Razvoj korpusa slovenskih spletnih uporabniških vsebin Janes

Tomaž Erjavec[2] Darja Fišer[1], Nikola Ljubešić[2, 3]

[1]Oddelek za prevajalstvo, FF UL
[2]Odsek za tehnologije znanja, Inštitut "Jožef Stefan"
[3]Oddelek za informatiko in družboslovje, FF UZG

Konferenca
"Slovenščina na spletu in v novih medijih"
25.–27. november 2015

Pregled

- Ozadje
- Korpus Janes
- Podkorpus tvitov
- Tekoče delo
- Prihodnje delo

Ozadje

Korpus kot osnova za

- Jezikoslovne raziskave
- Izdelavo orodij za procesiranje

1. Gradnja
2. Avtomatsko označevanje
3. Ročno označevanje (testna in učna množica za 2.)
4. Nazaj na 1.
   - **JANES v0.3**
   - **JANES Tviti v0.3.4**

Korpus JANES v0.3

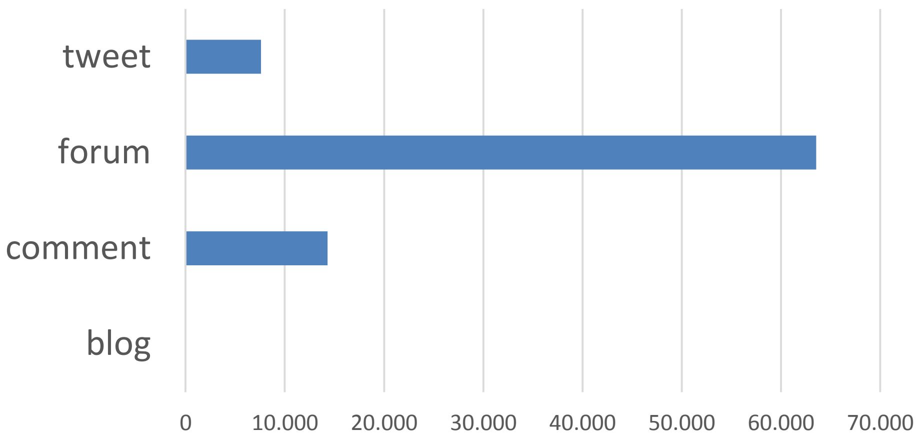
- Tviti
  - Velik del slovenske tvitosfere
  - TweetCat (Ljubešić et al. 2014), sprotni zajem
  - metapodatki: uporabniško ime, čas objave, št. všečkov in posredovanj
- Forumska sporočila
  - 3 forumi: med.over.net, avtomobilizem.com, kvarkadabra.net
  - namenski ekstraktorji besedila, enkratni zajem
  - metapodatki: uporabniško ime, tema, URL, čas in ID objave
- Komentarji na novice
  - 3 portali: RTV Slo, Mladina, Reporter
  - namenski ekstraktorji besedila, enkratni zajem (tudi novice)
  - metapodatki: uporabniško ime, čas in ID komentarja, URL in ID novice
- Blogi
  - slWaC 2.0 (Erjavec in Ljubešić 2014)
  - ime domene vsebuje "blog"
  - brez metapodatkov, mešano besedilo bloga in komentarji nanj

Obdelava korpusa

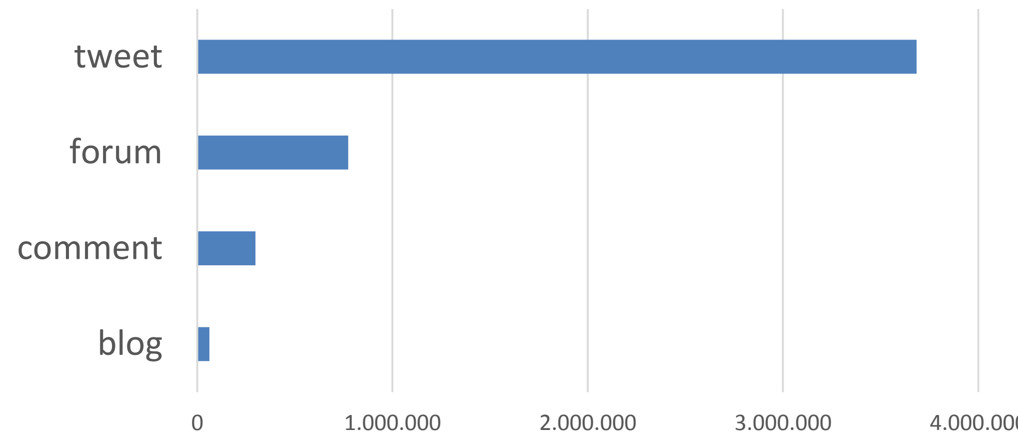
- Označevanje
  - (skoraj standardna) tokenizacija in stavčna segmentacija
  - standardizacija leksike z metodo CSMT (Ljubešić et al. 2014)
  - standardno oblikoskladenjsko označevanje in lematizacija
- Zapis
  - vertikalni format (za konkordančnik)
  - XML TEI P5 (delno)
- Konkordančnik
  - (no)Sketch Engine

