih dobilo motorno vozilo, ki se je razvilo iz prvotnih vozov na konjsko vprego, današnja oblika, ni nobenega dvoma, da se bo razvijalo in spreminjalo tudi v prihodnosti. Gre za ne tako oddaljeno prihodnost, na katero številni konstruktorji z vsega sveta mislijo že danes, da ne bi zamudili razvoja. V razvojnih oddelkih nekaterih velikih avtomobilskih tovarn so avtomobilu prihodnosti posvetili obsežna preučevanja. Na tem področju vodi ameriški Ford, ki dela načrte svojih avtomobilov za nekaj let vnaprej. Sličice na tej strani prikazujejo nekaj Fordovih modelov prihodnosti. Kot vidimo, imajo ti avtomobili kaj malo skupnega s temi, ki se vozimo z njimi danes: »solarus« predvideva za pogon uporabo sončne energije, »geminie« nakazuje popolnoma nova merila v oblikovanju, »deltair« pa naj bi bil idealno vozilo za mesta prihodnosti. Ne bo preteklo ravno veliko časa in ta čudna vozila bomo videli krožiti okrog nas.

Ljubljana v prihodnosti (okoli 1905)?



"Da ti odgovarjam od verste do verste, ne misli si, da me spreobernó tam ob Dravi takó mir nix... v bigottneža – ergo: Tvoj Tone ti ostane vedno še – ἐπίσκοπος (če ne papež) /.../ – malo (malo?) svojeglaven prostomišljec. (Darwin, Lyell, Haeckel in dr.) Narava, narava! Ovo ti v spomin! – Sicer pa, da kar tu omenim, še ni čisto sigurno, ostanem-li 'notri' ali ne, that is the question! Baje se mora človek strašno veliko 'na pamet' učiti – kailen! latinske citate iz svatago pisma, kakor veš, in to me straši: 'hinc illae lacrimae'. /.../"

Pismo Antona Aškerca, ki ga je enaindvajsetleten pisal prijatelju Požarju 29. avgusta 1877 (M. Boršnik, Aškerc. Življenje in delo. Ljubljana, 1939. 16-17):



Vir: Wikizidlo

Se jezik spreminja?

Se?

Na bolje ali na slabše?

➡ ➡ ➡ ➡ ➡ ● ➡ ➡ ➡ ➡ ➡

Niedna reč delavskega gospodarstva ne plača na se storjenega zatroska, mara ino truda tak bogato, kak tselna reja, ino le je ona blizo po vsih navih krajih skoro celó zapuvena ino ne s' tako marljivostjo ino skerbjo ravnanja, kak si zasluxi. Malokda najdemo v' ednem strani več kak dva ali tri tselinjake, ino se tote naj berxej le redko ino slabo orojene, zato je tudi na pridobitve ino hasek malo misliti. Mi imamo tselno rejo več za miloradost, kak pa za istok priprave. Ino ču se popravamo, odkod pride, da toto izdačno reč delavskega gospodarstva tak malo maramo ino oskerbujemo, to najdemo pri bliznevi izvedenji vseli, da je natem le nava nevednost naj več kriva.

Niedna reč deželškega gospodarstva ne plača na se storjenega zatroška, mara ino truda tak bogato, kak čelna reja, ino le je ona blizo po vsih naših krajih skoro celó zapušena ino ne s' tako marlivostjo ino skerbjo ravnana, kak si zasluži. Malokda najdemo v' ednemi strani več kak dva ali tri čelinjake, ino še tote naj beržej le redko ino slabo orojene, zato je tudi na pridobiček ino hasek malo misliti. Mi imámo čelno rejo več za miloradost, kak pa za istok priprave. Ino či se poprašamo, odkod pride, da toto izdačno reč deželškega gospodarstva tak malo maramo ino oskerbujemo, tè najdemo pri bližnešemi izvedavanji vseli, da je na tem le naša nevednost naj več kriva.

Knjižnica IMP IJS E8

Dajnko, Peter: Čebelarstvo. (1831) [Wikivir]

<http://nl.ijs.si/imp/wikivir/dl/WIKI00299-1831.html#pb.156>

Se jezik spreminja?

Kaj pojmuje kot jezik?

Kaj pojmuje kot jezikovno spremembo?

William Labov, Principles of Linguistic Change

1. del: Internal Factors (1994)

2. del: Social Factors (2001)

3. del: Cognitive and Cultural Factors (2010)

Spreminjanje:

Jezika samega ($\approx$ jezikovnega *sistema*)

Jezikovnih praks (rabe jezika)

Jezikovne zmožnosti ($\approx$ znanja jezika)

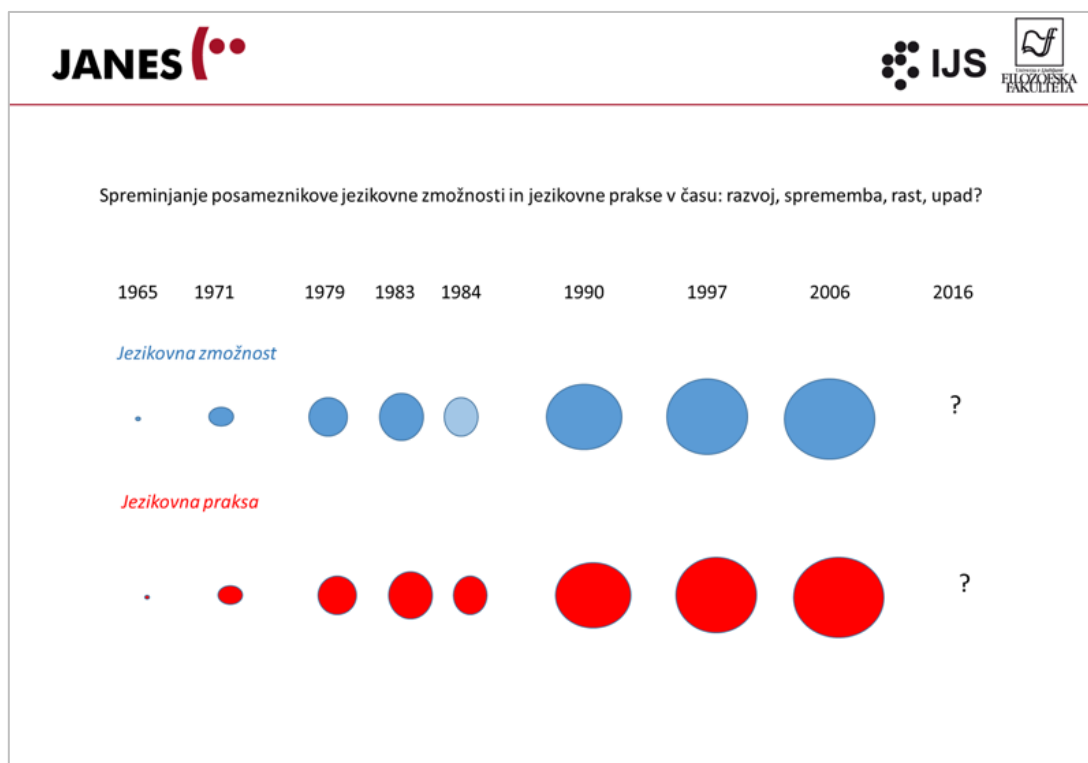
Jezikovnih stališč, jezikovnih ideologij ...

Jezikovne in govorne skupnosti



JEZIK

Krovno poimenovanje
za vse navedeno in še
kaj ...



Spreminjanje : raznolikost (variantnost)

Je jezikovna variantnost nekaj dobrega ali nezaželenega?

Jezikovna : besedilna (diskurzivna) zvrstnost

Jasna (in toga) delitev ali zveznost in hibridnost?

Spreminjanje : raznolikost (variantnost)

Viri za preučevanje enega in drugega:

- „surovi,“ primarni viri (jezikovne prakse – korpusi in drugo gradivo)
- sekundarni viri (jezikovni priročniki)

Metode za preučevanje enega in drugega

Nataša Logar

4 INTERPRETACIJA KORPUSNIH PODATKOV

4.1 Interpretacija korpusnih podatkov – izročki

JANES

IJS

FILOZOFSKA
FAKULTETA

Interpretacija korpusnih podatkov

Nataša Logar
Fakulteta za družbene vede

Poletna šola Janes, 5. 7. 2016

JANES

IJS

FILOZOFSKA
FAKULTETA

UVOD

0.1

Korpusno jezikoslovje izhaja iz spoznanja, da je jezik v prvi vrsti družbeni pojav, kot tak pa se manifestira izključno v besedilih, ki jih je mogoče **opisati in analizirati** (Teubert 2005/1999: 108).

Središče korpusnega raziskovanja je predvsem **performanca** (in manj ali pa sploh ne kompetenca) in **opazovanje jezika v rabi**, ki **vodi k teoriji** (in ne obratno) (Kennedy 1999: 7; Leech 1992: 107).

Korpusnih jezikoslovcev ne zanima le to, katere besede, strukture ali rabe so v jeziku mogoče, ampak predvsem to, kaj se bo v jezikovni rabi pojavilo kot **bolj verjetno, pogosto in tipično ter kaj kot individualno, posebno in enkratno**.

Korpusi so vir podatkov za **jezikovne opise in utemeljitve**.

(Nav. po: Logar 2015: 219.)

➡ Vse zgornje (v rdeči barvi) že pomeni tudi interpretacijo korpusnih podatkov.

0.2

a) Korpus = **vzorec**.

Vzorec je podmnožice celotne populacije, ki je bila na predviden način (i)zbrana za raziskavo. Raziskovalci morajo pri posploševanju podatkov na podlagi vzorca upoštevati način, na katerega je bil ta narejen.

(Peck, Olsen, Delore 2009: 8, 33; nav. po Logar, Dobrovoljc, Arhar Holdt 2015: 467.)

b) Korpus = veliko podatkov

– tudi **številskih podatkov** –
in številске podatke se da **vizualizirati**.

1. Korpus = vzorec

Dejavnikov, ki povzročajo napačno interpretacijo statističnih podatkov, je več:

- Pristranskost raziskovalca
 - Nekonsistentnost v definicijah
 - Primerjave in povezave med neprimerljivimi oz. nepovezljivimi spremenljivkami
 - Napačno sklepanje na celotno populacijo ←
- itd.

Primer:

„Število avtomobilskih nesreč, ki so jih v nekem mestu v določenem letu povzročile voznice, je 10. Število avtomobilskih nesreč, ki so jih v istem mestu in istem letu povzročili vozniki, je 40.

Torej: voznice so vozile bolj varno.“

Zakaj je sklep napačen?

(Vir: Jogi 2014: 2.)

Pri korpusih so za interpretacijo rezultatov poizvedb pomembni dvoji podatki:*

- **predočeni podatki:** podatki, ki so prikazani hkrati z rezultati poizvedb
- **zaledni podatki:** podatki, ki so povezani z izdelavo korpusa

* + poznavanje konkordančnika (v primeru iskanja po spletno dostopnem korpusu)

Anketa med uporabniki korpusa Gigafida (Logar, Dobrovoljc, Arhar Holdt 2015):

Pri interpretaciji podatkov iz korpusa Gigafida večinoma izhajam iz ...



Pogovor ob primerih

PRIMER 1: V čem je šibkost naslednje tabele /op.: primer je avtentičen/?

Tabela 1: Primerjava absolutne pogostnosti pojavnic za navedene pare v korpusih FidaPLUS in Gigafida

	FIDA+ (621 milijonov besed)	GIGAFIDA (1 milijarda besed)
informacija/podatek	157.306/307.641	317.803/594.487
ideja/zamisel	78.346/35.291	149.996/56.857
original/izvirnik	5.513/6.140	11.009/10.824
situacija/položaj	35.871/165.981	82.306/293.253
dialog/dvogovor	20.813/474	43.155/708
internet/medmrežje	73.222/2.827	121.394/5.265
aspekt/vidik	2.400/34.898	4.699/74.898

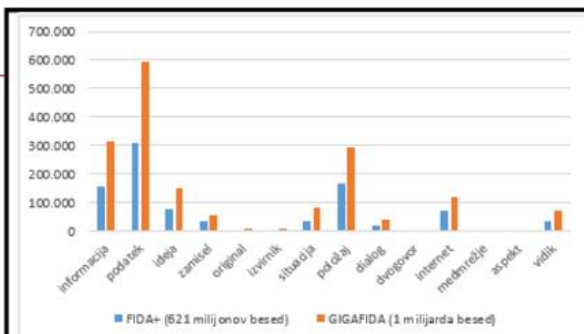
>>

Po preračunu na milijon besed:

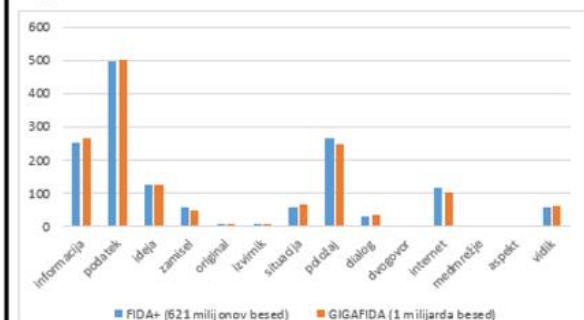
	FidaPLUS (621 milijonov besed)	GIGAFIDA (1.187.002.502 besedi)
informacija	253	268
podatek	495	500
ideja	126	126
zamisel	57	48
original	9	9
izvirnik	10	9
situacija	58	69
položaj	267	247
dialog	33	36
dvogovor	1	0,5
internet	118	102
medmrežje	5	4
aspekt	4	4
vidik	56	63

JANES

>>
Ali v sliki:



Po preračunu na mio besed



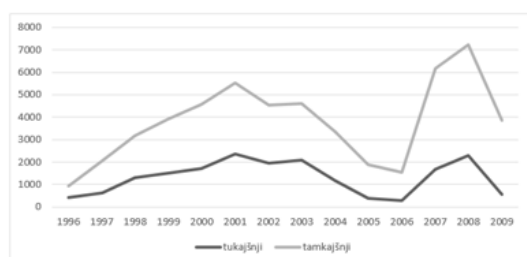
IJS

FILOZOFSKA
FAKULTETA

JANES

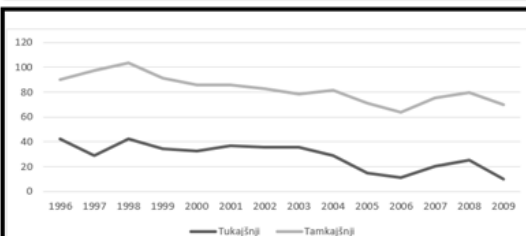
IJS

FILOZOFSKA
FAKULTETA



Pridevnika *tukajšnji* in *tamkajšnji* v časopisnem delu Gigafide v letih 1996–2009 po absolutni frekvenci.

Pozor: absolutna :
relativna frekvenca –
ČAS



Pridevnika *tukajšnji* in *tamkajšnji* v časopisnem delu Gigafide v letih 1996–2009 glede na število vseh besed iz časopisov v teh letih (v številu besed na milijon).

(Vir druge slike:
Kalin Golob, Logar 2014.)

PRIMER 2 /op.: izhodišče je avtentično/

Radijska oddaja Izvidnica (Val 202, 25. 12. 2015), voditelj Damjan Zorc:
„Ali je spletna materinščina vse bolj anglizirana, pravzaprav bi najbrž moral reči
amerikanizirana?“

V čem je problematičnost naslednjega pristopa in posledično interpretacije pridobljenih
podatkov?

Pristop: primerjava seznama besed v korpusu spletne slovenščine JANES s seznamom besed v
korpusu splošne slovenščine KRES > gl. naslednjo sliko.

Seznam besed

Korpus: JANES v0.4

Referenčni korpus: KRES (uravnoveženi)

Spremeni cilj in referenčni (pod)korpus

Stran 1 Pojdi Naslednja >

lemma	JANES v0.4		KRES (uravnoveženi)		Rezultat
	Frekvenca	Frekvenca/mil	Frekvenca	Frekvenca/mil	
t.co	2,773,177	12906.5	0	0.0	12907.5
The	66,929	311.5	31	0.3	248.5
IS	137,826	641.5	218	1.8	228.6
via	107,203	498.9	143	1.2	228.6
posted	44,411	206.7	10	0.1	191.8
Slovensc	50,619	235.6	55	0.5	162.4
he	28,713	133.6	1	0.0	133.5
Facebook	59,691	277.8	164	1.4	118.1
photo	38,625	179.8	75	0.6	111.4
kea	32,693	152.2	70	0.6	96.9
automatically	20,238	94.2	3	0.0	92.9
fora	108,790	506.3	552	4.6	90.9
OF	158,740	738.8	918	7.6	85.8
checked	20,555	95.7	19	0.2	83.5
Wikipedija	18,449	85.9	5	0.0	83.4
Tuta	32,385	150.7	100	0.8	82.9
followed	19,606	91.2	15	0.1	82.0
Us	30,663	142.7	100	0.8	78.5
tvit	16,569	77.1	3	0.0	76.2
ne vedeti	16,132	75.1	0	0.0	76.1
po moj	14,572	67.8	0	0.0	68.8
I'm	21,496	100.0	67	0.6	64.9
How	13,441	62.6	3	0.0	62.0
tkati	59,716	277.9	424	3.5	61.7
IT	107,206	498.9	875	7.3	60.5
hehe	23,143	107.7	123	1.0	53.8
Justa	39,725	184.9	299	2.5	53.4
my	61,845	287.8	553	4.6	51.7
unfollowed	10,858	50.5	0	0.0	51.5
liked	11,102	51.7	3	0.0	51.4
lp	33,428	155.6	269	2.2	48.4
Lp	26,438	123.0	195	1.6	47.4
NSI	9,864	45.9	0	0.0	46.9
What	14,297	66.5	57	0.5	45.8

1. JANES : KRES

>> Problematičnost pristopa je v neupoštevanju **zgradbe** korpusa JANES. Gl. naslednjo sliko.

JANES: zgradba

Korpus: JANES v0.4

Ime novega podkorpusa:

[Število besedil](#) [Število pojavnic](#)

COLLECTION.TYPE	#
☐ blog	34,535,479
☐ forum	47,067,665
☐ news	21,442,813
☐ tweet	107,053,232
☐ wikipedia	4,766,697

>> Če upoštevamo zgradbo korpusa JANES, dobimo naslednje slike.

Kakšna je torej pravilna interpretacija podatkov (s previdnostjo, da smo si ogledali le vrhnji del rezultatov)?

Seznam besed
Korpus: JANES v0.4
Podkorpus: Blogi

Referenčni korpus: KRES (uravnoveženi)
[Spremeni cilj in referenčni \(pod\)korpus](#)

Stran [Pojdi](#) [Naslednja >](#)

lemma	JANES v0.4 : Blogi		KRES (uravnoveženi)		Rezultat
	Frekvenca	Frekvenca/mil	Frekvenca	Frekvenca/mil	
Slovinc	10,602	307.0	55	0.5	211.4
Kouvrar	6,073	175.8	20	0.2	151.7
he	4,661	135.0	1	0.0	134.8
androma	3,005	87.0	0	0.0	88.0
DI	2,569	74.4	0	0.0	75.4
Zzzzzz	2,342	67.8	0	0.0	68.8
Androma	2,140	62.0	0	0.0	63.0
The	2,683	77.7	31	0.3	62.6
blog	21,271	615.9	1,155	9.6	58.3
Lp	5,112	148.0	195	1.6	56.9
clnik	3,632	105.2	106	0.9	56.5
branka22	1,936	56.1	5	0.0	54.8
Andro	2,175	63.0	21	0.2	54.5
komsomolec	2,156	62.4	23	0.2	53.3
www.youtube.com	1,786	51.7	2	0.0	51.9
Tebe	2,813	81.5	92	0.8	46.7
hehe	2,825	81.8	123	1.0	41.0
lp	4,338	125.6	269	2.2	39.2
IS	3,678	106.5	218	1.8	38.3
johanbank	1,276	36.9	1	0.0	37.6
politkomisar	1,975	57.2	68	0.6	37.2
OF	10,663	308.8	918	7.6	35.9
dajan	1,514	43.8	39	0.3	33.9
sailor	1,252	36.3	13	0.1	33.6
Zzzzz	1,061	30.7	0	0.0	31.7
Kobilica-poet	950	27.5	0	0.0	28.5
Tupamaross	975	28.2	4	0.0	28.3
www.publishwall.si	921	26.7	0	0.0	27.7
d.o.o.	907	26.3	0	0.0	27.3
sintrati	1,021	29.6	15	0.1	27.2
Us	1,608	46.6	100	0.8	26.0
BIH	857	24.8	0	0.0	25.8
tupamaross	883	25.6	4	0.0	25.7
lea199	823	23.8	0	0.0	24.8
gromovnik	904	26.2	15	0.1	24.2

1. JANES – blogi : KRES

Seznam besed
Korpus: JANES v0.4
Podkorpus: Forumi

Referenčni korpus: KRES (uravnoveženi)
[Spremeni cilj in referenčni \(pod\)korpus](#)

Stran 1 [Naslednja >](#)

lemma	JANES v0.4 : Forumi		KRES (uravnoveženi)		Rezultat
	Frekvenca	Frekvenca/mil	Frekvenca	Frekvenca/mil	
ne vedeti	11,252	239.1	0	0.0	240.1
lp	25,564	543.1	269	2.2	168.3
kea	11,977	254.5	70	0.6	161.6
feltna	8,924	189.8	23	0.2	160.2
po moj	7,282	154.7	0	0.0	155.7
Tuta	13,337	283.4	100	0.8	155.4
Lp	18,579	394.7	195	1.6	151.1
tkati	22,811	484.6	424	3.5	107.4
LP	26,291	558.6	563	4.7	98.6
kW	4,233	89.9	0	0.0	90.9
ops	4,786	101.7	27	0.2	83.9
The	3,802	80.8	31	0.3	65.0
he	2,983	63.4	1	0.0	63.8
cca.	12,641	268.6	389	3.2	63.7
v glaven	2,625	55.8	0	0.0	56.8
hehe	5,275	112.1	123	1.0	55.9
IS	7,220	153.4	218	1.8	54.9
Us	4,624	98.2	100	0.8	54.2
Tebe	4,337	92.1	92	0.8	52.8
bencinar	3,482	74.0	56	0.5	51.2
slička	4,857	103.2	126	1.0	50.9
oglasnik	3,752	79.7	73	0.6	50.3
šarlatanski	2,490	52.9	10	0.1	49.8
xenon	2,465	52.4	11	0.1	48.9
homologirati	2,313	49.1	7	0.1	47.4
kJ	2,129	45.2	0	0.0	46.2
pozdravljen	30,556	649.2	1,614	13.4	45.2
GET	2,446	52.0	23	0.2	44.5
pasati	6,769	143.8	279	2.3	43.7
x4	2,151	45.7	9	0.1	43.5
homologacija	3,903	82.9	114	0.9	43.1
mašina	6,479	137.7	277	2.3	42.0
amortizer	2,557	54.3	43	0.4	40.8

2. JANES – forumi : KRES

Seznam besed
Korpus: JANES v0.4
Podkorpus: Novice

Referenčni korpus: KRES (uravnoveženi)
[Spremeni cilj in referenčni \(pod\)korpus](#)

Stran 1 [Naslednja >](#)

lemma	JANES v0.4 : Novice		KRES (uravnoveženi)		Rezultat
	Frekvenca	Frekvenca/mil	Frekvenca	Frekvenca/mil	
Slovenec	9,527	444.3	55	0.5	305.7
NSi	1,960	91.4	0	0.0	92.4
JJ	4,107	191.5	211	1.8	70.0
bratuškov	1,472	68.6	0	0.0	69.6
Bratušek	1,643	76.6	15	0.1	69.0
The	1,671	77.9	31	0.3	62.8
www.youtube.com	1,147	53.5	2	0.0	53.6
DeSUS	1,115	52.0	0	0.0	53.0
Us	1,963	91.5	100	0.8	50.6
KPK	1,235	57.6	25	0.2	48.5
AB	981	45.7	0	0.0	46.7
he	973	45.4	1	0.0	46.0
Zoki	962	44.9	6	0.0	43.7
ne vedeti	898	41.9	0	0.0	42.9
Cerar	5,325	248.3	615	5.1	40.8
Juncker	983	45.8	24	0.2	39.1
IS	2,303	107.4	218	1.8	38.6
SMC	987	46.0	32	0.3	37.2
bravo	6,409	298.9	1,036	8.6	31.2
MIRNČAN	634	29.6	0	0.0	30.6
po moj	620	28.9	0	0.0	29.9
ZL	1,071	49.9	89	0.7	29.3
Ukrajina	5,298	247.1	940	7.8	28.2
zapatist	617	28.8	8	0.1	27.9
Patria	1,849	86.2	266	2.2	27.2
kea	853	39.8	70	0.6	25.8
haha	4,804	224.0	934	7.8	25.7
K_ris	527	24.6	0	0.0	25.6
law1	521	24.3	0	0.0	25.3
LOL	517	24.1	0	0.0	25.1
Tuta	948	44.2	100	0.8	24.7
neoliberalen	1,240	57.8	170	1.4	24.4
hehe	1,002	46.7	123	1.0	23.6
BIH	473	22.1	0	0.0	23.1
Dandot	463	21.6	0	0.0	22.6

3. JANES – novice : KRES

Seznam besed
Korpus: JANES v0.4
Podkorpus: Wikipedija

Referenčni korpus: KRES (uravnoveženi)
[Spremeni cilj in referenčni \(pod\)korpus](#)

Stran 1 Pojdi [Naslednja >](#)

lemma	JANES v0.4 : Wikipedija		KRES (uravnoveženi)		Rezultat
	Frekvenca	Frekvenca/mil	Frekvenca	Frekvenca/mil	
Wikipedija	17,766	3727.1	5	0.0	3579.5
CEST	4,047	849.0	0	0.0	850.0
Wikipedist	3,733	783.1	0	0.0	784.1
Welcome	3,781	793.2	13	0.1	716.8
wikipedija	5,050	1059.4	67	0.6	681.4
Дрвоо	3,226	676.8	0	0.0	677.8
пвоалжтоаь	3,225	676.6	0	0.0	677.6
Wilkommen	3,225	676.6	0	0.0	677.6
Welk	3,225	676.6	0	0.0	677.6
Bonvenon	3,225	676.6	0	0.0	677.6
Bem-vindo	3,225	676.6	0	0.0	677.6
Benvenuti	3,226	676.8	5	0.0	650.8
enciklopedičen	3,130	656.6	65	0.5	427.1
wikilettremini	1,962	411.6	0	0.0	412.6
peskovnik	6,472	1357.8	305	2.5	384.7
tilda	2,784	584.1	68	0.6	373.9
Wikipedi	1,620	339.9	2	0.0	335.3
vadnica	2,478	519.9	101	0.8	283.3
ČET	1,235	259.1	0	0.0	260.1
orakelj	2,448	513.6	144	1.2	234.4
vandalizem	2,764	579.9	197	1.6	220.4
ping	1,715	359.8	82	0.7	214.7
WP	1,361	285.5	46	0.4	207.3
GFLY	982	206.0	0	0.0	207.0
blokiranje	1,933	405.5	139	1.2	188.7
netname	885	185.7	0	0.0	186.7
wikipedist	901	189.0	4	0.0	183.9
inetno	866	181.7	0	0.0	182.7
nepristranski	3,606	756.5	382	3.2	181.6
Slovinc	1,175	246.5	55	0.5	169.9
IRC-kanal	790	165.7	0	0.0	166.7
Slovene	1,138	238.7	55	0.5	164.6
Ziga	766	160.7	5	0.0	155.3
IS	2,051	430.3	218	1.8	153.5
navaanje	3,776	792.2	524	4.4	148.2

4. JANES – Wikipedija : KRES

Seznam besed
Korpus: JANES v0.4
Podkorpus: Tviti

Referenčni korpus: KRES (uravnoveženi)
[Spremeni cilj in referenčni \(pod\)korpus](#)

Stran 1 Pojdi [Naslednja >](#)

lemma	JANES v0.4 : Tviti		KRES (uravnoveženi)		Rezultat
	Frekvenca	Frekvenca/mil	Frekvenca	Frekvenca/mil	
t.co	2,773,160	25904.5	0	0.0	25905.5
via	106,884	998.4	143	1.2	456.9
The	58,121	542.9	31	0.3	432.6
IS	122,574	1145.0	218	1.8	407.8
posted	44,357	414.3	10	0.1	383.5
Facebook	58,606	547.4	164	1.4	232.2
photo	38,434	359.0	75	0.6	221.9
automatically	20,201	188.7	3	0.0	185.1
he	19,842	185.3	1	0.0	184.8
Slovinc	26,988	252.1	55	0.5	173.8
checked	20,537	191.8	19	0.2	166.6
followed	19,557	182.7	15	0.1	163.3
fora	94,899	886.5	552	4.6	159.0
tvit	16,346	152.7	3	0.0	150.0
OF	129,044	1205.4	918	7.6	139.9
I'm	20,924	195.5	67	0.6	126.2
How	13,071	122.1	3	0.0	120.1
Us	21,945	205.0	100	0.8	112.5
kea	18,576	173.5	70	0.6	110.4
IT	93,191	870.5	875	7.3	105.5
Justa	38,239	357.2	299	2.5	102.9
unfollowed	10,858	101.4	0	0.0	102.4
liked	11,079	103.5	3	0.0	102.0
my	58,862	549.8	553	4.6	98.5
Tuta	17,063	159.4	100	0.8	87.6
What	13,688	127.9	57	0.5	87.5
rt	54,377	507.9	582	4.8	87.3
twitter	12,909	120.6	55	0.5	83.5
It's	12,518	116.9	54	0.4	81.4
photos	9,396	87.8	14	0.1	79.5
day	20,598	192.4	183	1.5	76.8
AB	7,913	73.9	0	0.0	74.9
Haha	7,671	71.7	0	0.0	72.7
good	20,318	189.8	207	1.7	70.2
LOL	7,367	68.8	0	0.0	69.8

5. JANES – tviti : KRES

PRIMER 3: Kje je razlog, da smo ob pogoju, ki ga kaže prva slika, dobili med rezultati tudi drugo konkordanco na drugi sliki?

