# The JANES project: Tools and resources for linguistic analysis and automatic processing of user-generated content in Slovene

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

[1]University of Ljubljana
[2]Jožef Stefan Institute
[3]University of Zagreb

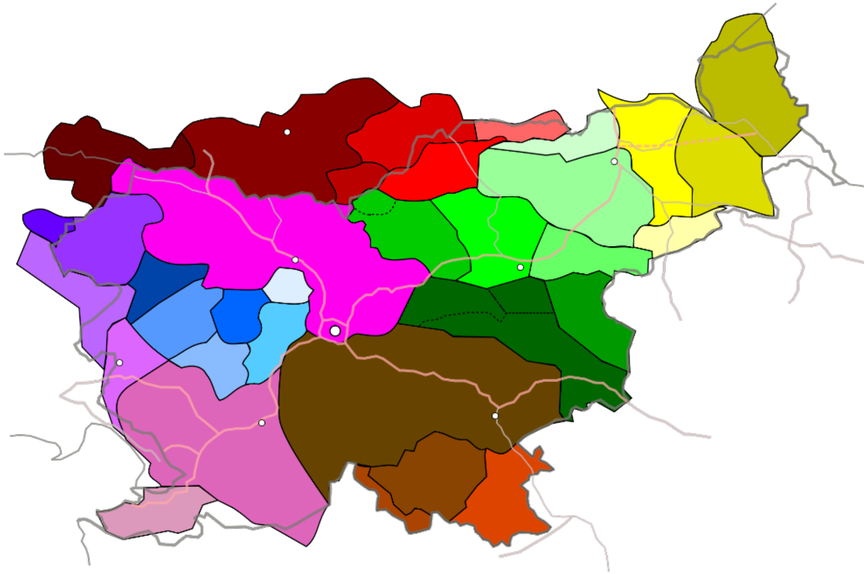
Rennes, 24 October 2015

# Outline

- Project background and goals
- Corpus construction
- Corpus annotation
- Project activities

# Background

- Basic facts
  - 2 mio inhabitants
  - Mediterranean, Alpine, Pannonian and Dinaric regions
  - 7 dialectal groups, 42 dialects
  - prescriptive, normativist culture
- Language resources
  - plenty of corpora (CLARIN.SI & CC)
  - standard language
- The Janes project
  - national basic research project
  - 3 yr, 2014-2017
  - 2 institutions, 8 team members
  - development of resources, tools and methods for the analysis of UGC
  - [http://nl.ijs.si/janes](http://nl.ijs.si/janes)



