“S kje pa si?”
Metapodatki o regionalni pripadnosti uporabnikov družbenega omrežja Twitter

Konferenca “Slovenščina na spletu in v novih medijih”

Jaka Čibej[1] in Nikola Ljubešić[2,3]

[1]Filozofska fakulteta, Univerza v Ljubljani
[2]Filozofska fakulteta, Univerza v Zagrebu
[3]Institut “Jožef Stefan”

Ljubljana, 26. november 2015

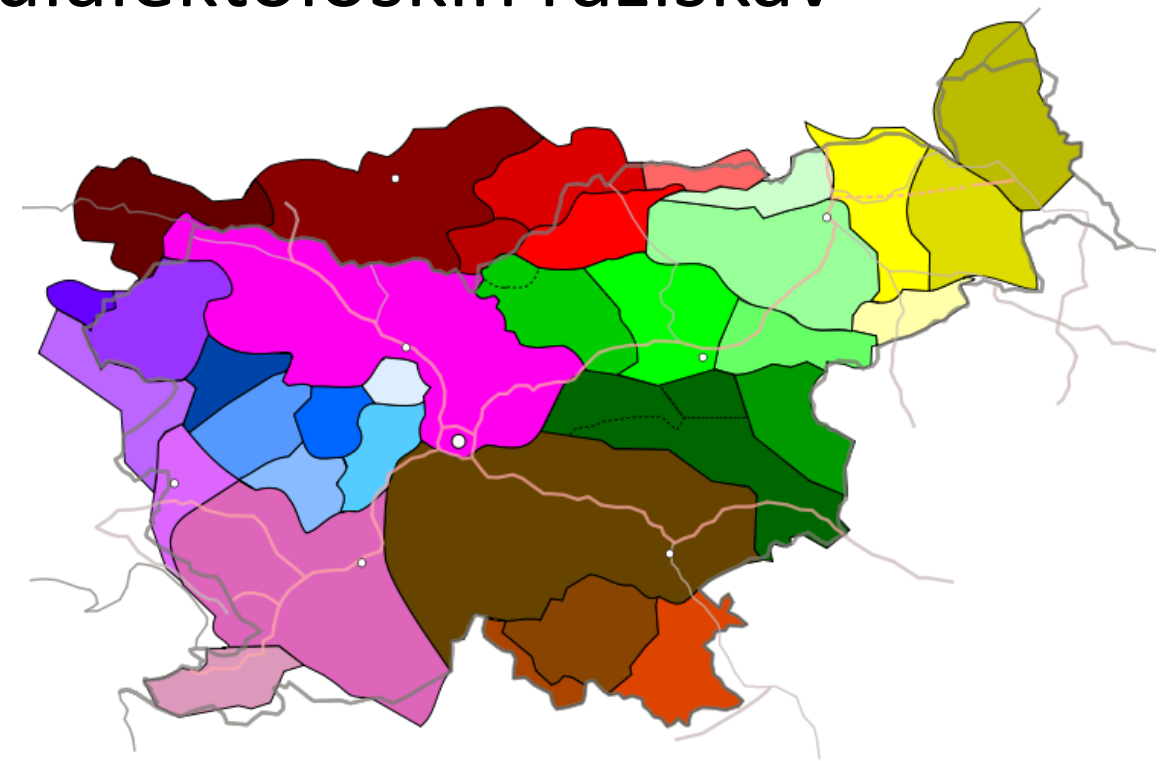
Pregled predstavitve

1. Ozadje
2. Motivacija in sorodne raziskave
3. Dodajanje regionalnih metapodatkov v podkorpus tvitov JANES
4. Regionalni podkorpusi tvitov
5. Zaključek in prihodnje delo

Ozadje

Ozadje [1/4]

- 7 narečnih skupin, >30 narečij (Ramovš 1931)
- Veliko dialektoloških raziskav



Ozadje [2/4]

- Narečja izumirajo (Kolarič 1954, Ramovš 1951)
- "Čista narečja" (idealnih govorcev)
  - standard *NORM (Non-Mobile Older Rural Male)* (Chamber & Trudgill 1994)
  - starost, izobrazba, spol, govorica staršev (Logar 1959)