Gigafida | Iskanje | Okolica | Seznam

Napredno iskanje

Uporabljaš napredno iskanje. [Vrni se na enostavno iskanje.](#)

Iskana beseda

Način iskanja ☒ vse oblike besed ☐ samo vpisana oblika

Besedna vrsta [Podrobnosti](#)

nam razdelili bone za malico. Možicelj se mi je **prijazno** smehljaj in rekel: "Bomo šli pa v naslednji

-Uporabniku ponuditi izredno **prijazno** in preprosto rešitev

Torej pozor: označevanje (različni označevalniki, napake pri označevanju)

JANES

PRIMER 4: korpusni šum

Gigafida, Sketch Engine: celica

celica (samostalniki)

Gigafida frekvenca = 84,338 (59.82 na milijon)

Constructions	O_s_števili	S_kakšen?	S_s-koga-česa
O_s_števili	1,594	0.50	
S_kakšen?	59,409	3.60	
S_s-koga-česa	23,241	2.90	
goriven	2,407	10.30	obnavljanje
siv	2,683	9.75	delitev
izvoren	1,743	9.72	skupek
jajčen	1,617	9.64	presaditev
rakav	1,635	9.63	membrana
živčen	1,742	9.53	rast
možganski	1,666	9.42	odmiranje
matičen	1,953	9.27	nastajanje
maščoben	1,292	9.18	razmnoževanje
sončen	2,564	9.12	jedro
kožen	1,108	8.96	vrhina

Vrsta besedila

- Časopisi (1,984)
- Revije (391)
- Internet (217)
- Drugo (29)
- Stvarna besedila (22)
- Več

Vir

- Dnevnik (893)
- Dolenjski list (534)
- drugo (493)
- Gorenjski glas (99)

17.00Male sive celice, kviz (VPS 17.00)

16.10Male sive celice, kviz: Mega abeceda, pon.

: Kultura; Odmevi; Mostovi; Risanka; Male sive celice; Zgodbe iz školjke; Nosorog in družina,

voz, film 16.00 Mostovi - 16.30 Poročila 16.45 Male sive celice - 17.45 Dolina krokarjev, odd. - 18.45 Risanka

8.00 Tedenski izbor: Mostovi; Male sive celice; Razjamikovi v prometu; Dolina krokarjev, odd.

TREBA KAKŠEN DAN DOMA DODOBRA OGRETI STOL IN NAPETI MALE SIVE CELICE, O TEM NI DVOMA. TEMU SE NE

to še ne pomeni, da bi morale počivati tudi sive celice. Razbajte jih ob božičnem kvizu...

tudi otiki ne pomagajo, pobagali bi kaki mehaniki za sive celice, a kaj ko so ti pod patronatom LDS

16.45Male sive celice, kviz (VPS 16.45)

strokovnjaki, kateri objavljajo te famozne članke imajo ekstremno veliko sivih celic v možganih, da se upajo iz teorije prejeti

JANES

Samo časopisi:

IJS

FILOZOFSKA
FAKULTETA

Vrsta besedila

- Časopisi (1,984)

Vir

- Dnevnik (893)
- Dolenjski list (534)
- drugo (284)
- Gorenjski glas (99)
- Delo (92)
- Več

Leto

- 2001 (271)
- 2004 (256)
- 2002 (225)
- 2000 (216)
- 2003 (197)
- Več

8.00 Odmevi - 8.30 Zgodbe iz školjke - 9.00 Male sive celice - 9.55 Dok Kihot iz La Manche, film

7.35 Male sive celice, kviz

Ta miselni organ je hudo požrešen: pri odraslem porabijo sive celice 25 odstotkov vse energije, pri novorojenčku pa kar

8.40 MALE SIVE CELICE

9.05 Male sive celice, kviz

9.10 Male sive celice, kviz

tako razlikujejo, da so se za zdaj le kirurgove sive celice sposobne sproti prilagajati nepričakovanim in nepredvidljivim situacijam, ki

strast, strokovnjaki pa trdijo, da je tovrstna telovadba sivih celic ne le koristna, ampak celo nujna za ohranjanje

-17.00 Obzornik - 17.10 Po Sloveniji - 17.30 Male sive celice - 18.20 Tretje oko - 19.10 Risanka - 19.30

; Poslovni barometer 16.00 Mostovi - 16.30 Poročila 16.45 Male sive celice - 17.45 Svet narave, nan. - 18.45 Risanka

Odmevi; Zgodbe iz školjke; Radovedni Taček; Male sive celice - 10.15 Švedski, film - 11.35 Srebrnogrni konjič

Naslov: Siva ekonomija, ki ne potrebuje sivih celic

Špeta Kuder, članica zmagovalne ekipe letošnjega televizijskega kviza Male sive celice

Tedenski izbor: Kultura; Odmevi; Mostovi; Male sive celice; Zgodbe iz školjke; Nosorog in prijatelji.

16.45 Male sive celice, kviz

8.35 MALE SIVE CELICE

16.05Male sive celice, kviz (VPS 16.10)

PRIMER 5: Kaj in kako iščem?

Korošec (1998: 16): poročevalski avtomatizmi, ki so po vsebini sklic na drug vir:

Kot poročajo tuji viri ...

Kot poroča agencija ...

Iz dobro obveščenih krogov ...

Iz krogov, ki so blizu vlade, se je izvedelo ...

(Več v Logar 2008.)

18 poglobljenih intervjujev z novinarji 11 slovenskih medijev (Logar, Kalin Golob 2015)

> več kot 30 enot, največ z jedrnim samostalnikom *vir*:

viri blizu X

naši viri

neuradni viri

uradni viri

zanesljivi viri

zelo zanesljivi viri

preverjeni viri

neimenovani vir

zaupni vir

anonimni vir

vir, ki ne želi pred kamero

vir, ki ne želi biti imenovan

vir, ki mu zaupamo

vir, ki ga poznamo

iz dobro obveščenih virov

od neimenovanega vira

več virov (nam je potrdilo)

ime vira je znano v uredništvu

s pomočjo virov



vir

(samostalni)
Gigafida Časopisi frekvenca = 104,526 (133.37 na milijon)

Constructions	S_kakšen?	58,939	13.10	S_v_ridil-s	30,508	12.70	S_s-koga-česa	15,647	12.10	S_osebik_od	13,091	11.90
O_s_števíli	3,763	10.30										
unarne relacije												
O_zanikanje	744	8.00										
O_količina	736	8.00										
obnovljiv	6,678	11.61	energija	7,133	11.00	izraba	472	9.42	trditi	1,605	9.89	
voden	4,439	9.94	financiranje	2,796	10.72	izkoriščanje	513	9.20	navajati	678	8.93	
neuraden	1,741	9.41	dohodek	1,372	9.76	trditve	478	9.15	praviti	2,291	8.65	
zanesljiv	1,535	9.11	zaslužek	881	9.58	raba	720	9.09	dodajati	291	8.38	
človeški	2,592	8.97	informacija	1,493	9.22	navedba	465	8.60	omenjati	272	8.14	
alternativen	1,270	8.94	navdih	431	8.75	upravljanje	427	8.02	zatrjevati	210	7.99	
energetski	1,409	8.81	okužba	419	8.34	varovanje	330	7.58	namigovati	121	7.92	
naraven	2,993	8.74	preživljanje	281	8.12	pridobivanje	275	7.39	sporočiti	307	7.90	
obveščen	791	8.63	prihodek	647	8.11	iskanje	459	7.33	prišepniti	85	7.68	
nelimenovan	713	8.49	onesnaževanje	272	8.04	zaščita	315	7.32	poročati	359	7.60	
pisen	981	8.15	voda	1,086	7.89	zagotavljanje	242	7.22	usahniti	54	6.97	
nelizčrpen	419	7.83	preživetje	251	7.74	uporaba	878	7.15	potrditi	216	6.80	
diplomatski	576	7.72	sevanje	230	7.70	delež	427	7.10	presahniti	44	6.72	
finančen	2,219	7.64	sredstvo	863	7.51	razvoj	861	6.93	zaupati	93	6.70	
zgodovinski	808	7.42	onesnaženje	180	7.50	pas	73	6.79	opozarjati	188	6.67	
edin	1,264	7.22	toplota	191	7.38	razpoložljivost	50	6.59	zatrčiti	51	6.40	
dodaten	1,404	7.18	vitamin	213	7.36	usposabljanje	71	6.55	predstavljati	338	6.34	
razpoložljiv	322	7.15	zlo	171	7.27	struktura	125	6.49	pričati	73	6.31	
dolgoročen	450	7.11	svetloba	221	7.24	onesnaženje	50	6.46	potrjevati	85	6.20	
pomemben	2,036	6.99	hrup	140	7.06	poraba	194	6.45	dvoniti	36	6.15	
glaven	1,837	6.89	ogrevanje	135	6.89	zagotovitev	74	6.43	ocenjevati	82	6.11	
lasten	1,129	6.82	beljakovina	140	6.88	omejenost	43	6.40	pojasnjevati	67	6.06	
zaupen	243	6.81	podatek	542	6.65	izčrpavanje	36	6.15	zagotavljati	122	6.06	
neusahljiv	199	6.77	zdravje	161	6.55	uvajanje	76	6.12	zanikati	48	6.04	
dragocen	304	6.73	napajanje	83	6.43	navajanje	39	6.07	meniti	175	6.04	



Povzetek 1. dela

Da bom pravilno interpretiral korpusne podatke, moram dobro vedeti:

- po čem iščem:
 - predloženi podatki
 - zaledni podatki
- kako iščem:
 - kaj sem dobil
 - česa nisem dobil (pa bi moral)

Nekaj primerov dobrih sklepanj na celoto

(Vse iz: Slovenščina na spletu in v novih medijih, 2015, ur. D. Fišer.)

1.

Prispevek preverja vrednost novega korpusnega gradiva za normativistične raziskave, in sicer z analizo pogostosti in zapisovanja zvez samostalnika z nesklonljivim levim prilastkom (*solo petje*, *RTV prispevek*) v korpusih Janes in Kres.

Gradivo korpusa Janes razkrije jasnejše trende zapisovanja tovrstnih zvez narazen, v primerjavi s korpusom Kres, kjer je število zvez znatno nižje, trendi v zapisu pa so bolj heterogeni. /.../ Na drugi strani izbrana metodologija razgali težave, ki jih pri luščenju podatkov povzroči neenotnost avtomatskega označevanja, in s tem pokaže na zadrege s kategorizacijo nesklonljivih levih prilastkov, ki se ob samostalniku pojavljajo.

(Arhar Holdt, Dobrovoljc 2015: 4.)

2.

Iz korpusne analize je razvidno, da je tako pri uradnih kot zasebnih računih močan trend zapisovanja imen industrijskih izdelkov z veliko začetnico. Z izjemo kategorije mobilne telefonije je pri uradnih računih povsod odstotek zapisa z veliko začetnico vsaj 70-odstoten, pri zasebnih pa vsaj 54-odstoten.

Med vsemi dvajsetimi obravnavanimi lemami je izstopala le lema *twitter*, ki jo zasebni uporabniki pogosteje pišejo z malo začetnico (54,9 %), medtem ko v zapisih uradnih računov še vedno prevladuje raba velike začetnice (76,0 %). Glede na to, da sicer splošen trend zapisovanja imen industrijskih izdelkov z veliko začetnico odstotkovno izstopa predvsem pri uradnih računih, lahko sklepamo, da proizvajalci in podjetja stremijo k doslednemu zapisu takih izdelkov z veliko začetnico oziroma z morebitnimi izvirnimi zapisi tipa *iPhone*.

(Goli, Popič, Fišer 2015: 32.)

3.

V preučevanih spletnih besedilih, zlasti v tvitih in komentarjih, se pisci pogosto izražajo na način, ki želi biti drugačen, učinkovit in odmeven. Ena od možnosti, kako to doseči pri hitrem, impulzivnem načinu pisne komunikacije /.../, je uporaba frazemov. /.../ Učinek, kot kažejo analizirani primeri, želi biti predvsem humoren, prenovitveni postopki pa so vezani na izrabljanje pomenskih lastnosti frazemov /.../ in vpletanjem aktualnih družbenih in političnih pa tudi osebnih dogodkov.

Zgledi primerjave med korpusoma Janes in Kres potrjujejo predvidevanje, da je raba frazemov tako v izhodiščnem pomenu kot v prenovitvah za spletna besedila zelo značilna, posledično pa je mogoče sklepati tudi na specifične lastnosti spletne komunikacije (hitra odzivnost, drugačnost, učinkovitost, neformalnost), ki se tudi sicer kažejo na vseh ravneh njenega jezikovnega opisa.

(Justin, Hirci, Gantar 2015: 36–37.)

2. Korpus = veliko številskih podatkov >> vizualizacija

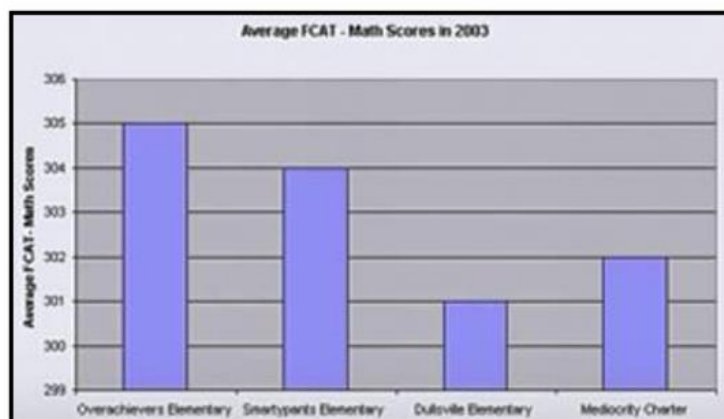
1.

Exposed: How Fox News Lies With Statistics, 29. 11. 2012

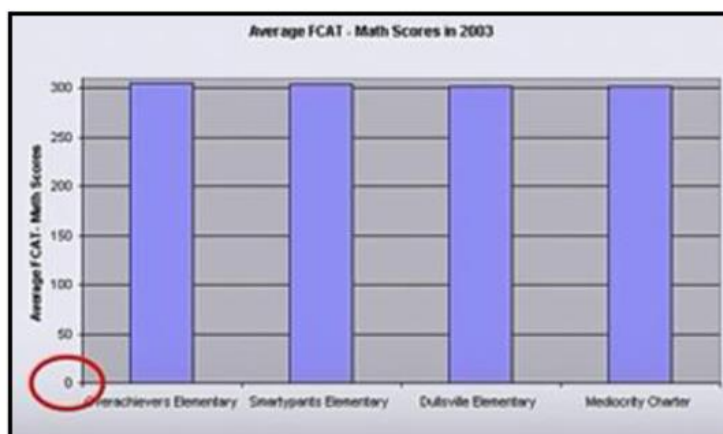
<https://www.youtube.com/watch?v=w7EvBxRYNME>

2. Kaj je narobe z naslednjo predstavitvijo podatkov?

(Vir: Konst 2011.)



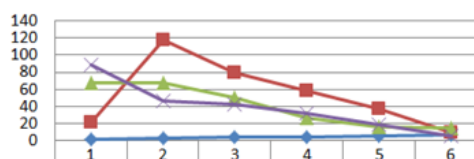
>>



Pogovor ob primerih

PRIMER 1: V čem vse je problematičnost
naslednje slike
/op.: primer je avtentičen/?

Nestandardno stavljene vejice zaradi atraktorjev "glede", "zaradi" in "kljub"



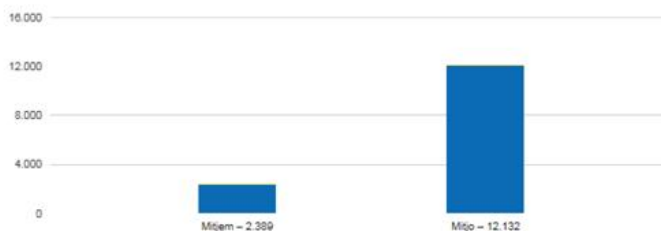
Št. pojavnic med lažnim atraktorjem in vejico	2	3	4	5	6	7
Št. primerov napačno stavljene vejice zaradi lažnega atraktorja "glede"	21	118	80	58	37	10
Št. primerov napačno stavljene vejice zaradi lažnega atraktorja "zaradi"	68	67	50	27	17	15
Št. primerov napačno stavljene vejice zaradi lažnega atraktorja "kljub"	89	46	43	32	19	6

PRIMER 2: S pomočjo naslednje slike odgovorite na vprašanje: Pišete uradno vabilo. Katero končnico pri imenu Mitja v orodniku ednine bi uporabili?

Kateri podatek na sliki je na odločitev najbolj vplival? Kako hitro ste prišli do odločitve?

(Vir: Slogovni priročnik.)

Na grafu si lahko ogledate podatke o rabi oblik lastnega imena **Mitja** v korpusu Gigafida. Pogostost je prikazana za niz oblik po variantah, npr. **Mitja**, **Mitja**, **Mitju** itd. in **Mitja**, **Mitje**, **Mitji** itd. Obe varianti sta skladni s trenutnim pravopisnim standardom.



NA DOLGO IN ŠIROKO

Samostalniki moškega spola, ki se v **imenovalniku** končajo na **-a** (nekateri tudi na **-e**, denimo **kamikaze**), v **rodilniku** pa na **-e**, spadajo v 2. moško sklanjatev (npr. **vajvoda** – **vajvode**, **sluga** – **sluge**, **Miha** – **Mihe** itd.). Vse te samostalnike lahko sklanjamo tudi po 1. moški sklanjavi (prim. **rodilnik** – **vajvoda**, **sluga**, **Miha**, **Aljoša**). V nestandardni rabi se pojavljajo tudi oblike, podaljšane s **-t** (npr. ***Mihata**). Če želimo preveriti, katere oblike so pri posameznih primerih pravilne in katere pogostejše, lahko to preverimo v **leksikonu besednih oblik**.

Standardni sklanjatveni vzorec za moško osebno ime **Luka** je sledeč:

- kdo ali kaj? – **Luka**
- koga ali česa? – **Luke/Luko**
- komu ali čemu? – **Luki/Luku**
- koga ali kaj? – **Luko/Luka**
- pri kom ali pri čem? – **pri Luki/Luku**
- s kom ali s čim? – **z Luko/Lukom**

Končnice pri drugi moški sklanjavi so identične končnicam 1. ženske sklanjave. To pomeni, da lahko tovrstna imena sklanjamo tako po "moški" kot "ženski" obliki, seveda pa oblika ni povezana z biološkim spolom nanašalnice, kar se kaže tudi v rabi – dajalniška oblika **Mihni** je denimo pogostejša kot oblika **Mihu**. Izjema pri tvorjenju pa je tvorba **svojitnega pridevnika**, ki ga tvorimo izključno po moški obliki (torej **vajvodov**, **slugov**, **Mihov**, ne ***vajvodin**, ***slugin**, ***Mihin**). Posebno pozornost je treba posvetiti **prekrivnim imenom** za moški in ženski spol, denimo **Jaša** (ž. sp.) – **Jašin** proti **Jaša** (m. sp.) – **Jašev**.

Povzetek 2. dela

Ko vizualiziram korpusne podatke, moram:

- upoštevati pravila in dobre prakse statistikov
- razumeti moč vizualiziranih podatkov

SKLEP

- Ne samo da se korpusne podatke **da** interpretirati, ampak **ne gre drugače**, kot da jih interpretiramo – interpretacija (preprosta ali kompleksna) je sestavni del vsake uspele korpusne poizvedbe.
- **Dolžnost sestavljalcev** korpusa je priprava opisov korpusne zgradbe in predstavitev odločitev v zvezi z njo; **dolžnost uporabnikov** korpusa pa je, da te opise in predstavitve poznajo. Samo tako bodo lahko rezultate svojih poizvedb interpretirali na pravilen način.

Logar, Nataša (2008): Poročevalstvo kot del jezika v medijih: kot poroča naša dopisnica, v domovini ni rednega dovoda transijade. V: PEZDIRC BARTOL, Mateja (ur.): Slovenski jezik, literatura, kultura in mediji: zbornik predavanj. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete. 157–163. Dostopno prek: http://centerslo.si/wp-content/uploads/2015/10/ssilk_44_zbornik.pdf.

Logar, Nataša (2015): Gradnja referenčnih korpusov na novo: nadgradnja Gigafide. V: GORJANC, Vojko, idr. (ur.): Slovar sodobne slovenščine: problemi in rešitve. Ljubljana: Znanstvena založba Filozofske fakultete. 218–240.

Logar, Nataša, Dobrovoljc, Kaja, Arhar Holdt, Špela (2015): Gigafida: interpretacija korpusnih podatkov. V: SMOLEJ, Mojca (ur.): Slovnica in slovar - aktualni jezikovni opis (Obdobja 34). Ljubljana: Znanstvena založba Filozofske fakultete. 467–477. Dostopno prek: <http://www.centerslo.net/files/file/simpozij/simp34/zbornik%202/Logar-Dob-Arh-Hol.pdf>.

VIRI:

- Exposed: How Fox News Lies With Statistics (29. 11. 2012). Dostopno prek: <https://www.youtube.com/watch?v=w7Ev8xRYNME> (26. 6. 2016).
- Izvidnica (25. 12. 2015). Val 202. Dostopno prek: <http://val202.rtvslo.si/2015/12/izvidnica-37/> (26. 6. 2016).
- Jogi, Usha (2014): Interpretation of data and statistical fallacies. Research hub 1/3. Dostopno prek: <http://www.slideshare.net/rhimri/interpretation-of-data-and-statistical-fallacies> (25. 6. 2016).
- Kalin Golob, Monika, Logar, Nataša (2014): Prostor v poročevalskem skupnem sporočanjem krogu. Slavistična revija 62/3. 363–373.
- Korpus Gigafida. Dostopno prek: <http://www.gigafida.net> (24. 6. 2016).
- Korpus JANES 0.4. Dostopno prek: <http://nl.ijs.si/janes/> (27. 6. 2016).
- Konst, Bill (2011): Identifying Misleading Graphs – Konst Math. Dostopno prek: <https://www.youtube.com/watch?v=ETbc8GJhfHo> (25. 5. 2016).
- Logar, Nataša, Kalin Golob, Monika (2015): Jezikovne izbire pri upovedovanju zaupnih virov informacij: iz zgodovine v sodobnost. Teorija in praksa 52/4. 651–669.
- Slogovni priročnik. Dostopno prek: <http://slogovni.slovenscina.eu/> (25. 6. 2016).

4.2 Interpretacija korpusnih podatkov – delavnica

Potek delavnice in vsebina raziskovalnih nalog

Študenti so razdeljeni v 5 skupin. Vsaka skupina si izbere en vsebinski sklop. Za vsako od tem (1., 2.) študenti kot izhodišče prejmejo opis iz Slovenske slovnice (J. Toporišič, 1991).

SKUPINE IN NALOGE:

1. Modifikacijsko izpeljevanje samostalnikov s pomenom manjšalnosti

- a) moški spol: -ič, -ček
- b) ženski spol: -ica, -ka
- c) srednji spol: -ece, -iče

2. Feminativi: samostalniki ž. spola iz samostalnikov m. spola

- a) -ica, -inja
- b) -ka, -ična

Namen raziskave vsake skupine je priprava **novega opisa izbrane tematike** na podlagi podatkov iz korpusa Janes 0.4. Celotno nalogo skupaj s 5-minutno predstavitvijo je treba pripraviti v 60 minutah. V zadnjih 30 minutah bo sledila predstavitev dela in ugotovitev.

Skupine v okviru izbrane tematike pripravijo besedilo z naslednjo strukturo:

- I. Kratka predstavitev obstoječih opisov (slovnica, SSKJ ...)
- II. Načrt pridobitve podatkov iz korpusa Janes 0.4
- III. Raziskava in beleženje ugotovitev (tudi tega, kaj ste dobili med podatki, a za temo ni bilo relevantno)
- IV. Interpretacija podatkov in nov opis izbrane tematike (oblikujte svojo strukturo, lahko dodaste tipične korpusne zglede, izstopajoče primere opremito z dodatnim komentarjem ipd.)
- V. Vizualizacija ene od ugotovitev

Viri, s katerimi si lahko pomagate (lahko pa dodaste še katerega po lastni izbiri):

- Slovenski pravopis 2001 – slovar: <http://bos.zrc-sazu.si/sp2001.html>
- Slovar slovenskega knjižnega jezika: <http://bos.zrc-sazu.si/sskj.html>
- Odzadnji slovar slovenskega jezika: <http://bos.zrc-sazu.si/odzadnji.html>

Iskalni koraki:

1. Izberete ISKANJE.

2. Pod vrsto iskanja izberete CQL. V okence vpišite: [lemma=".+ič" &tag="So.*"]], privzeti atribut: **lemma**.

POZOR

↓
ič se ustrezno zamenja z drugim obrazilom

3. Kliknite: IZDELAJ KONKORDANČNI NIZ (čisto spodaj).

uporabnik: janes korpus: JANES v0.4

Iskanje
Seznami
O korpusu

Korpus: JANES v0.4

Enostavno iskanje:

[Vrste iskanj](#) [Kontekst](#) [Lastnosti besedil](#) [?](#)

Vrsta iskanja ☐ enostavno ☐ lema ☐ zveza ☐ bes. oblika ☐ znak ☒ CQL

Lema:

Zveza:

Besedna oblika:

Znak:

CQL: [lemma=".+ič" &tag="So.*"]

[Seznam oznak](#)

☐ mala/velika začetnica

Privzeti atribut: lemma

Izdelaj konkordančni niz

4. Ko je konkordančni niz izdelan, pojdite na FREKVENCE (levi stolpec).

iskani niz	Podatki	Premešaj	Vzorec	Filter	Prekrivanja	1. zadetek v dokumentu	Frekvence	Oznake niza	Oblike niza	Dokumenti
wikipedia.pagetalk,slv,comment,XJamRastafire,neutral	Hvala za opozorilo. Pri očetovem imenu pa potem r									
wikipedia.pagetalk,slv,comment,XJamRastafire,negative	jezik od nemščine. Prav nasprotno. Ali kakor je v :									
wikipedia.pagetalk,slv,comment,romanm,neutral	stranka, ki se uradno imenuje Komunistična partij									
wikipedia.pagetalk,slv,comment,XJamRastafire,negative	Vau XJam! Hudo! Zlati rez pa to. Kje jih najdeš? Že l									
wikipedia.pagetalk,slv,comment,XJamRastafire,negative	in nemogočimi igrami me pa nadvse zanima. Pa									

Stran 1 od 10,002 [Pojdi](#) [Naslednja](#) [Zadnja](#)

5. V okencu namesto 0 določite 2 (lahko tudi višjo številko, npr. 5), v prvem stolpcu pa izberite atribut: **lemma**. Nato kliknite Izdelaj frekvenčni seznam.

Iskanje
Seznam
O korpusu
Konkordance
Vzorec
Filter
Prekrivanja
1. zadetek v dokumentu
Frekvence
Oznake niza
Oblike niza
Dokumenti
Kolokacije
Opis niza
Vizualiziraj

Večnivojska razporeditev po frekvenci

Minimalna frekvenca: 2

☒ prvi nivo ☐ drugi nivo ☐ tretji nivo ☐ četrti nivo

Atribut: lemma Atribut: word Atribut: word Atribut: word

Ne ločuj m/v črk ☐ Ne ločuj m/v črk ☐ Ne ločuj m/v črk ☐ Ne ločuj m/v črk ☐

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

iskani niz iskani niz iskani niz iskani niz

Položaj: 6D Položaj: 6D Položaj: 6D Položaj: 6D

Izdelaj frekvenčni seznam

Frekvenčna razporeditev po lastnostih besedil

Minimalna frekvenca: 0

Prikaži kategorije brez zadetkov: ☐

6. V seznamu, ki ga dobite, izberite besede, ki so tvorjenke in ustrezajo vaši nalogi (pomanjševalnice/feminativi), interpretirajte podatke ter zaključite raziskavo.

Tadeja Rozman

5 JE JANES ŠOLAR?

5.1 Je Janes Šolar? – izročki



JE JANES ŠOLAR?

Tadeja Rozman
Filozofska fakulteta Univerze v Ljubljani
Trojina, zavod za uporabno slovenistiko

Ljubljana, 6. 7. 2016



Internet slabo vpliva na jezik.
Internet kvari slovenščino.

Mladi ne znajo več pravilne slovenščine.

Kako naj se otrok nauči pravilne slovenščine?

Je to danes v poplavi jezikovnih zmazkov sploh mogoče?

Ko smo mi hodili v šolo, je bila večina pisane besede s katero smo se srečevali, napisana pravilno. Knjige, časopisi, televizija, ravno tako plakati, slogani ...

Zdaj pa napak kar mrgoli in to v kao lektoriranih tekstih. Da o slogu ne govorimo. Luštna zbirka knjig za otroke, bralce začetnike. Vsak drugi stavek se začneja z in.

Prijateljica je lektorica pri eni od založb. Pravi, da je obup, da ima čisto premalo časa in da gre lahko samo enkrat skozi knjigo. Tekst dobi direktno od prevajalca in pravi, da je večina prevodov obupnih. /.../

Skratka, kako in kje naj se otrok nauči pravilne slovenščine, če je pa dnevno v stiku z neko spakedranščino.

Pa pustimo internet in SMS sporočila obstrani.

(med.over.net, Al' prav se piše ...?)

Pozitivni vpliv interneta na otroke in mladino

- izobraževanje (e- učilnice, Moodle, Cobiss, Wikipedia ...),
- **pridobivanje jezikovnega znanja**,
- ohranjanje stikov s prijatelji in sorodniki iz tujine,
- ozaveščenost o socialnih vprašanjih (npr. poznavanje aktualnih zadev, drugih kultur ...),
- **ustvarjalnost in sposobnost izražanja**,
- razvijanje sposobnosti za ohranjanje pozornosti,
- sprostitev in zabava,
- spremljanje dnevnih dogodkov doma in po svetu,
- potovanje po svetu iz domačega naslonjača,
- hitro in preprosto spletno nakupovanje,
- elektronsko bančništvo ...

(e-priročnik: Internet in varnost otrok v osnovni šoli, Fakulteta za naravoslovje in matematiko, UM)

- »Večina jezikoslovcev meni, da internet slabo vpliva na jezik oz. na njegovo prihodnost.« (T. Burič, Internetni slog, diplomsko delo)
- David Crystal - [How is the internet changing language today?](#)

Kar se Janezek nauči, to Janezek zna.