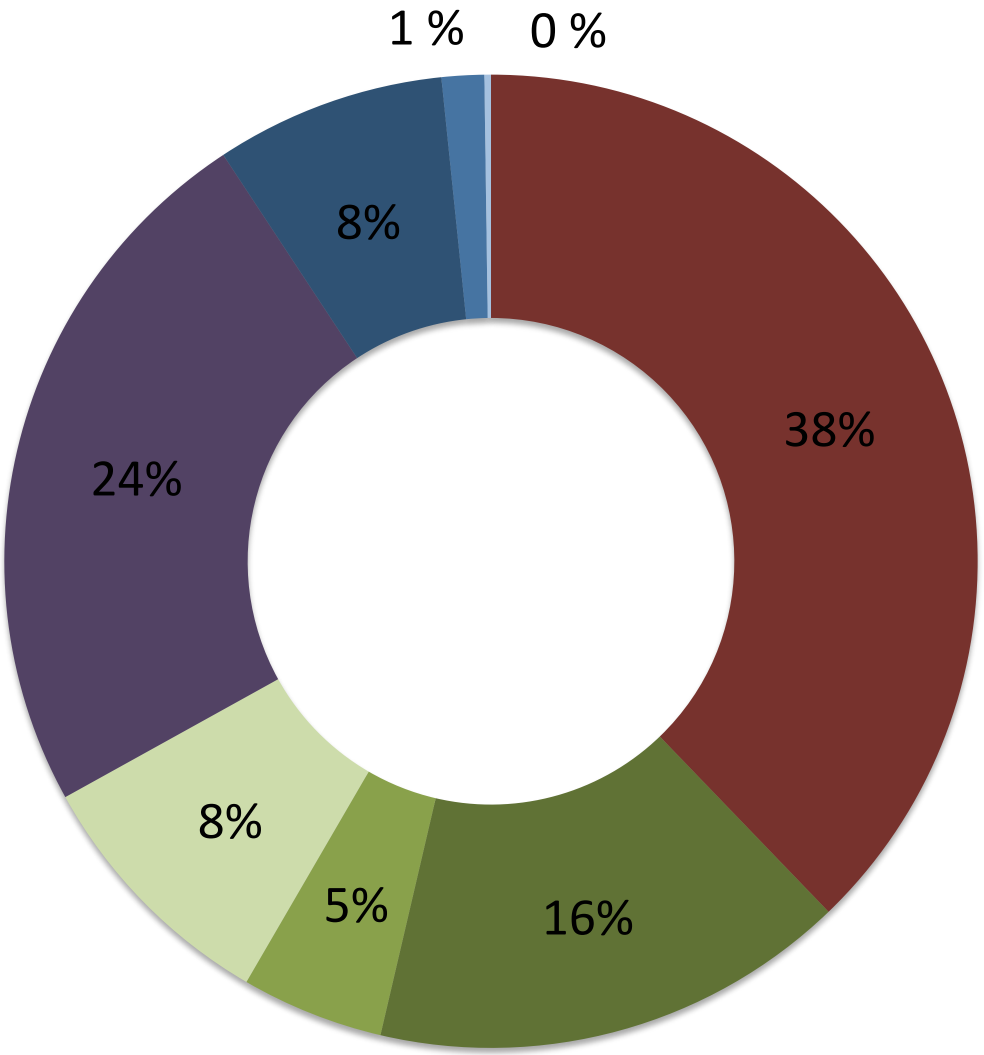
# Research goals

- Work Package 1: Corpus construction
- Work Package 2: Linguistic analysis
  - Task 1: Janes vs. standard Slovene
    - orthography (punctuation, capitalization, spelling)
    - regional variation (PhD)
    - syntax
  - Task 2: Janes vs. spoken Slovene
    - discourse markers
    - interactive elements
  - Task 3: offensive language
  - Task 4: collocations
  - Task 5: terminology
  - Task 6: semantic shifts
- Work Package 3: NLP tools

# Corpus JANES 0.3

- Tweets
  - TweetCat (Ljubešić et al. 2014)
  - Slovene-specific seed words -> Slovene users -> their network
  - metadata: username, time stamp, no. of retweets & favourites
- Forum messages
  - 3 forums: med.over.net, avtomobilizem.com, kvarkadabra.net
  - customized extractors
  - metadata: topic, post URL, post time stamp, username, post id
- News comments
  - 3 news portals: RTV Slo, Mladina, Reporter
  - customized extractors
  - metadata: article url, article id, username, post time stamp, post id
- Blogs
  - slWaC 2.0 (Erjavec and Ljubešić 2014)
  - “blog” in domain name
  - no metadata, mixed blog text and comments

