

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 ... ?

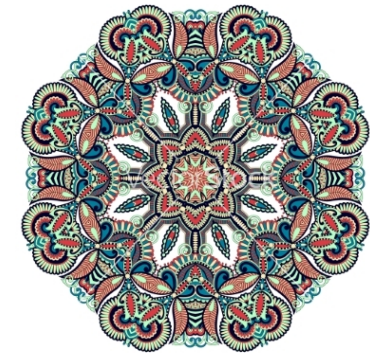


- Viri, orodja in metode za analizo nestandardne spletne slovenščine
 - nacionalni temeljni projekt ARRS
 - 2014-2017
 - <http://nl.ijs.si/janes/>
- 2 instituciji, 9 raziskovalcev
 - Filozofska fakulteta
 - Darja Fišer
 - Jaka Čibej
 - Špela Arhar Holdt
 - Ana Zwitter Vitez
 - Damjan Popič
 - Polona 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 govorjeno 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, vseč,
ravnokar, 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 materijal?

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

- 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

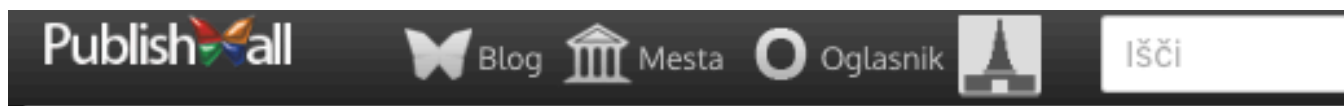
+1

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.

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



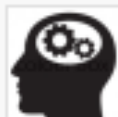
Pametnjakoviči

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



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

Taki osebki se nekaj časa potuhnjeno



abram1b2 Cej pred 3 dnevi



pametnjakovič

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

#38

- 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:Discussione:Lingua slovena](#) se z Boraczekom sprašujeva 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 izgovoriš 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 bega 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 v šolo in potem iz šole, ter na 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 stežka razumejo. Zato se uporablja takoimenovana knjižna slovenščina, ki je nekakšna slovenska različica esperanta in se jo večinoma srečuje le v knjižni obliki ter javnih obcih. Najbližja knjižni slovenščini je bojda novomeška dolensščina.

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 naredil jezikovne napake. --**Boraczek** 10:46, 17 jul 2004 (CEST)



Ljubljana



Slovenija



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 zavaja 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 (Ljubešić 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: hešteg #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 %

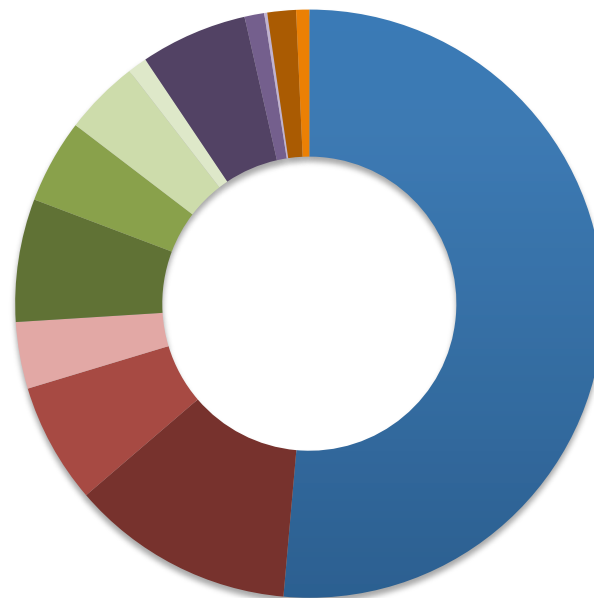
- metapodatki: lastni XML
- anotacije: TEI P5

```
<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="#Agpmpn">mali</w>
    </reg>
  </choice>
  <pc ana="#Z">.</pc>
</s>
```