Sestava korpusa JANES

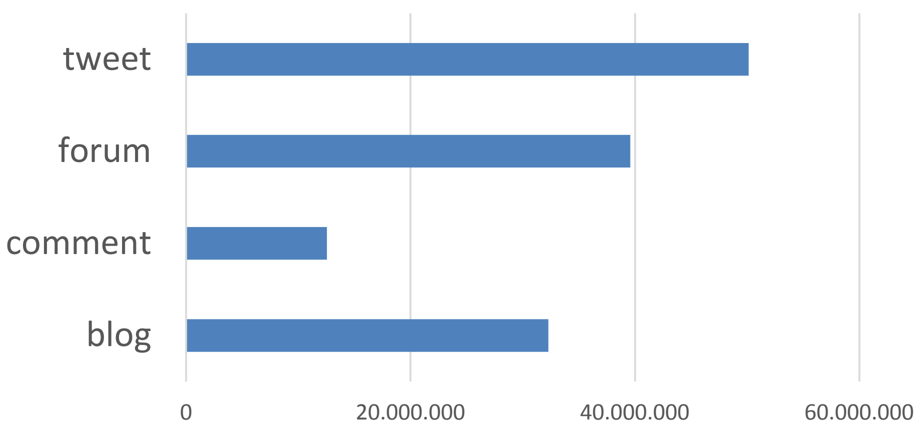
85.429 avtorjev



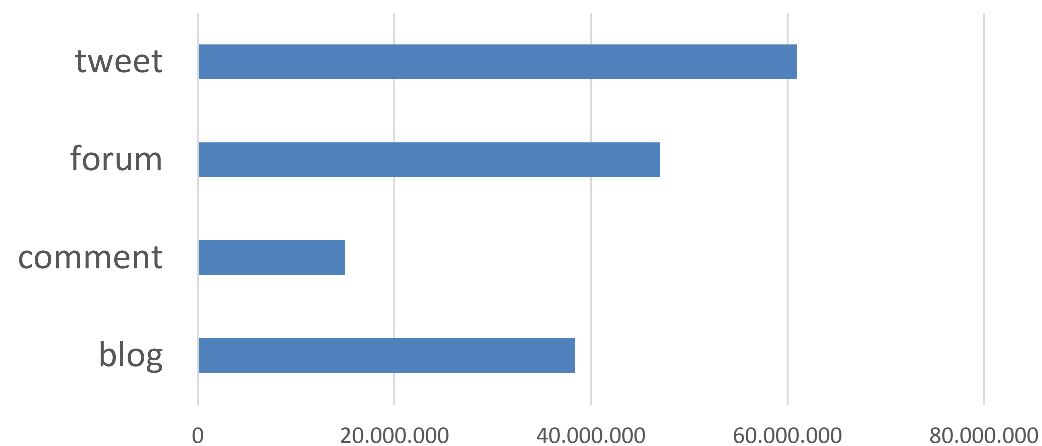
4.819.558 besedil



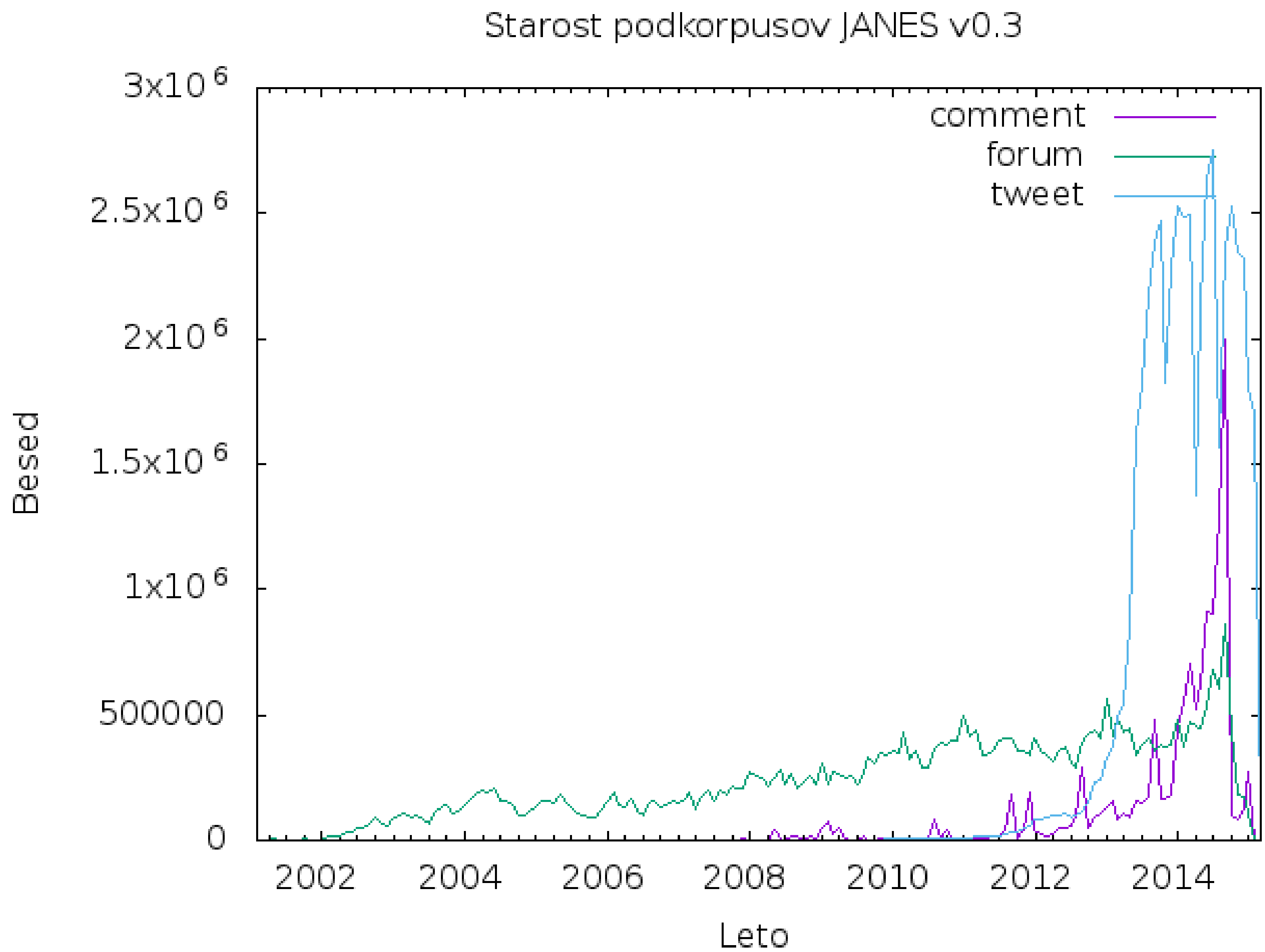
134.543.613 besed



161.289.153 pojavnic



Starost besedil



noSketchEngine: “boljše”

Iskalni niz boljše **27,257** > Premešaj **27,257** (169.0 na milijon)

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

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

Prva | Prejšnja Stra

text.type	forum
text.author	Goggy
text.title	VW Passat BKP trese - NUJNO POMOČ
text.date	2011-03-15
text.url	http://www.avtomobilizem.com/forum/viewtopic.php?f=6&t=84268&start=20#p1423667
text.id	janes.forum.avtomobilizem.6.84268.1423667