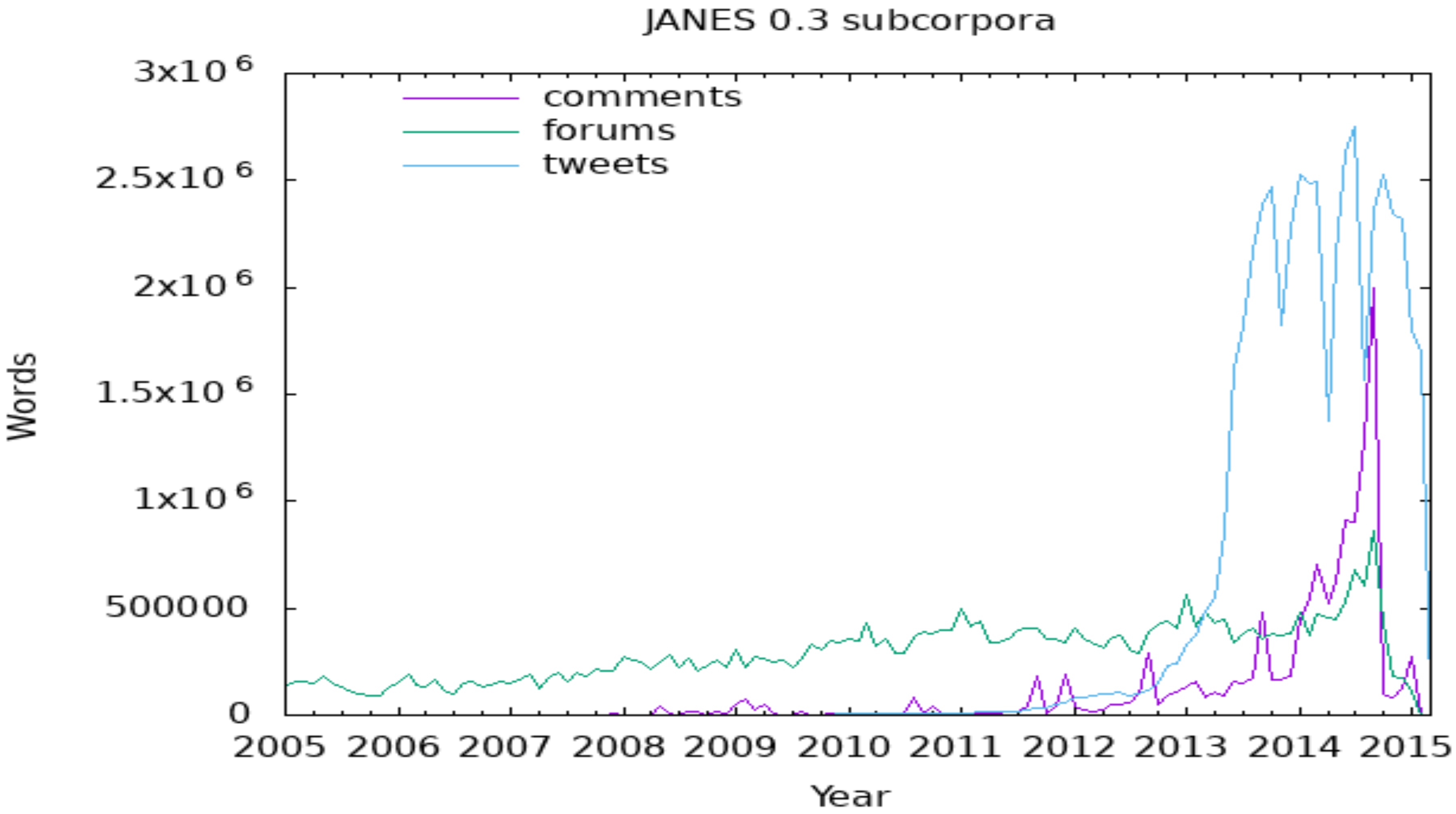
# Corpus composition



Total: 161 M tokens

| | |
|---|---|
| 🟥 Twitter | 61 MT |
| 🟩 Forumi: Avtomobilizem<br>🟩 Forumi: Kvarkadabra<br>🟩 Forumi: Medover.net | 47 MT |
| 🟪 Blogi | 38 MT |
| 🟦 Komentarji: RTV Slo<br>🟦 Komentarji: Mladina<br>🟦 Komentarji: Reporter | 15 MT |