Korpus JANES

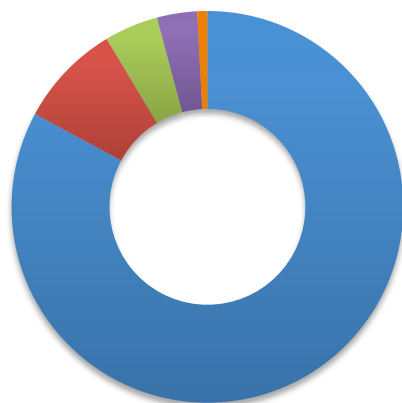
ANALIZA KORPUSA

Št. pojavnic



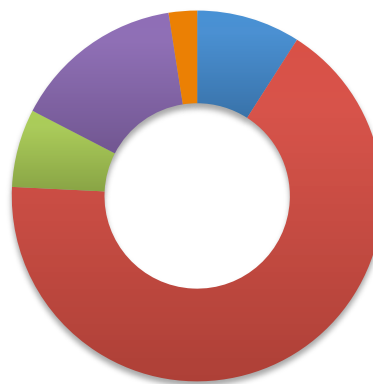
- tweet
- f.avtomobilizem
- f.medovernet
- f.kvarkadabra
- b.rtv slo.comment
- b.rtv slo.post
- b.publishwall.post
- b.publishwall.comment
- n.rtv slo.comment
- n.mladina.comment
- n.reporter.comment
- w.usertalk
- w.pagetalk

Št. besedil



- tweet
- forum
- blog
- news
- wikipedia

Št. avtorjev



- tweet
- forum
- blog
- news
- wikipedia

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

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

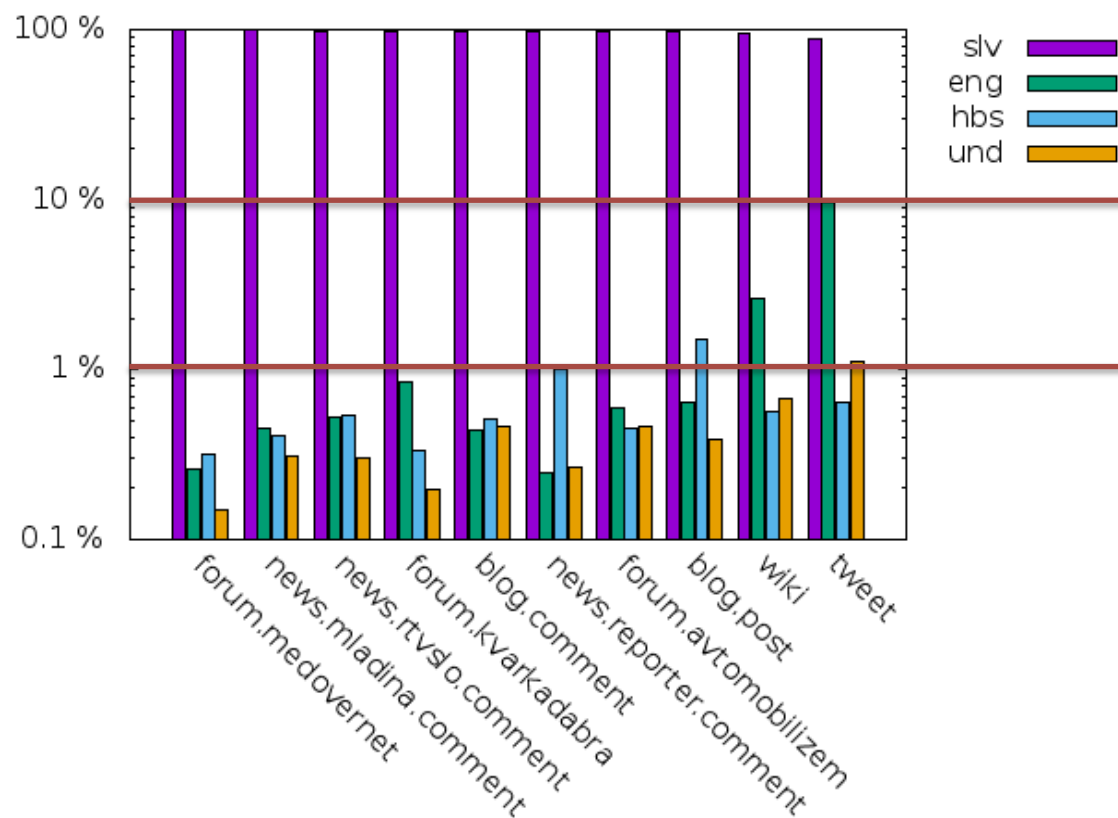
- na nivoju uporabnika

- spol
 - tip
 - regija (samo za tvite)