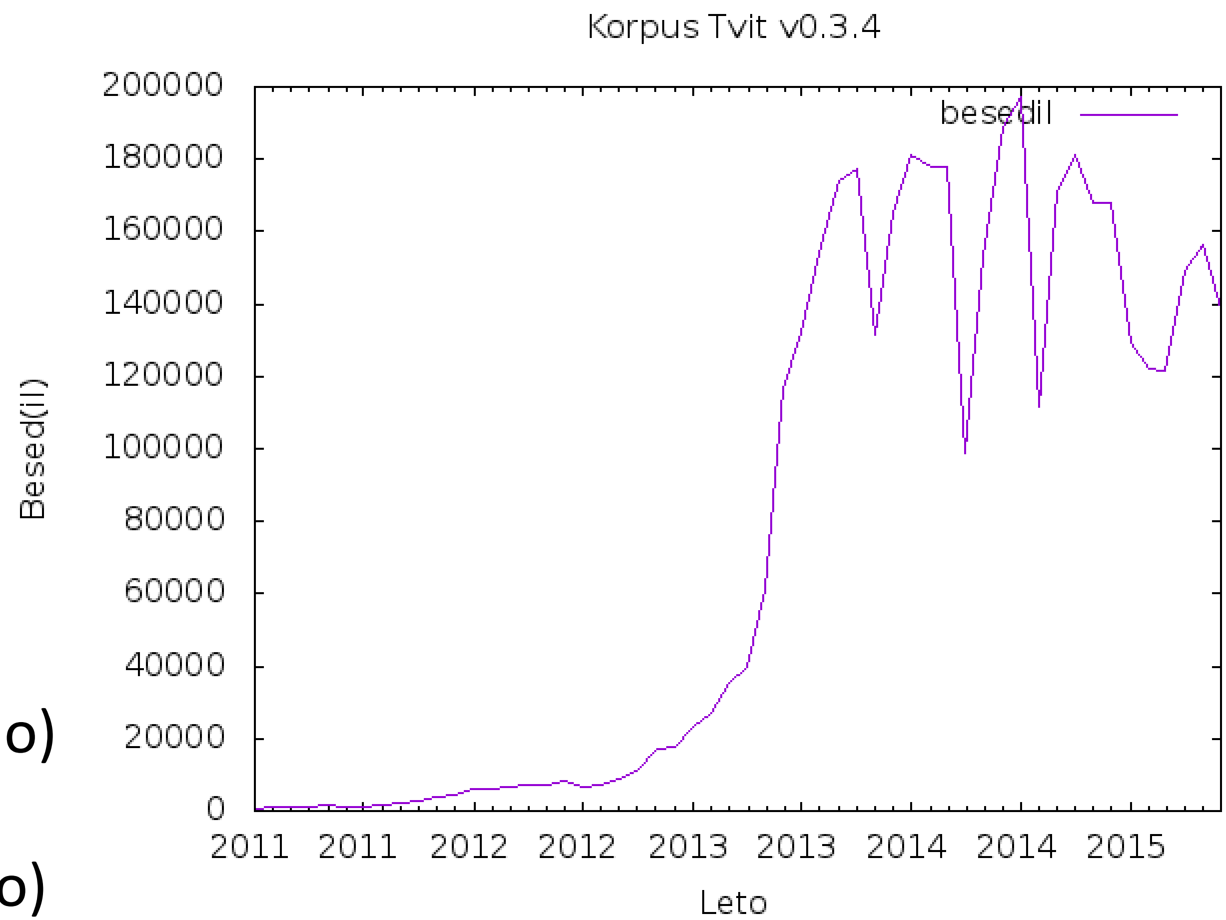
noSketchEngine: “jaz”

	word	Frekvenca
p | N	jaz	224,166
p | N	jst	15,654
p | N	js	11,256
p | N	jes	1,034
p | N	jez	308
p | N	vaz	208
p | N	jezt	62
p | N	ges	22
p | N	jaaz	10
p | N	jzt	9
p | N	ioz	8
p | N	jaaaaz	7
p | N	naz	6
p | N	jaaaz	6
p | N	jiz	4
p | N	joz	2
p | N	jaaaaaaz	2
p | N	iaz	2
p | N	iez	1

Tviti v0.3.4

Novi tviti:

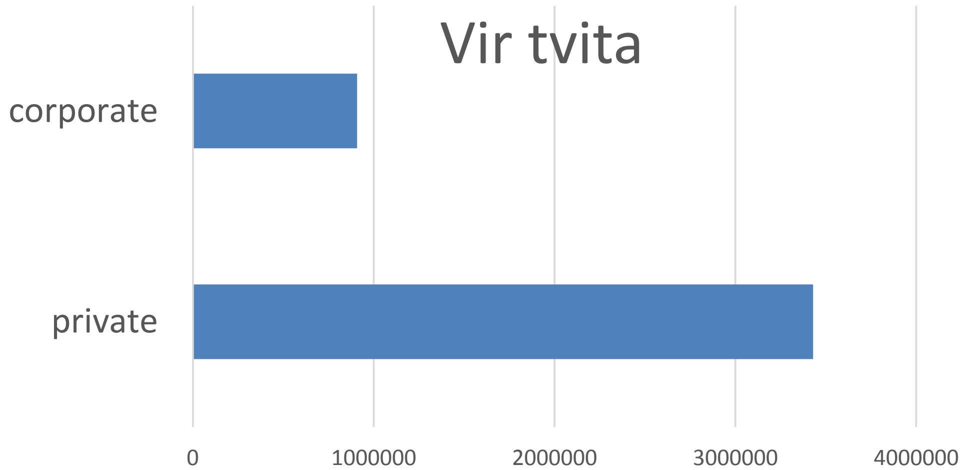
Pojavnic	70.104.566
Besed	52.774.593
Tvitov	4.337.767