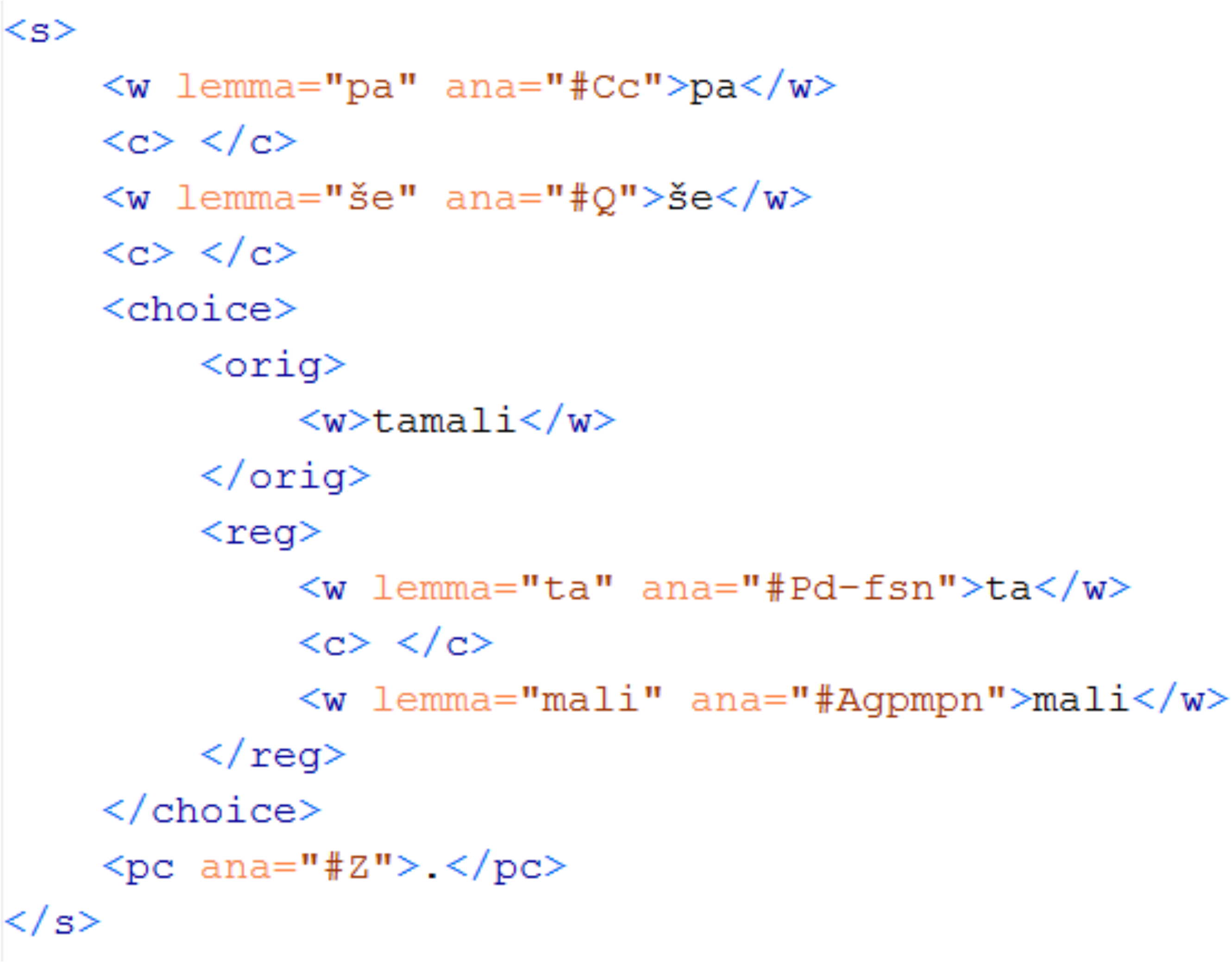
# Distribution through time



# Corpus processing

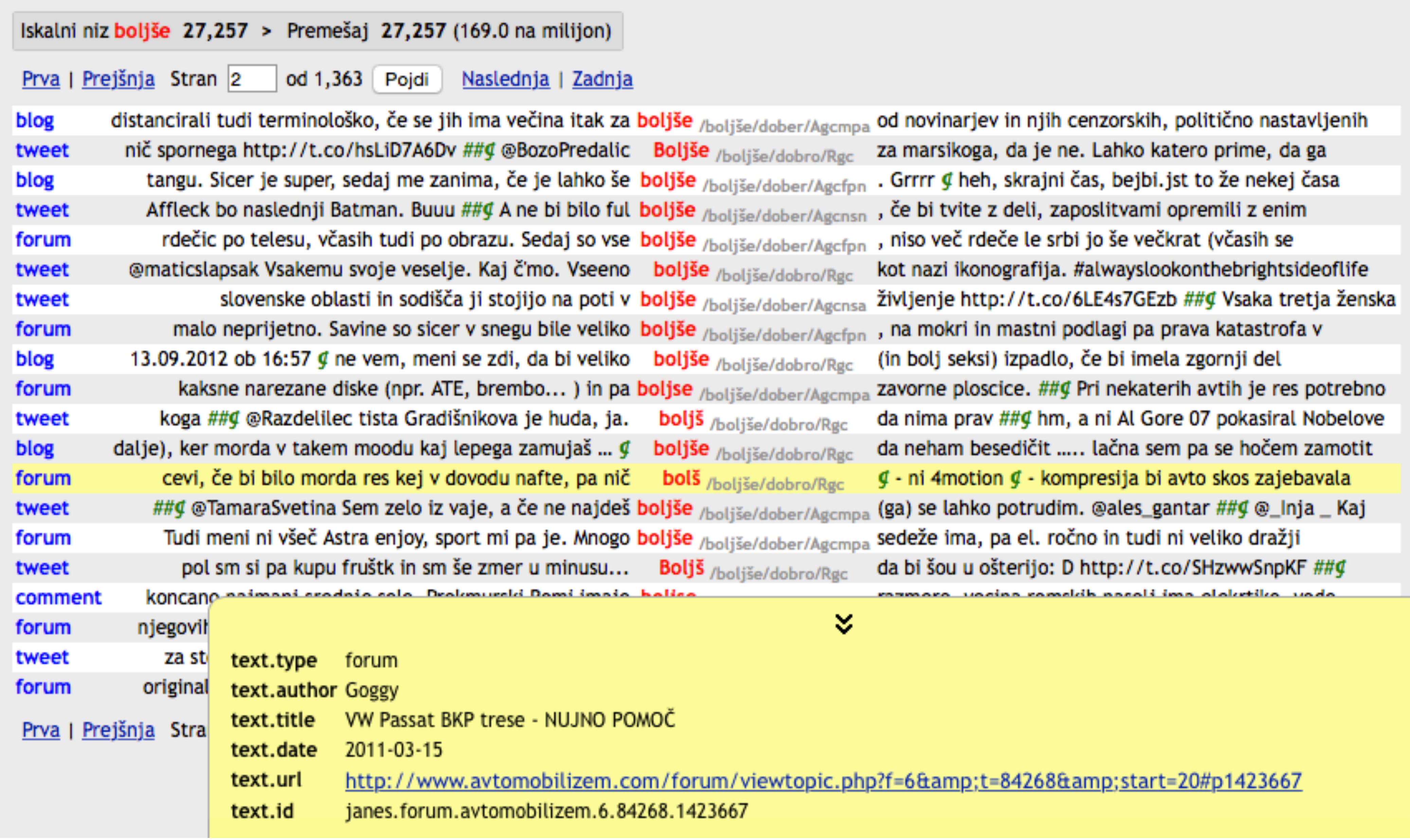
- Annotation
  - (almost standard) tokenization & sentence segmentation
  - **lexical normalisation with CSMT**
  - standard MSD tagging & lemmatization
- Encoding
  - (currently) bespoke XML for metadata
  - **annotated text in TEI P5**
- Exploration
  - **(no)Sketch Engine**