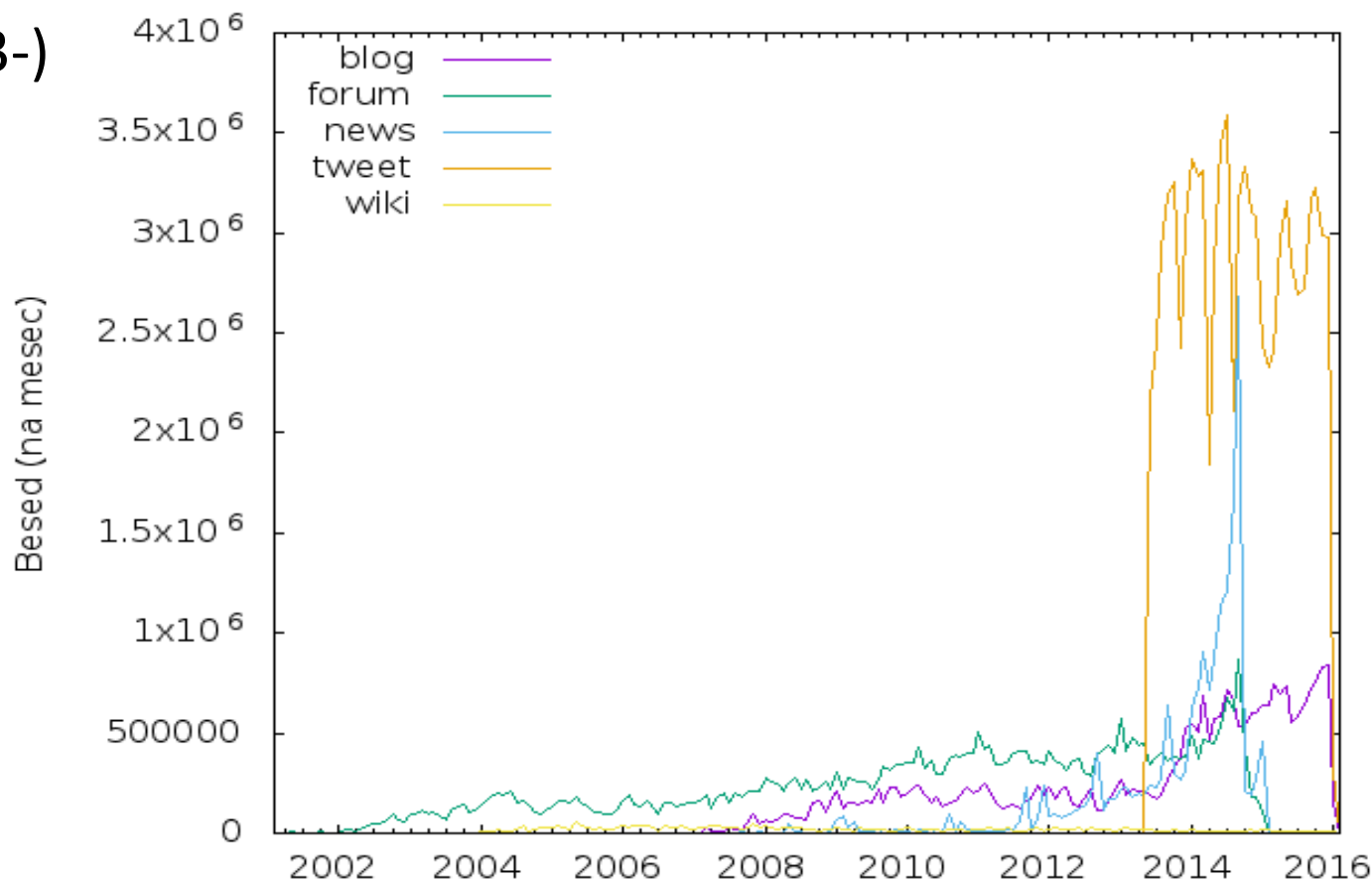
- na nivoju besedila

- jezik
 - sentiment
 - stopnja standardnosti

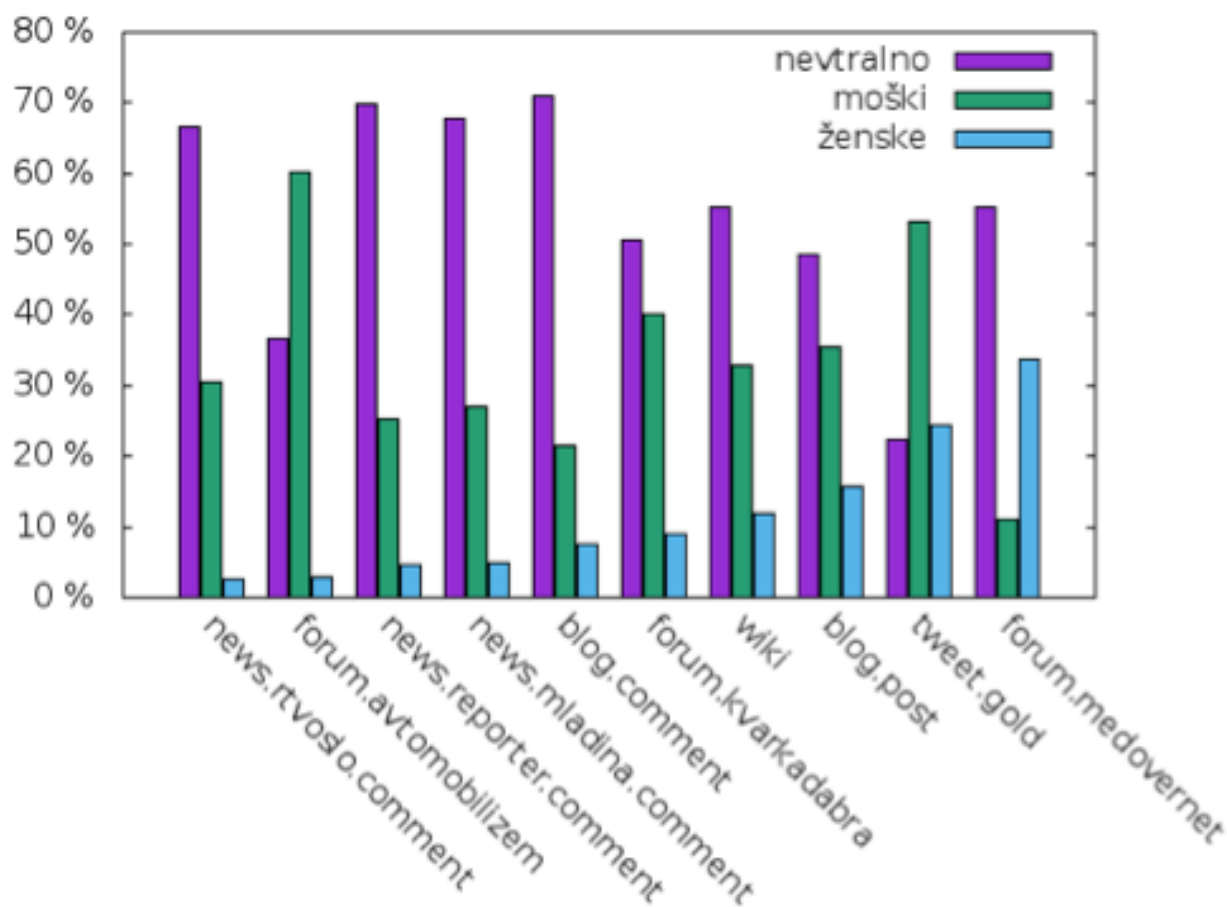
- detekcija:
 - *langpy*
 - slv, eng, hbs, und
- > 1 % tujejezičnih besedil samo wiki & tviti
- wiki: 2,6 % ang besedil
- tviti:
 - 9,6 % ang besedil
 - 1,1 % drugo



