Darja Fišer

1 PREDSTAVITEV PROJEKTA IN KORPUSA JANES

1.1 Projekt in korpus JANES – izročki

JANES IJS

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

JANES IJS

Predstavitev udeležencev

JAZ JANES, TI ... ?

Projekt JANES

- 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

Delovni sklopi

- 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 tegerja in lematizatorja
  - N3: Izdelava spremljevalnega korpusa

Korpus JANES

IZBOR IN ZAJEM BESEDIL

Izbor virov

Izbor virov

- 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

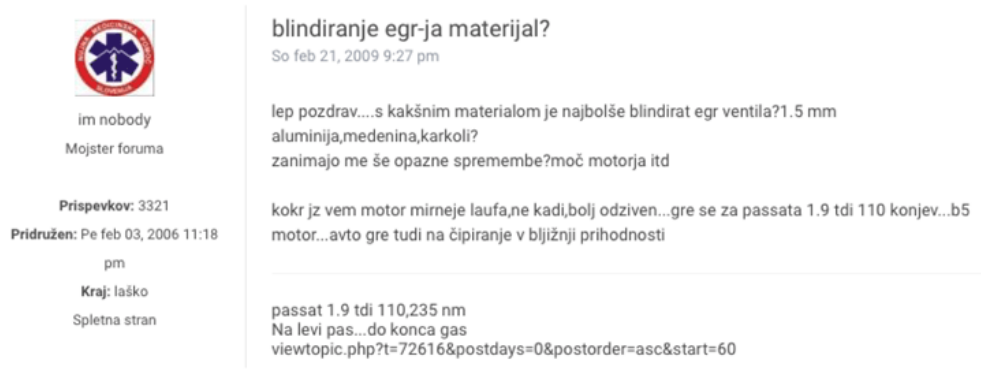
Tviti

- 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, ravnokar, končno, kdaj, preveč, očitno

Forumi

- 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



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

Tweet Share 20 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 zasluzijo. 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

Publish all 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" osebki. Je to neka nova okupacija izven zemeljskih energij, ki okupirajo ubogi um navadnega osebka? Imajo se za strašno vsevedega. Najhuje je, da je tak pametnajkovič sveto prepričan, da ima edini prav .

Taki osebki se nekaj časa potuhnjeno

abram1b2 Cej pred 3 dnevi 0

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

#38

Uporabniške & pogovorne strani na Wikipediji

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

Korpus JANES

OBDELAVA BESEDIL

Koraki

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

Stavčna segmentacija & tokenizacija

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 začenja 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)

Rediakritizacija

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 zasluzijo. 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*

Normalizacija

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 bljižnji 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)

Tegiranje & lematizacija

- 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: heštegi #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 %