# TEI text



```
<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>
```

# noSketchEngine: “better”



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

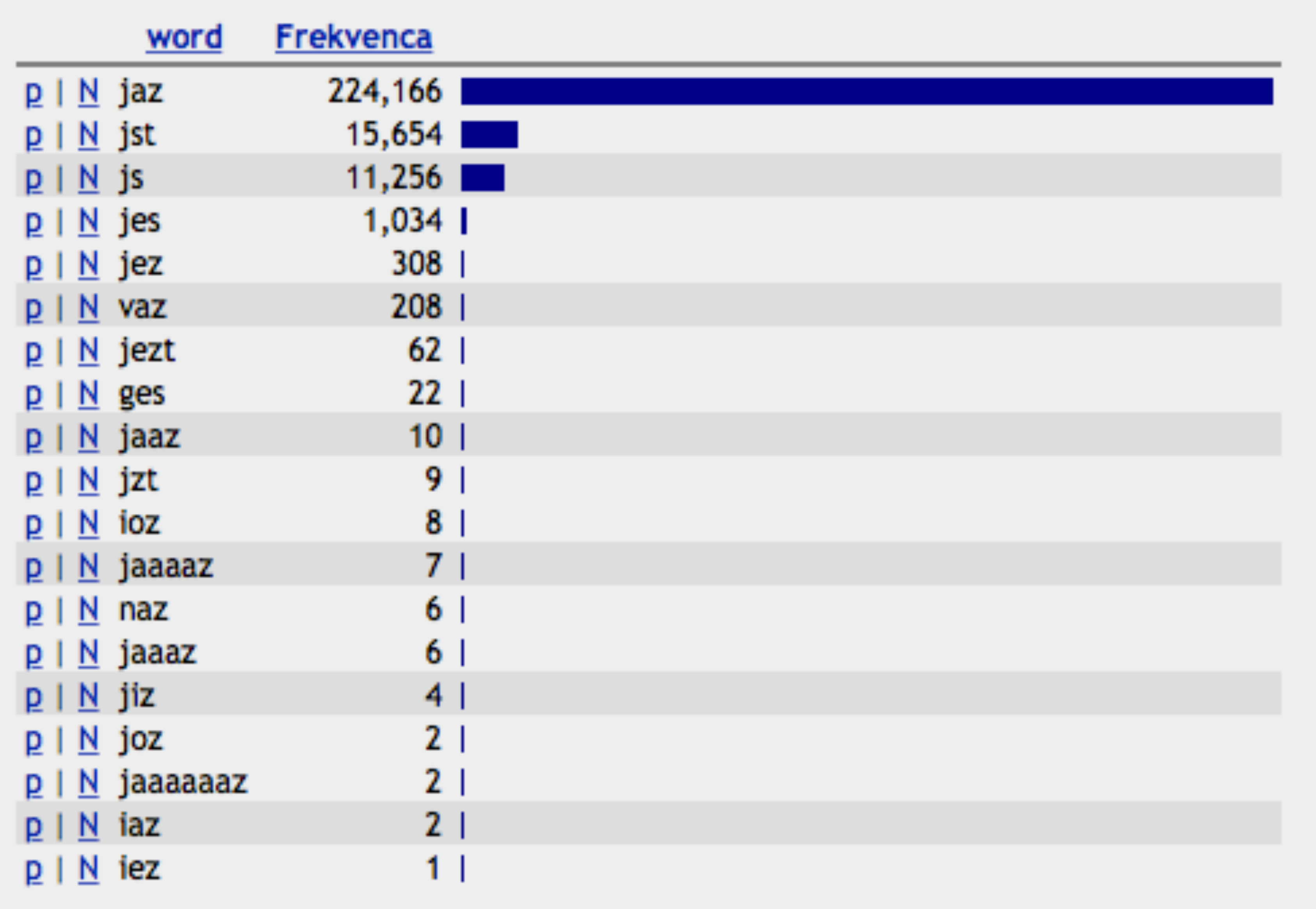
Prva | Prejšnja Stran 2 od 1,363 Pojdi Naslednja | Zadnja

| | | | | |
|---|---|---|---|---|
| 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 ##₲ @_Inja _ 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 | njegovil | | | |
| 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&amp;t=84268&amp;start=20#p1423667 |
| **text.id** | janes.forum.avtomobilizem.6.84268.1423667 |