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

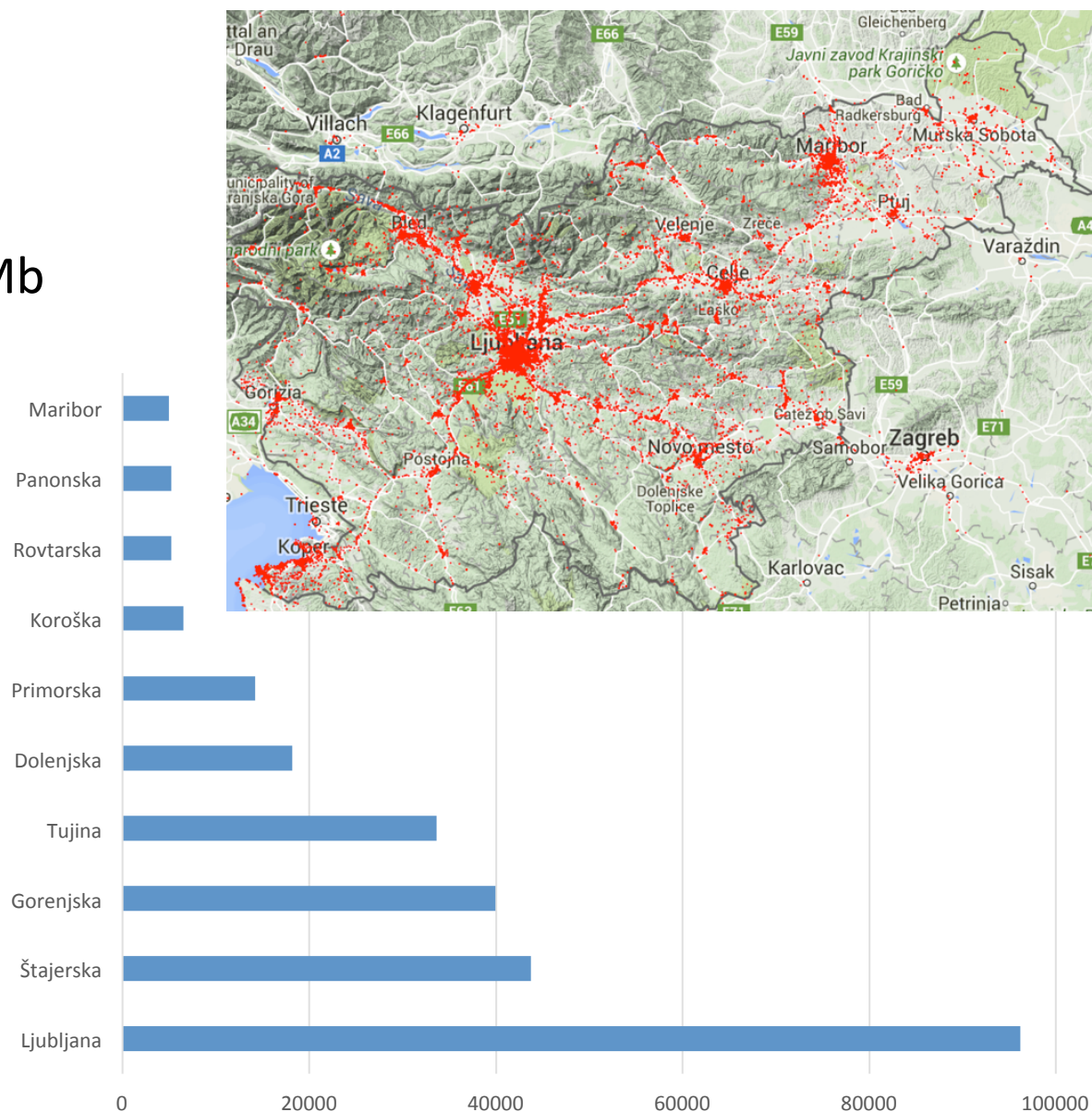


- detekcija:
 - 1. os. ed. pom. gl. + deležnik na -l (*sem/nisem/bom mislil/a*)
 - > 0.7 odkritih ž/m
 - > 1 % besedil
- evalvacija (tviti):
 - 76 % natančnost
 - 5 % napačni spol
 - 19 % nevtralni spol
- komentarji prevladuje N
- tviti & avtomobilizem prevladuje M
- medovernet prevladuje Ž