Zapis korpusa

- metapodatki: lastni XMl
- anotacije: TEI P5

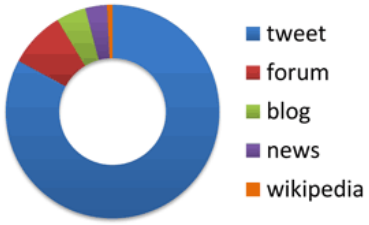
```
<s>
    <w lemma="pa" ana="#Cc">pa</w>
    <c> </c>
    <w lemma="še" ana="#Q">še</w>
    <c> </c>
    <choice>
        <orig>
            <w>tamali</w>
        </orig>
        <reg>
            <w lemma="ta" ana="#Pd-fsn">ta</w>
            <c> </c>
            <w lemma="mali" ana="#Agpmpn">mali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

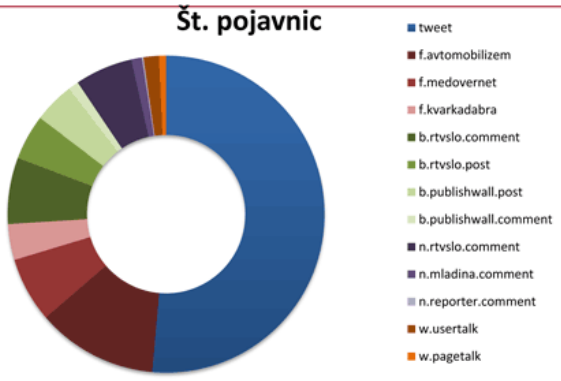
Korpus JANES

ANALIZA KORPUSA

JANES v številkah

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





JANES v številkah

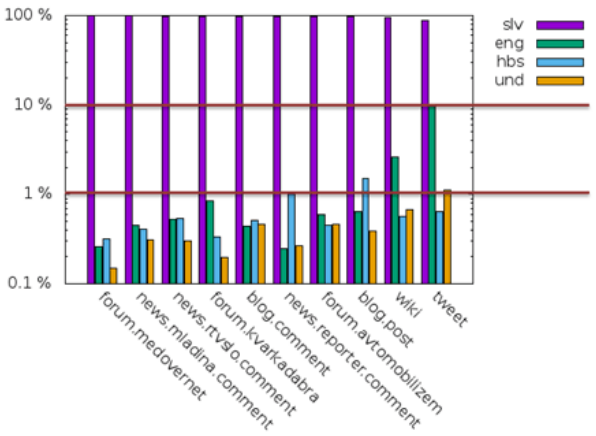
Podkorpus	Pov. št. besed / besedilo	Št. besed/uporabnika	Št. besedil/uporabnika
tweet	**12**	10.307,5	**857,6**
forum	51,4	616,5	12,0
avtomobilizem	38,5	1.713,5	44,5
medovernet	94,7	**234,7**	**2,5**
kvarkadabra	77,1	2.814,1	36,5
blog	71,3	4.373,8	61,3
rtvslo.comment	35,8	3.705,5	103,4
rtvslo.post	343,7	**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

Metapodatki

- 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

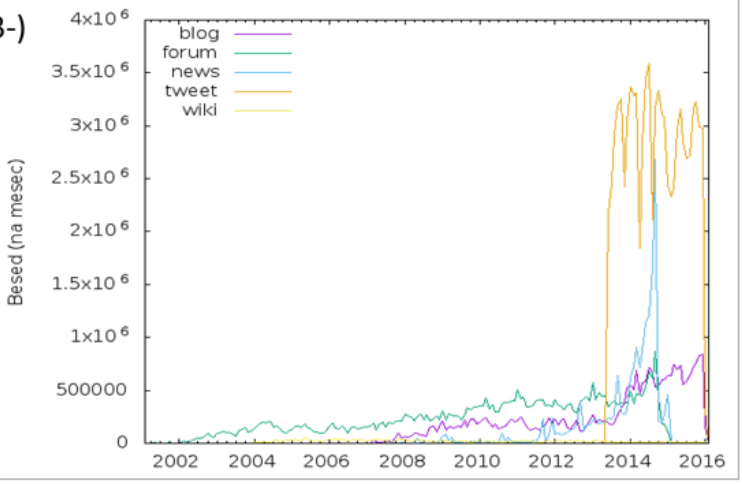
Jezik besedil

- 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



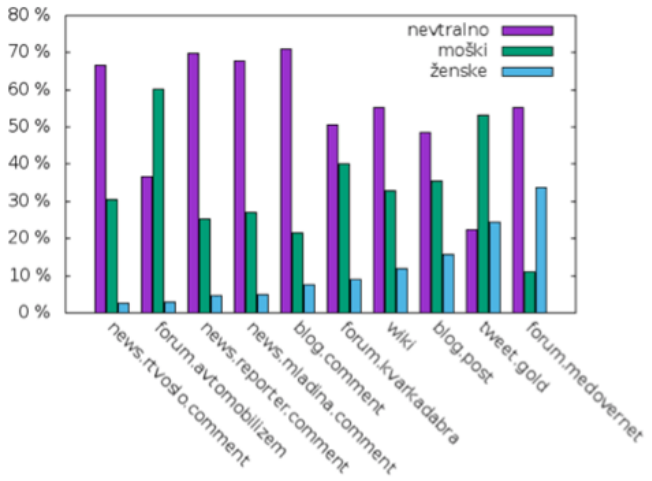