# noSketchEngine: "I"



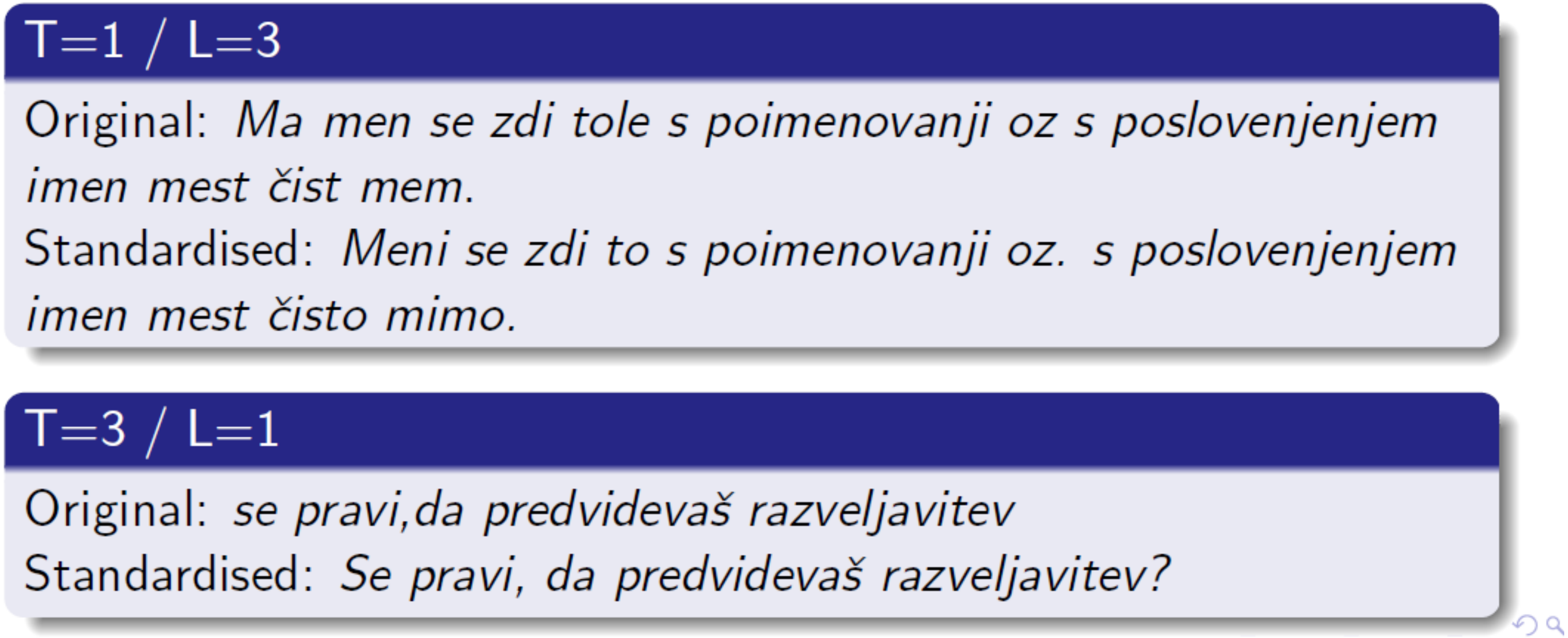
| | word | Frekvenca |
|---|---|---|
| p \| N | jaz | 224,166 |
| p \| N | jst | 15,654 |
| p \| N | js | 11,256 |
| p \| N | jes | 1,034 |
| p \| N | jez | 308 |
| p \| N | vaz | 208 |
| p \| N | jezt | 62 |
| p \| N | ges | 22 |
| p \| N | jaaz | 10 |
| p \| N | jzt | 9 |
| p \| N | ioz | 8 |
| p \| 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 |

# Corpus Tweet-sl 0.3.4

- Additional tweets
- Richer metadata
  - for authors:
    - Private/Corporate (manual)
    - Male/Female (manual)
    - Region (automatic)
  - for tweets:
    - Standardness (automatic)
    - Sentiment (automatic)

# Standardness

- Ljubešić et al.: Predicting the level of text standardness in user-generated content. RANLP 2015
- Technical and linguistic: T1 – T3, L1 – L3
- Annotation campaign
- Regressor training
- Evaluation: mean error 0.377 T and 0.424 L