Ozadje [3/4]

- Razvoj narečij?
  - omejevanje na ruralno okolje (Sgall et al. 1992)
  - dialekte zamenjajo sociolekti (Sgall et al. 1992)
  - strnjevanje narečij (Niebaum & Macha 1999)
  - nova dialektizacija (Labov 1994)
- Ohranjanje narečij?
  - zavestno kultiviranje – Bavarska (Reichan 1999)
  - naklonjenost rabi narečja – Norveška (Jahr 1997)

Ozadje [4/4]

- Sodobne korpusne dialektološke raziskave
  - nizozemščina → DynaSAND (Kunst & Wesseling 2010)
  - nordijski jeziki → Nordic Dialect Corpus (Johanessen et al. 2009)
  - angleščina → Freiburg Corpus of English Dialects (Hernández 2006)

Motivacija in sorodne raziskave

Motivacija

- Računalniško posredovana komunikacija
  - npr. Facebook, Twitter, Snapchat
  - vseprisotna (SURS)
    - družbena omrežja: 60 %
    - (video)telefoniranje: 41 %
- Spletni jezik / netspeak (Crystal 2001, Ueberwasser 2013, Erjavec & Fišer 2013)
  - mdr. regionalne jezikovne prvine

Sorodne raziskave

- Regionalne jezikovne prvine v spletnem jeziku
  - do zdaj domena jezikovnih tehnologov
- Razvoj orodij za šumna besedila
  - avtomatska detekcija jezikov/različic jezika
    - arabščina (Cotterell & Callison-Burch 2014, Harrat et al. 2013)
    - ameriška angleščina (Eisenstein et al. 2015)
    - hrvaščina/srbščina/bosanščina/črnogorščina (Ljubešić & Kranjčić 2014)
  - oblikoskladenjsko označevanje
    - švicarska nemščina (Ruef & Ueberwasser 2013)
    - alzaščina (Bernhard & Ligozat 2013)
  - strojni prevajalniki
    - regionalna <-> sodobna standardna arabščina (Harrat et al. 2014)
    - dunajščina <-> standardna avstrijska nemščina (Haddow et al. 2013)

JANES

IJS

Univerza v Ljubljani
FILOZOFSKA FAKULTETA

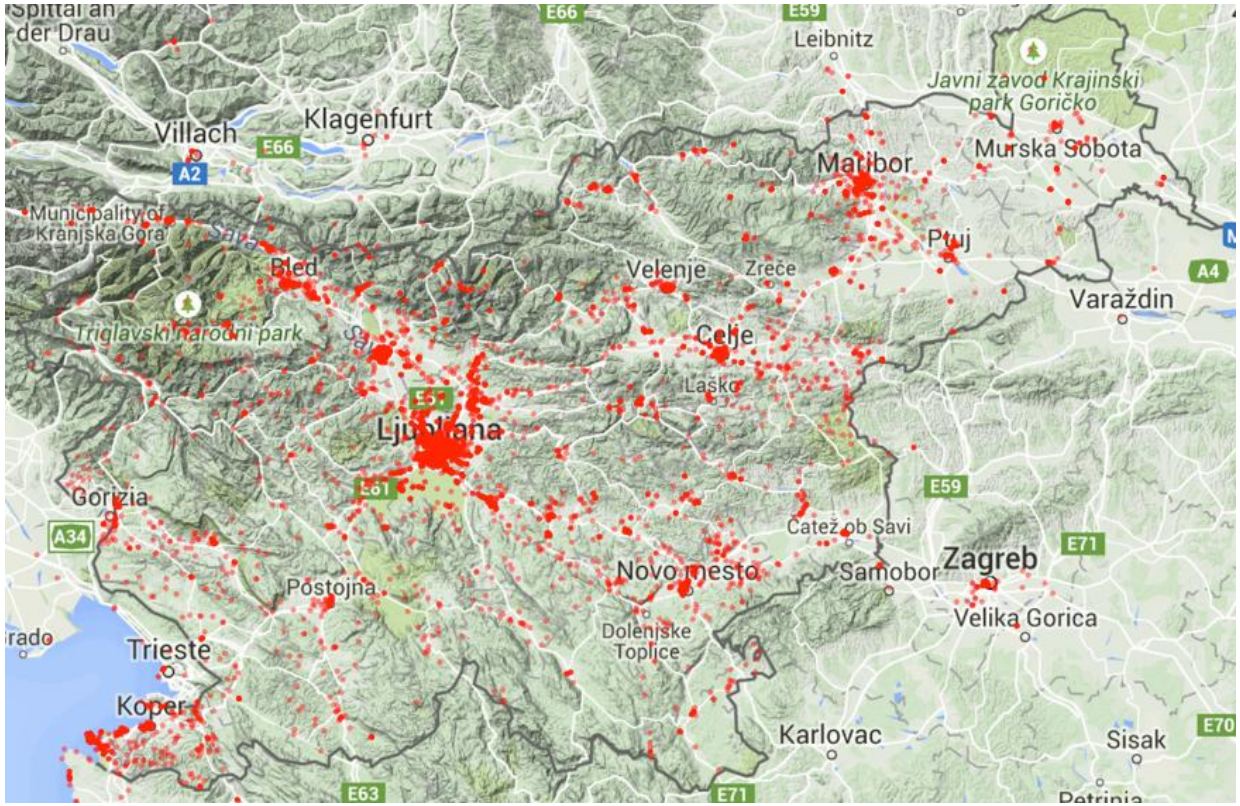
Regionalne jezikovne prvine v spletni slovenščini