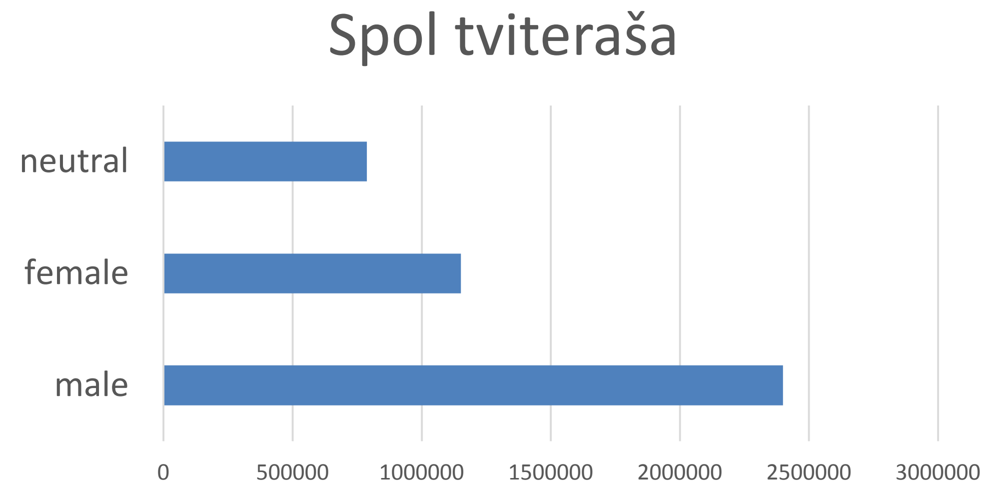


Novi metapodatki

- za avtorje:
  - zasebni oz. organizacija (ročno)
  - spol (ročno)
  - regija (avtomatsko)
- za tvite:
  - stopnja standardnosti (avtomatsko)
  - sentiment besedila (avtomatsko)

Tvitov po avtorjih





Geolokacija:
6,2 % označenih
Regije: J. Čibej, FF

Stopnja standardnosti

- Ljubešić et al., RANLP 2015
- Tehnična in jezikovna: T1 – T3, L1 – L3
- Ročno označevanje
- Učenje linearnega regresorja
- Evalvacija: povp. abs. napaka = 0.377 T in 0.424 L