Janes - Šolar

- Janes – korpus spletne slovenščine
- Šolar – korpus šolskih pisnih izdelkov

Sporazumevanje v slovenskem jeziku

- 2008–2013
- Evropski socialni sklad EU in Ministrstvo za izobraževanje, znanost, kulturo in šport RS
- www.slovenščina.eu
- aktivnost: Nova didaktika poučevanja slovenskega jezika (2009–2010)

Korpus Šolar

korpus šolskih pisnih izdelkov:

- učenci 3. triletja OŠ in dijakov SŠ
- samostojno tvorjeni pri pouku
- (materni govorci)

namen:

- raziskovanje (pisne) jezikovne zmožnosti
- priprava učnih gradiv, vsebin
- korpusna pedagoška slovnica

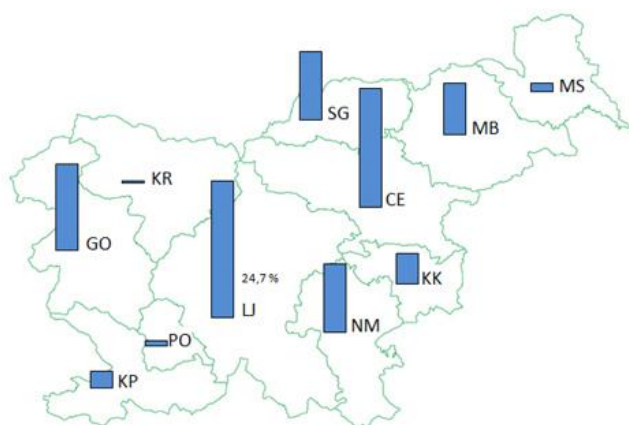
Zbiranje besedil

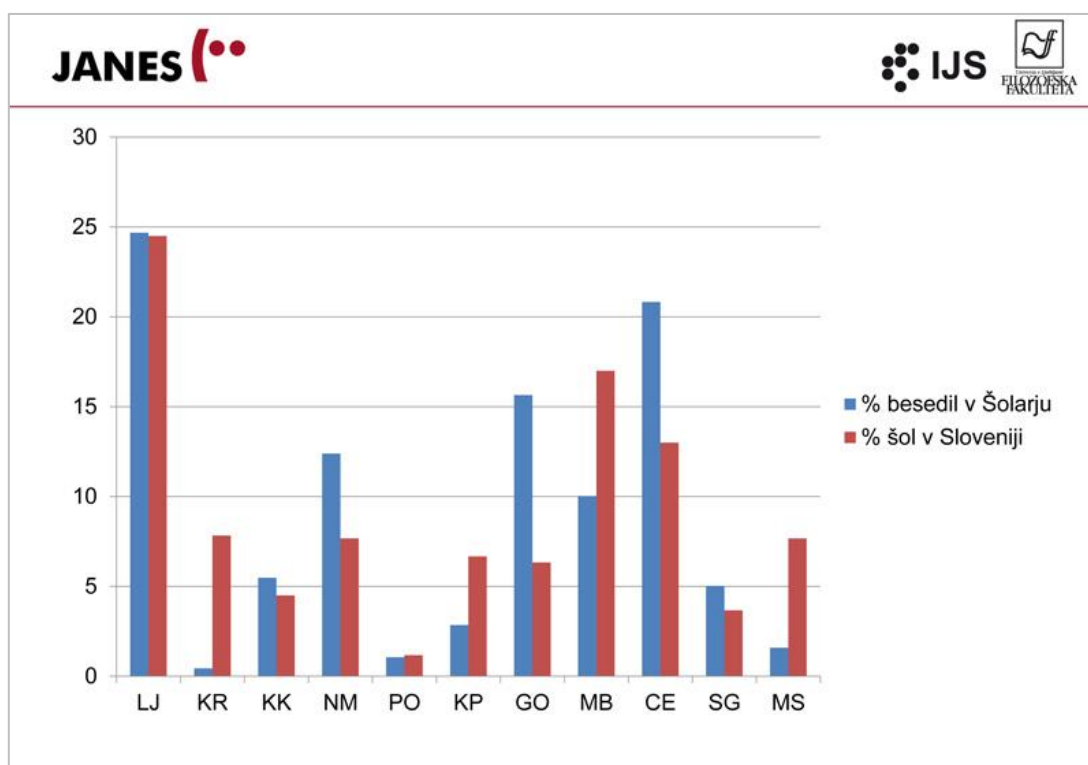
- šolsko leto 2009/2010
- avtorske pravice (prosta dostopnost korpusa)
- 39 (40) šol (vse slovenske regije):
 - JZ: 30 (31), SV: 9
 - LJ 13, NM 6 (7), KK 2, PO 2, GO 4, KP 2, KR 1
 - MB 2, MS 1, CE 4, SG 2
- 8594 besedil

Izbor besedil

- 2703 besedil, 967477 besed
- ! regijska uravnoveženost: JZ 60 %, SV 40 %
- uravnoveženost po stopnjah šolanja, razredih ...
- slovenščina (84,2 %) + drugi predmeti (11)
- ! učiteljski popravki (56 % korpusa)

Delež besedil po regijah





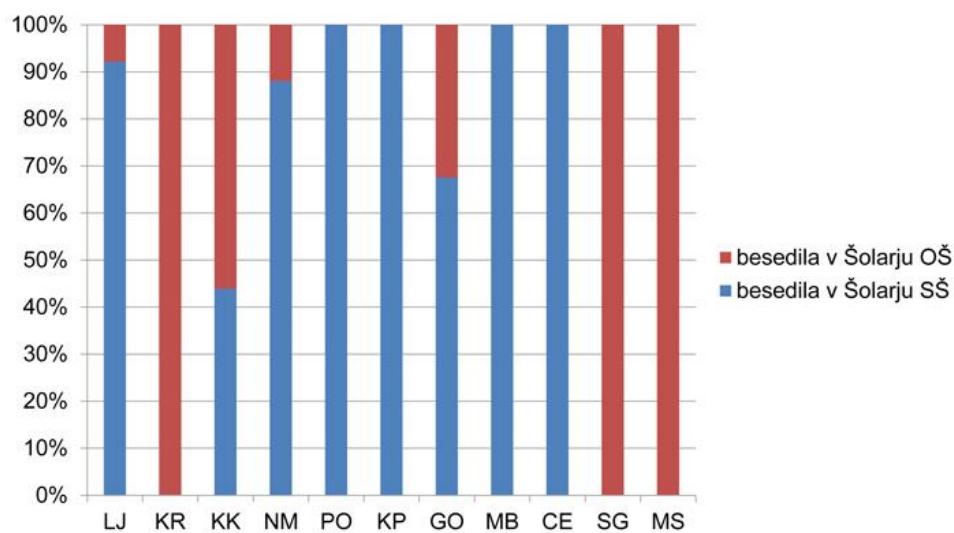
JANES

IJS

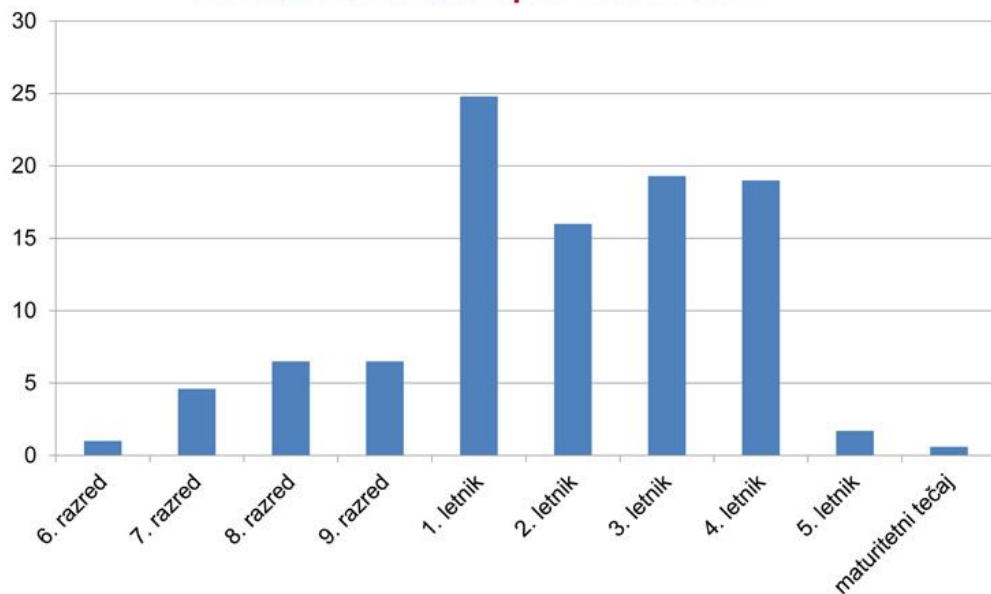
Delež besedil po stopnji šolanja

Šola	Število besedil	Delež	Število besed	Delež
osnovna šola	505	18,7 %	133423	13,8 %
gimnazija	1172	43,3 %	520136	53,8 %
strokovno izobraževanje	843	31,2 %	261496	27 %
poklicno izobraževanje	183	6,8 %	52422	5,4 %

	% besedil v Šolarju	% šol v Sloveniji
OŠ	18,61	75,2
SŠ	81,39	24,8



Delež besedil po razredih



Nadgradnja korpusa Šolar

- 2015–2018
- Ministrstvo za kulturo
- <http://solar.trojina.si/>

Šolar 2.0

Ročna transkripcija

drugim pa nastavijo zanke, za katere se ni težko uloviti.
Ženske je Linhart predstavil kot ljudi, ki imajo glavo
na pravem mestu in jih v pravi meri vodita enkrat
razum, drugič srce. Ženske tudi pomagajo nekaterim
moškim, da se nebi preveč zapletli ali si škodili. Konkretno!

drugim pa nastavijo zanke, **za** katere se ni težko uloviti. Ženske je Linhart predstavil kot ljudi, ki imajo glavo na pravem mestu in jih v pravi meri vodita enkrat razum, drugič srce. Ženske tudi pomagajo nekaterim moškim likom, da se **nebi** preveč zapletli ali si škodili.

Vnos popravkov

drugim pa nastavijo zanke, za katere se ni težko uloviti.
Ženske je Linhart predstavil kot ljudi, ki imajo glavo
na pravem mestu in jih v pravi meri vodita enkrat
razum, drugič srce. Ženske tudi pomagajo nekaterim
moškim, da se nebi preveč zapletli ali si škodili. Konkretno!

drugim pa nastavijo zanke, **za**<p>**v**</p> katere se ni težko uloviti. Ženske je Linhart predstavil kot ljudi, ki imajo glavo na pravem mestu in jih v pravi meri vodita enkrat razum, drugič srce. Ženske tudi pomagajo nekaterim moškim likom, da se **nebi**<p>**ne bi**</p> preveč zapletli ali si škodili.

Tipologija napak

- osnova: označevanje napak (Mojca Stritar - KUST)
- 4 tipi napak s podptipi:
 - ZAPIS (podtipi), OBLIKA, BESEDIŠČE, SKLADNJA (podtipi)
- 35.035 napak

Označevanje popravkov

drugim pa nastavijo zanke, za katere se ni težko uloviti.
Ženske je Linhart predstavil kot ljudi, ki imajo glavo na
pravem mestu in jih v pravi meri vodita enkrat
razum, drugič srce. Ženske tudi pomagajo nekaterim moškim
likom, da se nebi preveč zapletli ali si škodili. Konkretno!

drugim pa nastavijo zanke, <u tip="B">za<p>v</p></u> katere se ni
težko uloviti. Ženske je Linhart predstavil kot ljudi, ki imajo glavo na
pravem mestu in jih v pravi meri vodita enkrat razum, drugič srce.
Ženske tudi pomagajo nekaterim moškim likom, da se <u tip="Z"
podtip="SN">nebi<p>ne bi</p></u> preveč zapletli ali si škodili.

Označevanje popravkov

drugim pa nastavijo zanke, za katere se ni težko uloviti. Ženske je Linhart predstavil kot ljudi, ki imajo glavo na pravem mestu in jih v pravi meri vodita enkrat razum, drugič srce. Ženske tudi pomagajo nekaterim moškim likom, da se ne preveč zapletajo ali ne škodijo. Konkretno!

drugim pa nastavijo zanke, za katere se ni težko uloviti. Ženske je Linhart predstavil kot ljudi, ki imajo glavo na pravem mestu in jih v pravi meri vodita enkrat razum, drugič srce. Ženske tudi pomagajo

konkretno! nekaterim moškim likom, da se nebi preveč zapletli ali si škodili.

Napake zapisa

- črkovanje
- pisanje skupaj/narazen
- mala/velika začetnica
- ločilo
- krajšava
- številka

Napake besedišča

- zamenjave besed (polnopolnomenških, nepolnopolnomenških!)
- zamenjava besednih zvez
- zamenjava presežne in osnovne oblike pridevnika
- zamenjava glagolskega vida in predpon
- ...

Napake oblike

- neustrezne končnice
- neustrezen nedoločnik, namenilnik
- neustrezno tvorjen svojilni pridevnik
- neustrezen pretekli deležnik
- ...

Napake skladnje

- besedni red
- izpuščeni podatki
- odvečni podatki
- neustrezna struktura

- <http://www.slovenscina.eu/korpusi/solar>
- <http://solar.trojina.si/>
- T. Rozman, I. Krapš Vodopivec, M. Stritar, I. Kosem: Empirični pogled na pouk slovenskega jezika
- I. Kosem idr.: Analiza jezikovnih težav učencev: korpusni pristop
- www.korpus-solar.net
- <http://projekt.slovenscina.eu/Vsebine/SI/Kazalniki/K15.aspx>

JANES

IJS

FILZOVSKA
FAKULTETA

Šolar

trojina

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA IZOBRAŽEVANJE,
ZNANOST, KULTURO IN ŠPORT

O korpusu Pomoč

Išči Seznam besed

Navadno iskanje Kontekst Lastnosti besedil

Koda napake :

Napaka:

Popravek:

[oznake napak](#) [označevanje zadetkov](#)

Lexical Computing
2.35.1-SKE-2.137.1-3.86.10

www.korpus-solar.net

JANES

IJS

FILZOVSKA
FAKULTETA

Uporaba Šolarja

- Možnosti uporabe jezikovnih tehnologij za določanje težav pri rabi vejice, Uporaba strojnega učenja za postavljanje vejic v slovenščini (Holozan)
- Postavljanje vejic v slovenščini s pomočjo strojnega učenja in izboljšanega korpusa Šolar (Krajnc in Robnik-Šikonja)
- Analiza jezikovnih težav učencev: korpusni pristop (I. Kosem idr.)
- Možnosti uporabe podatkov iz korpusa Šolar za pripravo slovarskih priročnikov (Arhar Holdt in Rozman)

Tip napake	Podtip	Skupno število napak ²¹	Število povezanih napak
Napaka zapisa	Črkovanje	2672	18
	Pisanje skupaj/narazen	1179	
	Mala/velika začetnica	2125	872
	Ločilo	15371	775
	Krajšava	23	
	Številka	50	
	Skupaj	21420	1665
Napaka besedišča		3807	434
Napaka oblike		3618	1241
Napaka skladnje	Besedni red	1265	212
	Izpuščeni podatki (izpust)	1607	240
	Odvečni podatki (odveč)	2665	588
	Neustrezna struktura	653	61
	Skupaj	6190	1101

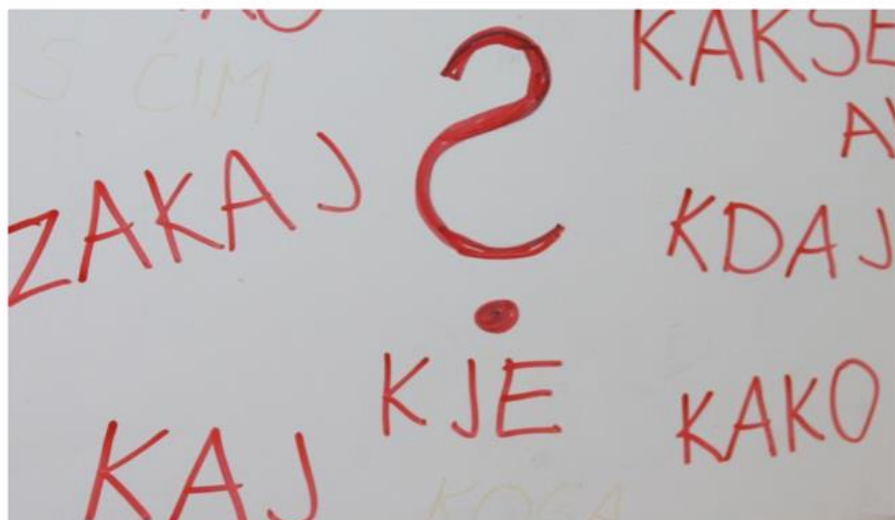
Analize Šolarja

- vejice
- zanikani glagoli (nebi), predložne zveze (nakoncu)
- zapis glasu (medvet), sklopa (življenski), govornjeni zapis (skos), predlogi (k/h, z/s)
- pogovorne besede, splošne besede, zaimki (ki/kateri, moj/svoj), predlogi (v/na – iz/z/s)
- nedoločnik – namenilnik, ujemanje samostalnika, ujemanje povedka z osebkom
- zaporedje stavčnih členov, slogovne napake, izražanje svojine
- ...

Analize Šolarja

- Pedagoški slovnični portal
<http://slovnica.slovenscina.eu/>
- slovarski priročniki

Spletna slovenščina in Šolar



Je Janes šolar?

- raziskovalni cilji
- korpusne in druge raziskovalne metode
- različni korpusi
- nadgradnja korpusov

Helena Dobrovoljc

6 KNJIŽNOJEZIKOVNI STANDARD IN SPLET

6.1 Knjižnojezikovni standard in splet – izročki



Izrazje

knjižni jezik = standardni jezik knjižni jezik > standardni jezik
knjižni jezik ≠ standardni jezik

Zakaj dva termina?

- **knjižni jezik** prvotno povezan z grafičnim sporazumevanjem, danes poimenovan s stališča kakovosti
- **standardni jezik** poimenovan s stališča postopka, ki ga opravimo v procesu standardizacije
- **knjižnojezikovni standard** spomensko: knjižnojezikovna norma

Zakon o javni rabi slovenščine se sklicuje na knjižnojezikovni standard.

SSKJ

POMENSKO USTREZNEJŠI

stándarden -dna -o prid. (ã)

1. **nanašajoč se na standard, predpis**: standardne mere oblačil / standardni predpisi

2. **splošno uporabljan, uveljavljen**: standardna metoda raziskovanja; **temeljno // navaden, običajen**: standardno pohištvo / standardna rešitev problema

◊ ekon. standardna deviacija številčni izraz, ki kaže na razlike od povprečja določenih podatkov; kor. standardni plesi družabni plesi, nastali pod vplivom angleškega načina plesanja;

lingv. standardni jezik knjižni jezik;

ŠIROKO SPREJET IN RAZŠIRJEN

knjižen -žna -o prid. (i)

1. **nanašajoč se na knjigo**: knjižni hrbet, ovitek; knjižna omara, polica / knjižni sejem; novosti s knjižnega trga njegove pesmi so izšle v knjižni obliki

• ekspr. knjižni molj *kdor zelo veliko bere, študira* // redko **književen, literaren**: izdajati knjižni časopis / delo je njegov knjižni prvenec

2. lingv. **nanašajoč se na kodificirani jezik jezikovne, narodnostne skupnosti**: knjižni jezik; knjižna izreka

3. **ki se rabi za knjiženje**: knjižni stroj

◊ tisk. knjižni tisk **tehnika visokega tiska s črk in klišeev**

Različno izrazje, različni nazori



Knjižni jezik moramo razumeti kot **rezultat nadzorovanega procesa** – kot umetno regulirano tvorbo ... V prenesenem pomenu bi se lahko izrazili, da imajo jezikoslovci vrtnarji različne predstave o rastlinah, ki jih urejajo.

Kaj pomeni kultiviranost?

Kaj pomeni nestandardno?

Pojmovanje norme in predpisa

- implicitna norma (nezavednost: kolektivna navada jezikovne skupnosti)
- eksplicitna norma (dogovornost: pravila, raba začetnice ...)
- norma posameznih jezikovnih ravnin (pravorečna, oblikoslovna ...)

Kaj pa spletna norma?

- implicitna norma (interpunkcijska pravila na internetu @, stičnost pike ...)
- eksplicitna norma (netiketa)

Prožna ustaljenost

Prožna ustaljenost je načelo, s pomočjo katerega rešujemo vprašanja, povezana z živostjo govorjenega jezika, ki je nosilec sprememb in razvoja, in željo po ustaljenosti njegove pisane podobe.

(Mathesius 1932)

Pri usmerjanju knjižnega jezika je pomemben podatek o tem, kako se glede inovativnih prvin v jeziku odločajo govorniki intuitivno (nelektorirana besedila).

S katerimi problemi se srečujemo pri naslednjih dvojnicah?

- a) Stožice: stožiški/stoženski
- b) Braille: brajlca/brajica
- c) pletna/pletnja

Kaj pravijo o omenjenih prvinah besedila različnih korpusov?

	Janes	Kres	Gigafida
brajica	11	15	176
brajlca	1	5	78
pletna	101	36	506
pletnja	8	5	181
stoženski	3	0	1
stožiški	117	35	481

Spreminjanje knjižnojezikovnega standarda kot dejstvo



Čas spremeni vse. Nobenega razloga ni, da bi jezik ušel temu univerzalnemu zakonu.
(F. de Saussure, 1912)



Slepo sklicevanje na enkrat zajeto normo je torej nazadnjaško, saj jeziku preprečuje nujno potreben razvoj. (Toporišič, ESJ, 1992)



Svet nespreminjajoče se jezikoslovne odličnosti, ki temelji na sijaju preteklih literarnih oblik, obstaja le v domišljiji. Edini jeziki, ki se ne spreminjajo, so mrtvi jeziki. (D. Crystal, 2005)

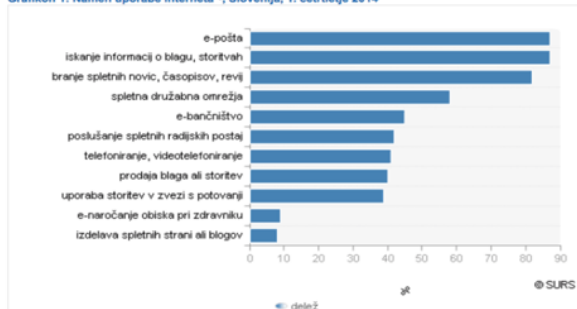
Dinamika jezikovnih sprememb

... na Slovenskem v odvisnosti od družbenega dogajanja:

- konec Avstro-Ogrske (1918): polnofunkcijska raba slovenščine
- konec 2. svetovne vojne (1945): sprememba družbene ureditve
- 60. leta: razmah radia in televizije (2. komunikacijska revolucija)
- 2000: splet (3. komunikacijska revolucija): razširitev skupnosti pišočih v jeziku

Razširjenost interneta

Grafikon 1: Namen uporabe interneta¹⁾, Slovenija, 1. četrtletje 2014

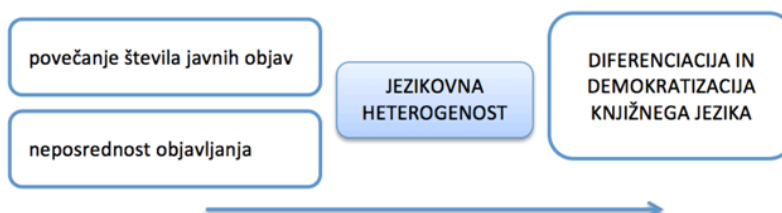


1) Osebe, stare od 16 do 74 let, ki so internet uporabljale v zadnjih 3 mesecih.
Vir: SURS

Ocena: ca 38.000
uporabnikov interneta
(= potencialnih piscev
besedil na spletu) v
Ljubljani

Za primerjavo:
Mladinska knjiga je v
letu ustanovitve izdala
šest knjig.

Diferenciacija in demokratizacija



Kvantitativna in kvalitativna norma

Ena glavnih predpostavk kvantitativnega raziskovanja je, da obstajajo v stvarnosti objektivne zakonitosti, predvsem v obliki vzročno-posledičnih zvez, ki jih je mogoče odkrivati, razložiti, s tem pa tudi napovedati in nadzirati. (Vogrinc 2008)



En Prešernov ali Župančičev zapis, ki vsebuje neko jezikovno rabo, nam mora biti tehtnejši od velikega števila zapisov ljudi, pa če hočete tudi pisateljev. Josip Vidmar (Jezik in slovstvo, 1964)

Posodabljanje normativnih priročnikov in korpus Janes

Neustaljenost knjižnojezikovnega standarda

Neustaljenost v zapisu

Podomačitev v ZAPISU je hitrejša, a tudi "vzratna"

1. **prečkovanje** (hitrejša prilagajanje soglasnikov kot samoglasnikov, razen podvojenih):

(a) podvojeni soglasniki: *clipping* – *kliping*, *mobbing* – *mobing*; *aquafitness* – *akvafitnes*; (b) podvojeni samoglasniki: *googlanje* – *guglanje*; *geek* – *gik*; (c) samoglasniki, ki jih slo. ne pozna: *casual* – *kežuel; *wifi* – *vajfaj; *lionizem* – *lajonizem; *camp* > *kamp/kemper*; *dandy* > *dendi/dandi*; (č) vzratna zamenjava: *textanje*, *qrc*

2. **zapiranje zeva ali hijata**:

(a) *amoniak* – *amonijak*; *revialen* – *revijalen*; (b) *borelia* – *borelija* : *borelioza*; *olimpijada* : *olimpionik*; *diamant*; *violina*;

3. **prilagoditev zapisa v besedah, prevzetih v starejših obdobjih**:

(a) *sprint* – *šprint*; *stopati* – *štopati*; *avtostop* – *avtoštop*; *stadion* – *štadion*; *instruirati* – *inštruirati*; (b) *institucija* – *inštitucija*; *instrument* – *inštrument*; (c) *specifičen*; *inspiracija*; (č) *konzerva/konserva*; *insulin/inzulin*; *disertacija/disertacija*

Neustaljenost v oblikoslovju

Spremembe števila in spola:

- *starš, brokoli, copat/copata, cigaret/cigareta, prsi/prsa; karitas, vodja, vilice, ara, panda ... tapeta/tapet?*

Sklanjatvene neustaljenosti:

- latinsko-grška imena: *Aristotel in Aristoteles* TODA *Aristotela*; *unikum unika (unikuma)*
- lastna imena, homonimna z občnimi: *Požar Požarja* – *požar požara*; *Sever Severja* – *sever na severu*; *gejzir ...*
- imena z neobstoječim polglasnikom: *biftek bifteka, avstralopitek avstralopiteka*
- sklanjanje nedoločnih števnikov: z *več/večimi*
- podaljšave osnove s t: *Mark(ot)a ... Silvo; medo, nono*

Glagolske dvojnice:

- *neso, reko* IN *nesejo, rečejo; vedo/vejo, bodo/bojo*

SMERNICE: delovanje jezika
po analogiji oddaljuje
jezikoslovje od iskanja
historičnih utemeljitev

Smernice v besedotvorju

Praktičnosporazumevalno poenobesedenje (= Bogata besedotvorna pestrost izkazuje usmeritev h gospodarnemu izražanju. Tako se pojavljajo nove, tudi tvorjenke, ki niso le odraz pomanjkanj prostora v računalniško posredovani komunikaciji, temveč tudi povsem evolucijskega poenobesedenja.)

- digitalni fotografski aparat > digitalec
- halogenska žarnica > halogenka
- anonimno pismo > anonimka
- eksterni izpit > eksterec

Nesistemske tvorbe (= tvorjenke, ki so nepredvidljive, ki jim ni mogoče določiti podstave, ne glede števila podstavnih besed ne glede njihovih kategorialnih lastnosti):

- e-denarnica, filmofil, ksenofob, iglomat, šopoholik

Ženski par moškemu:

- esejistka, velemojstrica, politologinja, žirantka

Svojilni pridevniki iz priimkov ob ženskih imenih:

- metoda Marije Montessori, metoda po Montessorijevi nasproti *Montessorijina metoda
- lestvica Viktorije Apgar, metoda po Apgarjevi nasproti *Apgarjina metoda

Zloženke, ki to niso

angl. potato salad – *krompir solata – krompirjeva solata

angl. New York marathon – Newyorški maraton

piknik prostor, kasko zavarovanje, pleksi steklo, mentol bombon, narko kartel, evro kovanec, pingpong miza, last minute ponudba

Lego kocke, Hama perlice, Petrol klub kartica, Poli salama, Adidas majica, Denis Avdič šov, Mercator center

Priporočila:

Hama perlice: Igram se s perlicami Hama

kamp oprema: oprema za kampiranje

evro kovanec: evrski kovanec

Purizem v skladnji

- *ne rabiš biti pravnik, samo pismen ...*
- *pred in po predavanjih ...*
- *delati na ...*
- *kjer v ozirlnih odvisnikih*
- *treba je / potrebno je*

V SSKJ označeno z *neustalj.* / *publ.*

Kakšen je nestandardni jezik v korpusu Janes?

Ugotovitve:

- heterogenost nestandardnega jezika
- močna želja po opaznosti (arhaizmi, neologizmi, rekontekstualizacija ...)
- mešanje prvin govora in zapisa
- napoved potencialnih jezikovnih sprememb prihaja iz nestandardnih besedil (centrifugalne in centripetalne sile – Bahtin; prožna ustaljenost – Mathesius), zanimajo pa nas sistemski premiki, ki vplivajo na celotno skupino istovrstnih jezikovnih enot (npr. raba ločilnega števnika ob množinskih samostalnikih)

Sklep

Štirje tipi knjižnojezikovnega standarda glede na to, čigavo jezikovno rabo upoštevamo pri normiranju jezikovnih prvin (Dolnik 2011):

- **univerzalistični tip**, ki je odprt jezikovni rabi v optimalni meri (je liberalen do rabe)
- **intencionalni tip**, ki se usmerja na prestižne prvine (iz rabe sprejema prvine, ki jih uporabniki občutijo kot prestižne)
- **integracijski tip**, ki se usmerja na ožjo knjižno rabo (njegova težnja je usmerjanje jezikovne prakse v duhu te rabe)
- **rezistenčni tip**, ki se usmerja na ohranitev danega sistema (**konservativni tip**)

Kam se bo uvrstila slovenščina?

6.2 Knjižno in spletno – dva jezika, dva sveta? – delavnica

6.2.1 Jezikovna neustaljenost

V sodobnem knjižnojezikovnem standardu govorimo o sinhroni dinamiki sistema takrat, ko v jeziku soobstajajo enakofunkcijske variante ali dvojnice, uporabnik pa lahko izbira med njimi brez posledic za razumevanje, interpretacijo, pomen ... Variante spremljamo. Ob posodobitvi priročnikov knjižnojezikovnega standarda pogosto spremenimo kodifikacijo, če v rabi izkazujemo dovolj visoko pogostnost ene od različic/variant. Pri spremljanju knjižnojezikovnega standarda smo ugotovili, da govorica mladih (sleng) pogosto prinaša v jezik progresivne jezikovne prvine, inovacije, ki sčasoma postanejo del knjižnojezikovnega standarda. Pri posodobitvi pravopisa posebno pozornost namenjamo (1) novim jezikovnim prvinam (npr. *piconošča*) in tistim, ki jih priročniki, nastali v obdobju 2. polovice 20. stoletja, obravnavajo kot (2) knjižnopogovorne. Manj očitne, a za uporabnika pomembne, so (3) spremembe slovničnih lastnosti obstoječih leksemov (*starši, vodja, ara* ipd.).