- Še neraziskane
- Korpus spletne slovenščine JANES
- Metapodatki o regionalni pripadnosti

Dodajanje regionalnih metapodatkov v podkorpus tvitov JANES

Zajem slovenskih tvitov z geolokacijo

- Orodje TweetCat (Ljubešić et al. 2014)
- začetek: januar 2015



Izbor ustreznih avtorjev

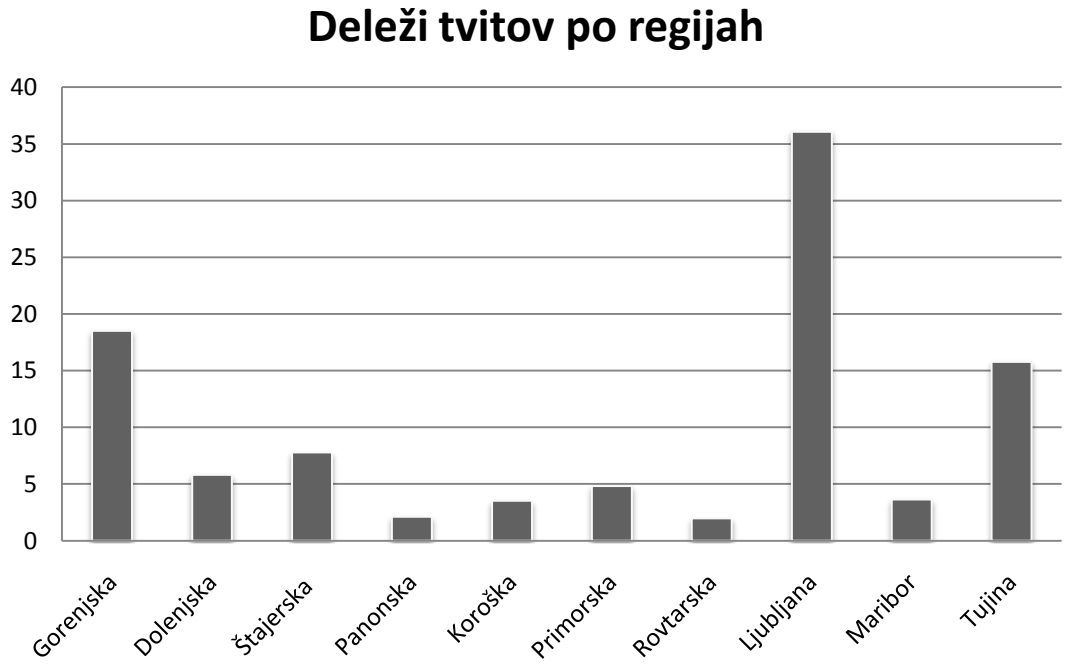
- ~130.000 tvitov, ~1.600 avtorjev
  - izločitev računov organizacij na Twitterju
  - izločitev uporabnikov, ki niso vključeni v korpus JANES
- preostanek: 119.236 tvitov, 1.461 (zasebnih) avtorjev