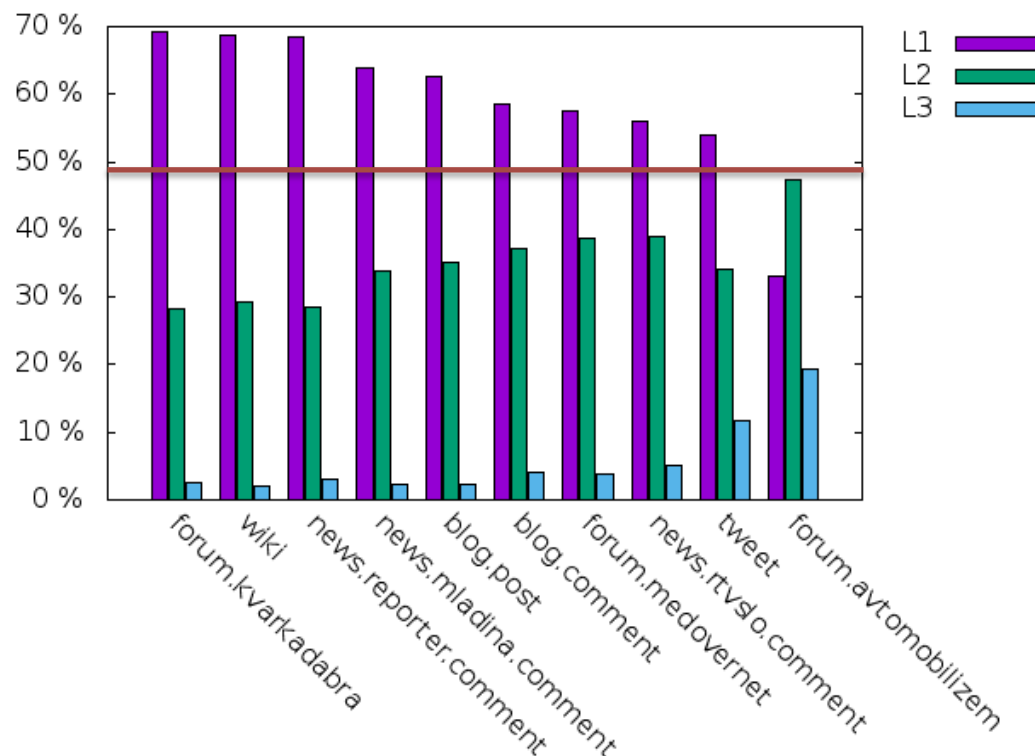
- tip avtorja
 - osebni računi posameznikov (prosti čas)
 - uradni računi medijskih hiš, institucij, podjetij (profesionalna raba)
- označevanje
 - ročno (analiza profila uporabniškega računa & zgodovino objav)
 - tviti (blogi kmalu)
- rezultati
 - 76 % zasebnih uporabnikov
 - 24 % korporativnih uporabnikov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

