# Razvoj učne množice za izboljšano označevanje spletnih besedil

Konferenca Jezikovne tehnologije in digitalna humanistika 2016

Jaka Čibej, Špela Arhar Holdt, Tomaž Erjavec, Darja Fišer

Filozofska fakulteta, Univerza v Ljubljani
Inštitut "Jožef Stefan"

Ljubljana, 30. september 2016

# Pregled predstavitve

1. Uvod in motivacija
2. Priprava in vzorčenje podatkov
3. Smernice za označevanje in specifike označevanja spletnih besedil
4. Označevalska kampanja
5. Kvantitativni pregled označene množice
6. Zaključek

# Uvod in motivacija

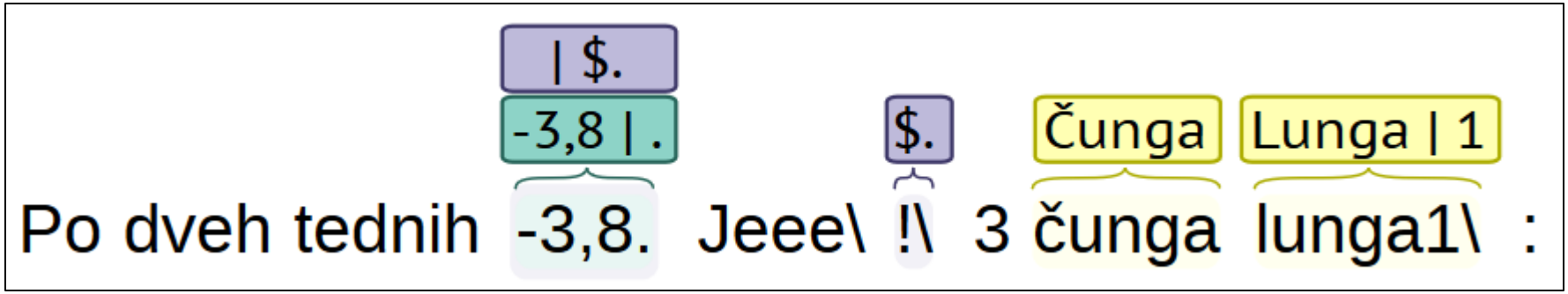
- jezik spletnih žanrov → težja obdelava z obstoječimi orodji
- prilagoditev in nadgradnja
- učni korpus spletne komunikacije
  - vzorčen iz korpusa JANES (Fišer et al., 2015)