PRIMER 1

Motivacija:

Preglejte naslednje izpise iz Slovenskega pravopisa 2001 in ugotovite, kakšna je normativna razlaga danih besed:

*instrumènt*¹ -ênta m (è é) publ. izmenjati ratifikacijske ~e sporazuma *listine*

*instrumènt*² -ênta m (è é) kirurški, merilni, optični ~i; publ. ~i delitve dohodka |predpisi|; uglaševati ~e *glasbila*

inštrumènt -ênta m (è é) uglaševati ~e *glasbila*; kirurški, merilni, optični ~i *instrumenti*

Vaja:

S pomočjo pregleda besednih skic in konkordanc ugotovite, katera različica prevladuje v Gigafidi/Janesu? Ali korpus Gigafida/Janes kaže na pomensko razdelitev v skladu s pravopisom (1) *instrument* 'listina, predpis' – (2) *instrument/inštrumènt* 'kirurški, merilni, optični pripomoček, glasbilo'?

PRIMER 2

Motivacija:

Preglejte naslednje izpise in ugotovite, kakšna je normativna razlaga glagolskih oblik na *-jo* namesto na *-do* (*bojo, vejo, jejo, dajo* : *bodo, vedo, jedo, dado*)

bíti² sem [sə] nedov. si je, sva sta sta, smo ste so (í), poudarjeno sèm sì jè, svà stà stà, smò stè sò (ě ĭ è, à, ò è ò); bódi -te in -íte (ô; î); bíl -à -ó in -ò (î à ô; î à ò), bilí in -ì -è -à (î; ĭ è à); prihodnjik bom boš bo, bova bosta, pokr. bota, bosta, pokr. bota, bomo boste, pokr. bote, bodo, knj. pog. bójo, zastar. bodem itd., poudarjeno bóm itd., zastar. bódem itd.; °bom bil

bom (ô; ô); pogojnik bi --, poudarjeno bì (î); nikalno níssem [sə] nísi ní itd. (í); ne bóm itd.; nê bi (é) in ne bì;

dáti dām dov. dáš dá; dáva dásta dásta; dámo dáste dájó, privzdig. dadó, dadé; dàj dájte; dál -a, dát, dán -a; (dàt) (á)

íti grém in íti grèm dvovid. grémo in grêmo, star. gremò, gréste in grêste, star. grestè, gredó in gredò in grêjo, star. grejò, bom šèl in pójdem itd.; pójdi -te tudi -íte, pokr. ídi -te, nedov. gredóč, star. idóč; šèl [šəy] šlà šlò tudi šló (íti/íti grém/grèm, grémo/grémo, star. gremò, gréste/gréste, star. grestè, gredó/gredò/gréjo, star. grejò, pójdem; pójdi/pôjdi; šàl šlà šlò/šlò)

védeti vém nedov., 2. in 3. os. dv. vésta; 2. os. mn. véste, 3. os. mn. vèjo in vedó, védi -te, vedóč; védenje; (é)

Raziskava:

Ali lahko na gradivu iz korpusov Janes/Gigafida potrdite prodiranje pogovornih oblik na *-jo* v primerjavi z *-do*?

1. poiščemo vse občne samostalnike z variantno končnico *-lec* oz. *-vec*

CQL: [word=".+[jd]o" & tag="G.*"]

2. izdelamo frekvenčni seznam wod – lemma_lc (v primeru velikega števila zadetkov nastavite ustrezno minimalno frekvenco, npr. 10)

word	lemma_lc	Frekvence
p N bodo	biti	199,174
p N imajo	imeti	92,213
p N nimajo	imeti	29,360
p N delajo	delati	26,125
p N morajo	morati	20,958
p N pravijo	praviti	15,375
p N morejo	moči	15,067
p N znajo	znati	14,100
p N pridejo	priti	13,623
p N vedo	vedeti	13,394
p N gredo	iti	12,019
p N majo	meti	11,182
p N obstajajo	obstajati	10,706
p N želijo	želei	10,652
p N mislijo	mislii	9,395
p N dajo	dati	9,118
p N dobijo	dobiti	8,930
p N govorijo	govoriti	8,856
p N živijo	živeti	8,197
p N vidijo	videti	7,799
p N grejo	iti	7,364
p N hočejo	hoteti	6,991
p N kažejo	kazati	6,743
p N uporabljajo	uporabljati	6,524
p N naredijo	narediti	6,417
p N prodajajo	prodajati	6,216
p N pišejo	pisati	6,080
p N delujejo	delovati	6,040

3. seznam uredimo po lemma_lc, identificiramo ustrezne kandidate in preverimo pogostost ter rabo

p N bodo	biti	199,174
p N bojo	biti	923

p N Gredo	iti	147
p N Grejo	iti	133
p N gredo	iti	12,019
p N grejo	iti	7,364

p N izvedo	izvedeti	137
p N izvejo	izvedeti	102

PRIMER 3

Motivacija:

Preglejte naslednje izpise iz Slovenskega pravopisa 2001 in ugotovite, kakšna je normativna razlaga danih besed:

volívec -vca m z -em člov. (i)

volívka -e ž, člov. (i)

volívčev -a -o (i)

volílec -lca [ɥc] m z -em člov. (i) *volivec*

volílka -e [ɥk] ž, člov. (i) *volivka*

volílčev -a -o [ɥč] (i)

Preberite odgovor iz Jezikovne svetovalnice:

Vprašanje:

Kakšno je vaše stališče do dileme *volilec/volivec*?

Odgovor:

Člena 946 in 949 v pravopisnih pravilih SP 2001 določata naslednje:

- Priponsko obrazilo *-lec* ima prosto varianto *-vec*, če je na koncu korena / ali *lj*: *voliti* – *volivec* ali *volilec*, *ponavljati* – *ponavljavec* ali *ponavljalec*.
- Enaka pravila veljajo tudi za pisanje tvorjenk s končnim *-ka* (*bralka* – *pevka*, *volivka/volilka*), *-ski* (*bralski* – *pevski*, *volivski/volilski*), *-stvo* (*bralstvo*, *pevstvo*, *volivstvo/*

volilstvo) ter za izpeljanke iz tvorjenk na *-ec*, *-ka*, *-ski* in *-stvo* (*bralčev*, *pevčev*, *volivčev/volilčev*, *bralkin*, *volivkin/volilkin*).

Zveza "**prosta varianta**" verjetno pomeni, da lahko poljubno izbiramo med eno ali drugo različico, torej *volivec* **ali** *volilec*.

V pravopisnem slovarju pa je na prvem mestu *volivec* kot **nadrejena pisna različica**.

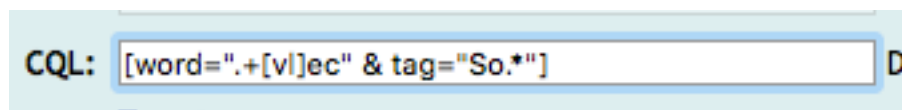
Če se glede pogostnosti rabe obeh variant posvetujemo z gradivom, kakršnega prinaša besedilni korpus Gigafida, ugotovimo, da se beseda najpogosteje pojavlja v časopisih in da je – zanimivo – slovarska prioriteta najpogosteje upoštevana v časniku Delo (16.401), *volilca* pa dopuščajo lektorji časnika Dnevnik (17.816). Razmerje med obema variantama je v prid *volivcu* (46.127). *Volilec* se pojavi 30.451-krat.

Raba torej potrjuje določilo v pravopisnih pravilih.

Raziskava:

Ali lahko ugotovitve za *instrument/instrument* posplošimo na vse podobne primere, v katerih se pisne variante razlikujejo v sklopih šp/sp in št/st, npr. *štadion/stadion*, *štopati/stopati*, *šprinter/sprinter*?

1. poiščemo vse občne samostalnike z obraziloma -vec ali -lec



2. izdelamo frekvenčni seznam (v primeru velikega števila zadetkov nastavite ustrezno minimalno frekvenco, npr. 10)

3. seznam uredimo po abecedi, identificiramo ustrezne kandidate in preverimo pogostost ter rabo, pri tem smo pozorni na:

- a) pravilo za dodajanje obrazil -lec in -vec
- b) ločevanje med vršilci dejanja in napravo/sredstvom

p N solzilec	83	■
p N solzivec	193	■

DODATNI PRIMERI

Raba refleksivnih glagolov:

iztíríti -im dov. iztírjen -a; iztírjenje (í î) *kaj* Diverzanti so iztirili vlak; poud. *iztiriti koga* Otrokova smrt jo je iztirila |spravila iz duševnega ravnotežja|; Lokomotiva je iztirila *se je iztirila*

iztíríti se -im se (í î) Zaradi plazu se je lokomotiva iztirila

dopásti se -pádem se dov., nam. dopàst se; dopádenje; drugo gl. pasti¹ (á; á â) nekuj. pog. *biti všeč, ugajati: komu ~ ~ dekletu, sam sebi*

lemma_lc	Frekvence
p N ajati še	67
p N bati se	4,963
p N bliskati se	50
p N bližati se	1,385
p N bojevati se	84
p N boriti se	961
p N braniti se	148
p N brati se	375
p N brigati se	138
p N bruhati se	57
p N cediti se	53
p N dajati se	53
p N dati se	3,699
p N dati še	903
p N dejati se	200
p N delati se	951
p N delati še	172
p N deliti se	116
p N dobiti se	838
p N dobiti še	361
p N dodati se	53
p N dodati še	296
p N dogajati se	888
p N dogovoriti se	133
p N dojiti se	67
p N dokazati se	56
p N dopasti se	95
p N dotakniti se	187
p N dotikati se	115
p N dozdevati se	123
p N dreti se	184
p N drogirati se	64
p N družiti se	200
p N držati se	2,734
p N dvigati se	52
p N dvigniti se	138
p N fučkati se	52

Raba množinskih samostalnikov:

stárš -a m s -em člov. (á) neknoj. pog. *eden od staršev*

prsa prs s mn. (ř; ř) neknoj. pog. *prsi*

word	Frekvenca	
p N očala	3,084	
p N očal	806	
p N očali	581	
p N ocala	371	
p N Očala	140	
p N očalih	114	
p N ocal	75	
p N očale	60	
p N ocali	60	
p N OČALA	32	
p N Ocala	12	
p N očalah	11	
p N ocalih	11	
p N očalc	9	
p N ocale	7	
p N očalom	6	
p N Očale	6	
p N očalami	5	
p N OČAL	5	
p N učala	4	
p N očalmi	4	
p N Očal	4	
p N očarla	2	
p N očalam	2	
p N ocalom	2	
p N OČALI	2	
p N očalo	1	
p N očalE	1	
p N ocalo	1	
p N Očalom	1	
p N Očalnika	1	
p N OČALE	1	
p N OCALA	1	

Mija Michelizza

7 TIPOLOŠKA ANALIZA SPLETNIH BESEDIL

7.1 Tipološka analiza spletnih besedil – izročki



Tipologija spletnih besedil

Mija Michelizza

Inštitut za slovenski jezik Frana Ramovša ZRC SAZU



Tipi spletnih besedil

- Katere tipe spletnih besedil poznamo?
- Zakaj bi nas lahko zanimali različni tipi besedil?
- Glede na katere kriterije bi lahko razdelili spletna besedila?
- <https://www.1ka.si/main/survey/uporabnost.php?anketa=100973>

Mejniki v zgodovini računalniško posredovane komunikacije (RPK)

- elektronska pošta (1971)
- spletni forum (1980)
- IRC (1988)
- SMS (1992)
- blog (1997)
- Google (1999)

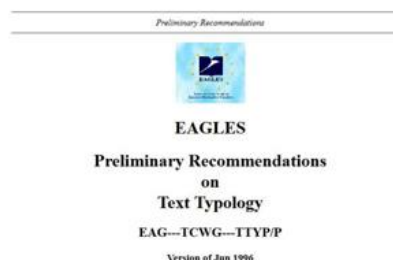
Mejniki v zgodovini računalniško posredovane komunikacije (RPK)

- Wikipedija (2001)
- Skype (2003)
- Facebook (2004)
- Twitter (2006)
- Snapchat (2011)

Zastopanost spletnih besedil v korpusih

- Priporočila *EAGLES*
(junij 1996)

*V sklopu korpusnega
jezikoslovja priporočila
uvajajo trdelno delitev
besedil; poleg tradicionalne
delitve na govorna in pisna
dodajajo tretjo skupino
elektronskih besedil.*



Spletna besedila v tujih korpusih

- *Cambridge English Corpus*
- *CORPES XXI*
- Predvidena so spletna besedila v referenčnih korpusih nizozemščine, poljščine in hrvaščine.

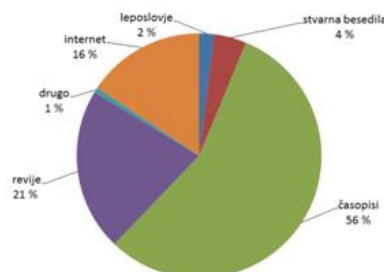
Podrobneje o korpusih za posamezne jezike v Logar Berginc, Ljubešić (2013): *Gigafida in slWaC: Tematska primerjava*.

Besedilni korpusi (samo) iz spletnih besedil

- *Oxford English Corpus*
- Korpus za svahili (1999)
- *Janes*

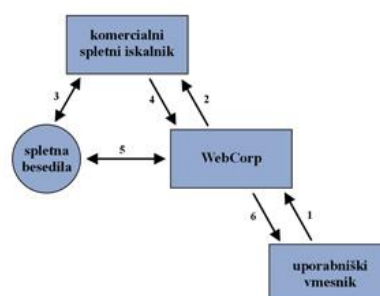
Spletna besedila v slovenskih korpusih

- *Korpus FIDA* (1998): poskusno vključenih 11 dokumentov spletnih besedil oz. 20.999 pojavnic
- *Korpus FidaPLUS* (2006): manjši delež besedil z interneta (1,24 %)
- *Korpus Gigafida* (2012): 16 % internetnih besedil



Spletni korpusi (1/2)

- Splet kot korpus
 - Raba komercialnega spletnega iskalnika
 - Dodajanje posebnega procesorja komercialnemu spletnemu iskalniku za jezikoslovno rabo, npr. *WebCorp*.



Spletni korpusi (2/2)

- Splet za korpus
 - Za razliko od spleta kot korpusa, kjer neposredno uporabimo spletne iskalnike za jezikovne raziskave, gre v primeru spleta za korpus za zbiranje spletnih strani (navadno s pomočjo *pajkanja*) in oblikovanje posebnega korpusa iz spletnih vsebin.
 - Primeri: *BootCat*, *WebBootCat*, *WaC*, *TenTen*, *Janes* ...

Problematika poimenovanja teh besedilnih enot

- (zvrsti (Toporišič))
- žanri (tudi kiberžanri, spletni žanri), tipi besedila, besedilni tipi, besedilne vrste
 - Definicija spletnega žanra s tremi elementi: (1) vsebina, (2) oblika in (3) funkcionalnost (Shepherd in Watters (1998))
 - Crowston in Williams (2000)
 - Pet tipov žanrov: (1) reproducirani, (2) adaptirani, (3) novi, (4) spontano nastali, (5) neklasificirani (Santini (2007))

Problematika poimenovanja teh besedilnih enot

- Značilnosti spletnih besedil: (1) hibridnost, (2) individualizacija, (3) evolucija
- od-nič-do-večžanrska klasifikacija (angl. *Zero-to-Multi-Genre Classification*)

Podrobneje v Santini (2007): *Characterizing Genres of Web Pages*.

- ... tipi računalniško posredovane komunikacije (RPK) (Crystal 2011: *Output*)

Poskus klasifikacije “besedilnih vrst”

- pogovarjalne BV
- predstavitvene in promocijske BV
- oglasne in komercialne BV
- poročevalske in širšepublicistične BV
- programerske BV
- akademske BV
- uradne in uradovalne BV
- literarne in polliterarne BV

Podrobneje v Mikolič, Rolih (2015) in *Slovar sodobne slovenščine: Problemi in rešitve* (2015).

Fasetna klasifikacija Susan Herring (1/3)

- Pristop, ki se je uveljavil v bibliotekarskih in informacijskih znanostih
- S. Herring je sestavila (dopolnil D. Crystal) zbirko kontrastivnih parametrov, s katerimi lahko definiramo tipe RPK
- Vse kategorije niso nujno relevantne za posamezne enote, ki jih analiziramo

Fasetna klasifikacija Susan Herring (2/3)

- Tehnični parametri
- (1) sinhronost
- (2) razdrobljenost enot
- (3) trajnost
- (4) dolžina
- (5) kanali
- (6) identiteta sodelujočih
- (7) javnost
- (8) prilagoditev
- (9) format

Fasetna klasifikacija Susan Herring (3/3)

- Družbeni parametri
- (1) struktura sodelujočih
- (2) značilnosti sodelujočih
- (3) namen
- (4) aktivnosti
- (5) tema
- (6) ton
- (7) pravila organizacije
- (8) pravila družbene primernosti
- (9) jezikovna pravila
- (10) kod

Žanrska analiza poslovnih e-sporočil

- Prikaz modificiranega načina komuniciranja in nastajanja novih oblik žanrov:
 - Primer rabe e-pošte kot transportnega sredstva za uveljavljeni žanr
 - Primer novonastajajočega žanra v komuniciranju preko e-pošte – veriga sporočil, ki predstavlja dalj časa trajajočo komunikacijo

Podrobneje v Nidorfer Šiškovič (2009): *Žanrski pristop k analizi poslovnih e-sporočil*.

Nekaj primerov (1/5)

Trenutno najraje berete



Ameriška osuplost po britanskem referendumu

50



Denacionalizacija na Hrvaškem: Zadnji »ne« v čakalni vrsti za ...

24



PRIHRANI za dopust! Preveri poletno znižanje prodajnih uspešnic na mimovrste=)

Široka ponudba primerna za vsak žep – vse kar potrebuješ za nepozabne poletne počitnice.



ZDA-Rusija: Zbledela magičnost Magnitskega

23



Nemčija išče poti prihodnjega razvoja EU

22

storitev omogoča L^N



Otrok padel v občestni jašek z meteorno vodo

V soboto ob 18.54 je v kraju Gradišče pri Vojniku otrok padel v občestni jašek v katerem ga je zalivala meteorna voda. Otroka so rešili gasilci,...



Raziskave so pokazale: otroci v Evropi si ne znajo pravilno čistiti zob

15 % otrok že pred dopolnjenim sedmim letom starosti dobi prvo zalivko, kar 9 % otrok pa v tem času že izpujijo zob. Gre za resne zobozdravstvene...

Nekaj primerov (2/5)

Potovanja

Sreda, 15. 6. 2016, 12:43, ovišeno: 3 dni, 18 ur

0.525

Potovalni priročnik: kako spakirati malo, a hkrati vse

Pripravljanje kovčka za daljše potovanje je lahko stresno, če se tega lotite nepravilno in v zadnjem hipu. Sledite našim nasvetom, da bodo počitnice minile brez skrbi in grenkega priokusa.



Počitnice vas niso pravi čas za stresno situacijo in hiro, zato se ne boj, da jih prej umetno vse pripravite.

[Izberite okna Priba](#)

[www.priba-okna.si](#)

Z uporabo pravega stila in klasične elegancije okrasite vsako hišo.

#1 Pripravite seznam

Potrebujete prazen list papirja, svinčnik in nekaj preudarnosti. Preverite, kakšno bo vreme v kraju, kamor potujete, dopišite, s kom potujete in kaj na počitnicah nameravate početi. Glede na to pripravite seznam in ne pozabite: manj spakirate, bolje bo. Seznam naj bo preprost, pomembno je, da ne pozabite na osnove:

#6 Zavarovanje za tujino



Zavarovanje se ti ni zagotovo pravi odločitev ali ne.

Za počitnice brez skrbi ne pozabite na zavarovanje. Seveda pred odhodom na pot razmisli na to, da bo vse v redu, a veliko bolje se boste počutili, če boste vedeli, da bo tudi v primeru kakšne nesreče poskrbljeno za vas. Zato pravočasno sklenite tudi zavarovanje za tujino.

Zavarovanje za tujino vam je namenjeno vsem, ki se odpravljajo v tujino, pa naj bo to izlet za en dan, počitnice, daljše potovanje, službeno pot ali študij.

Zakaj bi sklenili zavarovanje za tujino? Ker mudi več!

S sklenjenim zavarovanjem za tujino vi zagotovite:

- 24-urno existenco v slovenskem jeziku,
- informacije o najbližji zdravstveni ustanovi,
- nudijo najprej medicinske pomoči tudi v zdravstvenih ustanovah, ki niso v javni zdravstveni mreži običajne države,
- najprej zdravstvene stroške do bolnišnice ali klinike ter najprej zdravstvene,
- najprej zdravstvene stroške,
- zdravstvene stroške posledic kumulativnih bolezni,
- skrbijo za informiranje posameznikov oseb,
- preveza v domovino,
- obvezne stroške,
- v sklopu dodatkov za neugodno zavarovanje pa si zavarovanje nadgradite tudi z neugodnim zavarovanjem za čas trajanja potovanja.

Nekaj primerov (3/5)

ZRC SAZU @ZrcSazu · 45m

Naj ti prišepnem, da si ... pametna.
Naj ti rečem, da si ... lep.
Povej drugače. #SSSJ



bab.la @babla 2m

#Danish is easy! :-)



Nekaj primerov (4/5)

Družba za avtoceste @DARS_SI - 13s
DARS KONEC dogodka: (A2 Obrežje - Ljubljana: Nesreča) #dars #promet DARS.si /?tweet=-213657

Alpaka Retweeted
Frane @dratpina - 5m
Morda kdo ve, če bi se dalo kje sposoditi / rentat teleskop? RT zelo zazeljen, hvala!

Nelly Fox @gromska_strela - 7m
Ko sponzoruje brez joba govorijo o "no stress", stfu bitch, resno. Ne bom te spoštovala, ever.

ZF @ovtva - 5m
@gromska_strela enako vzdrževane "gospodije" s polnimi ritmi in armado pomočnic.

Nelly Fox @gromska_strela - 43s
@ovtva a nisva sam fauš?

View conversation

Nekaj primerov (5/5)

Miselne vijuge 39

Iztrebani relevanci
Za njegovo izkušnjo "Nadpisane" 2012 vseh pravih izkušnjo. Med pravih izkušnjo se jih izkušnjo...

S šalo na stran
Franci Kari

na twitterju so povedali
@Alisa: Včasih polni za v njegovo obdelavo, da izgleda, kot da gre na morje. Potem kupim hiš alkoholov in vladarji. Ker nam gre vrt.

@Radio Če v teli dveh vidih jehovica 4. Odločno, verjetno gre na morje.

@Tukot Brez: "Če bi daga gledal v mojo punco, lahko videl kako je drugi odprta" - Dnevnik 318 pr. n. št.

@Boris Formalič Če sem pravi izkušnjo, Slovenija je 12 let ni izgubila niti ene tekme na evropski prvenstvu v nogometu.

@Radivoj Aliberti Sodimo od 10 Anglije je sodilo, da bodo uspešni v prvenstvu na Euro2016. Ostali so pa le niso spili divji piva.

@Alisa Gantar Zagreb barobar na 4. etazi.

OD SREDE PRI NAŠEM PRODAJALCU ČASOPISOV.
Naročila: 080 2007 ali naročila@bivevniki.si

NELJESKI

HEY EUROPE!
YOU LOOK GREAT.
HAVE YOU
LOST A FEW
POUNDS?

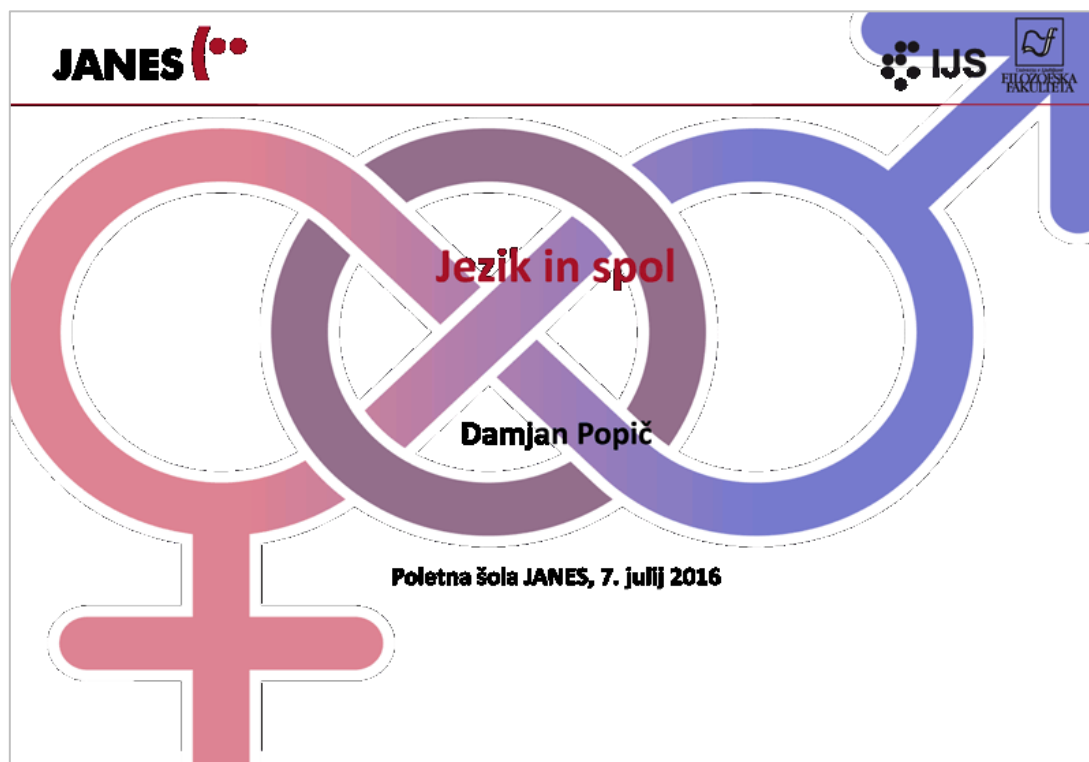
Literatura:

- Crystal, David (2011): *Internet Linguistics*. London in New York: Routledge.
- Gatto, Maristella (2009): *From Body to Web. An Introduction to the Web as Corpus*. Bari: Editori Laterza.
- Herring, Susan (2007): A Faceted Classification Scheme for Computer-Mediated Discourse. *Language@Internet*, 1/2007. Spletni vir: <<http://www.languageatinternet.org/articles/2007/761>>.
- Logar Berginc, Nataša, Ljubešić, Nikola (2013): Gigafida in slWaC: Tematska primerjava. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, št. 1. 78–110. Spletni vir: <http://www.trojina.org/slovenscina2.0/flipping_book/2013/01/05-LogarLjubescic/files/assets/basic-html/#1>.
- Michelizza, Mija (2015): *Spletna besedila in jezik na spletu: Primer blogov in Wikipedije v slovenščini*. Ljubljana: Založba ZRC.
- Mikolič, Vesna, Rolih, Maša (2015): Besedilna zvrstnost v novih medijih kot slovarska vsebina. *Slovnica in slovar – aktualni jezikovni opis. Obdobja 34* (ur. M. Smolej). Ljubljana: Univerza v Ljubljani, Filozofska fakulteta. 511–518.
- Nidorfer Šiškovič, Mojca (2009): Žanrski pristop k analizi e-sporočil. *Infrastruktura slovenščine in slovenistike. Obdobja 28* (ur. M. Stabej). Ljubljana: Univerza v Ljubljani, Filozofska fakulteta. 271–277.
- Santini, Marina (2007): Characterizing Genres of Web Pages: Genre Hybridism and Individualization. *Proceedings of the 40th Hawaii International Conference on System Sciences – 2007*. Spletni vir: <http://www.nltg.brighton.ac.uk/home/Marina.Santini/HICSS_07.pdf>.

Damjan Popič

8 JEZIK IN SPOL

8.1 Jezik in spol – izročki



(vir: Skok et al. 2016)

JANES

IJS

FILZOVSKA
FAKULTETA

Moški govor – ženski govor

ženska.si

Seks & Partnerstvo » Partnerstvo

Kako za vruga razumeti moškega?!?

Intimate medicine marec 6, 2014 / 2 leta ago

Ta članek morajo nujno prebrati vse ženske, ki se jim še vedno
zdi, da so moški z drugega planeta!

KOMENTIRAJ

f SHARE

SHARE

g+ SHARE

NOVICE

Vpiši svoj email naslov:

(vir: Korbar in
Milovanović 2016)

JANES

IJS

FILZOVSKA
FAKULTETA

Moški govor – ženski govor

moški.si

Sodoben moški » Partnerski odnos

Kako razumeti žensko med prepirom?

Melita Kuhar marec 6, 2016 / 1 mesec ago

Prispevek strokovnjaka

Odgovor: nikakor. Hecam se. Dejstvo pa je, da se moški in ženske
prepiramo drugače...

KOMENTIRAJ

f SHARE / 14

SHARE

g+ SHARE

NOVICE

Email:

(vir: Korbar in
Milovanović 2016)

JANES



Politična (ne)korektnost v akciji
(vir: Skok et al. 2016)

Janez Janša @JansaSDS
Na neki FB strani javne hiše ponujajo poceni usluge odsluženih prostitutk Evgenije C in Mojce PŠ. Eno za 30€, drugo za 35€. #ZvodnikMilan.



JANES

*„Ajde, miška mala, gremo!
To mi deli, to mi deli!“*



(vir: Skok et al. 2016)

*»Poenostavite za ženske,
lepo prosim.«*



(vir: Mladina; Skok
et al. 2016)

*"Kako rad imam te
novinarke,
ki tako zlahka
pokleknejo."*



Bratislav Gašić

Preklinjanje

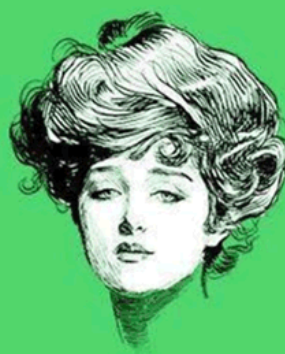
- Preklinjati kot star dedec
- Preklinjati kot star pijanec
- Preklinjati kot mornar + star mornar, pijan mornar
- Preklinjati kot kočijaž (furman) + star, pijan
- Preklinjati kot star knap

- Preklinjati kot Turek
- Preklinjati kot Madžar
(vir: Gigafida)

*vrnil se je pijan in preklinjajoč;
preklinjajoči moški (SSKJ)*

(vir: Gigafida, Korbar in
Milovanović 2016)

I do not spew
profanities.
I enunciate them
clearly, like a
fucking lady.



your eCards
someecards.com

Smerokaz

- Moški in ženski jezik
- Jezik o moških in ženskah
- Spolno nevtrarno izražanje

Petra je najboljši kemik na svetu.