Razdelitev Slovenije na regije

- 9 regij → 7 narečnih skupin, Ljubljana, Maribor

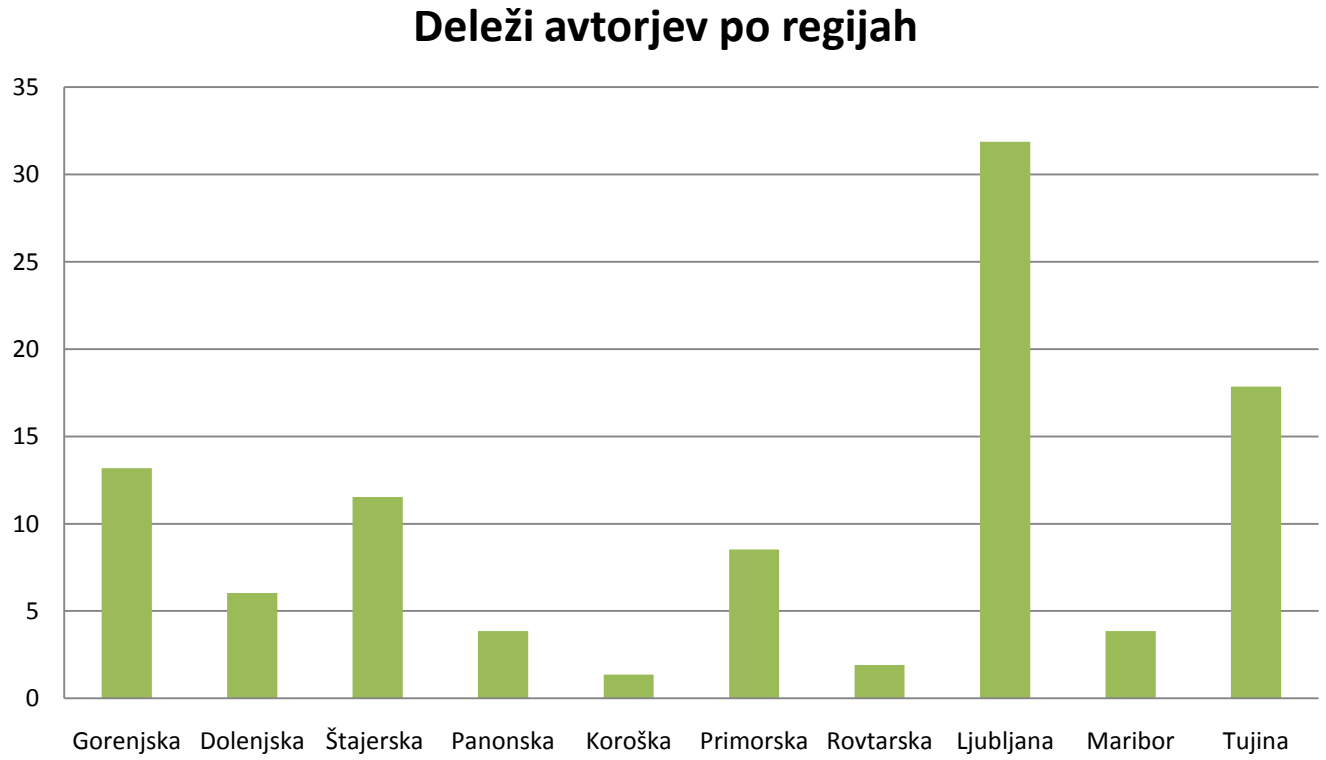
Razvrščanje tvitov po regijah

- metoda metanja žarka (*ray-casting method*)