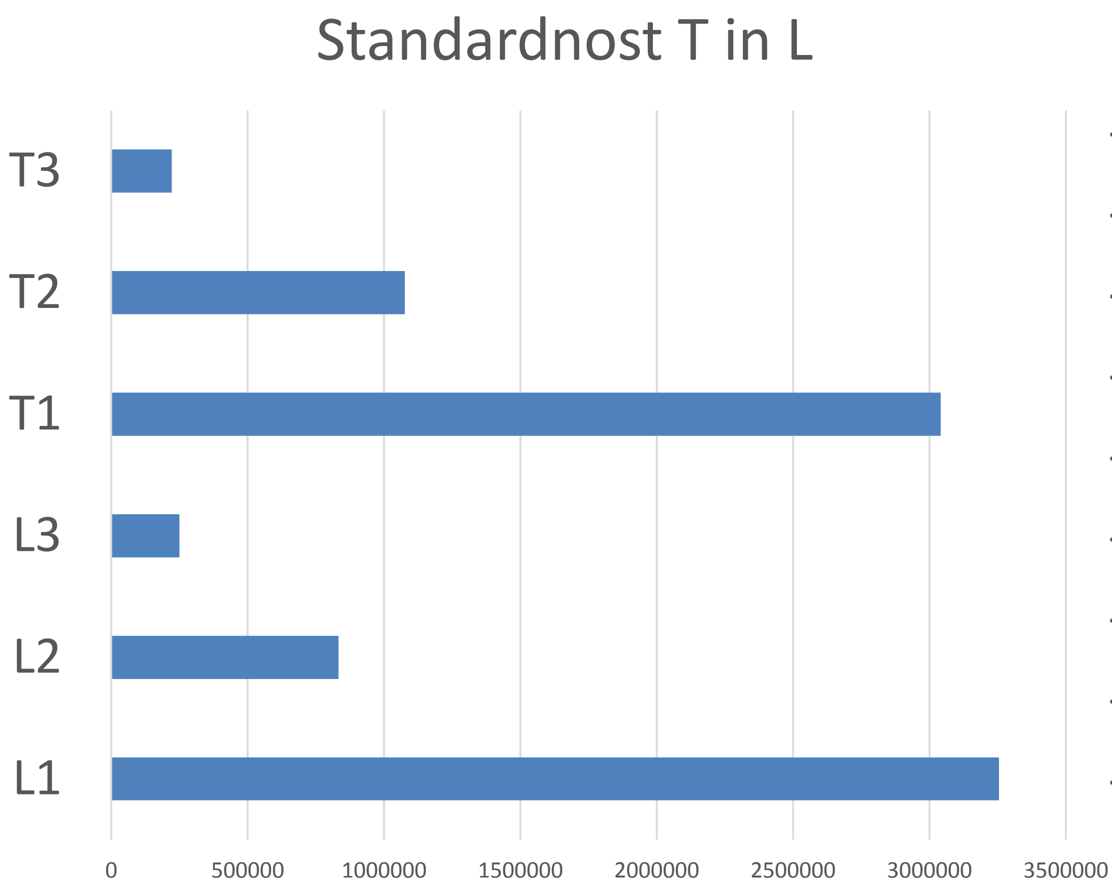
T=1 / L=3

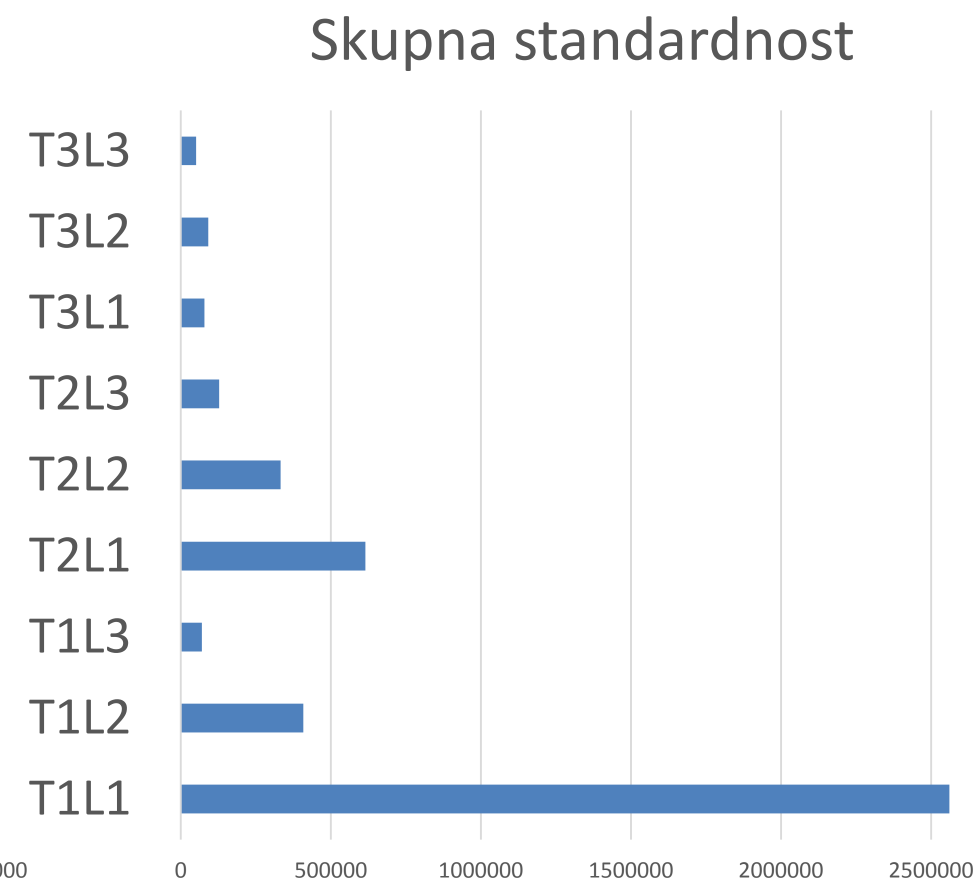
Original: *Ma men se zdi tole s poimenovanji oz s poslovenjenjem imen mest čist mem.*
Standardised: *Meni se zdi to s poimenovanji oz. s poslovenjenjem imen mest čisto mimo.*

T=3 / L=1

Original: *se pravi,da predvidevaš razveljavitev*
Standardised: *Se pravi, da predvidevaš razveljavitev?*