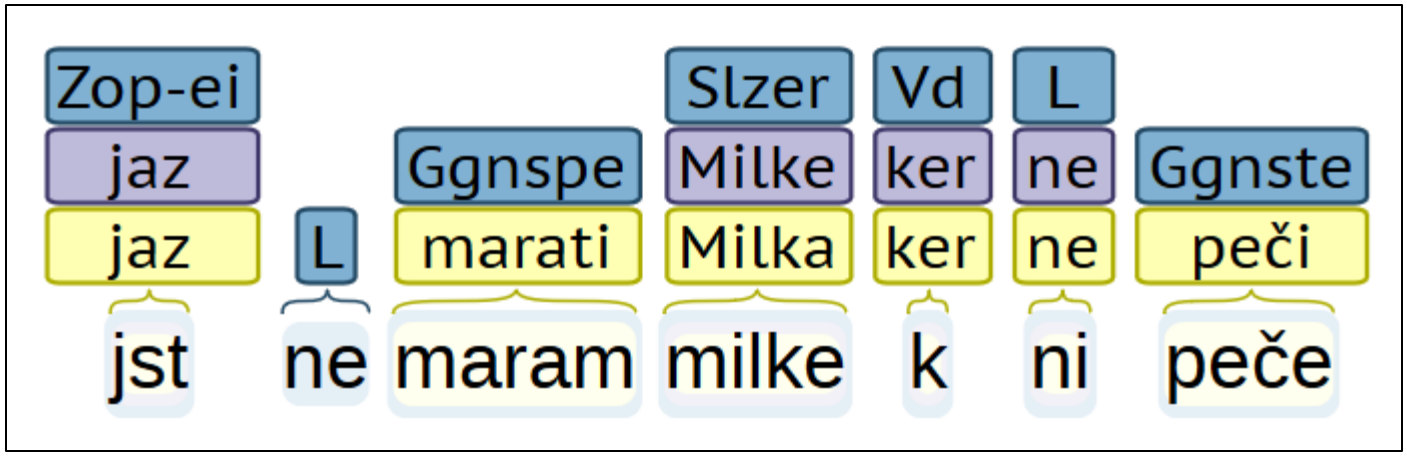
# Priprava in vzorčenje podatkov

- Kons1 → 4000 tvitov
- Kons2 → 4000 forumskih sporočil in komentarjev na novice
- vzorčenje po tehnični/jezikovni standardnosti besedila (Ljubešić et al., 2015)
  - L1–L3: (ne)standardno besedišče/zapis
  - T1–T3: (ne)pravilna raba ločil, presledkov, velikih začetnic
- avtomatsko označevanje (Erjavec, 2011; Ljubešić et al., 2014)
  - stavčna segmentacija
  - tokenizacija
  - normalizacija
  - lematizacija
  - oblikoskladenjsko označevanje

# Označevalna platforma

- WebAnno (Eckart de Castilho et al., 2014)
  - pretvorba podatkov (TEI – TSV – TEI)



# Smernice za označevanje

ročna analiza 200 naključnih tvitov

**tehnične smernice** (anotacijska shema ipd.)

**jezikoslovne smernice** (odločitve v zvezi z jezikom)

Smernice: nl.ijs.si/janes/viri

# Specifike označevanja spletnih besedil

- Stavčna segmentacija
  - določanje meje med povedmi
    - emotikoni in emojiji (:D 😉 )
    - ključniki (#jtdh2016)
    - URL-naslovi (rtvslo.si)
    - sklici na uporabniška imena (@jaka)

Liverpool zasluženo owna Twitter, ampak na vrhu je pa fucking Iago Aspas hahaha :) #nogomet #LFC #SOULIV http://t.co/LCyEvyoVD7

# Specifike označevanja spletnih besedil

- Tokenizacija
  - ločevanje/združevanje napačno združenih/razdeljenih pojavnic
    - fruktoza-glukoza → fruktoza - glukoza
    - žensk ( e / o ) → žensk(e/o)

# Specifike označevanja spletnih besedil

- Normalizacija
  - načelo minimalne intervencije
    - pofarbat → pofarbati, ne *pobarvati
    - ne normaliziramo npr. izbire besedišča (rabiti – *potrebovati)
  - Problem 1: nestandardne besede brez neposredne standardne ustreznice (*ornk*, *orng*, *orenk*, *oreng*)
  - Problem 2: tujejezične prvine z različnimi stopnjami prevzetosti *(sharati, sherati, šerati)*
  - Rešitev: normalizirana oblika je najfrekventnejša različica v korpusu JANES

# Specifike označevanja spletnih besedil

- Lematizacija
  - smernice ssj500k (Holozan et al., 2008)
  - Problem: tujejezične prvine
    - če ni jasno razvidnih znakov prilagoditve → lema je enaka obliki (jailbreak, hrvatskog)
    - sicer upoštevamo slovenska oblikoslovna načela → chataš – chatati,  spoilerji – spoiler

# Specifike označevanja spletnih besedil

- Oblikoskladenjsko označevanje
  - dopolnitev sistema JOS z novimi oznakami:
    - @jaka → Na
    - #jtdh2016 → Nh
    - rtvslo.si → Nw
    - :D → Ne

# Označevalska kampanja

**Usposabljanje označevalcev**

Dvodnevna delavnica za 11 študentov jezikoslovja (MA)