Starost besedil

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



Spol avtorjev

- 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

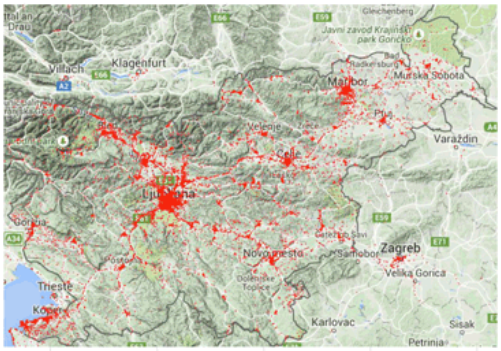


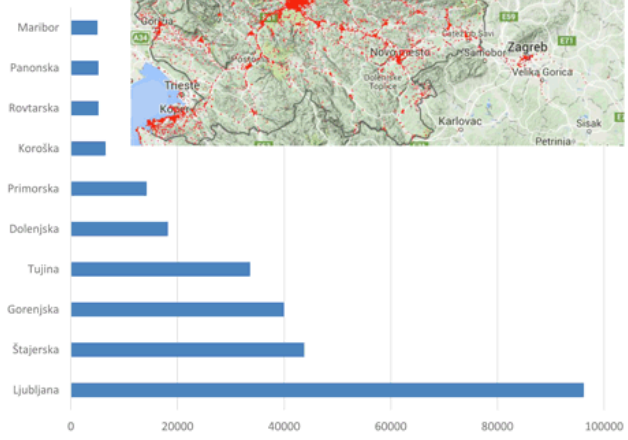
Tip avtorjev

- 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

Regija avtorjev

- 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

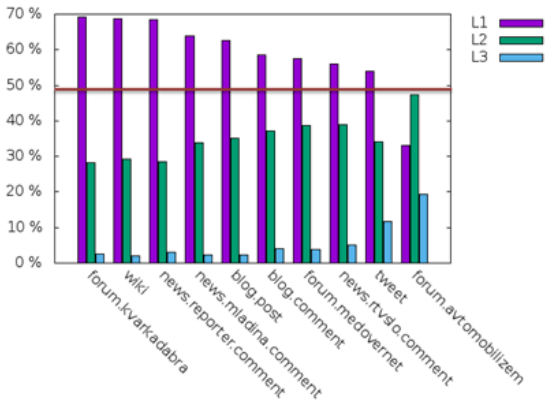




Stopnja standardnosti

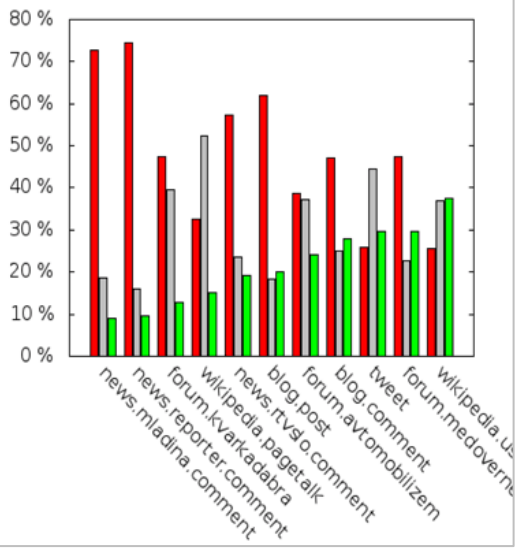
- stopnje standardnosti (Ljubešić et al. 2015):
  - tehnična & lingvistična 1-3
- učenje modela:
  - ročno označenih 1.200 besedil
- evalvacija:
  - povp. abs. n. 0,38 T / 0,42 L
- rezultati:
  - najbolj nestandarden avtomobilizem (20 % L3)
  - precej nestandardni tudi tviti (12 % L3)
  - najbolj standardni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

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



Sentiment besedil

- 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 podkorpusov



Korpus JANES

NERAZREŠENI PROBLEMI

Izboljšave korpusa

- Sestava korpusa
  - reprezentativnost
  - celovitost
  - uravnoteženost
- 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

Objava korpusa

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