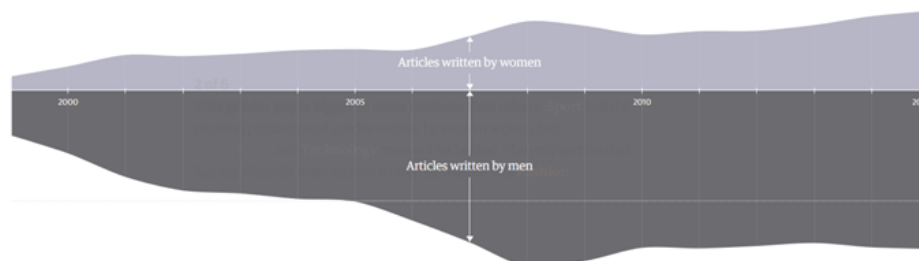
vs.

Petra je najboljša kemičarka na svetu.

Nevtralno?

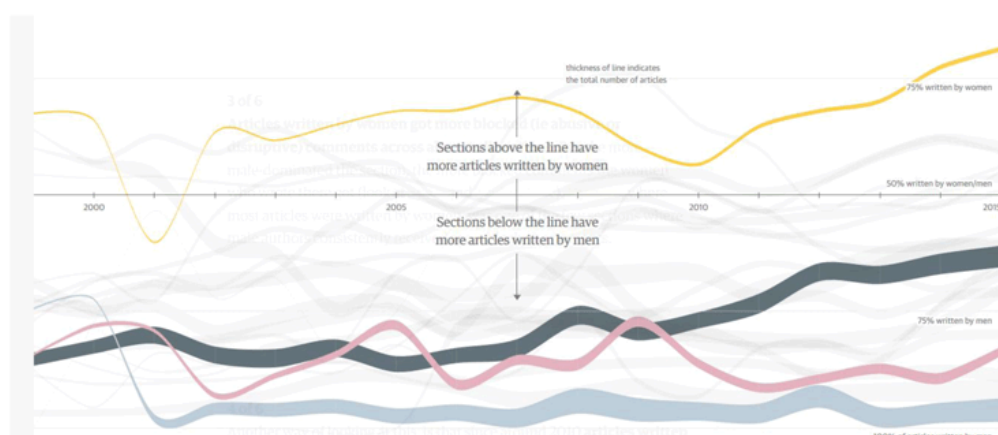
- Kot pravi Žigon, ...
- Kot pravi Žigonova, ...
- Kot pravi T. Žigon, ...
- ...

Z jezikom nad spol ...



Vir: The Guardian

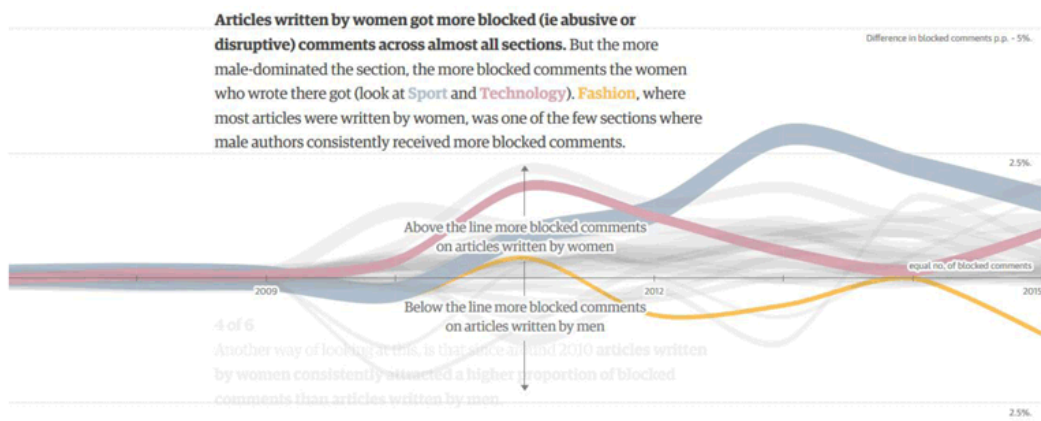
Z jezikom nad spol ...



Vir: The Guardian

Z jezikom nad spol ...

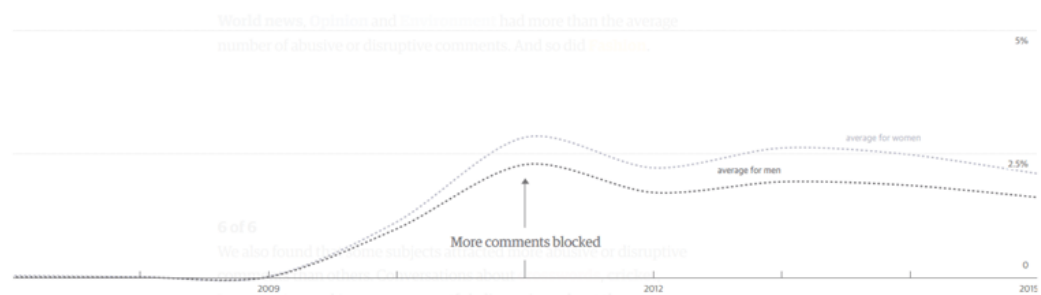
Articles written by women got more blocked (ie abusive or disruptive) comments across almost all sections. But the more male-dominated the section, the more blocked comments the women who wrote there got (look at **Sport** and **Technology**). **Fashion**, where most articles were written by women, was one of the few sections where male authors consistently received more blocked comments.



Vir: The Guardian

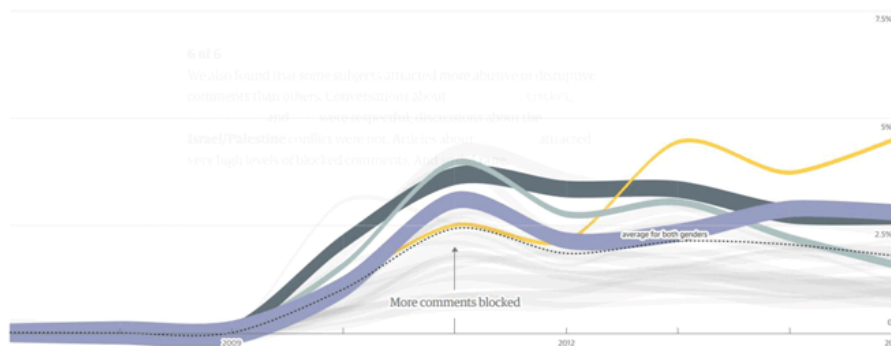
Z jezikom nad spol ...

World news, Politics and Environment had more than the average number of abusive or disruptive comments. And so did **Business**.



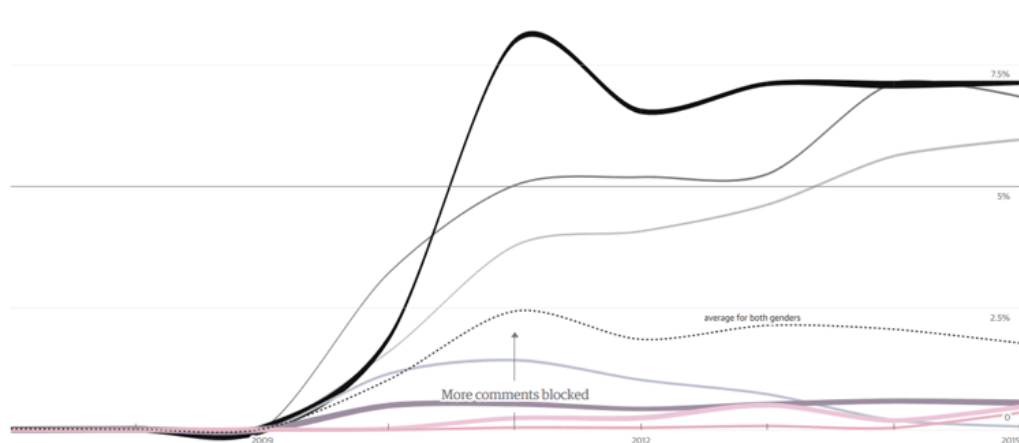
Vir: The Guardian

Z jezikom nad spol ...



Vir: The Guardian

Z jezikom nad spol ...



Vir: The Guardian

On išče njo

- 60 letni moški, iz okolice MB, želim spoznati urejeno žensko za resno zvezo in skupno življenje. Leta niso važna. Bodi resna, poštena in ljubeča.
- 53 letni moški, iz Prekmurja, zaposlen, urejen, nekadilec in nealkoholik, išče prijateljico za resno zvezo.
- Sem Robert iz Celja, star 39 let, rad bi spoznal punco za skupno zvezo, sem pa tudi na fb pod Robert Kreber, če si resna, se mi lahko javiš.
- Fredi, 41 letni, simpatičen moški iz okol.MS, 175/75, modrih oči, športno aktiven, kolesarjenje, plavanje, pohodi v hribe, obožujem morje in vse kar je povezano z njim
- Pošten, dobrasrčen, simpatičen, razumevajoč, samostojen moški, nealkoholik in nekadilec, želim spoznati žensko nekadilko, lahko tudi tujko, od 40 do 45 let

(vir: Čermelj et al. 2016)

Ona išče njega

- Maribor, 21-letna svetlolaska, dolgih las, lepega nasmeha, katera se rada poljublja, rada se carta, rada se igra in veliko več še zna, išče moškega.
- 57 letna, iz okolice Ptuja, ljubiteljica plesa in gobarjenja, išče resnega moškega, svojih let.
- 31 letno preprosto dekle. Si želiš družbe, druženja? Morda se ujameva. Sem močnejše postave iz okolice Maribora. Iščem fanta za skupno preživljanje prostega časa in raznovrstna razvajanja
- 27, IVANKA ; Kar pokam od energije :) Sem izobražena in zaposlena. Obožujem potovanja in šport, v ljubezni razočarana
- 23, SARA ; Ne spadam ravno med lepotičke, sem pa dobrega srca, če ti to kaj pomeni. Upam, da končno spoznam nekoga ki ne gleda samo na zunanost

(vir: Čermelj et al. 2016)



HackerFactor
Gender Guesser

<http://www.hackerfactor.com/GenderGuesser.php>

M ali Ž?

Do kod torej še lahko seže zabitost, preden jo bomo prepoznali kot družbeno škodljivo? Do kdaj se nam bo zdelo normalno, da moramo po diktatu večine vsi po božje častiti gasilce; se protokolarno klanjati predsedniku, ki kaže mišice na koledarjih; poslušati govejo muziko, ki jo Mario opravičuje z ratingom; opuščati dvojino in rodilnik, ker je lažje? Do kdaj?

M ali Ž?

Danes pa uvodna poglavja v gerontologijo. Jekleni Karl Erjavec se je s svojim, danes že ponarodelim pojmovanjem humorja lotil koalicijskih partneric.

Čim se Karl Erjavec začne šaliti iz koalicijskih partneric, celo politično apatični državljani sklepa, da so volitve blizu. Menda državna volilna komisija svoje dejavnosti intenzivira takoj, ko prične Erjavec žaltave razdirati na račun koalicijskih partnerjev. Pričnejo čistiti volilne skrinjice in podobno. Kakorkoli, Karl se je tokrat lotil starosti svojih političnih partnerjev. Za stranko SMC, ustanovljeno komaj pred dvema letoma, je ugotovil, da je »dveletni otrok« in kako je hecno, da tako mlad otrok vodi državo.

M ali Ž?

V zadnjem času se v medijih zelo pogosto pojavlja gospa Vesna Vuk Godina, socialna in kulturna antropologinja. Pojavlja se predvsem v kontekstu družinskega zakonika. Tragično je, da jo zaradi dejstva, da družinskega zakonika ne zagovarja, mediji, ki DZak niso naklonjeni, nekritično povzemajo kot predstavnico strokovne javnosti. (Še bolj me moti to, da ni proti DZak zaradi zares zgrešenih vsebin, temveč zaradi tega, ker DZak prepoveduje fizično kaznovanje otrok! Kako lahko kdorkoli trdi, da se mu gre za otroke, če podpira fizično nasilje?!)

M ali Ž?

Pretekli teden je dal nekaj domislekov, za katere je mogoče z gotovostjo trditi, da bodo postali del slovenske folklore. V širšem smislu. Folklor bo najprej postal Cerarjev domislek o tem, da je Slovenec. Ki ne jamra. Ampak išče rešitve. Ta je že postal tržno blago. Seveda v posrečenih predelanih različicah. Je že natisnjen na majice. Ki bodo šle za med, to sploh ni dvoma. Gre za eno od značilnosti slovenske politične kulture: da neidentificirani posamezniki z inteligentnim humorjem posegajo v politiko in politično življenje tako, da izpostavijo bistvo političnih dogodkov.

M ali Ž?

Prevca in Kopitarja ne boš videl po diskačih ob ta najhujših manekenkah pa dret se čist nažgana: "Midva sva carja! Ne može nam niko ništa!" Ker sta ostala, kar sta. Tko k Ronaldo. In ziher se vsi trije zavedajo, kaj so. Da niso kr eni. Da so jebal ježa, sami samcati, nobena mami ni namest njih trenirala. Zato so dans pač tam, kjer so.

Ker ljudje nismo vsi isti. In zakaj se jaz zdej kle s frendom pogovarjam, kok se spodobi objavljat in se hvalit? Ker skromnost je vrlina, pa še na živce greš ljudem, če je nimaš. No, vrline si pa jaz nekam vtaknem.

M ali Ž?

Ležala je vest človeštva, prasica lena zavaljena, v svojem brlogu in se še malo ni zmenila, kadar je kdo potrkal na njena vrata. Tudi nešteti voditelji humanitarnih organizacij, društev, gibanj, forumov, milijoni demonstrantov z vsega sveta so zaman kričali o krivicah, mučenju, klanju nedolžnih, umiranju sestradanih in bolnih - ona se je samo obrnila na drugo stran in spet zadrnjohala. A glej, nekega dne je nekdo potrkal na njena vrata na čisto nov način, tok tok, iz špranje pod vrati pa je začela pritekati blago dišeča lužica čiste pozitivne energije. Odprla je eno oko in vprašala, kdo je pred vrati. Dobri Janez z Zaplane, se je glasil odgovor.

M ali Ž?

Draga /.../
pravkar sem prebral/a tvoje vabilo in /.../ "razmišljanje" o palatalizaciji in sem ŠOKIRAN/A! Nič nimam proti sinhroniji, nič proti novim pristopom, ampak tole pa iziničuje vse, kar so k slovenistiki (in slavistiki) prispevali jezikoslovci od Miklošiča do Ramovša in naprej. Izkazuje popolno nepoznavanje pojma palatalizacija, kaj šele njenih zakonitosti. Tako igračkanje z glasovi je gola statistika in ni vredno akademske sfere, oprosti. Meni ni prav nič zanimivo, ampak žalostno, neskončno žalostno. Povezuje stvari, ki nimajo vzročne povezave oz. jih razlaga na popolnoma nerelevanten, da ne rečem neumen način. Zakaj ni žažica - halo? kdaj pa je bila tabeseda sprejeta v slovenščino oz. kdaj se je zgodila palatalizacija?! Prosim, ne poneumljaj naših študentov, naj to delajo z Američani.
Lp,

M ali Ž?

Kolegica /.../, Vi se v svoji 'dobronamernosti' tega najbrž niti ne zavedate —, ampak Vaše pismo je 'condescending'. Sem pač malo prestar/a, pa tudi naredil/a sem že kaj, da bi me hvalile — kako so 'z veseljem prebrale' — in me potem za nagrado povabile gospodične, ki morda ne bodo nikoli ustvarile kaj izvirnega. Tudi to, da ste slučajno na Oxfordu danes nikogar več ne impresionira. Danes pač že vsi vemo, kako se kandidatke regrutirajo [sic!]. Če bi me povabil kakšen resen profesor, to bi pa že bilo nekaj drugega.

M ali Ž?

Ni vse slabo. Tudi v politiki se kje začuti upanje. Res, koliko navdušencev nad idejami senatorja Sandersa o pravičnejši Ameriki, da ne omenjam Frančiška in njegovih somišljenikov. Pa še kaj bi se našlo.

Vendar tam, kjer se odloča, povezano z velikimi posledicami, ni dobro. Slábo, ki zadobiva ogromne razsežnosti, je delo posameznikov ali skupin, katerih učinek na politično-ekonomsko stanje planeta je bistven. K tem skupinam in posameznikom bi štel/a tudi brezimne subjekte, ki uživajo v nasilju, ksenofobiji, sovraštvu, egoizmu ali totalni brezobzirnosti do svojega bližnjega in daljnega ter volijo ljudi s slabim psihološkim profilom.

M ali Ž?

Brexit je zmaga populizma in praznih puhlic, kakršna se verjetno obeta še kje po pretreseni Evropski uniji. Nigel Farage je bil včasih le človek z idejo. Z gorečimi nastopi in vnetljivo retoriko (z)rušiti sistem od znotraj.

Njegova ideja je iz skrajnosti vstopila v zmernost, ko je skrajna retorika o "lenih Grkih" in nevarnih migrantih (ali za delovna mesta ali za varnost) postala sredinska. Medtem ko so se standardi javnega komuniciranja, padanja etičnih in socialnih meril prek grške krize, migrantskih vprašanj in terorizma nižali, je evropski poslanec Farage postajal "glas razuma". Številnim je zelo všeč udrihanje po Bruslju, pa četudi je vezano na mite in izmišljotine. Bil je "šala", popestritev. Zanimivo, mar ne, da številni izmed teh, ki se trudijo razbiti Evropsko unijo, dobro živijo na račun evropskih davkoplačevalcev, ki jim omogočajo poslansko plačo ter dnevnice za bivanje v Bruslju in Strasbourgu. Marine Le Pen vključena.

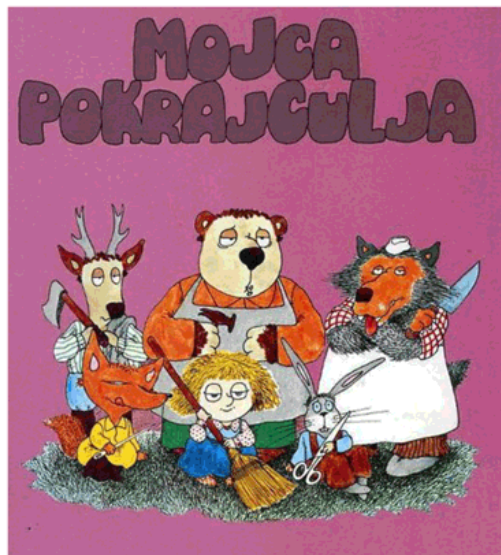
Podoba spola – nekoč in danes



(vir: Reven in Cigan 2016)



Podoba spola – nekoč in danes



(vir: Reven in Cigan 2016)

Pogled nazaj ...

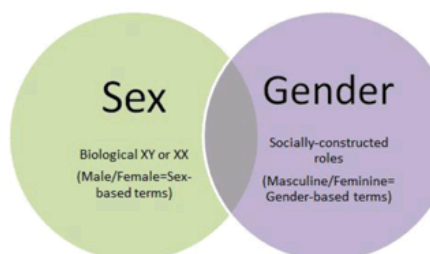


Ali je jezik le odsev sveta okoli nas?

- Povezava med predmetom in njegovim imenom
- Sapir-Whorfova hipoteza: Nismo pasivni posnemovalci sveta okoli nas, skozi jezik vanj projiciramo svoje ideje
- Pomemben vpliv kulture

Spol – biološka in družbena kategorija

- Moški in ženski spol: biološke različnosti; kulturni in družbeni pomen
- Različna pričakovanja o vedenju, vlogi in dejavnosti vsakega spola
- Spolna dihotomija
- *Sex vs gender*



Spol – družbena kategorija

- **Spolna identiteta:** zavedanje lastnega spola in odzivanje nanj
- **Spolna vloga:** pričakovanje o vedenju glede na biološki spol
- **Spolni značaj:** lastnosti, ki naj bi bile predpisane določenemu spolu

(vir: Priimek 2016)

Teorije o različnosti spolnih vlog

1. „Moško“ in „žensko“ vedenje je genetsko določeno.
2. Človeško vedenje je kulturno določeno in naučeno.



JANES

IJS

SSKJ

oseba ženskega spola, navadno doroala

človek moškega spola, navadno doroael

Lastnostni pridevniki

ženska
majhna
mlada
modrooka
temnolasa
vitka
mirna
prijazna
noseča
poročena

moški
bradat
lep
mlad

Zgledi rabe
gospodinja pere in lika (iztočnica *likati*)
negovala ga je v boleznì (iztočnica *negovati*)
njegova žena je slab šofer (iztočnica *šofer*)

JANES

IJS

nimfománka -e ž (â) ženska z bolezensko povečano spolno slo: bila je prava nimfomanka; nenasitna nimfomanka / pevec je v njej prebudil nimfomanko

andromanija -e ž (i) bolezensko povečana spolna sla žensk; nimfomanija: trpi za andromanijo

(vir: Skok et al. 2016)

Kaj počnejo fantje in kaj dekleta

SSKJ

Fant se je:

izučil
napravil
naredil
priučil
razvil v moža
streljal in začel študirat
z leti zbrihtal
razjaril
razšopiril

Dekle se je:

dušilo v solzah
kot mačkica privila k fantu
nosilo mimo gruč fantov
že zgodaj prostituiralo
lepo razraslo
skurbalo
sprostituiralo
teplo z jokom
brnilo pred vsiljivci
zatrapalo vanj
zlajdralo
zonegavilo s tujcem
zvlačugalo

SSKJ 2

Glagol

Število pojavitev

prostituirati (se) 575
sprostituirati (se) 5
skurbati (se) 40
zlajdrati (se) 0
zonegaviti (se) 3
zvlačugati (se) 0

(vir: Šutar et al. 2016)

O ženskah

Babe se zbirajo, dež bo.

Če baba nima metle, pa z
jezikom pometa.

Nevesta bogata je rada
rogata.

Fatalna ženska → ki zaradi
svoje lepote in nenavadnosti
odločujoče, navadno
negativno vpliva na moškega
Če lepo dekle šepeta, se
sliši dalj kot glas zvona.

Ženska je kakor aprilsko
vreme.

Dobra žena je božji dar.

Dobro ni vzeti iz mlina
praseta in iz krčme
dekleta.

Ženska in hudič

Kamor si hudič sam ne
upa, pošlje babo.

Babe se še hudič boji.

Če vrag omaga, mu
baba pomaga.

Kjer hudič ne zmore,
ženska pripomore.
Hudič babi brusi jezik.

Baba je še hudiča v
flašo stlačila.

Babe se še hudič boji v
babariji, zato jo pošlje
k lastni hudobiji.

O zakonu

Kdor hoče hvaljen biti,
mora umreti, kdor hoče
opravljan biti, se mora
oženiti.

Mož je glava družine, žena
pa krona – zlata ali trnova.

Žena drži pri hiši tri vogle
in še pri četrtem možu
pomaga.

Uboga hiša, kjer kokije
pojejo in petelin molči.

Na nebu ni dveh
sonc, v trdni hiši ne
dveh gospodarjev.
Žena pusta, mož gotovo
pijanec.

Moža je lažje dobiti kot
obdržati.
Žena se v možu vidi.

Babe, pipca in konja ne
posojaj sosedu

O moških

Boječ fant ni še
nikoli pri lepi
punci spal.

Ni že vsak, ki nosi
hlače, mož.

Biti copata

Brez moške glave
ni prave naprave.
Brez moške glave
pri hiši je težko
življenje.

Moški : ženska

Kjer so babe na konju, tam
so možje na psu.

Ljubezen se pri moškemu
začne v očeh, pri ženski v
ušesu.

Če se moškemu kaj pove,
gre pri enem ušesu noter,
pri drugem pa ven; če
ženski, gre pri obeh ušesih
noter, pri ustih pa ven.

Varuj se ženske, ki govori o
svoji kreposti, in moškega,
ki se hvali s svojo
poštenostjo.

Prilagojeno po Milovanović in Korbar
(2016)

S kakšen?	7,540	1.80
samski	117	8.76
poročen	130	8.71
odrasel	129	8.68
pravi	976	8.59
mlad	386	8.41
oborožen	71	7.91
heteroseksualen	57	7.83
poženščen	49	7.70
sanjski	56	7.51
možat	41	7.43
normalen	124	7.24
sodoben	72	7.18
zrel	40	7.12
neznani	55	7.07
povprečen	70	7.04
kastriran	31	7.04
privlačen	33	6.98
seksi	32	6.89
ženski	67	6.85
oblečen	29	6.80
sposoben	44	6.77
uspešen	67	6.76
mrtev	41	6.75
nasilen	33	6.74
zamaskiran	25	6.72

S kakšen?	12,421	1.60
samski	312	9.54
feminilen	198	8.98
moški	302	8.86
ženstven	158	8.67
mlad	484	8.56
emancipiran	133	8.42
faličen	134	8.35
sodoben	182	8.13
noseč	92	7.83
odrasel	94	7.67
pameten	147	7.53
gol	100	7.50
poročen	77	7.39
uspešen	127	7.37
sanjski	72	7.30
normalen	143	7.20
ubog	82	7.20
sposoben	81	7.20
močen	141	7.09
namenjen	101	7.08
pogumen	58	7.08
muslimanski	61	7.07
samozavesten	50	6.99
posiljen	47	6.91
izobražen	50	6.87

Kako pa je potem v resnici?

- Moški in ženski govor
- Govor o spolih
- Spolno neobčutljivo izražanje

Barve po moško in žensko

Table 8.14 Male and female authors' use of colour terms in the BNC

	Male authors			Female authors			Female/ male quota
	Rank	N	Per million	Rank	N	Per million	
black	1	7,527	247.32	2	4,815	332.30	1.34
white	2	6,713	220.57	1	5,101	352.04	1.59
red	3	4,250	139.65	3	3,061	211.25	1.51
green	4	3,927	129.03	5	2,560	176.68	1.40
blue	5	2,770	91.02	4	2,787	192.34	2.11
brown	6	2,279	74.88	6	1,828	126.16	1.68
yellow	7	1,299	42.68	8	890	61.42	1.44
pink	8	653	21.46	7	963	66.46	3.10
purple	9	339	11.14	9	335	23.12	2.08
crimson	10	118	3.88	10	159	10.97	2.83
mauve	11	44*	1.44	12	56	3.86	2.68
beige	12	38	1.25	11	92	6.35	5.08
indigo	13	25	0.82	13	40	2.76	3.37
cerise	14	9	0.30	14	5	0.35	1.17

(vir: Lundquist 2009)

Hedging

Table 8.16 Distribution of *you know* by function and by sex of speaker in a corpus of spoken New Zealand English

Function	Number of occurrences	
	Female	Male
<i>Speaker certain</i>		
Conjoint knowledge signal	12	13
Emphatic	22	10
Attributive	29	20
Subtotal	63	43
<i>Speaker uncertain</i>		
Appealing	16	18
Linguistic impression signal, lexical	17	22
Linguistic impression signal, qualifying	1	10
Linguistic impression signal, false start	8	9
Subtotal	42	59
Grand total	105	102

(vir: Lundquist 2009)

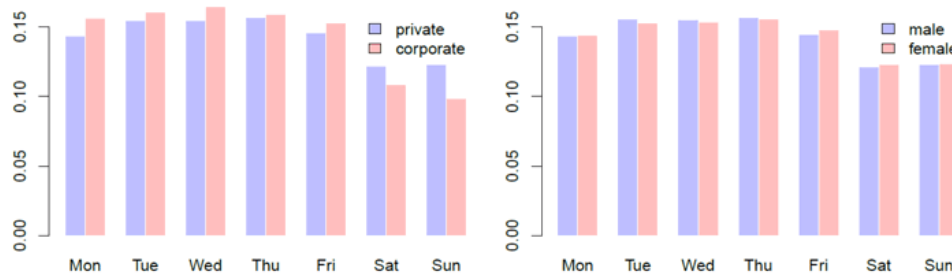


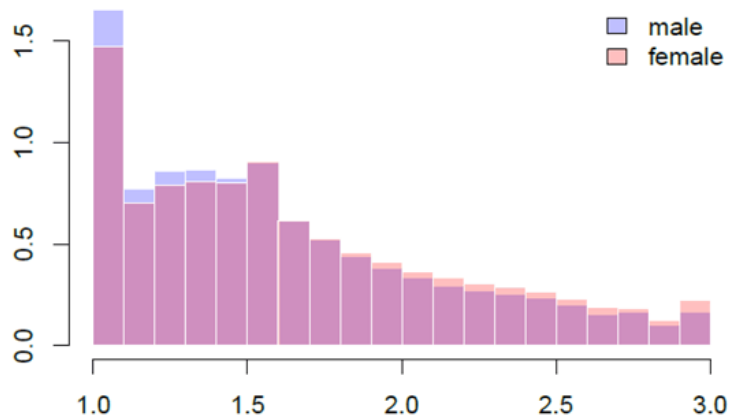
Figure 1: Probability distribution of tweets by day of week, separate by source (left) and gender (right)

(vir: Fišer et al. 2016)

	retweeted	favorited
private	8.5%	30.2%
corporate	16.3%	18.0%
male	9.4%	29.2%
female	6.8%	32.9%

(vir: Lundquist 2009)

Standardnost besedila



(vir: Fišer et al. 2016)

Sentiment

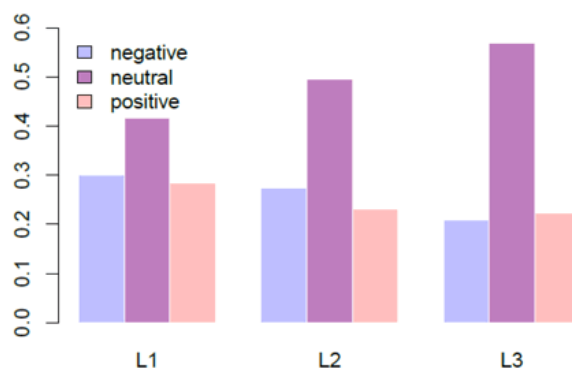


Figure 7: Distribution of the sentiment by language standardness (L1 - completely standard, L2 - slightly non-standard, L3 - very non-standard)

(vir: Fišer et al. 2016)

Standardnost po času

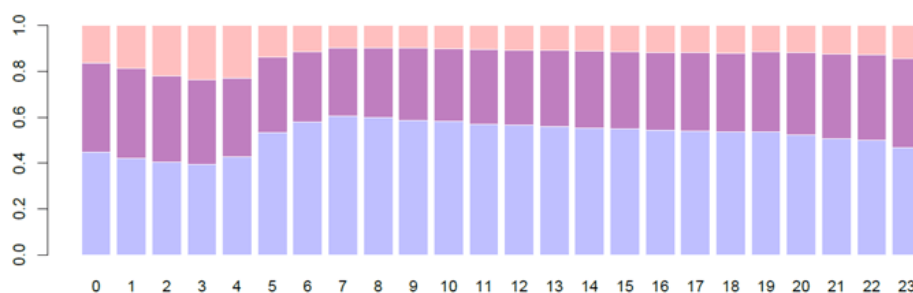


Figure 4: Standardness by hour of day, standard represented with blue, very non-standard with red

(vir: Fišer et al. 2016)

Kako in kaj govorijo spoli

- Standardnost
- *Hedging*
- Zgovornost
- Čas objavljanja
- Sentiment
- Odnosi do določenih pojavnosti
- Preklinjanje
- Vljudnost
- Barve

Govor o spolu

- Kaj (najpogosteje) počnejo moški in kaj ženske?
- Kakšni so moški in kakšne ženske?
- Kdo
 - pije,
 - vara,
 - lenari,
 - spletkari,
 - tečnari itd.