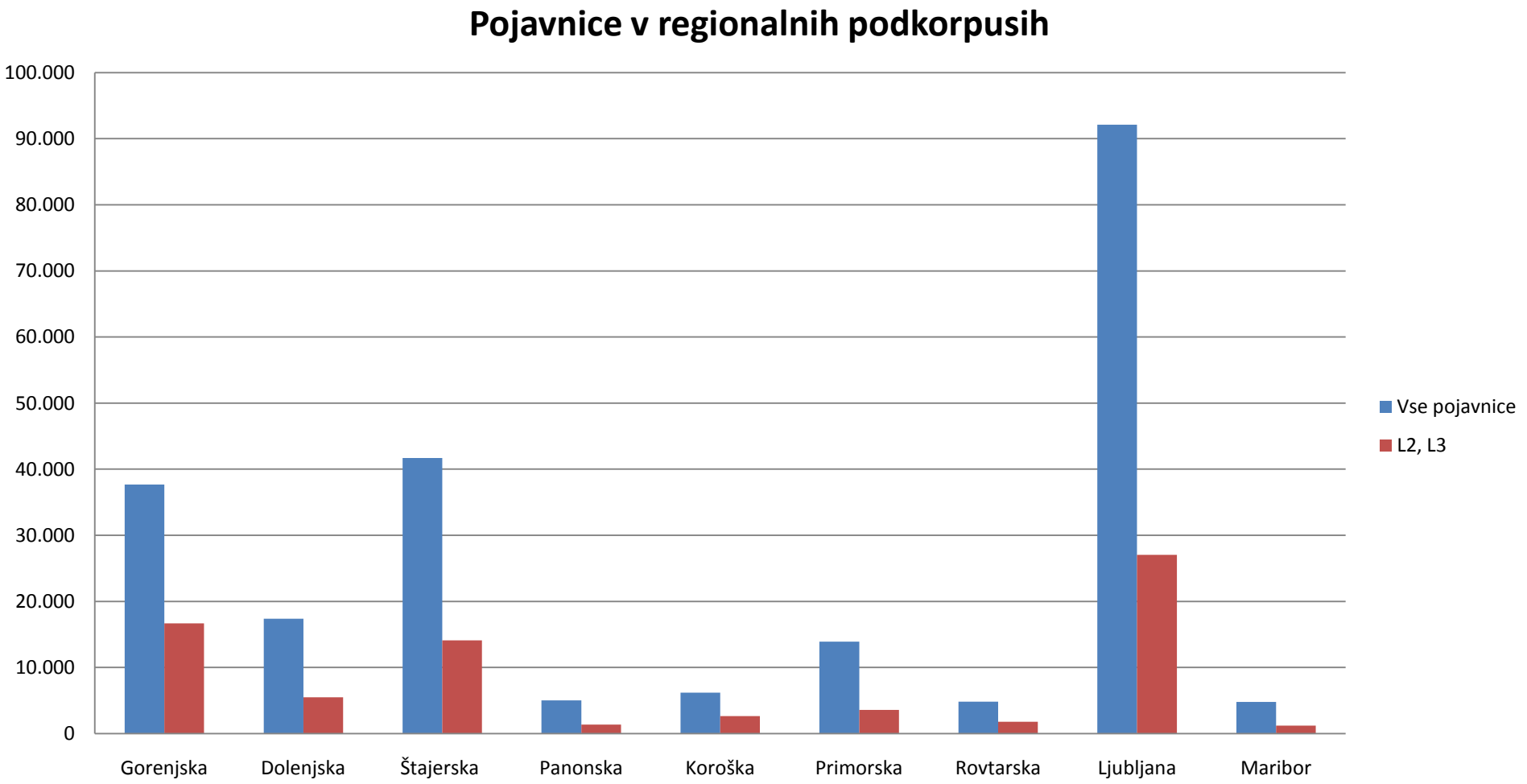
Pripis metapodatkov

- Pogoji:
  - 90+ % tvitov iz ene regije
  - min. 3 tviti
- 364 avtorjev
  - povp. 74
  - med. 18
  - max. 1188