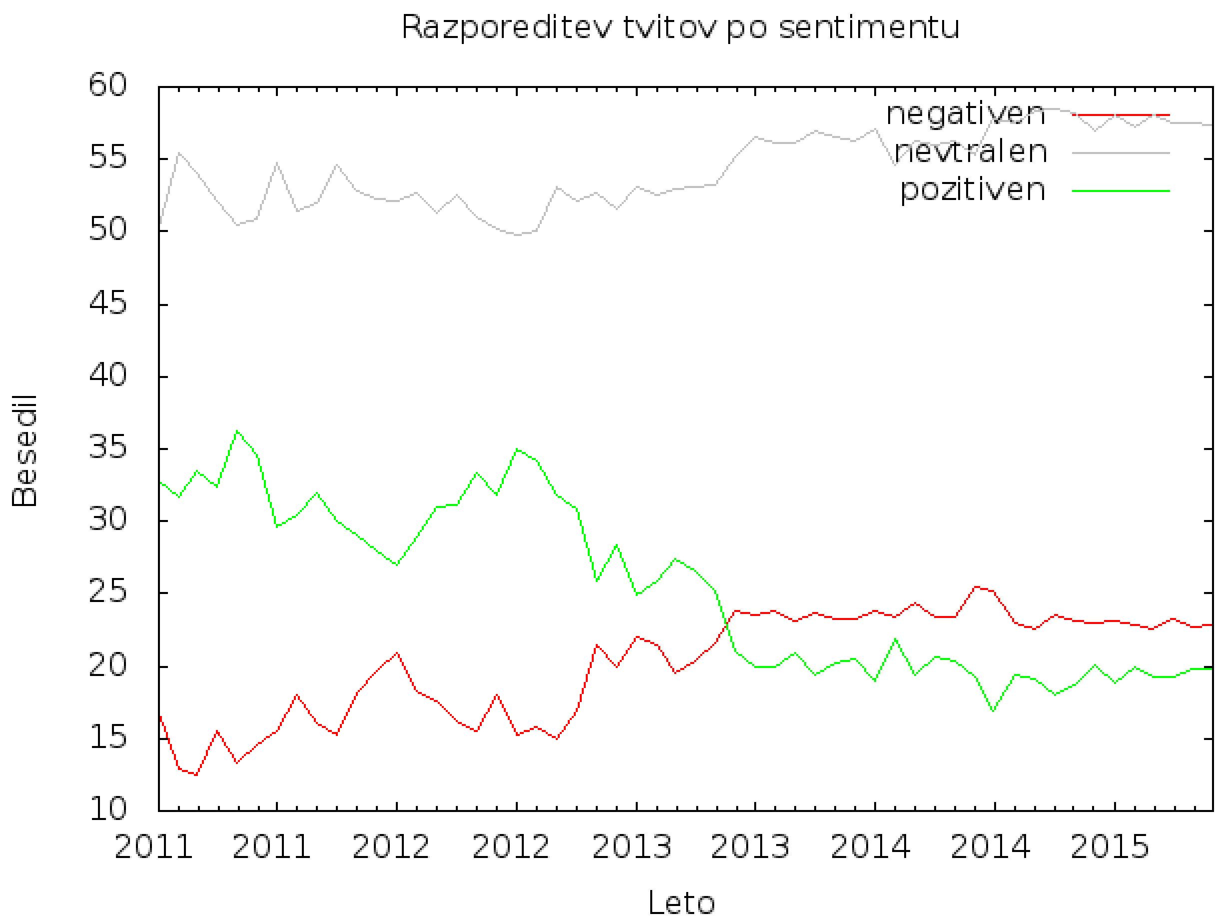
Standardnost





Sentiment besedila

- Avtomatsko anotiranje: SVM + označeni sl tviti (Jasmina Smailović, IJS)
- Evalvacija na 1.000 tvitih (šport in politika, 2 anotatorja)
  - naključno = 37,7 %
  - vse anotacije = 57,3 %
  - oba anotatorja = 62,1 %
  - IAA = 76,5 %



V delu: rediakritizacija

- Neproblematično zagotoviti veliko učno množico, vendar v splošnem problem ni trivialen: *„Problem je resen"*
- *Corpus-based diacritic restoration for South Slavic languages*. Ljubešić et al., oddano na LREC'16:
  - uporaba Wikipedije, splošnega spleta in Twitterja za hr, sr in sl
  - evalvacija nad Wikipedijo in Twitterjem
  - rediakritizacija: slovar + kontekst *„Kaj se smejes?"*
  - slovenščina: osnovnica 88 %, charlifter 93 %, leksikon 98,2 %, leksikon + kontekst 99,1 %
  - tviti-sl: 37 % dvoumnost, 31 % tokenizacija, 6 % ime, …
- Potrebno še uporabiti za korpus JANES v0.4

V delu: označevanje tvitov

- Cilj: korpus, v katerem so tviti pravilno:
  1. **tokenizirani in stavčno segmentirani**
  2. **standardizirani**
  3. oblikoskladenjsko označeni in lematizirani
- Referenčni korpus za testiranje in šolanje
- Navodila za označevanje
- Šolanje anotatorjev
- Izvedba tečaja
- Resno delo