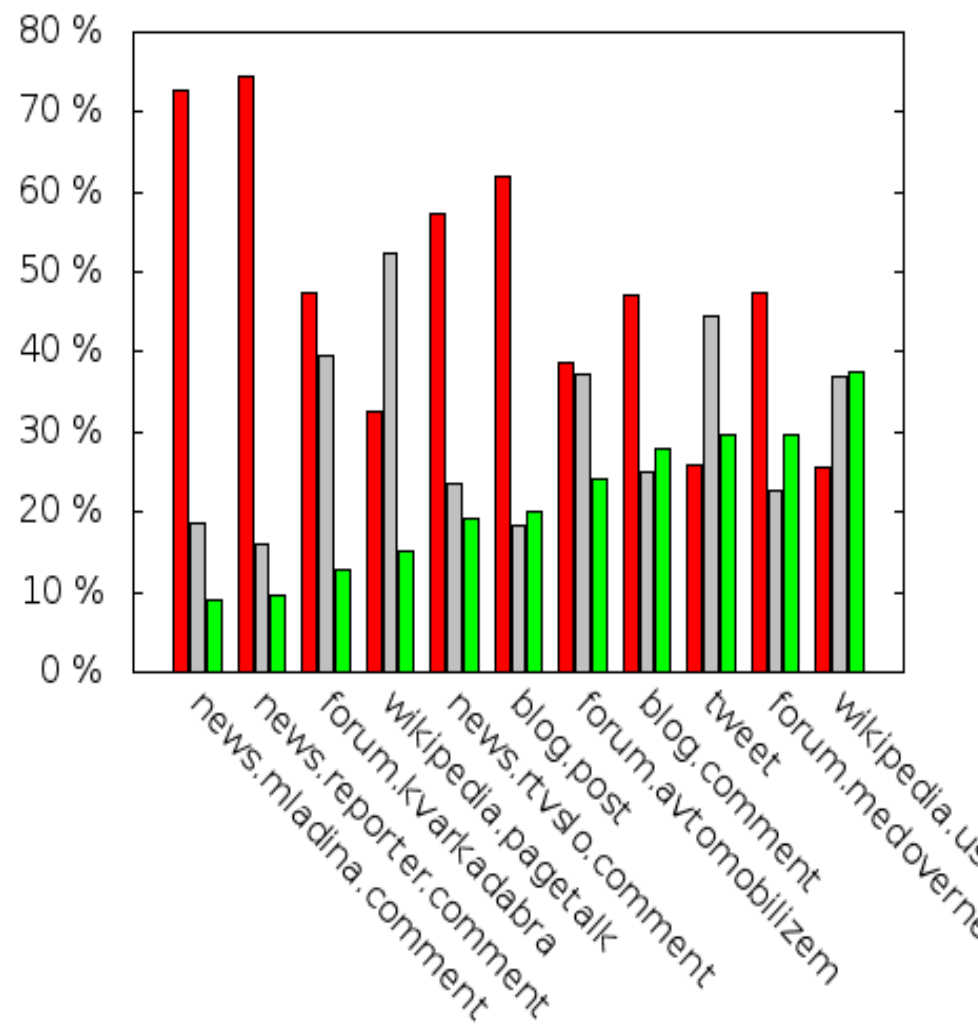


- stopnje standardnosti (Ljubešić et al. 2015):
 - tehnična & lingvistična 1-3
- učenje modela:
 - ročno označenih 1.200 besedil
- evalvacija:
 - povp. abs. n. 0,38 T / 0,42 L
- rezultati:
 - najbolj nestandarden avtomobilizem (20 % L3)
 - precej nestandardni tudi tviti (12 % L3)
 - najbolj standardni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

Podkorporus	T=1 / L=3	T=3 / L=1
tweet	<i>A nis bla včer na Bledu?</i>	<i>komunistična ideologija ubijaj,kradi laži.....zelo primerna za aktualno vlado,,,,,</i>
news.comment	<i>Men so drugač vsi ful lepi, ampak zver je pa ekstra kjut. Pa ful lep nasmešek ma. Pa obrvi..</i>	<i>Zadeva je nerodna in zgled zelo slab ,kar se tiče ostalih članic ,ki prav tako visij (m)o na nitki !</i>



- sentiment (Smailović 2014):
 - + / - / 0
- učenje modela:
 - 5000 ročno označenih tvitov
- 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 podkorporusov



Korpus JANES

NERAZREŠENI PROBLEMI

- Sestava korpusa
 - reprezentativnost
 - celovitost
 - uravnoveženost
- Metapodatki
 - dodajanje starosti uporabnikov
 - podkorpus politikov, mikrozvezdnikov, 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

[**http://nl.ijs.si/janes/**](http://nl.ijs.si/janes/)

tenks 😊