Table 8.9 The most frequent adjectives associated with *man* and *woman* in the frame *a + adjective + man/woman* in the BNC

Rank	<i>a + adjective + man</i>		<i>a + adjective + woman</i>	
1	young	873	young	361
2	good	172	married	117
3	#-year-old*	133	#-year-old*	100
4	married	107	beautiful	67
5	big	100	middle-aged	48
6	black	83	pregnant	42
7	rich	83	good	35
8	tall	77	single	32
9	local	76	real	22
10	great	66	black	21
11	dead	65	fat	21
12	white	62	tall	20
13	nice	61	white	19
14	middle-aged	60	naked	18
15	new	57	different	16

Note: * The hash sign # indicates any number, so included here are 2-year-old, 3-year-old, 49-year-old etc.

Vir: Lundquist 2009

Moški so ..., ženske so ...

Man: *big, rich, local, great, dead, nice, new*

Woman: *beautiful, pregnant, single, real, fat, naked, different*

Vir: Lundquist 2009

Kaj počnemo ...

Table 8.13 Most frequent verbs in the frames *she + verb + him* and *he + verb + her* in the BNC

Rank	she + verb + him		he + verb + her	
1	told	559	told	581
2	gave	219	gave	283
3	saw	201	kissed	162
4	watched	134	took	143
5	heard	134	asked	125
6	loved	132	loved	123
7	wanted	100	wanted	118
8	asked	93	saw	117
9	followed	92	held	100
10	found	73	watched	93
11	let	68	pulled	91
12	reminded	68	left	88
13	felt	60	led	82
14	left	60	handed	75
15	kissed	57	reminded	66

Vir: Lundquist 2009

she + verb + him: heard, followed, found, let, felt
he + verb + her: took, held, pulled, led, banded

Figure 8.2 Unique verbs in the top fifteen in the frames *she + verb + him* and *he + verb + her* in the BNC.

Vir: Lundquist 2009

Smernice za nezaznamovano rabo

- Splošno načelo je, da se pri poimenovanju oseb, poklicev itd. praviloma izogibamo rabe generičnega moškega spola
- Predlog za uporabo množinskega samostalnika: na ta način se izognemo spolno diskriminatorni samostalniški obliki v množini; ta pristop lahko uporabimo
 - v tekstu, ki je formalen (npr. delavstvo namesto delavci);
 - v tekstu, ki vključuje osebni slog (npr. študentarija namesto študenti).ga moškega spola.

(vir: Komisija za ženske v znanosti)

Smernice za nezaznamovano rabo

- Predlog za uporabo glagolnika kot samostalniške besede ali uporabo nadomestnega samostalnika, ki ni v »nezaznamovanem« (tj. generičnem) moškem spolu; predlagana raba je možna le, kadar to ustreza vsebini (npr. »S historične perspektive je arhivarstvo pomembno opravilo.« in »Arhivarjenje je s historične perspektive pomembno opravilo.« namesto »S historične perspektive opravlja arhivar pomembno opravilo.« ali »S historične perspektive opravljajo arhivarji pomembno opravilo.«)

(vir: Komisija za ženske v znanosti)

Smernice za nezaznamovano rabo

- Predlogi za rabe samostalnikov, ki označujejo osebe, v ženski in moški obliki:
 - v tekstu, ki je formalen, a hkrati vključuje zahtevo po stilu, uporabimo obe obliki v celovitem zapisu (npr. ženske in moški/moški in ženske) ali najdemo izraz, ki je množinski (npr. poslušalstvo);
 - v tekstu, ki vključuje osebni slog, lahko poleg zgoraj navedenih različic uporabimo obe spolno zaznamovani obliki v ednini (npr. profesor in profesorica, bralka in bralec);
 - v tekstu, ki je administrativne narave (vključno z zakonskimi besedili), lahko uporabimo okrajšave za drugo spolno zaznamovano obliko samostalnika (npr. ministrica/-er ali minister/-ca).

(vir: Komisija za ženske v znanosti)

- Pri rabi samostalnika človek in ostalih slovničnih pokazateljev spola, ki se nanj naslanjajo (»kdor«, »vsak«) v stavkih, s katerimi želimo tretjeosebno in s tem posplošeno opisati določeno mnenje ali situacijo, nastopi moška skladnja in s tem z vidika ozaveščanja razmerij spola absurдна situacija, ko začne ženska govoriti kot moški (npr.: »Ana pravi: 'Človek je pač prisiljen ukrepati v skladu s situacijo.'«) Temu se lahko izognemo z opuščanjem takšnega načina izražanja ali, v primeru ženske govorne, s posplošeno rabo samostalnika ženska (npr. »Ženska je pač prisiljena ukrepati v skladu s situacijo.«) Ob opisani rabi se sicer vsiljujejo dodatni (konotativni) spolno zaznamovani pomeni, a z redno aplikacijo pravila se začnejo izgubljati.

(vir: Komisija za ženske v znanosti)

Nezaznamovana sintaksa

- V primerih, ko se glagolska oblika ravna po samostalniku, lahko menjamo vrstni red spolno zaznamovanih samostalnikov tako, da je glagolska oblika enkrat ženska, drugič moška (npr. »informacije, ki so jih podale respondentke in respondenti« in »informacije, ki so jih podali respondenti in respondentke«). Ta način spolno občutljive rabe jezika ustreza slovničnemu načelu, da se glagolska oblika lahko ravna po »naravnemu« spolu, in sicer tistemu, ki je v stavku bližje glagolu.

(vir: Komisija za ženske v znanosti)

Nezaznamovana sintaksa

- V primerih, ko gre za subjekt, ki je spolno raznoliko sestavljen in se v skladu s pravili slovnice glagol spolno diskriminatorno izrazi z moško končnico (npr. »Parlamentarke Zares in parlamentarec SDS so se odločili, da ne bodo glasovali.« ali, s primerom iz vsakdanje govorice, ki vključuje še /simptomatični/ pasji subjekt: »Tina, Maja in Švrk so šli na sprehod.«), je priporočljivo običajno sintakso razgraditi, tako da se diskriminatornemu pravilu izognemo (npr. »Parlamentarke Zares so se odločile, da ne bodo glasovale, enako velja za parlamentarca SDS.« ali »Tina in Maja sta šli na sprehod, s sabo sta vzeli Švrka.«)

(vir: Komisija za ženske v znanosti)

Nezaznamovana sintaksa

- V uporabi dvojine, ki se nanaša na ženski stavčni subjekt, uporabljamo končnico glagola -i, saj se specifičnost slovenščine za izražanje dvojine ob nanašanju na dve ženski tudi pri glagolu ne bi smela izgubiti (npr. »Šli sva v kino.«).

(vir: Komisija za ženske v znanosti)

Toda ...

Tabela 1: Raba moškega slovničnega spola v generični obliki.

		Uporaba moškega slovničnega spola v generični obliki	Drugo
Eva	63 %	19	11
Mateja	39 %	14	22
Andrej	50 %	5	5
Jasna	29 %	4	10
Nika	74 %	14	5
Maja	50 %	8	8
Bojan	68 %	15	7
Tjaša	38 %	5	8
Katarina	39 %	20	31
Janez	74 %	40	14
Oliver	77 %	10	3
Skupaj	56 %	154	119

(vir: Perger 2016)

Toda ...

Tabela 2: Raba ženskega slovničnega spola pri samonaslavljanju.

		Samonaslavljanje v moškem slovničnem spolu	Samonaslavljanje v ženskem slovničnem spolu
Eva	63 %	55	92
Mateja	78 %	25	93
Jasna	94 %	13	188
Nika	74 %	13	38
Maja	75 %	13	40
Tjaša	99 %	2	205
Katarina	87 %	17	123
Skupaj	87 %	118	758

(vir: Perger 2016)

Nevtralnost – kaj z njo?

- „Ženske“ oblike
- (Tipično moški/ženski) poklici
- Poimenovanja skozi čas

Poimenovanja skozi čas

Table 8.4 Frequency of *firemen* and *fire fighters/firefighters* in the Time Corpus:
per million words

	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
firemen	11.4	14.0	12.0	9.3	9.1	8.3	5.9	2.3	0.7
fire fighters/ firefighters	0.24	0.95	1.34	1.04	0.84	1.70	7.75	3.66	10.70

(vir: Lundquist 2009)

Poimenovanja skozi čas

Table 8.6 Frequency of *poet* and *poetess* in the Time Corpus: absolute figures

	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
poet	537	871	815	1,052	1,147	611	397	335	127
poetess	36	44	54	50	23	6	1	0	0

Table 8.7 Frequency of *actor* and *actress* in the Time Corpus: absolute figures

	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
actor	463	639	685	1,246	1,122	932	832	732	730
actress	601	697	646	1,077	1,048	726	609	461	387
Total	1,064	1,336	1,331	2,323	2,170	1,658	1,441	1,193	1,117

(vir: Lundquist 2009)

Špela Arhar Holdt

9 SKLADENJSKA ANALIZA (NE)STANDARDNE SLOVENŠČINE

9.1 Skladenjska analiza (ne)standardne slovenščine – izročki



Skladenjska analiza (ne)standardne slovenščine

Špela Arhar Holdt

Zavod za uporabno slovenistiko Trojina
Filozofska fakulteta Univerze v Ljubljani

Poletna šola Janes, 8. julij 2016



Naloga za današnje predavanje

- Spoznamo korpusni pristop k skladenjskim vprašanjem, metode in vire, ki so za slovenščino trenutno na voljo.
- Ogledamo si nekaj izbranih skladenjskih korpusnih raziskav: zakaj je bila izbrana določena tema, kako je bila zastavljena metodologija – kar je uporabno za morebitno izbiro in zasnovo magistrskih nalog.
- Skupaj razmislimo o uporabnosti skladenjskih podatkov za različne vrste nalog (jezikoslovne raziskave, razvoj jezikovnih tehnologij, prevajanje, lektoriranje ipd.)
- **Vaša naloga:** med predavanjem si zapišite eno skladenjsko idejo/problem/vprašanje, ki se vam zdi zanimiva za korpusno raziskavo. Zadnji del predavanja bo namenjen debati o zbranih temah.

Skladnja

- „Skladnja ali sintaksa je poseben del slovničnega nauka o jeziku. Uči, kako se delajo (oz. kako so narejene) pravilne povedi in njihovi deli. Povedi in njihove dele skladamo in prepoznavamo na podlagi skladenjskih vzorcev iz besed in stalnih besednih zvez (in iz delov, ki so že sami lahko povedi), pri tem pa zmeraj upoštevamo tudi položaj povedi v danem besedilu oz. v njegovem delu.“ (Toporišič 2004: 487)
- V korpusnem jezikoslovju analiziramo velike količine realnega jezika, da identificiramo in opišemo tipične (in atipične) jezikovne vzorce in značilnosti njihove rabe v danem kontekstu. Zelo poenostavljeno rečeno so skladenjske analize tiste, ki se ukvarjajo z vprašanjem, kako se besede sestavljajo v večje enote (besedne zveze, stavke, povedi, besedila), pri čemer nas zanima tako oblikovna kot pomenska raven.

Tipične skladenjske teme

- **Besedne zveze:** katere se pojavljajo v rabi, kakšne so njihove oblikovne in pomenske lastnosti ...
- **Stavčni členi:** kako se pojavljajo v različnih vrstah stavkov, v kakšna razmerja stopajo, kako se izražajo pomenske sestavine ...
- **Tipologija stavkov in povedi:** katere vrste stavkov poznamo, kako so notranje sestavljeni, kako se povezujejo v širše enote (oblikovno in pomensko) ...
- **Raziskave značilnosti upovedovanja:** raba trpnika/tvornika; različne vrste naklona; zanikanje; stopnjevanje; nanašanje; izražanje poročanega govora; besedni red ...
- **Specifike skladnje v različnih besedilnih vrstah/žanrih:** znanstvena besedila, leposlovje ipd. v primerjavi s ‚splošnim jezikom‘ / primerjava pisnega in govorjenega jezika / primerjava standardne in nestandardne skladnje ...
- ...

Označenost korpusnih podatkov

- Korpusne analize so kvantitativne in kvalitativne. Bistveni predpogoj za oboje je kvalitetna označenost podatkov.
- Označevanje poteka po ravlinah: segmentacija, tokenizacija, lematizacija, oblikoskladnja, skladnja, pomenske lastnosti (Polona bo govorila o SRL); + druge raziskovalnospecifične kategorije.
- Večina slovenskih korpusov je označenih do vključno ravni oblikoskladnje, konkordančniki so zasnovani tako, da je po teh podatkih mogoče zelo dobro iskati in jih uporabljati.
- Največji skladiščno označen korpus je korpus ssj500k, orodja za analizo obstajajo, manjka možnost sintetičnih primerjav (kvantitativni del).
- Določene skladiščne raziskave je mogoče opraviti tudi na neoznačenih podatkih (npr. distribucija členka *pa* v različnih žanrih).

Oblikoskladnja

- Vsaka beseda ima pripisano besedno vrsto in pripadajoče kategorialne lastnosti. Večina večjih korpusov (npr. Gigafida, Kres, GOS, Šolar, Lektor, Janes) je označenih po sistemu JOS: <http://nl.ijs.si/jos/>.
- Oblikoskladenske oznake so temelj za večino korpusnih analiz, zato je pomembno, da **sistem označevanja dobro poznamo** – katere oznake so na voljo, kako jih program pripisuje, kje so močne in šibke točke označevanja, kje so razlike med označevanjem in jezikoslovnimi kategorijami, na katerih sistem stoji.
- **Dobro razumevanje označevanja je temelj korpusne metodologije (in ključni del vašega metodološkega razdelka).**
- Kako torej spoznamo sistem označevanja? Pregledamo oznake sistema in specifikacije, seznanimo se s temeljno literaturo o označevanju konkretnega korpusa, preizkušamo različne iskalne pogoje, sumničavi smo do rezultatov. Včasih pomaga preveriti svoje predpostavke pri kom, ki je sodeloval pri korpusni gradnji.

mesto v kodi	atribut	vrednost	koda
0	besedna_vrsta	pridevnik	P
1	vrsta	splošni	p
		svojilni	s
		deležniški	d
2	stopnja	nedoločeno	n
		primernik	p
		presežnik	s
3	spol	moški	m
		ženski	z
		srednji	s
4	število	ednina	e
		dvojina	d
		množina	m
5	sklon	imenovalnik	i
		rodilnik	r
		dajalnik	d
		tožilnik	t
		mestnik	m
		orodnik	o
6	določnost	ne	n
		da	d

Oblikoskladnja: pridevnik po sistemu JOS

Čeprav je perunika na pogled bolj malo podobna **beli** liliji, si s to cvetlico deli kar nekaj skupnih simbolnih pomenov. Tako je razkošni **modri** cvet perunike, ki je bil priljubljen že v starem Egiptu, simbol vladarskega dostojanstva, odličnosti, zmagovalstva in slave. *Gigafida < Gea, 2001*

- **beli** = [Psnzem] - pridevnik, splošni, nedoločena stopnja, ženski spol, ednina, **mestnik**
- **modri** = [Psnmeid] - pridevnik, splošni, nedoločena stopnja, moški spol, ednina, imenovalnik, določna oblika

Primer raziskave 1

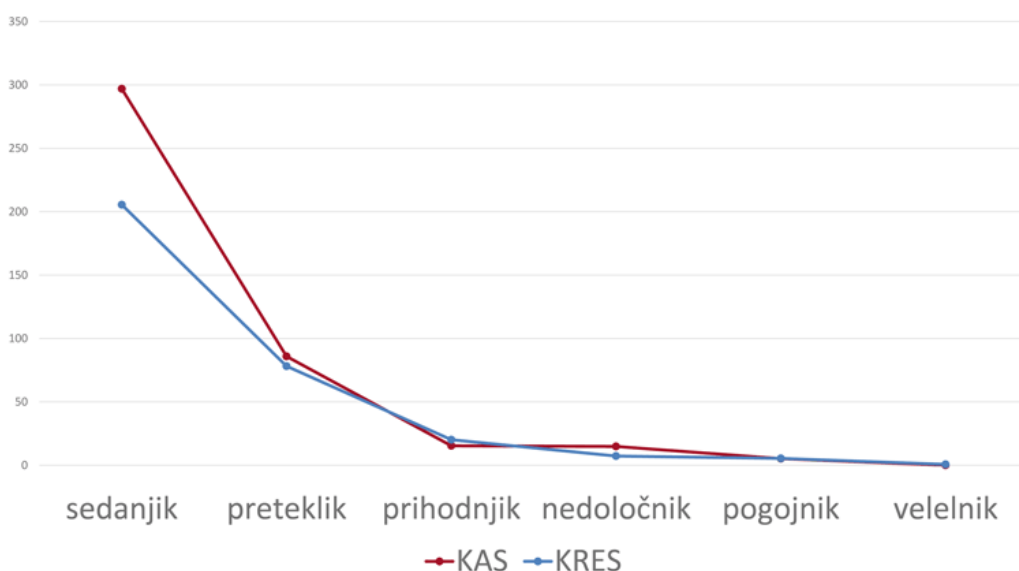
Špela Arhar Holdt, Nataša Logar in Tomaž Erjavec: Slovenska znanstvena besedila: korpusni opis – trpnik. *Toporišičeva Obdobja*, Ljubljana, 10.–12. november 2016. V pripravi.

- **Ideja:** Ko se v literaturi omenjajo značilnosti znanstvenega jezika, se običajno omenja raba trpnika. Raba trpnika se pogosto tudi vrednoti kot pretirana in slogovno slaba. Vendar te predpostavke niso empirično potrjene.
- **Metoda:** Izdelamo seznam trpniških skladenjskih vzorcev, ki jih je mogoče pridobiti z uporabo oblikoskladenjskih oznak. Z uporabo konkordančnika SketchEngine primerjamo vzorce v dveh korpusih: referenčnem korpusu Kres in korpusu znanstvenih besedil Kas.
- **Rezultat:** Podatki pokažejo, da se trpniški vzorci pojavljajo v znanstvenih besedilih pogosteje samo v sedanjiku, v ostalih primerih pa ne. Tako distribucija vzorcev kot njihove tipične leksikalne zapolnitve se v različnih časih in naklonih razlikujejo. V 'splošnem jeziku' je trpnik zelo prisoten, je pa morebiti manj opazen (prim. da je nekaj *razprodano* ali *operacionalizirano*).

Primer raziskave 1

	primer iz slovnice	CQL za korpusno iskanje na atributu »tag«	korpusni primer
sedanjik	spoštovan je	"Gp-s.*" "Pdn.*"	je prikazan
preteklik	spoštovan je bil	"Pdn.*" "Gp-s.*"	namenjen je (bil)
		"Gp-s.*" "Gp-d.*" "Pdn.*"	je bil ustanovljen
		"Gp-d.*" "Gp-s.*" "Pdn.*"	bil je prepričan
prihodnjik	spoštovan bo	"Gp-p.*" "Pdn.*"	bo uporabljen
		"Pdn.*" "Gp-p.*"	predstavljen bo
nedoločnik	spoštovan biti	"Gp-n.*" "Pdn.*"	biti izpolnjen
		"Pdn.*" "Gp-n.*"	pripravljen biti
pogojnik	spoštovan bi bil	"Pdn.*" "Gp-g" "Gp-d.*"	zaželen bi bil
		"Gp-g" "Gp-d.*" "Pdn.*"	bi bil namenjen
		"Gp-d.*" "Gp-g" "Pdn.*"	bil bi pripravljen
velelnik	bodi spoštovan	"Gp-v.*" "Pdn.*"	bodi pripravljen
		"Pdn.*" "Gp-v.*"	hvaljen bodi

Primer raziskave 1



Primer raziskave 1

	Kas	Kres
najpogostejše zapolnitve	ustanovljen, sprejet, izveden, uporabljen, opravljen, vključen, ugotovljen, namenjen, narejen, določen, objavljen, pripravljen, izbran, dosežen, postavljen	prepričan, sprejet, pripravljen, namenjen, ustanovljen, objavljen, navdušen, izbran, opravljen, rojen, zgrajen, presenečen, izdan, znan, narejen
visoko na seznamu pri obeh korpusih	ustanovljen, opravljen, pripravljen, objavljen, izbran, narejen, izveden, ugotovljen, določen, izdan, zgrajen, postavljen, vključen, prepričan, zaposlen	
primeri, znatno pogostejši v enem korpusu	zajet, zabeležen, predložen, opredeljen, viden, zastavljen, dodan, omogočen, obdelan, poudarjen, dopolnjen, izločen, realiziran, osredotočen, pregledan	navdušen, ranjen, presenečen, ukraden, narisan, naperjen, poražen, posiljen, odrezen, obarvan, prepleten, dvignjen, najavljen, razbit, spočet
korpusnospecifični primeri	optimiziran, dimenzioniran, odčitan, izpodbit, reorganiziran, kategoriziran, podcenjen, operacionaliziran, poiskan, oskrbovan, zmodeliran, pozicioniran	opustel, preklan, kostumiran, otrpel, zaljubljen, izžreban, upadel, razprodan, opevan, razkačen, očitán, pridušén, negovan, deponiran, izmozgan

Skladenjsko označevanje

- „Rezultat označevanja so skladenjsko označeni korpusi oz. drevesnice, ki predpostavljena skladenjska razmerja eksplicirajo na velikem vzorcu besedil dejanske rabe in omogočajo statistični pregled vzorcev distribucije skladenjskih struktur, zato so navadno izhodišče za nadaljnjo, bolj poglobljeno jezikoslovno analizo, hkrati pa so pomembni za razvoj jezikovnih tehnologij, kar je temeljnega pomena za ohranjanje konkurenčnosti ter polnofunkcionalnosti jezika med ostalimi jeziki“ (Ledinek 2014: 18).
- Skladenjsko označevanje (ročno) vs. razčlenjevanje (strojno).
- Sistemi za skladenjsko označevanje so zelo različni, od plitkih oz. skeletnih modelov (zamejitev stavčnih enot z določitvijo jedra) do popolnega oz. globinskega označevanja, kjer označujemo funkcijskoskladenjska in pomenskoskladenjska razmerja med elementi povedi (ibid: 19).

Skladenjsko označevanje

- **Odvisnostni sistemi** temeljijo na odnosih med elementi (kjer se predideva, da je eden od delov jedrni, drugi od njega odvisen). So običajna izbira za morfološko bogate jezike s prestejšim besednih redom. Drugi glavni princip je **frazna gramatika**.
- Za slovenščino: Slovenska odvisnostna drevesnica po vzoru Prague Dependency Treebank (Erjavec in Ledinek 2006); korpus ssj500k, ki je skladijsko označen z odvisnostnim sistemom JOS-SSJ (Dobrovoljc et al. 2012) in je bil pred kratkim tudi preveden v model Universal Dependencies (Dobrovoljc et al. 2016).

Ssj500k

- Korpus obsega 11.411 skladijsko označenih povedi (Krek et al. 2015). Ob pripravi podatkov je bil razvit razčlenjevalnik, ki ga lahko preizkusimo na <http://razclenjevalnik.slovenscina.eu/>.
- Podatke ssj500k lahko pregledujemo v **označevalniku SSJ** („Brank“), ki omogoča tudi preoznačevanje (torej je uporaben tako za analize obstoječega gradiva kot za označevanje novega).
- Tudi pri skladijski ravni je za kvalitetno izvedbo raziskav treba dobro poznati sistem označevanja.
- Na korpusu ssj500k še ni bilo opravljenih veliko raziskav. Pomembna izjema je **monografija N. Ledinek (2012)**, ki prinaša študijo **slovnično-pomenskih vzorcev za 20 slovenskih glagolov**.
- Korpus, označevalnik in razčlenjevalnik so prosto dostopni za uporabo (www.slovenscina.eu).

JANES		
IJS		
Skupina povezav	Tip povezave	Kaj povezuje
Povezave prvega nivoja označujejo razmerja znotraj besednih zvez.	dol	Jedro in določilo besednih zvez.
	del	Deli zloženega povedka.
	prir	Jedra v prirednih zvezah znotraj stavka.
	vez	Besede ali ločila v vezniški vlogi.
	skup	Nepolnomenenske besede, ki imajo zelo močno tendenco po sopoljavanju.
Povezave dr drugega nivoja označujejo <u>stavčne</u> člene.	ena	Osebek stavka.
	dve	Predmet stavka.
	tri	Prislovno določilo lastnosti.
	štiri	Ostala prislovna določila.
Povezava tretjega nivoja se uporablja za povezovanje <u>vseh</u> ostalih struktur.	modra	Hierarhično najvišje pojavnice, skladijsko manj predvidljive in oddaljene strukture, vrinki, ločila.

JANES

IJS

FILZOVSKA
FAKULTETA

C:\Users\Špela\OneDrive\Documents\JANES\POZNAČEVANJE\500k.xml - Označevalnik stavkov

Izberi datoteko...

Prikaži vse stavke

150

Povezave...

Urejanje izbranega stavka

Nastavitve...

[5:19.67] Lahko si nekaj mislimo.

[5:20.68] Oblekla si je bila karbojka in pleten pullover, bane si je spetela v rep, ki ji je ljubko štrlel na zadnji; bila je drugačna, kot sem je bil vajen, zdela se mi je nekako veliko bolj vsakdanja in za to sem ji bil hvaležen.

[5:20.69] Pripoveda je brez pozdrava, kot ta se bala nadija se velikokrat in se z ustnicami nepriznava dostavila moji.

[5:20.70] Želela je najverjetneje, da bi bil to le blag dotik ob srečanju, toda moje diari so se oklenile njenih gladkih lic in je niso več spustile.

[5:20.71] Čutil sem njene polne nenavikane ustnice, ki so polagoma dojemale, kako zelo si jih želim.

[5:20.72] Vohal sem njen topli, najverjetneje še tudi od hitre hoje razgret dihanje, in se prepuščal občutku lepega.

[5:20.73] Spoznaval sem, morda prvič, da o vsej stvari želim in hotenju, ki ga je zganiša v meni, ni moglo povedati z besedami, zato nisem nič poslušal.

[5:21.74] Na voljo naj bi mi bili milijardi tolarjev (10 milijonov DEM) več kot pred petimi leti.

[5:22.75] V nedeljo, 23. maja, bo v kinotiah v Ljubljani nastopil Julio Iglesias - Cene vstopnic za koncert se gibljejo od 19 do 29 tisoč tolarjev, kar za slovenski žep nikakor ni malo - Zato smo mimoidoče Ljubljance popravljali, z

[5:23.76] V osnovni šoli v Bistrici pri Trbiču bodo v soboto, 8., in nedeljo, 9. maja, pripravili 27. Mednarodne dneve mineralov, fosilov in okolja.

[5:23.77] Predstavilo se bo sto razstavljencev iz petnajstih držav, ki bodo razstavljali in prodajali naravne in obdelane lepote skritega podzemlja.

[5:23.78] Razstavo bo v petek, 7. maja, ob 19. uri odprl Ivo Bizjak, varuh človekovih pravic.

[5:24.79] Tako je bil včeraj obiskani profesionalni teniški turnir ATP v Donžalihu, ko je nastopil Mariborčan Iztok Božič.

[5:25.80] Kot smo že poročali, je bila afera Holmec eden od razlogov, da so v parlamentu zamenjali notranjega ministra Mirka Bandja (LDS).

[5:25.81] Na predlog premiera Janeza Dmroviča so za naslednika na vrhu notranjega ministrstva imenovali Boruta Šušjeta.

172

JANES IJS FILZOVSKA
FAKULTETA

C:\Users\Spela\OneDrive\Documents\JANES\POZNAČEVANJE\ug500k.xml - Označevalnik stavkov

Iskanje povezav

MSD G* D* S*

Lema

Oblika

Pogoj za povezavo

Brš obstaja povezava od A do C tpa dve

Brš obstaja povezava od C do B tpa dol

Dodaj

Stavek	A	B	C	A → C	C → B
ssj1.1.4	Prilo	do	preizkusa	dve	dol
ssj2.3.13	prilo	do	rešitve	dve	dol
ssj3.4.15	trpo	za	alergijam	dve	dol
ssj3.6.19	govorimo	o	motnj	dve	dol
ssj3.6.19	umerjavajo	z	obdobj	dve	dol
ssj3.7.24	Opozorila	na	komunikacijo	dve	dol
ssj3.7.24	gre	za	zdravljenje	dve	dol
ssj3.8.30	upravlja	z	aparatom	dve	dol
ssj3.9.32	zaleže	Pri	bolečinah	dve	dol
ssj3.14.47	vplivala	na	okolje	dve	dol
ssj4.15.48	postregla	z	naključj	dve	dol
ssj4.15.50	sovpadal	z	števom	dve	dol
ssj4.15.51	izlital	z	načrs	dve	dol
ssj4.15.54	prikazljati	za	prenos	dve	dol
ssj5.19.64	gre	za	diskriminacijo	dve	dol
ssj5.20.69	dotaknila	z	utricami	dve	dol
ssj5.20.73	povedati	o	žetj	dve	dol
ssj5.25.81	menovali	za	naslednika	dve	dol
ssj5.26.83	Gre	za	vozilo	dve	dol
ssj5.26.84	primerja	z	avtomobilom	dve	dol

Dvojni klik = prikaz izbrane povezave v glavnem oknu.