Institut "Jožef Stefan" , Ljubljana, 13.–14. november 2015

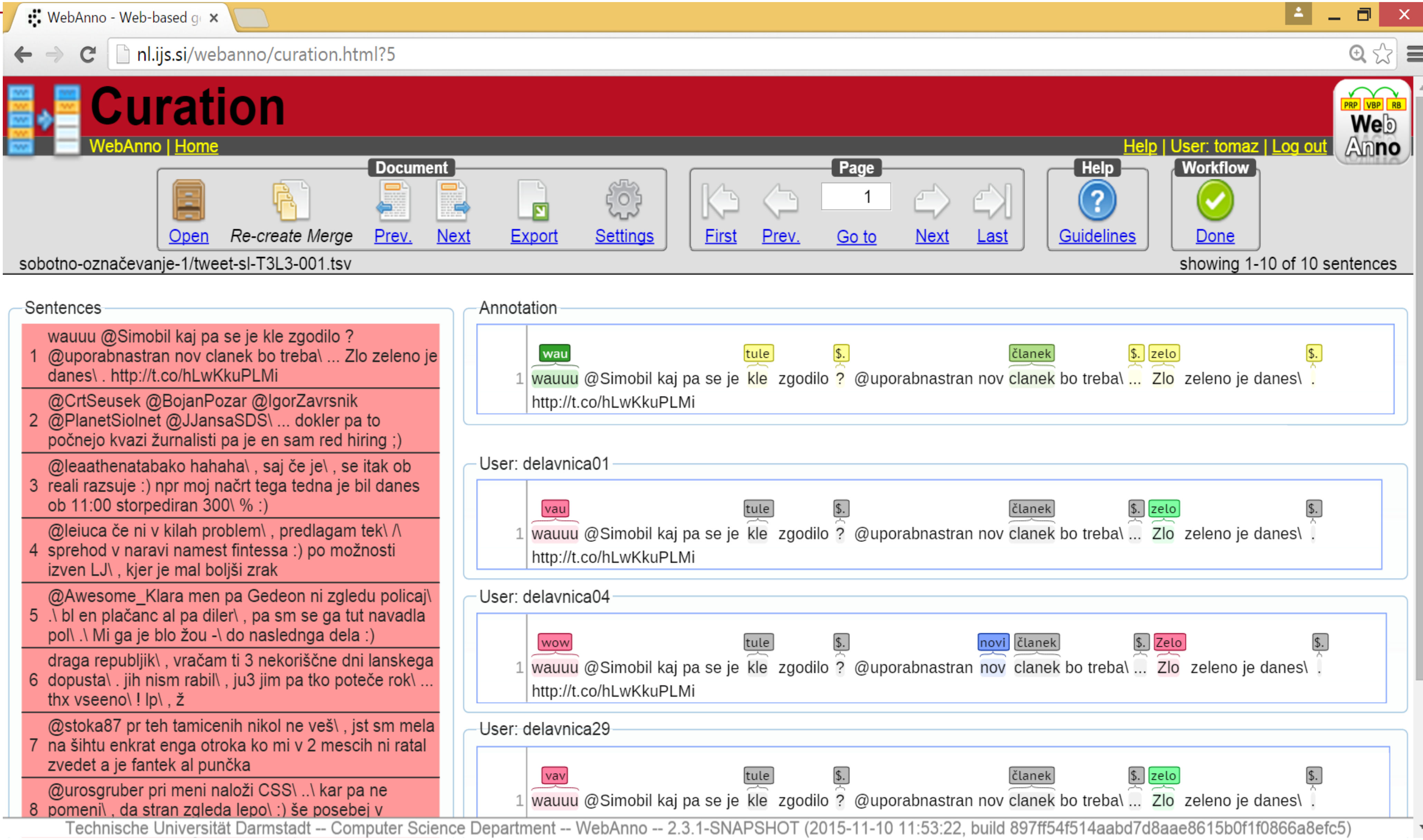
Petek, 13. november

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

Sobota, 14. november

10.00–10.30	**Tehnični vidiki označevanja nestandardnih jezikovnih prvin v WebAnno** (Tomaž Erjavec)
10.30–11.00	**Smernice za označevanje nestandardnih jezikovnih prvin v korpusu JANES** (Jaka Čibej)
11.00–11.30	**Koordinacija označevalcev** (Darja Fišer)
11.30–12.00	Odmor za kavo
12.00–13.30	**Označevanje, 1. del**
13.30–14.30	Kosilo
14.30–16.00	**Označevanje, 2. del**
16.00–16.30	Odmor za kavo

Platforma: WebAnno



Prihodnje delo: ročno označevanje 2

- Oblikoskladenjsko označevanje:
  - nabor oznak (emotikoni, emoti, ključniki)
  - oznake za nestandardne besede („nebom“)
- Lematizacija
- Tuje in „tuje“ besede?
- Besede za slovarček?
- Izvedba označevanja

Prihodnje delo: korpus

- Komentarji z Wikipedije
- Blogi (bolj zares)
- Forumov in novic verjetno ne
- = JANES v0.4 (začetek 2016)

- Tviti:
  - novi metapodatki z vprašalnikom, npr. starost
  - podkorpus javnih osebnosti/politikov

Prihodnje delo: avtomatsko označevanje

- Glavna naloga zadnjega leta projekta
- Izboljšati na vseh ravneh:
  - tokenizacija, normalizacija, oblikoskladenjsko označevanje, lematizacija
  - ročno označeni podatki
  - boljši programi (RE, CRF)
- = JANES v1.0

Prihodnje delo: distribucija

- Obljuba:
  - prosto dostopno na konkordančniku
  - odprto dostopno na repozitoriju CLARIN.SI
- Vendar:
  - avtorske pravice
  - zaščita zasebnosti
  - pogoji uporabe
- Rešitve:
  - anonimizacija, premešanje, vzorčenje
  - mogoče različni pogoji za različne uporabnike
  - + spremljevalni korpus