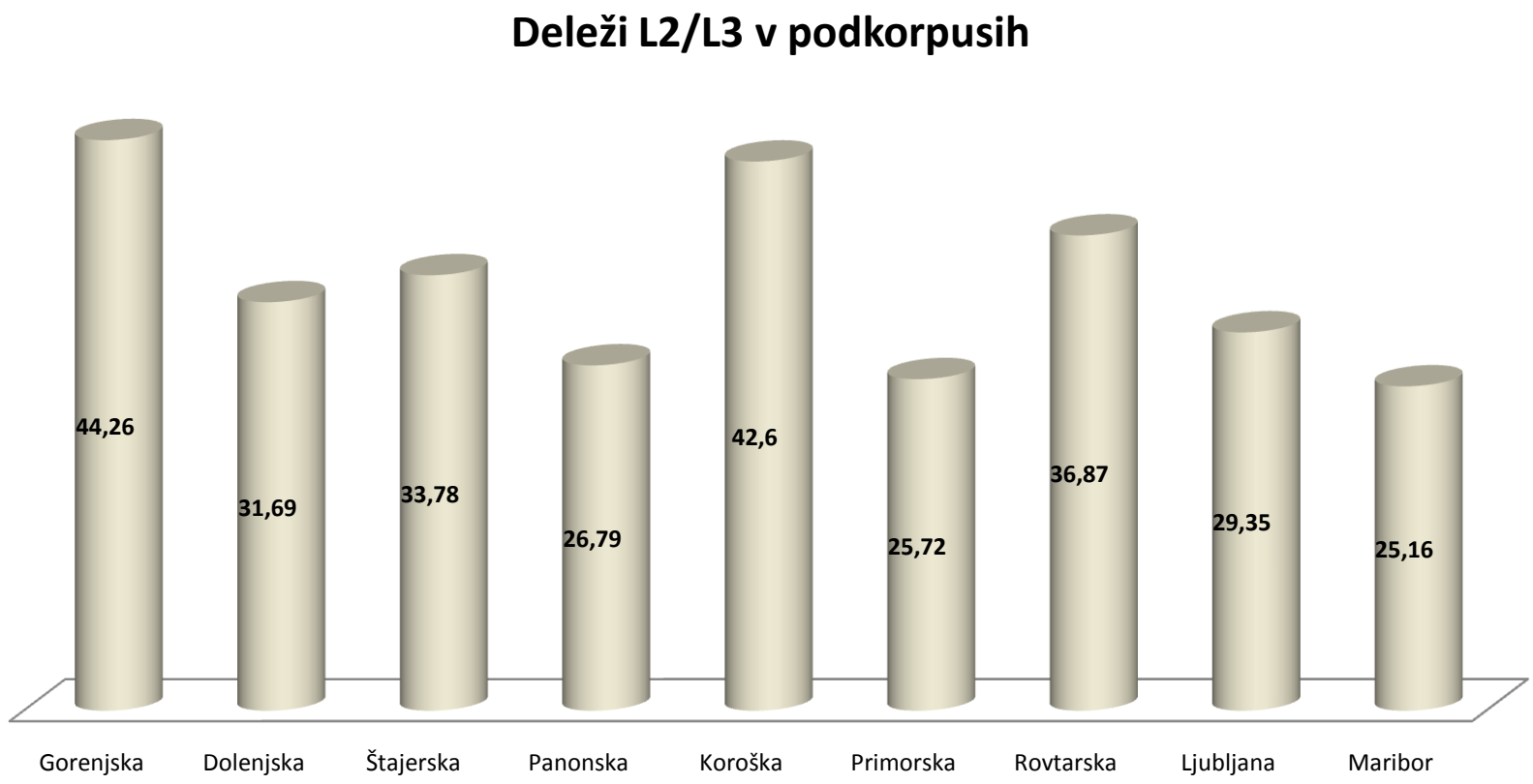


Regionalni podkorpusi tvitov JANES

Regionalni podkorpusi tvitov JANES



Regionalni podkorpusi tvitov JANES



Zaključek in prihodnje delo

Zaključek in prihodnje delo

- Postopek za avtomatski pripis metapodatkov o regionalni pripadnosti
- Prihodnje delo:
  - povečanje obsega podkorpusov (novi tviti)
  - izključitev premajnih podkorpusov
  - jezikoslovna analiza
  - primerjava s korpusom GOS
  - model za avtomatsko klasifikacijo regionalnih jezikovnih različic
    - primerjava z gručenjem

Hwala za pazurnast.