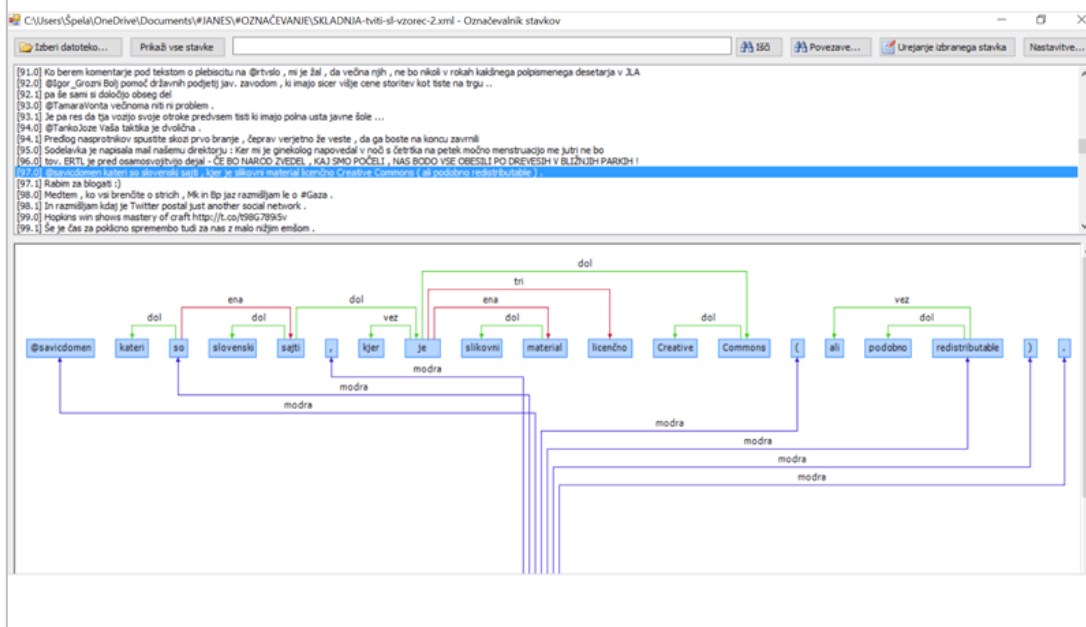
Raziskave nestandardne skladnje

- Nestandardna skladnja se v preteklosti pogosto omenja na liniji pravnarobe oz. s stališča vrednotenja sloga. Primer iz Nove slovenske skladnje (Toporišič 1982):

Ob veznikih *ne samo/le — ampak/temveč/marveč tudi* prestava v rodilnik ni obvezna.³⁰ *Tega/To nisi pravil le meni, ampak vsake-mu, ki te je hotel poslušati.* Izjemoma se tako govori še pri nič: *Nič/Ničesar ji ne manjka pri meni, Nič/Ničesar mi ne pripoveduj!, Nimam kaj/česa obžalovati.* — Neknjižno je, če postavljamo v rodilnik osebek drugih zanikanih povedkov, npr. *Ničesar (prav nič) se ni zgodilo.* Tudi pri trpniku je prestava osebk v rodilnik nedopustna: *Takih stvari se ne govori na glas* (namesto *Take stvari se ne govorijo na glas*), *Tega se mi ni naročilo nam. To se mi ni naročilo.*³¹ Napačno je tudi *Tega mi ni bilo naročeno.*³²

Raziskave nestandardne skladnje

- Prve analize, v katerih se opazuje značilnosti skladnje v računalniško posredovani komunikaciji na večji količini gradiva v Michelizza (2015).
- Širših korpusnih študij še ni.
- S sistemom JOS-SSJ smo označili 200 tvitov, na osnovi katerih bodo dopolnjenje smernice za označevanje.
- S tem delom bi bilo mogoče nadaljevati (označiti več primerov, morda tudi druge besedilne vrste, natrenirati razčlenjevalnik). Če koga zanima, naj se javi!
- Problem raziskovanja nestandardne skladnje: **manjka korpusno osnovani slovnici opis standardnega jezika**. Če ne vemo, kaj je v jeziku tipično standardno, tudi ne moremo empirično opredeljevati, kaj je nestandardno.



Primer – zaznamovan besedni red?

L1T1 (3/48)

- Nekateri zvesti podporniki.. Še vam ni jasno, da če bi želeli videti vsak tweet kandidatov, bi enostavno sledili njim? #predsednik12
- peter_pec Plus, premikanje na SD kartico sem probala že stokrat. Tudi telefon to ponudi kot možnost. Rezultat? "Ni predmetov za premikanje"
- thetide so Ready To Go. Se veselim dobre letine pristankov. #aritmija #koncert #landings #novaplata <http://t.co/mytIEbvTUz>

Primer – zaznamovan besedni red?

L3T3 (7/49)

- union_pivo pizda, zakaj morm zmer uniona pit z laško kozarca? A to je taka politika al nimate kozarcev al vam je vseeno za kulturo piva??!!
- merineseri pa če ni sniceksaaaa ;) zdej sm se spomnu, da mam doma doma arašide, k jih je treba res kopat iz zemlje ;)
- TaiaG ja no, men se to tud skooz dogaja! Se veckrat pa s senckami na instagramu. Nardim smokey zgleda pa cist neki neznega :/
- Njokifestival ma ja...butaste @SlovenskeNovice so se na @francikek spravlje...#kreteni...to je čist legalno... @TinoMamic @ErikaPlaninsec
- AlanBStard2 @MajdaSirca @stanka_d @SoMe_Meli jaz sem posnel na RŠ in prinesel na golf v Kr je folk čist znoru, danes malo nerodno heh

Primer raziskave 2

Arhar Holdt, Š. in Dobrovoljc, K. (2016): Vrednost korpusa Janes za slovensko normativistiko. *Slovenščina 2.0*, 4 (2): 1–37.

- **Ideja:** Primerjati, kako se pojavljajo zveze samostalnika z neujemalnim levim prilastkom (*solo petje*, *RTV prispevek*) v korpusih Janes in Kres.
- **Metoda:** Z uporabo konkordančnika SketchEngine primerjamo vzorce v dveh korpusih: referenčnem korpusu Kres in korpusu Janes.
- **Rezultat:** Analiza razkriva: da se referenčni korpus Kres in korpus Janes glede zapisa teh zvez pomembno razlikujeta; da je raba tovrstnih zvez v korpusu Janes pogostejša in bolj raznolika kot v korpusu Kres; da se v obeh korpusih pojavlja visok delež zvez, ki v rabi izkazujejo variantnost v zapisovanju, tudi na ravni posameznih prilastkov; in – vsaj na prvi pogled – presenetljivo, da je raba v korpusu Janes konsistentnejša, kar nakazuje, da jezikovna regulacija obravnavanega problema povečuje variantnost v jezikovni rabi.

Primer raziskave 2

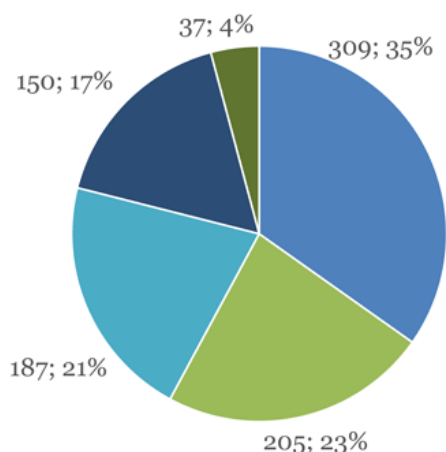
- Kot potencialne zveze samostalnikov z neujemalnim levim samostalniškim prilastkom smo z orodjem noSketchEngine iz obeh korpusov izluščili nize dveh zaporednih samostalnikov, samostalnika v imenovalniku in samostalnika s poljubnimi oblikoskladenjskimi lastnostmi, pri katerih se dana oblika prvega samostalnika ne glede na velikost črk v celotnem korpusu pojavi pred vsaj tremi različnimi oblikami leme jedrnega samostalnika (npr. *rtv prispevek*, *rtv prispevka*, *rtv prispevkom*). Če je bil ta pogoj izpolnjen, je bil niz oblike prilastka in leme jedra prepoznan kot potencialna zveza samostalnika z neujemalnim levim samostalniškim prilastkom (*rtv prispevek*).
- Pregled rezultatov je razkril, da tokenizacijske, besednovrstne in oblikoskladenjske interpretacije, kakršne so bile korpusnim besedilom pripisane v postopku strojnega oblikoskladenjskega označevanja, na podatke vplivajo tako zaradi neenotnosti označevalnikov kot zaradi nedoslednosti v obstoječih jezikovnih virih.

Primer raziskave 2

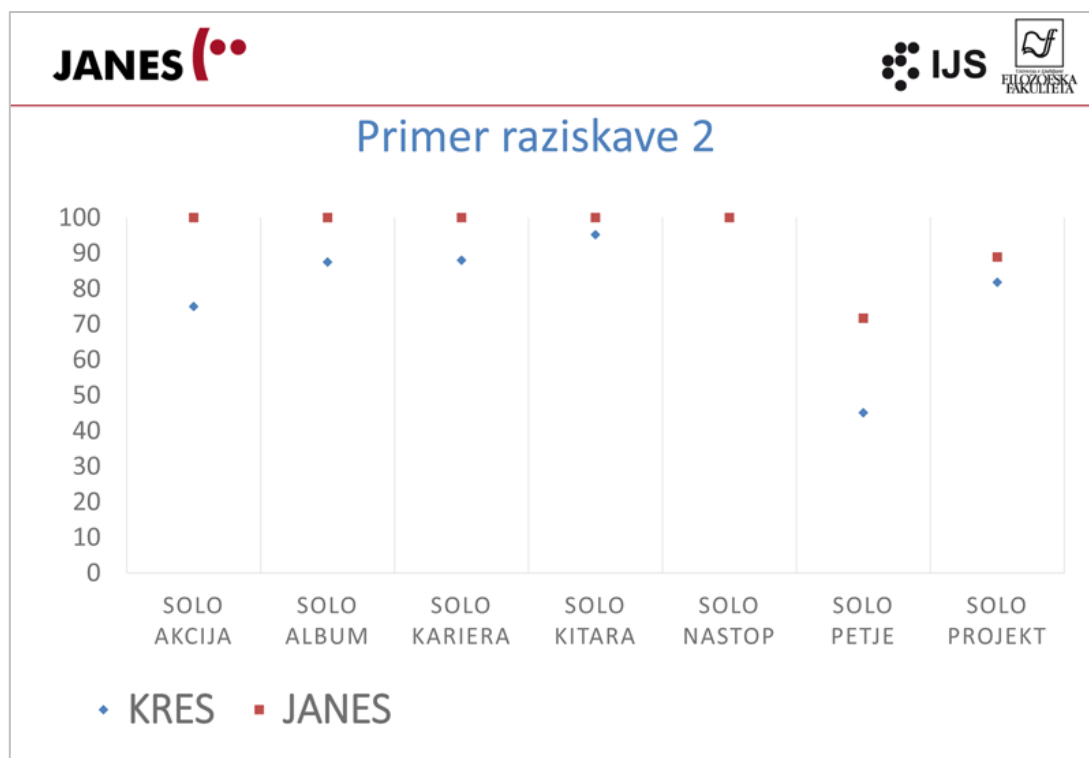
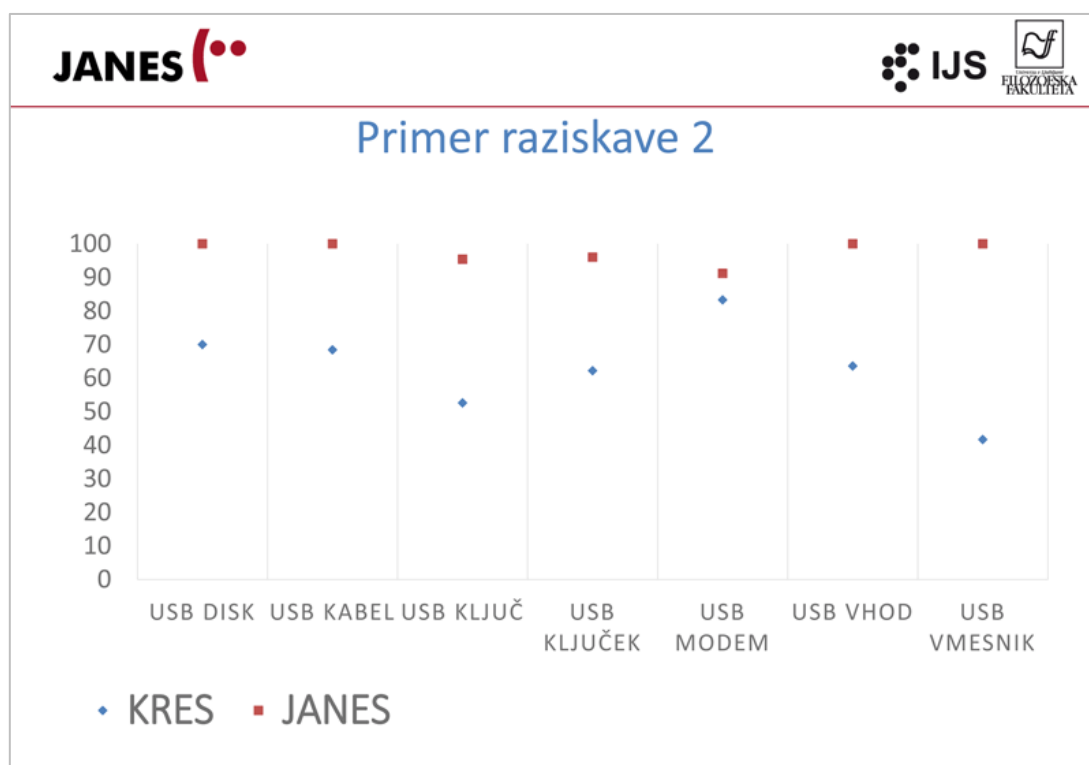
	Kres		Janes		Skupno
	Abs.	Rel.	Abs.	Rel.	
pojavnice (vse)	95.897	987	212.808	1.662	
pojavnice (pregibne)	49.047	505	108.303	846	
različnice – zveze	3.054	31	7.840	61	888
različnica – prilastki	1.432	15	2.851	22	719

Načini zapisa zveze	Kres	Janes
samo zapis narazen (npr. <i>loto številka</i>)	71 %	75 %
zapis narazen in z vezajem (npr. <i>tv film, tv-film</i>)	13 %	8 %
zapis narazen in skupaj (npr. <i>špas teater, špasteater</i>)	11 %	9 %
zapis narazen, z vezajem in skupaj (npr. <i>new york, newyork, new-york</i>)	5 %	7 %

Primer raziskave 2



- Zveze z neujemalnim prilastkom (joga studio, android telefon)
- Lastna imena (butan plin, indiana jones)
- Kratične zveze (usb ključek, c vitamin)
- Nerelevantni rezultati (york city, pearl jama)
- Citatna oz. polcitatna imena (after party, bad boy)



T'ain't what you do
it's the way
that you do it

that's what gets results.



"T'ain't What You Do (It's the Way That You Do It)" is a song written by [jazz](#) musicians [Melvin "Sy" Oliver](#) and [James "Trummy" Young](#). It was first recorded in 1939 by [Jimmie Lunceford](#), [Harry James](#), and [Ella Fitzgerald](#).^[1]

Literatura in nadaljnje branje

- Arhar Holdt, Š., Nataša Logar in Tomaž Erjavec: Slovenska znanstvena besedila: korpusni opis – trpnik. *Toporišičeva Obdobja*, Ljubljana, 10.–12. november 2016. V pripravi.
- Arhar Holdt, Š., Dobrovoljc, K. (2016): Vrednost korpusa Janes za slovensko normativistiko. *Slovenščina* 2.0, 4 (2): 1–37.
- Böhmová, A., Hajič, J., Hajičová, E. and Hladká, B. (2003). The Prague dependency treebank. In *Treebank: Building and Using Parsed Corpora*. Netherlands: Springer, pp. 103–127.
- Crystal, D. (2011). *Internet Linguistics: A Student Guide*. London, New York: Routledge.
- Dobrovoljc, K. and Nivre, J. (2016). The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '16)*. Portorož, pp. 1566–73.
- Dobrovoljc, K., Erjavec, T. and Krek, S. (2016). Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino. In *Proceedings of the Conference on Language Technologies and Digital Humanities*. Ljubljana (in print).
- Dobrovoljc, K., Krek, S. and Rupnik, J. (2012). Skladenjski razčlenjevalnik za slovenščino. V T. Erjavec, J. Žganec Gros (ur.), *Zbornik 8. konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, pp. 42–47.
- Erjavec, Tomaž in Ledinek, Nina, 2006: Slovenska odvisnostna drevesnica, prvi rezultati. V T. Erjavec, J. Žganec Gros (ur.), *Zbornik 5. konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, pp. 162–167.
- ERJAVEC, Tomaž, idr., 2008: *Specifikacije za leksikon besednih oblik: SSJ kazalnik 3*. Kamnik. <http://projekt.slovenscina.eu/Vsebine/SI/Kazalniki/K3.aspx>
- Erjavec, T., Fišer, D., Krek, S. and Ledinek, N. (2010). The JOS linguistically tagged corpus of Slovene. In: *LREC 2010, 7th International Conference on Language Resources and Evaluations*. Valletta, pp. 1806–1809.
- Holozan, P., Krek, S., Pivec, M., Rigač, S., Rozman, S. and Velušček, A. (2008). *Specifikacije za učni korpus*. Kamnik: Projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ.
- Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C. and Smith, N. A. (2014). A dependency parser for tweets. In *Proc. of EMNLP*. Doha, Qatar, pp. 1001–1012.
- Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N. and Holz, N. (2015). *Training corpus ssj500k 1.4, Slovenian language resource repository CLARIN.SI*, <http://hdl.handle.net/11356/1052>.
- Ledinek, Nina, 2014: Slovenska skladnja v oblikoskladenjsko in skladenjsko označenih korpusih slovenščine. Ljubljana: Založba ZRC, ZRC SAZU.
- Michelizza, M. (2015). *Spletna besedila in jezik na spletu*. Založba ZRC, ZRC SAZU, Ljubljana.
- Toporišič, Jože, 1982: *Nova slovenska skladnja*. Ljubljana: DZS.
- TOPORIŠIČ, Jože, 2004: *Slovenska slovnica: četrta, prenovljena in razširjena izdaja, 2. natis*. Maribor: Obzorja.

Polona Gantar

10 KORPUSNA FRAZELOGIJA

10.1 Korpusna frazeologija – izročki



Korpusna frazeologija

Povej mi s kom se družiš in povem ti ...

Polona Gantar
apolonija.gantar@ff.uni-lj.si

Poletna šola spletne slovenščine za študente
Ljubljana, 4. do 8. julij 2016



Vsebina

- **Korpusno jezikoslovje**
 - O korpusih in korpusnem jezikoslovju
 - O možnosti izrabe korpusov v leksikologiji in frazeologiji
- **Frazeologija kot jezikoslovna veda**
 - osnovni pojmi frazeološkosti
 - večbesednost
 - ustaljenost
 - pomenska (ne)prozornost
 - jezikovnosistemske »nepravilnosti«
 - ekspresivnost, metaforičnost
 - raziskava: frazeološke prenovitve v spletni slovenščini
 - zamejitev področja
 - frazeologija v ožjem in širšem smislu
- **Korpusna frazeologija**
 - leksikalna gramatika
 - sintagmatsko-paradigmatska razmerja (kolokacije)
 - Projekt Corpus pattern Analysis
 - orodja za analizo večbesednih enot v korpusu
 - Besedne skice
 - skladenjski in stavčni vzorci (LBS)
 - Kolokacijski slovar slovenskega jezika
 - Parseme shared task
 - Raziskovalna naloga: Korpusna analiza fraznih glagolov

KORPUSI IN KORPUSNO JEZIKOSLOVJE

Razvoj korpusov

- **Korpusi v preteklosti**
 - zbirke jezikovnih podatkov
- **Korpusi danes**
 - obsežne elektronske zbirke jezikovnih podatkov
 - računalniška berljivost besedil: možnost procesiranja jezikovnih podatkov
 - standariziran zapis: označevanje korpusnih dokumentov; podatki o besedilu ločeni od besedila samega; primerljivost in tipologizacija (TEI – Text Encoding initiative)
 - označenost:
 - zapis korpusa: dokumentiranost besedil
 - pripisovanje jezikoslovnih oznak: strukturna in jezikoslovna označenost

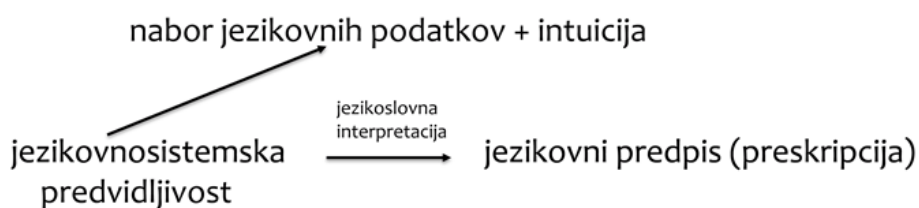
Označenost korpusov

- **strukturno označevanje**
 - tokenizacija osnovne korpusne enote – pojavnice (presledek)
 - segmentacija segmenti, npr. povedi (končno ločilo)
- **jezikoslovno označevanje** jezikoslovna interpretacija
 - **lematizacija**

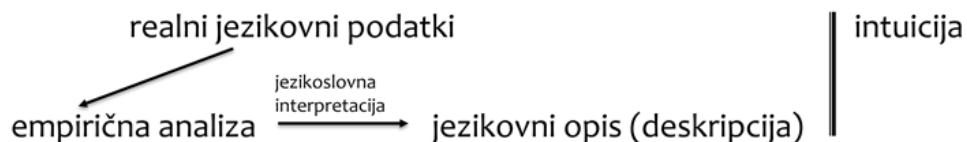
list - sam. m. sp. (list-a, list-u, list-Ø; list-om; list-i, list-ov, list-e, list-om ...)

lista – sam. ž. sp. (list-e, list-i, list-o; list-e; list-Ø, list-am)
 - **oblikoskladenjsko označevanje** sobesedilno pogojene morfološke oznake

Pred- (ne)korpusno jezikoslovje



Korpusno jezikoslovje



intuicija

- intuicija pri pridobivanju jezikovnih podatkov pove več o tem, kako deluje intuicija, kot o jeziku samem
 - **samospraševanje** (Kaj pomeni ... ? Ali lahko rečem ...? Tega nikoli ne bi rekel ... Koklikokrat rečem?)
 - **nesprejemljivi : sprejemljivi stavki**
- intuicija rojenih govorcev je nepogrešljiva pri jezikoslovni interpretaciji podatkov

empirična analiza

Empirična analiza jezika, kot se kaže v *dejanski rabi* (empirija – izkustvo), pomeni registriranje realnega stanja in ne zanašanje na intuitivno vedenje o jeziku.

“ ... če želimo o jezikoslovju govoriti kot o znanosti, potem morajo biti raziskave omejene izključno na to, kar je mogoče v jeziku opazovati ...”

Bloomfield (30. Leta 20. stol.) revija *Language*

Možnosti izrabe korpusnih podatkov

▪ Jezikoslovne raziskave

- leksikologija
- frazeologija
- leksikografija (jezikovni opis; priročniki)
- **oblikoslovje**: ty, tenkju, tenks, tn timer, thanks, thnx, tenx ; tvitersiti, tweetersiti, twittersiti ...
- **skladnja**: alfa samec; RTV prispevek
- **pomenoslovje**:
 - „I don't believe in word sense“ (Kilgarriff 2009)
 - „you shall know the word by the company it keeps“ (Firth 1975)

JANES Interdisciplinarne raziskave

▪ Korpusi in jezikovnotehnološke raziskave

- jezikovno rudarjenje
- strojno prevajanje
- analiza sentimenta
- prepoznavanje imenskih enitet
- skladenjsko označevanje
- semantično označevanje
- označevanje VLE

▪ Sociološke raziskave brexit – grexit... kdor ne skače ni Slovenci

▪ Kulturološke raziskave kaj je tipično slovensko ...

FRAZEOLOGIJA KOT JEZIKOSLOVNA VEDA

Frazeološke enote so večbesedne strukturno in pomensko ustaljene enote, ki glede na svoje sestavne elemente izkazujejo samostojni, od sestavnih elementov neodvisni pomen. V besedilih se uporabljajo predvsem za izražanje negativnega ali pozitivnega vrednotenja in so pogosto glede na nefrazeološke ustreznike stilno zaznamovane.

- večbesednost
- strukturna in pomenska ustaljenost
- pomenska neprozornost
- jezikovnosistemske „nepravilnosti“
- ekspresivnost in metaforičnost

JANES ena ali več besed?



tokenizacija in
lematizacija!

- Alfa romeo, alma mater
- al dente, at hoc, a priori
- d. o. o., omg! espe
- narodnozabavni, narodno-zabavni, narodno zabavni
- naprimer – na primer; navsezadnje – na vse zadnje
- bati se, smejati se, svitati se ...

- računati na koga : računati na pamet
- priti do denarja : priti do razglednega mesta
- cukati ga : cukati ga za rokav

JANES ena ali več besed?



- d. o. o., omg! espe
- al dente, at hoc, a priori
- Alfa romeo
- narodnozabavni, narodno-zabavni, narodno zabavni
- naprimer – na primer; navsezadnje – na vse zadnje
- bati se, smejati se, svitati se ...

- **računati na** koga : računati **na pamet**
- **priti do** denarja : priti **do** razglednega mesta
- **cukati ga** : cukati **ga** za rokav

frazni glagol

prosti glagolski morfem

zaimsek

predlog: za izražanje meje v
prostoru do katere sega dejanje

predložna
prislovna zveza

JANES ena ali več besed - Janes



frazni glagoli

- up loudati – uploudati, daunoloudati
- up date – updatati, updejtati

trenutno se **uploada** ampak bo še trajalo varno, da si vse slike **up loadal** na FaceBook

izrazi, nastali po frazeološki poti

- piš me v uh – pišmevuh; pišmevuhovstvo
- majke mi – majkemi, jebi ga – jebiga
- je bela cesta – jebela cesta – jebelacesta
- ne bodi ga treba – ne-bodi-ga-treba – nebodigatreba

Majke mi, toti ni čist diht...
@merineseri midve sva v žlahti, **majkemi**

drugo

- omb - o moj bog
- lp - lep pozdrav
- smsjati - sms-jati
- konec koncev - koneckoncev, kratko malo - kratkomalo
- stand up - standup; on line - on-line - online

O MOJ BOG kaki tviti
OMG, kir popizditis dobivam

JANES ena ali več besed - Janes



da danes zopet ne bo nič, odstraniš make	up	, daš v pižamo in skriješ pod kovter.
Najbolje pa je, ce to storis sam. Heads	up	. ¶ V Bauhausu so me gledali kot največjega
Čakam, da požegnajo, da lahko grem. Thumbs	up	za šefa, ki mi je prinesel svež rogljiček
zanimivo, omenja tudi Jacksona oz. njegovo back	up	vokalistko . ¶ La femme qui se poudre - čudovit
za novo leto ". Letni časi so res fucked	up	. ¶ Če imaš vrane v akustičnem dvorišču
če si muzikaličen te zna odpeljati daleč. Up	to you.	. ¶ @tjablonsky + Bojan Adamič.
pretiravali @GobaFunk ¶ @strupko sicer pa, shut	up	! :) končno najdem podobnomislečo punco...
letih. Žalostno, da je moral priti push	up	od drugod. ¶ @BKovacec @EvaKobe24ur
je bila totalna čudovitka, hvala za stand	up	comedy šov. #soocenjeLJ ¶ Kdo je stilistka
uporabil samo VIP ¶ Današnji sanjski line	up	bi bil: Dead Boys, Dead Kennedys in kot
pa politični noži v hrbtih... Luka, grow	up	. ¶ @petrasovdat V pričakovanju izjave, da
ne bi vedla:) ¶ Glavno, da imam še perfect	meikap	, ko v fejk kroksicah nesem smeti ven
http://t.co/SB6VzUFJTt ¶ @Pawla40 ej sam en mejhn	heads up	, kle ni lajkov tko k na FBju! #welcome ¶
odstopil, pa ga žal nima! ¶ DROLC go home and	šadap	! ¶ Hehe tipa ko sem videl na odmevih sem
@BanicGregor ¶ A majo ta tedn vremenarji kšn	Standup	comedy slučajno? Vrjetno bo zvečer bl
http://t.co/UfWN1M2tpT ¶ Si že videl tale delček	lineupa	za Tomorrowland? Noooro kajne? ! : D Kdo

Izrazna in pomenska ustaljenost?

- **Kurja koža** ol ouver! Rade #Šerbedžija je Zakon!
- Omojboh... 30 sekund #Sanremo2014 pa že **kurja koža**, solzne oči...
- včeraj sem prešlatala celo zimsko kolekcijo. **kurja koža** v trenutku!
- 3cm veliko **kurjo kožo** dobim že pri trailerju, kaj šele bo.
- pa ko čuješ zvok od stare impreze ti kr **kurja koža** rata
- omg, meni kar **kurja koža** postane, ko vidim ono fotografijo
- bemu sunac, **kurja koža**, pa napete bradavičke:)
- kadarkoli se ga spomnim, me oblije **kurja polt**
- Zakaj se mi je **kurja polt** naredila, ko sem prvič poslušala Luko Mesca.
- Rusom in Kitajcem se počasi **ježi koža**, dolgo ne bodo več prenašali psihopatov.
- Ko slišim za besedo strokovnjak, dobim **kurjo kožo**.
- Na čelu države bo Gospod Janez Janša, rdeči tk kr **kurjo kožo** dobite ko ga omenim

poz.

neg.

- **variantnost sestavin**
 - Nizozemci so pravi trn v **peti** Slovencev.
 - Ja res, ful so koristni ... kot **trn v riti** ...
 - Me zanima, zakaj jim je ravno Mišič tak trn v **očesu**.
- **vrivanje besed**
 - trn v **poslanski** peti
 - **Kosovski** trn v **črnogorski** peti
 - zasebniki v lekarnstvu so **hud** trn v peti javnim lekarnam
- **oblikovne in skladienske spremembe**
 - Slaba banka je trn v peti »oldboysom«
 - Alenka je trn v **Cerarjevi** peti
- **variantnost v besedilni realizaciji**
 - kadar se ji **zapiči v peto kak trn resnice**, ...
 - Flirtanje je lahko nedolžen klepet, včasih pa postane igra z **ognjem**.
 - Ne bom dovolil, da spet zaidemo v igračkanje z **ognjem**.
 - vseeno ne pozabi biti previden: z **ognjem** se ni dobro igrati!

JANES Jezikovnosistemske „nepravilnosti“ IJS



- **oblikoslovje**
- **besedna oblika**
Je skoda usakga twita. Kr nej bodo na ignor
Čakal sem te ko kreten, ti pa me daš na ignore.
- **besedna vrsta**
KrNeki test 😊
prej sem kr neki fejku z krnekimi copati pa so me včasih podplati dlje casa bolel
- **skladnja**
- **nepričakovan skladenjski vzorec**
Dvorni norci tudi iz sebe norce brijejo
gre za (kaj)
- **pretvorbena omejenost**
zdrava pamet *pamet je zdrava
star maček *maček je star
- besedilo**
- **elipsa**
mi o volku, kozel iz zelja
človek človeku volk
- **referenčnost**
delati iz muhe slona
tresla se je gora, rodila se je miš

JANES



Pomenska (ne)prozornost?