**T=1 / L=3**

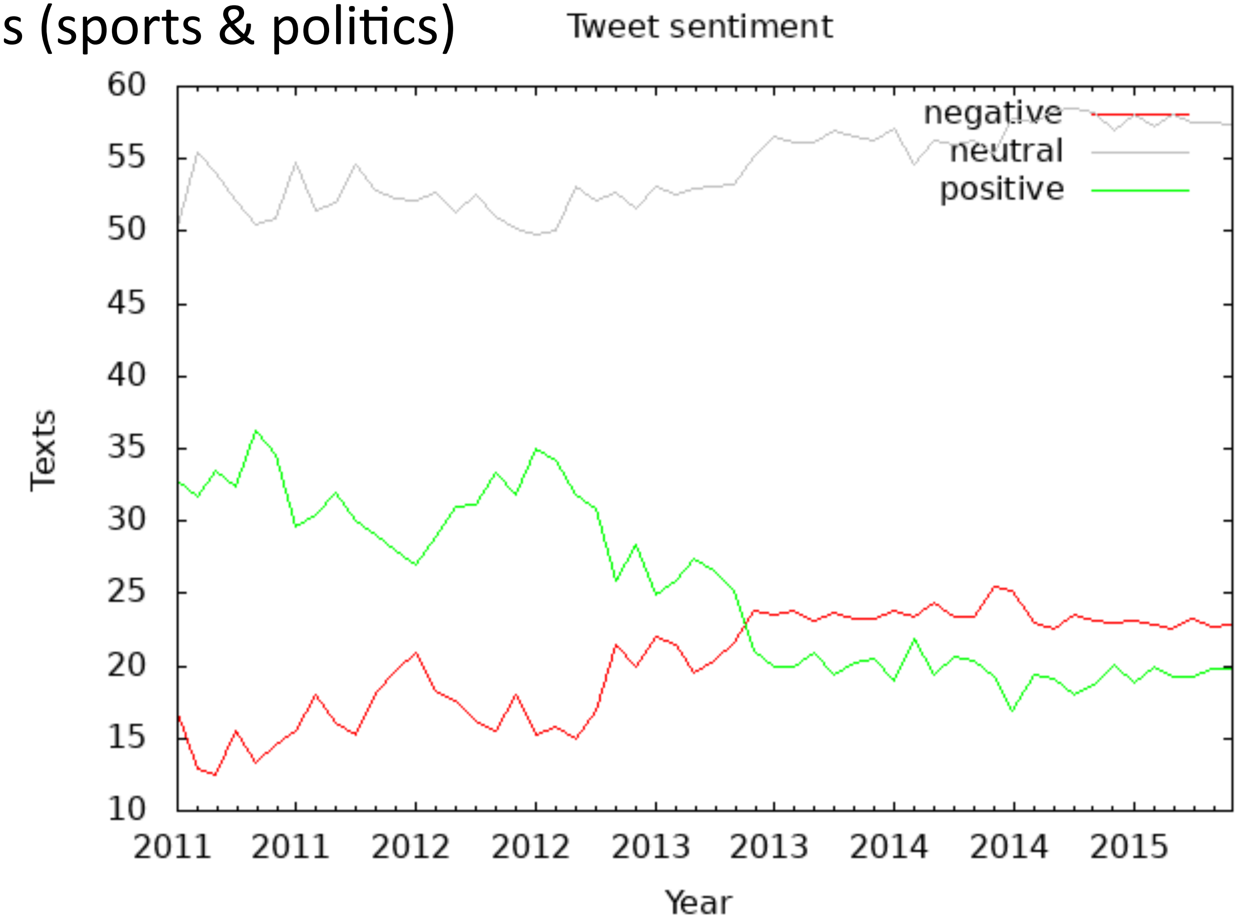
Original: *Ma men se zdi tole s poimenovanji oz s poslovenjenjem imen mest čist mem.*
Standardised: *Meni se zdi to s poimenovanji oz. s poslovenjenjem imen mest čisto mimo.*

**T=3 / L=1**

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

# Sentiment

- Jasmina Smailović, IJS:
  - PhD on SVM training for financial tweet sentiment prediction
  - Further work also on Slovene (large manually annotated dataset)
  - Annotated Tweet-sl 0.3.4 with sentiment
- Evaluation on 1,000 tweets (sports & politics)
  - Baseline = 0.377
  - One-annotator ~ 0.573
  - Both-annotator = 0.621
  - Interannotator = 0.765



# Tweet-sl 0.3.4 metadata

```
id="tid.392972411765018626"
name="007_delic"
created="2013-10-23T11:14:58"
retrieved="2013-10-24T05:20:32.036451"
favorited="0"
retweeted="0"
in_reply="tid.392965997352591361"
lang="sl" lang_prob="0.996788622457"
standard_tech="T1" standard_tech_n="1.1"
standard_ling="L1" standard_ling_n="1.2"
sentiment="neutral"
source="private"
sex="female"
geo="-"
```

# To-dos

- rediacritisation
- improve normalisation
- manually annotated datasets
- CMC tagset
- CLARIN.SI dissemination:
  - Technical (anonymisation etc.)
  - Legal
- monitor corpus

# JANES Summer Camp

- Uni Lj, 24 – 28 Aug, 2015
- 25 high school students from all over Slovenia
- format
  - 5 days, 5 topics
  - lecture, exercise, project
  - project presentation
- Invited talks and evening events
- On-line slides and other teaching materials
- Good media coverage

# WebAnno Tutorial & Hackaton

- JSI, 13 – 15 Nov 2015
- In cooperation with CLARIN.SI
- Day 1: lecture, tutorial
- Day 2: annotating tweets (standardness, corrections)
- First step in producing a gold-standard dataset:
  - normalisation, lemmatisation, tagging
  - (+ syntax + annotation of original)