**Preizkušanje označevalcev**

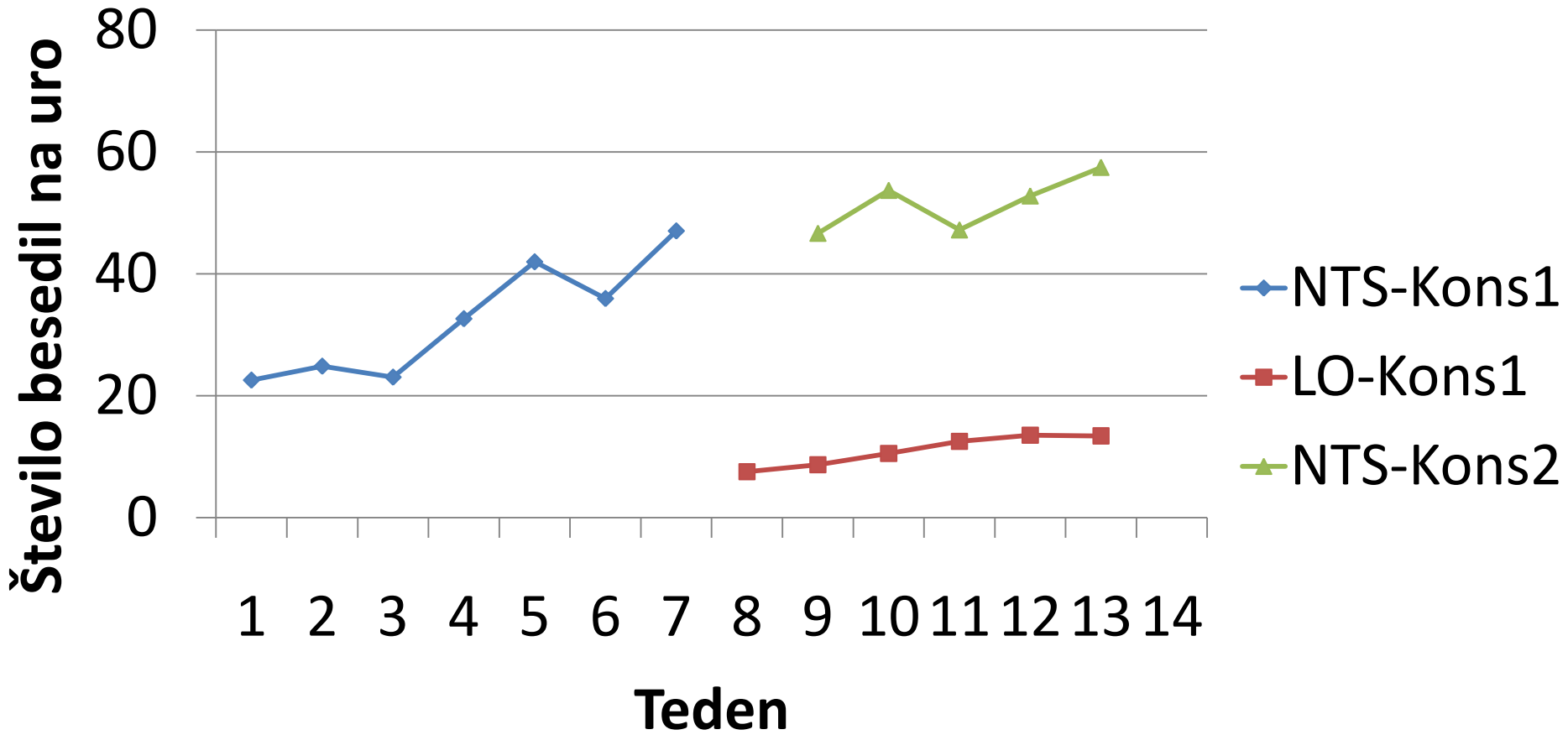
označevanje testne množice (100 tvitov)

**Označevanje podatkovne množice**

9 označevalcev, 2 razsodnika

# Delotok označevanja

# Učinkovitost označevalcev



# Kvantitativni pregled označene množice

| | Besedila | Stavki | Besede | Pojavnice | Norm. | Norm. % |
|---|---|---|---|---|---|---|
| Kons1-MSD | 880 | 2.365 | 20.537 | 23.958 | 4.888 | 20,4 |
| Kons1 | 3.940 | 9.976 | 86.593 | 102.719 | 11.881 | 11,6 |
| Kons2 | 1.927 | 4.473 | 34.583 | 41.056 | 4.728 | 11,5 |
| **Kons** | **5.867** | **14.449** | **121.176** | **143.775** | **16.609** | **11,6** |

# Zaključek

- priprava podatkov, smernice in delotok označevanja učnega korpusa
- prosti dostop (CLARIN.SI)
- cilji priprave korpusa:
  - izboljšano označevanje spletnih besedil
  - rešitve za problematiko označevanja spletnih besedil
  - dopolnitev leksikonov z nestandardnim besediščem
- Nadaljevanje:
  - nadgradnja orodij
  - evalvacija novih različic

Hvala za pozornost.