• star

- star model avta
- stara mama
- staro mestno jedro
- stari običaji
- star papir
- star prdec
- stara celina
- stara zaveza



▪ črn

- črna barva
- črn oblak
- Črno morje
- črna kava
- črn kruh
- črna kronika
- črna gradnja
- črna vdova



JANES Pomenska (ne)prozornost?

stara mama
dnevna soba
spalna srajca
prekmurska gibanica
pametni telefon

?

SLO	CZ	NEM	ANG
čez čas/sčasoma	po čase/začas	im Laufe der Zeit	in time
v hipu/v trenutku/kot bi trenil/takoj	na první pohled	im Augenblick	in (less than) no time
skrajni čas	nejvyšší čas	höchste Zeit	high time/about time

JANES

Ekspresivnost, metaforičnost,
jezikovna ustvarjalnost

FE kot kolektivne metafore – frazeološki
pomen

Nenevtrarno izražanje

spraviti v red – urediti
mazati si roke z (drekom) – ukvarjati
se s čim neprijetnim ali kaznivim
delati iz muhe slona – pretiravati

Iz tega ne more nastati nič dobrega, kajti zaradi **umiritve razmer** raje **prikrijejo dejstva**. In v celi državi ni **pogumnega politika**, ki bi dovolj **odločno** problematiziral takšno početje. To je posmeh vsem, ki so zaradi vladnih ukrepov prisiljeni **povsod** varčevati. Tudi sindikati in upokojenci **nočejo priznati resnice**, ne vidijo **dalnosežnosti problema**. Njihova **zmaga bo imela negativne posledice** za celotno Slovenijo. /.../ Skoraj neopazno je bil odstavljen prvi nadzornik HSE, eden redkih, ki **je upal opozoriti na nepravilnosti**.

Iz tega ne more nastati nič dobrega, kajti zaradi **ljubega miru** raje **pometejo smeti pod preprogo**. In v celi državi ni politika **z jajci**, ki bi **udaril po mizi** in problematiziral takšno početje. To je posmeh vsem, ki so zaradi vladnih ukrepov prisiljeni varčevati **na vsakem koraku**. Tudi sindikati in upokojenci **tiščijo glavo v pesek**, ne vidijo dlje **od svojega nosu**. **Ta zmaga bo Pirova**, poražena pa bo cela Slovenija. /.../ Skoraj neopazno je bil odstavljen prvi nadzornik HSE, eden redkih, ki je **dregnil v sršenje gnezdo**.

FE kot kolektivne metafore – frazeološki pomen

Nenevtrarno izražanje

Opis splošno človeških lastnosti, narodovih značilnosti, družbenih pojavov (narodovi stereotipi)

Sekundarna metaforizacija – FE kot povod za ustvarjalno in duhovito rabo.



six pack



one pack = [vampek]

spraviti v red – urediti
mazati si roke z (drekom) – ukvarjati se s čim neprijetnim ali kaznivim
delati iz muhe slona – pretiravati

krvav pod kožo
naj sosedu krava crkne
kdor ne skače ni Slovenc

ljubi svojega bližnjega ...
... samo pazi, da te ne dobijo

zrno do zrna ...

gluten

kamen na kamen ...
nepremičninski davek

JANES Frazeološke prenovitve



- **Konferenca Slovenščina na spletu in v novih medijih**
- Justin, Hirci in Gantar, 2015: Rana ura, slovenskih fantov grob: analiza frazeoloških prenovitev v spletni slovenščini <http://nl.ijs.si/janes/wp-content/uploads/2015/11/JANES15-06-Rana-ura.pdf>
- **Teoretična izhodišča raziskave** (Kržišnik 1990 idr.)
- **frazeološke enote**
 - raba FE je v vsakem besedilu **opazna**
 - višja pogostnost FE je lahko znak **bogatejšega in pomensko bolj razplastenega jezika** (Kržišnik 1990)
- **frazeološke prenovitve**
 - frazeološke prenovitve kot hotene in namerne inovacijske spremembe; niso sistemske in so zato enkratne
 - spremenijo obliko in/ali modificirajo pomen
 - učinkovita prenovitev mora vzpostavljati **razvidno razmerje do pomena izhodiščne FE**, hkrati pa prinašati **nove pomenske razsežnosti**.
- **računalniško posredovana komunikacija**
 - nekorrigirana sporočila
 - specifični žanri
 - elementi nestandardnosti, govornosti; inovativnost, individualnost, učinkovitost, opaznost ...

JANES



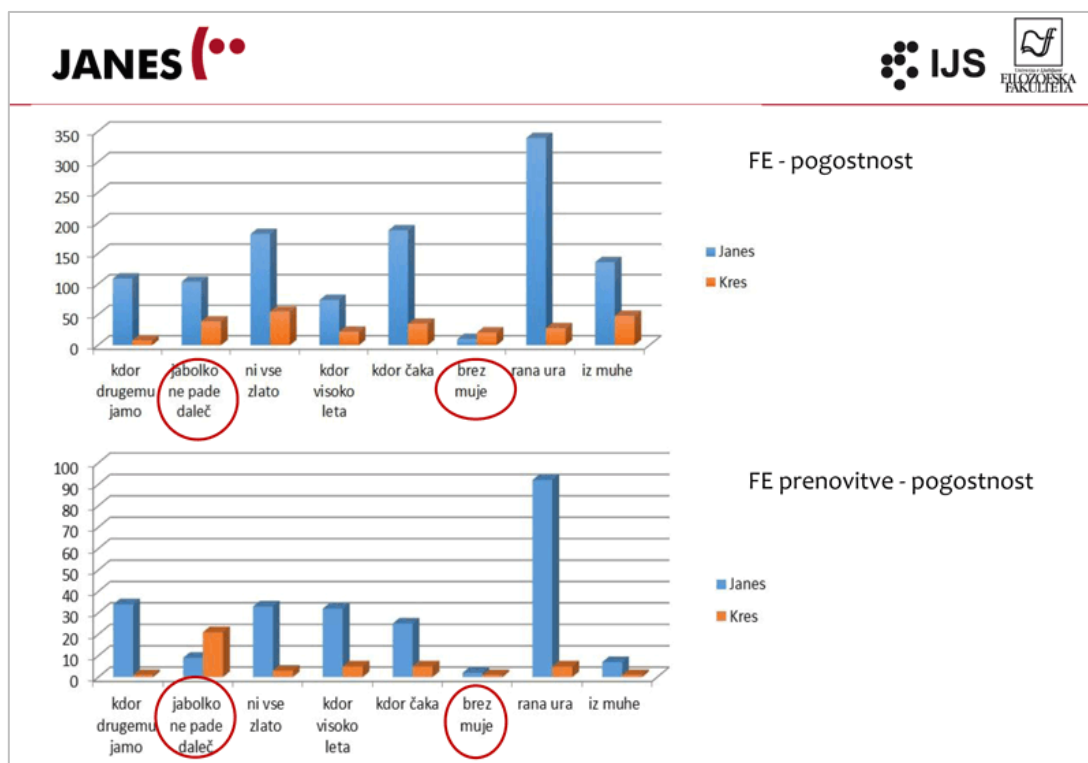
- **Namen raziskave**
 - ugotoviti zastopanost frazeoloških enot v spletno specifičnih žanrih
 - delež frazeoloških prenovitev
 - značilnosti ustvarjalne jezikovne rabe
- **Korpus**
 - izhajali smo iz podkorpusa Janes v0.3 Tweet (Fišer et al. 2014)
 - Za primerjavo smo uporabili korpus Kres (Logar et al. 2012); izločili smo internetna besedila.
- **Orodja in metode**
 - Sketch Engine (Kilgariff et al. 2004), in sicer funkcije
 - iskanje v konkordančniku
 - urejanje besedilnega okolja in filtriranje besedil
 - besedne skice in
 - frekvenčni prikaz obravnavane FE v določenem besedilnem tipu

Obravnavane FE

- *kdor drugemu jamo koplje, sam vanjo pade*
- *jabolko ne pade daleč od drevesa*
- *ni vse zlato, kar se sveti*
- *kdor visoko leta, nizko pade*
- *kdor čaka, dočaka*
- *brez muje se še čevelj ne obuže*
- *rana ura, zlata ura*
- *iz muhe delati slona*

Kvantitativna analiza

Iskalni pogoj	Korpus Janes (161 mio besed)				Korpus Kres (120,5 mio besed)			
	vse pojavitve	izho- diščne FE	pren- ovitve	nejasno /dobesedno	vse pojavitve	izhodi- ščne FE	preno- vitve	nejasno/ dobesedno
<i>kdor drugemu jamo</i>	109	59	34	12	8 (3)	7	1	0
<i>jabolko ne pade daleč</i>	104	91	9	4	39 (2)	18	21	0
<i>ni vse zlato</i>	182	146	33	3	55 (13)	48	3 (1)	4 (1)
<i>kdor visoko leta</i>	74	38	32	4	22 (3)	17	5	0
<i>kdor čaka</i>	188	155	25	10	35 (5)	27	5	3 (1)
<i>brez muje</i>	10	3	2	5	21 (0)	16	1	4
<i>rana ura</i>	339	215	92	32	28 (5)	17	5 (2)	6
<i>iz muhe</i>	136	128	7	1	48 (11)	45	1	2
skupaj	1142	835	234	71	256 (42)	195	42 (3)	19 (2)



JANES **Kvalitativna analiza**

IJS FILOZOFSKA
FAKULTETA

- **ponovna aktualizacija (dobesedni pomen)**
 - *kdor drugemu jamo koplje, je prvega že zasul*
 - *kdor drugemu jamo koplje, je grobar*
 - *kdor visoko leta, je pilot/ptič/stewardesa*
 - *kdor visoko leta, daleč vidi*
 - *kdor čaka, ima dobre živce*
 - *kdor čaka, postane star*
 - *rana ura, zaspana ura*
 - *rana ura, rano dan zjebe*
- **demetaforizacija -> ponovna vzpostavitev metaforičnega premika**
 - *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*

▪ križanje FE

- *kdor drugemu jamo koplje, je sam svoje sreče kovač*
- *ni vse zlato, kar se Janezek nauči*
- *kdor visoko leta, se najslajše smeje*
- *kdor visoko leta, ne pade daleč od drevesa*
- *rana ura, slovenskih fantov grob*

▪ družbena aktualizacija

- *kdor drugemu jamo koplje, vlada vanjo pade*
- *kdor drugemu jamo koplje... je jutri vabljen v Rovte*
- *ni vse zlato, kar je Žižek*
- *kdor visoko leta, je Peter Prevc*
- *rana ura, zlati Rolex*

DA

NE

• večbesednost

- ✓ popihati jo
- ? priti do (česa)

smejati se

• pomenska neprozornost

- ✓ star maček

stara mama
pametni telefon
potni list

• ekspresivnost, metaforičnost

- ✓ (imeti) dolg jezik

taščin jezik*
črna luknja*

frazeologija v ožjem smislu vključuje večbesedne enote, ki izkazujejo od sestavin neodvisen in hkrati metaforičen pomen.

- **glagoli se/ si** (prosti morfem : povratni zaimek)
smejati se (prosti morfem)
umivati koga – umivati se (be) (povratnost – zaimek)
predvidevati kaj – kaj se predvideva (tvornost/trpnost)
delati kaj : delati se (pomenska razlikovalnost)
- **frazni glagoli**
gre za (koga/kaj)
priti do (česa)
računati na (koga/kaj)
- **zveze s pomensko oslabljenimi glagoli**
delati družbo
dati mir
vzeti na znanje
- **zveze s slovničnimi/funkcijskimi sestavinami**
v zvezi z/s, v bistvu, z drugimi besedami ...
ne glede na, za razliko od ...
bolj ali manj, sem in/pa/ter tja ...
po eni strani - po drugi strani ...

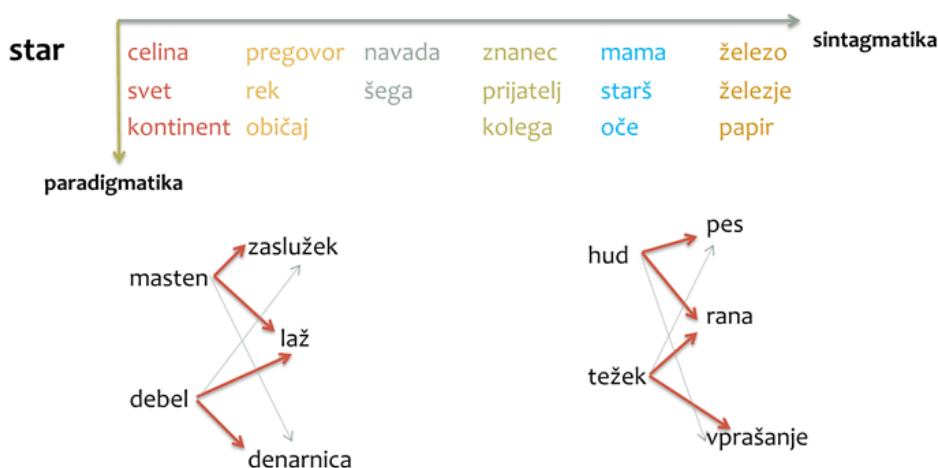
KORPUSNA FRAZELOGIJA

„you shall know a word by the company it keeps“ (Firth 1957: 11)

„I don't believe in word senses“ (Kilgarriff 1997: 91)

- Pomen kot zaključena enota ne obstaja, obstajajo zgolj **pomenske tendence**. (Hanks 2013)
- Pomenske tendence besed vplivajo na **izbire v svojem besedilnem okolju**.
- Vsak pomen se oblikuje **v konkretni govorni situaciji** in je odvisen od vsakokratnega jezikovnega in zunajjezikovnega **konteksta**.
- Pomeni so neposredno povezani z **vzorci, strukturami in frazeološkimi enotami**, v katerih se pojavljajo.
- Izbira pomena, kjerkoli v besedilu, ima odločilen vpliv na okoliške izbire. (J. M. Sinclair, *Corpus Concordance Collocation*, 1991)

„you shall know a word by the company it keeps“ (Firth 1957)



JANES **IJS** **FILOZOFSKA
FAKULTETA**

uporabnik: polona korpus: Kres

gramatična relacija

koža (samostalnik)
Kres frekvenca = 20,475 (170.0 na milijon)

Iskanje
Seznam
Besedne skice
Tezaver
Najdi X
Primerjalne
skice
O korpusu
Shrani
Spremeni
nastavitve
Grupiranje
Razvrsti
Relacije
Več podatkov
Manj podatkov

kolokator

frekvenca
št. pojavitev v
korpusu

statistična
jakost
(salinece)

Constructions	S_kakšen?	S_predi-pred	S_s-koga-česa	S_koga-kaj
S_biti_s5 237 2.2	občutljiv 263 9.79	pod 471 8.07	nega 262 9.71	vlažiti 44 9.03
	suh 451 9.73	skozi 153 7.62	staranje 153 9.6	dražiti 56 8.91
	gol 177 9.1	na 1,882 5.98	čvrstost 36 8.04	pomirjati 45 8.62
	masten 130 8.97	brez 83 5.36	tip 191 8.03	negovati 53 8.38
	gladek 112 8.68	prek 11 4.86	čiščenje 83 7.99	gladiti 26 8.2
	ovoj 84 8.63	iz 191 4.83	plast 116 7.98	zgladiti 26 8.08
	pomarančen 79 8.49	čez 23 4.77	vnetje 62 7.76	navlažiti 22 8.03
	razdražen 69 8.44	do 115 4.72	luščenje 29 7.75	ščiti 50 7.87
	nečist 42 8.37	preko 6 4.26	predelovalec 26 7.62	napeti 33 7.85
	lasten 308 8.29	z 423 4.19	prožnost 29 7.54	učvrstiti 18 7.83
	napet 68 7.91	po 124 3.82	elastičnost 26 7.54	izsušiti 25 7.79
	zrel 70 7.88	v 692 3.77	videz 104 7.52	predreti 16 7.58
	trd 102 7.87	za 259 3.57	barva 289 7.51	očistiti 45 7.47
	živalski 84 7.81	ob 30 3.11	vlaženje 23 7.45	odreti 12 7.4
	razpokan 44 7.79	od 21 2.19	draženje 25 7.43	razdražiti 12 7.35
	čvrst 53 7.76	k 2 1.85	obnavljanje 28 7.42	mehčati 12 7.27
	nežen 72 7.64	zaradi 6 1.61	povrhnjica 21 7.25	odstraniti 73 7.15
	izsušen 39 7.61	med 11 1.59	strojenje 19 7.23	razvajati 17 7.12
	kozji 39 7.55	o 2 0.28	površina 125 7.21	odrgniti 9 7.07
	debel 83 7.52		ten 20 7.17	osvežiti 16 7.05
	mehek 67 7.46		občutljivost 28 7.14	prebosti 10 6.93
	zdrav 113 7.42		srbenje 19 7.12	odnesti 37 6.78
	srbeč 30 7.29		tonus 18 7.09	poživiti 11 6.74
	utrujen 46 7.27		okužba 43 7.03	zaščiti 30 6.73
	poškodovan 47 7.23		vlažnost 20 7.02	hladiti 11 6.71

JANES **IJS** **FILOZOFSKA
FAKULTETA**

Baza kolokacijskega slovarja slovenskega jezika

http://bkssj.cjvt.si/

KSSJ:

- avtomatsko izluščeni podatki za 2.5000 iztočnic iz LBS + avtomatsko izluščeni podatki za novih 2.500 iztočnic

BKSSJ (trenutno):

- 2.500 iztočnic
- 2.310.100 kolokacij
- 72.117 struktur
- 5 zgledov/kolokacijo

Rezultati iskanja

Število rezultatov: 199

Iztočnice:

Študentski

Kolokacije:

[pretežno] pretežno študentski

[študentski] študentski servis

[študentski] študentski dom

[študentski] študentski bon

[blok] študentski blok

[mlad] mlad in študentski

Študentski

Študentski in

Študentski blok

Študentski center

Študentski dom

Študentski

Študentski pridevnik

študentski bon

pridevnik + samostalnik

- Študenti se bodo tako s študentskimi boni lahko še naprej prehranjevali tudi med počitnicami.
- Več kot 90 odstotkov anketiranih študentov mariborske univerze se prehranjuje s študentskimi boni, pri čemer večini zadošča 20 bonov na mesec.
- Pa vendar kljub manjšim težavam ostaja prehrana na študentske bone prijubljena in izkoriščena možnost.
- V zakon je bila dodana tudi sankcija prepovedi prodaje študentskih bonov za dobo najmanj enega in največ treh mesecev.
- Po skoraj dva meseca trajajoči gneči in zmedbi pred poslovalnicami za nakup študentskih bonov prodaja bonov sedaj poteka nemoteno.

Corpus Pattern analysis (Hanks)

- Projeciranje pomena iz sobesedila na posamezno besedo v korpusu -> Pattern Dictionary of English Verbs

- Theory of Norms and Exploitations* (Hanks 1994, 2013; Hanks in Pustojevsky 2004, 2005).

- osnovni vzorci (norms)
- skladenjske realizacije (alternations)
- ustvarjalne rabe (exploitations)

- **[[Human 1]] adore [[Human 2 | Deity | {Animal = Pet}]]**
[[Human 1]] has extremely strong positive emotions with regard to [[Human 2 | Deity | {Animal = Pet}]]
- **[[Human]] adore [[Entity | Eventuality]]**
[[Human]] enjoys [[Entity | Eventuality]] very much
- Pivovarna Laško bo vložila pritožbo.
[Human | institution]] vloží [[Entity = Document]]
[Human | institution]] predloži/da [[Entity = Document]] z namenom sprožiti določen postopek

*Teorija jezikovnih konvencij in možnosti njihove izrabe

<http://pdev.org.uk/#browse?q=f=C>

stavčni vzorci v leksikalni bazi za slovenščino

PATTERN: [[Human 1 = Health Professional]] vaccinate [[Human 2 = Patient]] {(against [[Disease]])}

IMPLICATURE: [[Human 1 = Health Professional]] injects a vaccine into [[Human 2 = Patient]] to produce immunity against [[Disease]]

cepiti *glagol*

če se **ČLOVEK** cepi, prejme CEPIVO proti določeni BOLEZNI |ali| če **ZDRAVNIK** cepi **ČLOVEKA**, mu v telo vnese CEPIVO proti določeni BOLEZNI

- kdo cepi
- kdo se cepi
- kdo se cepi proti čemu
- kdo cepi koga/kaj
- kdo cepi koga proti čemu
- kdo cepi koga s čim

avtomatsko pridobivanje stavčnih vzorcev izkorpusa

gramatična relacija	vzorec v LBS	glagol	korpusni primer
V_S_V_O3	<v>kdo/kaj G komu</v>	čestitati, zoperstaviti, dvoriti, gospodinjiti, pridružiti	Bushu je ob <k>izvolitvi</k> <i>čestital</i> tudi palestinski voditelj Jaser Arafat. Ne kar hitro, pošteno so si vzeli čas in mi dvorili, kot se spodobi.
V_S_V_O3_O2	<v>kdo/kaj G komu koga/česa</v>	ne dovoliti	»On mi tega ne dovoli.«
V_S_V_O3_O4	<v>kdo/kaj G komu kaj</v>	predlagati, preprečiti, nakloniti, prekrižati	Bela mačka mi je prekrižala pot, mogoče bi bilo dobro vplačati loto.
V_S_V_O4_predl	<v>kdo/kaj G koga/kaj + predlog</v>	obvarovati, odvrčati, predvidevati	Liz se ni pritoževala, niti ga ni skušala odvrčati od tega
V_lahko-G		privoščiti	»Naslednje logično vprašanje: si lahko privoščiva vzdrževanje dveh gospodinjstev?
V_gl_Cit		pojasniti	A tudi tam so nas zavrnili zaradi istega razloga, nam je pojasnil župan Jazbec.

čestitati (glagol) Gigafida

Constructions		
O_tretja_oseba	4.134	1.10
O_nedoločnik_cs	1.772	4.70
V_lahko_G	909	10.90
O_povratni_si	600	4.70
O_povratni_se	525	0.60
V_S_V_O3	516	47.80

O_tretja_oseba (glagolske zveze z sebkom v dajalniku)
boli koga kaj
svita se komu kaj
briga koga

O_nedoločnik_cs (zveze z faznimi in modalnimi glagoli)
morati, začeti, želei, hoteti, nameravati ...; treba je [opozarjati,
čestitati ...]

V_lahko_G (zveze z modalnimi prislovi)
lahko računate, predvidevate ...

O_povratni_si (zveze s povratnosvojilimi zaimki oz. morfemi)
smejati se; upati si/se

V_S_V_O3 vzorec: Sam-Glag-Sam3
kdo čestita komu

gojiti (glagol)

<u>kom-čem_g5</u>	4165	
<u>S_v-d_g5</u>	2579	5.1
<u>S_na-d_g5</u>	1192	5.9
<u>S_po-d_g5</u>	220	1.8
<u>S_ob-d_g5</u>	88	2.3
<u>S_pri-d_g5</u>	59	1.1
<u>S_o-d_g5</u>	27	0.6

S_v-d_g5

gojiti v [rastlinjaku, lončku nasadu, vrtnariji]
gojiti v [ribogojnici, akvariju, epruveti]
gojiti v [Sredozemlju]

S_na-d_g5

gojiti na [plantažah, gredicah, vrtu, balkonu]
gojiti na [farmi, posestvu]
gojiti na [Kitajskem, Primorskem]

S_po-d_g5

gojiti po [vrtovih]

S_ob-d_g5

gojiti ob [opori]

S_pri-d_g5

gojiti pri temperaturi [x] stopinj

S_o-d_g5

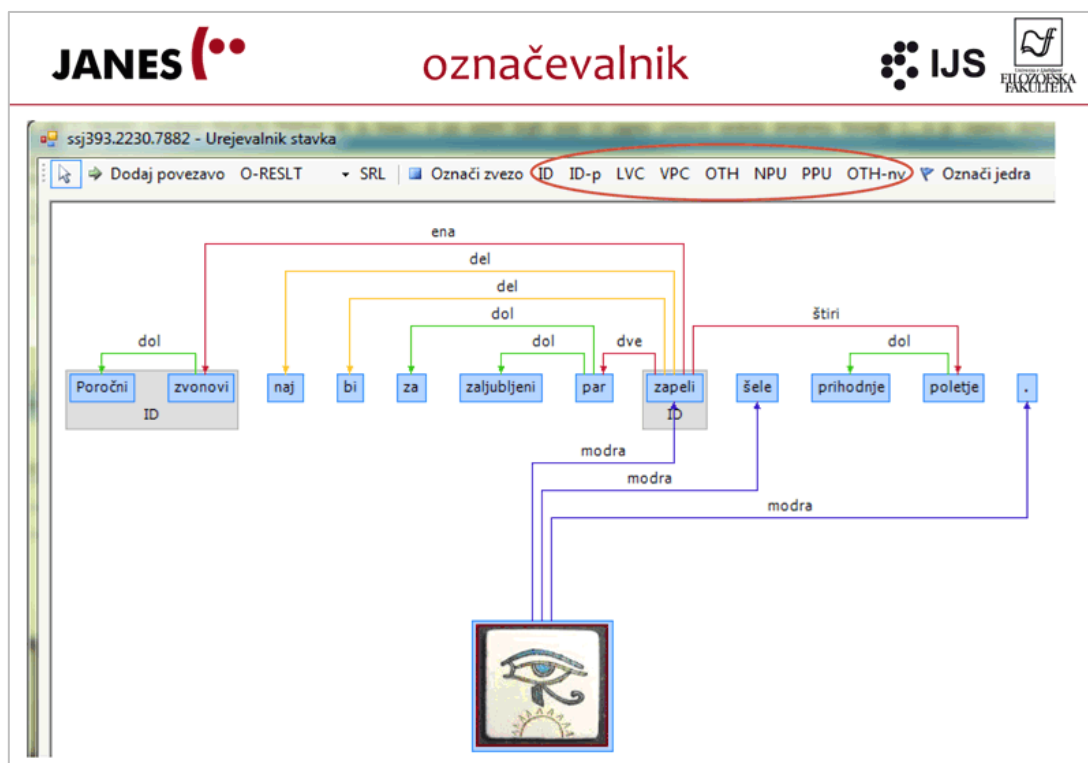
gojiti stereotipe o [harmoniki]
gojiti predsodke o [duševnem zdravju]

P A R S E M E

PARSING and Multi-word Expressions

<http://typo.uni-konstanz.de/parseme/>

- interdisciplinarna raziskovalna mreža v okviru evropske COST akcije
- računalniško procesiranje večbesednih enot
- standardizacija
- vključevanje v medjezikovno primerljive procesljive jezikovne vire
- PROJEKT: izdelava učnega korpusa z označenimi glagolskimi večbesednimi enotami za 20 jezikov (slovenščina, nizozemščina, grščina, bolgarščina ...)
 - mednarodni standardi in metode označevanja (Guidelines for annotators)
 - označevalnik: SentenceMarkupSystem
 - VLE:
 - glagolski frazemi (in pregovori) – ID (stopiti na žulj)
 - zveze s pomensko izpraznjenimi glagoli – LVC (sprejeti odločitev, biti v dvomih ...)
 - predložni glagolski frazemi/frazni glagoli – VPC (priti do – 'doseči; trpeti za - 'oboleti')
 - drugi glagolski frazemi - OTH



JANES IJS FILOZOFSKA
FAKULTETA

- **anliza predločnih glagolskih zvez z vidika prepoznavanja večbesednih leksikalnih enot**
- izbira korpusa/vzorca
- orodje za analizo: SkE: besedne skice
- izhodišče raziskave: glagol + (konkretni) predlog
- CQL: [tag="Gg.*"] [word="pod"] ali [tag="Gg.*"] [tag="D.*"]
- urejanje konkordančnega niza
- prenos v excelovo datoteko
- analiza ...

- sedeti pod borovcem
- [popustiti, sesesti se, klekniti ...] pod bremenom
- [dogajati, imeti, potekati ...] pod budnim očesom (koga/česa)
- podpisati se pod kaj – 'strinjati, podpirati,
- ...

- ČERMÁK, František, 2001: Substance of idioms: perennial problems, lack of data or theory? *International Journal of Lexicography* 14/1. Oxford: Oxford University Press. 1–20.
- ČERMÁK, František, 1985: Frazeologije a idiomatika. František Čermák, Josef Filipec: *Česká lexikologie*. Praha: Academia. 166–248.
- FIRTH, John, R., 1957: Modes of Meaning. *Papers in Linguistics*. London: Oxford University Press. 1934–51.
- GANTAR, Polona, Iztok KOSEM and Simon KREK, 2016: Discovering Automated Lexicography: The Case of the Slovene Lexical Database. *Int J Lexicography* (2016) 29 (2): 200–225 doi:10.1093/ijl/ecw014
- Gantar, Polona, 2007: Stalne besedne zveze v slovenščini: korpusni pristop. Ljubljana, Inštitut za slovenski jezik Frana Ramovša ZRC SAZU
- GANTAR, Polona, 2008: (Slovenska) leksika med leksikonom in slovnico. *Jezik in slovstvo* 53/5. 19–35.
- HANKS, Patrick, 2013: *Lexical Analysis: Norms and Exploitations*. Cambridge: MIT Press.
- KREK, Simon, 2012: New Slovene sketch grammar for automatic extraction of lexical data. *SKEW3, tretja mednarodna delavnica orodja Sketch Engine, Brno, Češka, 21–22. marec 2012*.
- KREK, Simon, 2015: Leksikografska orodja za slovenščini: slovnica besednih skic. Gorjanc, V., Gantar, P., Kosem, I. in Krek, S. (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL. 358–378.
- KREK, Simon in KILGARRIFF, Adam, 2006: Slovene Word Sketches. Erjavec, T., Žganec Gros, J. (ur.): *Proceedings of the 5th Slovenian and 1st International Languages Technology Conference*. Ljubljana, Slovenia. 62–67.
- KRŽIŠNIK, Erika, 1990: Tipologija frazeoloških prenovitev v Cankarjevih proznih besedilih. *Slavistična revija* 38/4. 399–420.
- KILGARRIFF, Adam, 1997: »I Don't Believe in Word Senses.« *Computers in the Humanities* 31(2). Str. 91–113.
- SINCLAIR, J. M., 1991: *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Verdonik, D. (2015): Jezikovnoteoretska načela v korpusnem jezikoslovju. *Slovenščina* 2.0, 3 (1): 1–27.
- ŽELE, Andreja, 2012: *Pomensko-skladenjske lastnosti slovenskega glagola*. Zbirka *Linguistica et philologica*. Ljubljana: Založba ZRC, ZRC SAZU