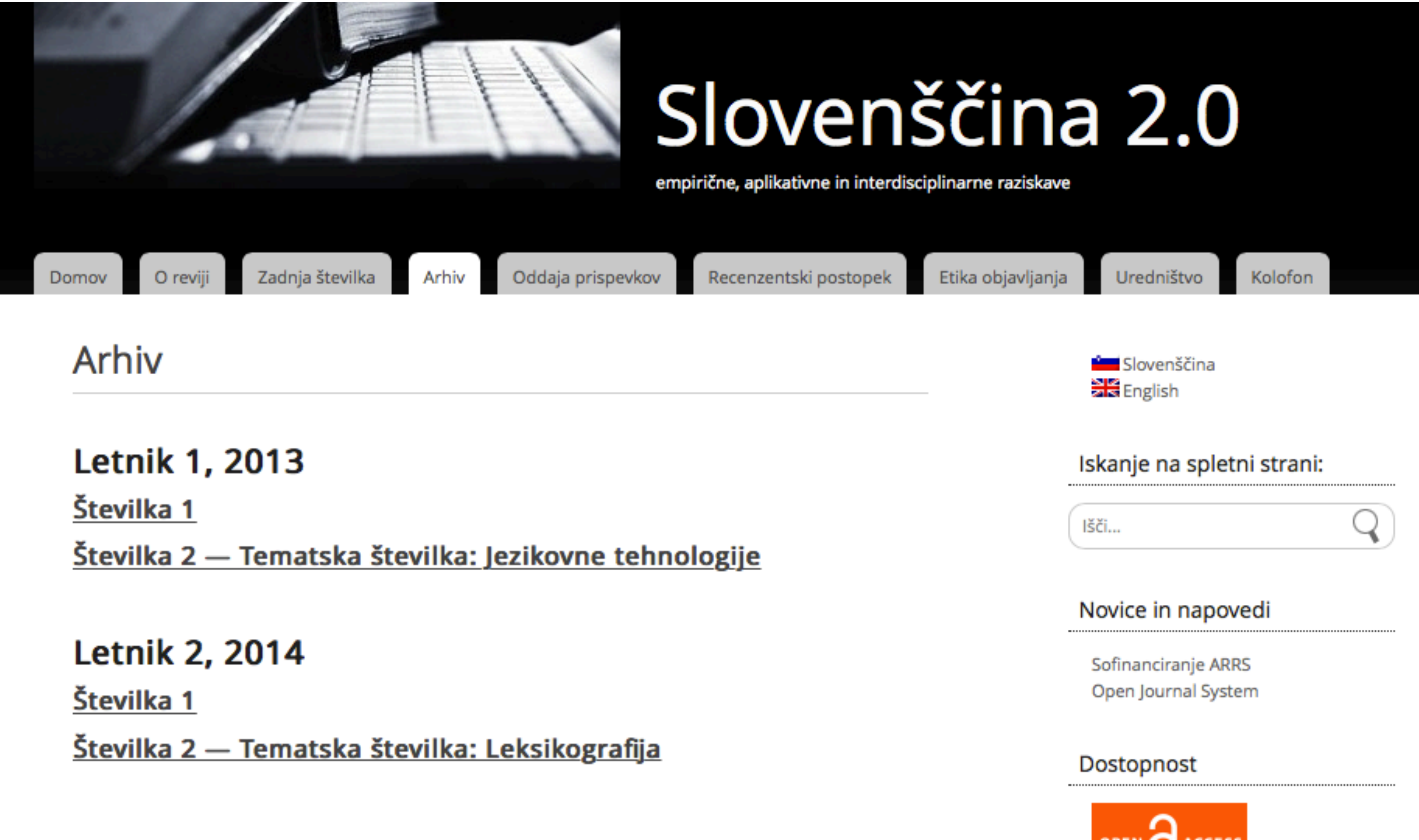
# JANES Conference

- Ljubljana, 25 – 27 Nov 2015
- 15 reviewed papers, 23 authors
- Best student paper award
- Amazing invited speaker!
(Michael Beißwenger, TU Dortmund)
- Panel discussion
- Tutorial on statistics and R for linguists
(Maja Miličević, Uni Belgrade)

# JANES Express

- Stops:
  - – Zagreb, Croatia (4 Dec 2015)
  - – Belgrade, Serbia (10 Dec 2015)
  - – Sarajevo, Bosnia (next year)
- 1 day event:
  - – student workshop (noSkE)
  - – annotation workshop (WebAnno)
  - – evening lecture (annotating UGC corpora)

# Journal Special Issue 2016



Slovenščina 2.0

empirične, aplikativne in interdisciplinarne raziskave

## Arhiv

### Letnik 1, 2013

Številka 1

Številka 2 — Tematska številka: Jezikovne tehnologije

### Letnik 2, 2014

Številka 1

Številka 2 — Tematska številka: Leksikografija

Slovenščina
English

Iskanje na spletni strani:

Novice in napovedi

Sofinanciranje ARRS
Open Journal System

Dostopnost

**http://nl.ijs.si/janes/**

tenks ☺