Framework for an Analysis of Slovene Regional Language Variants on Twitter

4th Conference on CMC and Social Media Corpora for the Humanities

Jaka Čibej

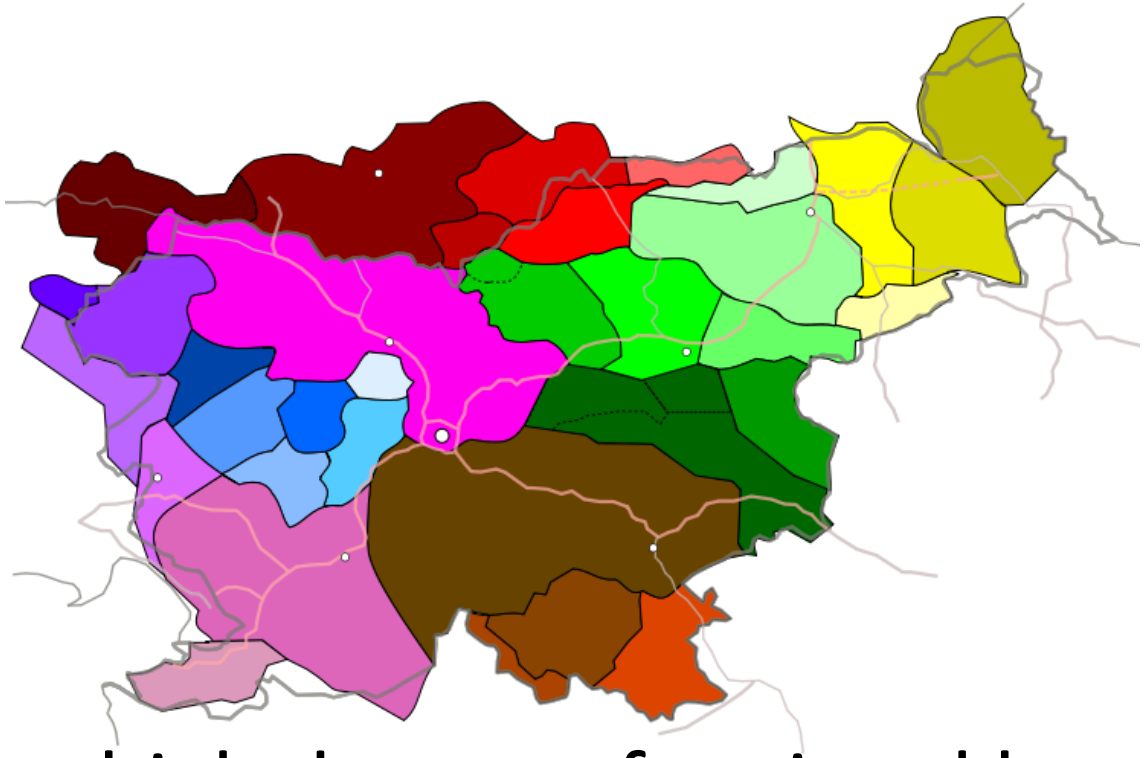
Faculty of Arts, University of Ljubljana

Ljubljana, 28 September 2016

Outline

1. Introduction and Motivation
2. Dataset Preparation
3. Typology of Non-Standard Slovene Language Elements on Twitter
4. Dataset Annotation
5. Results
6. Measures of Regional Specificity
7. Conclusion

Introduction



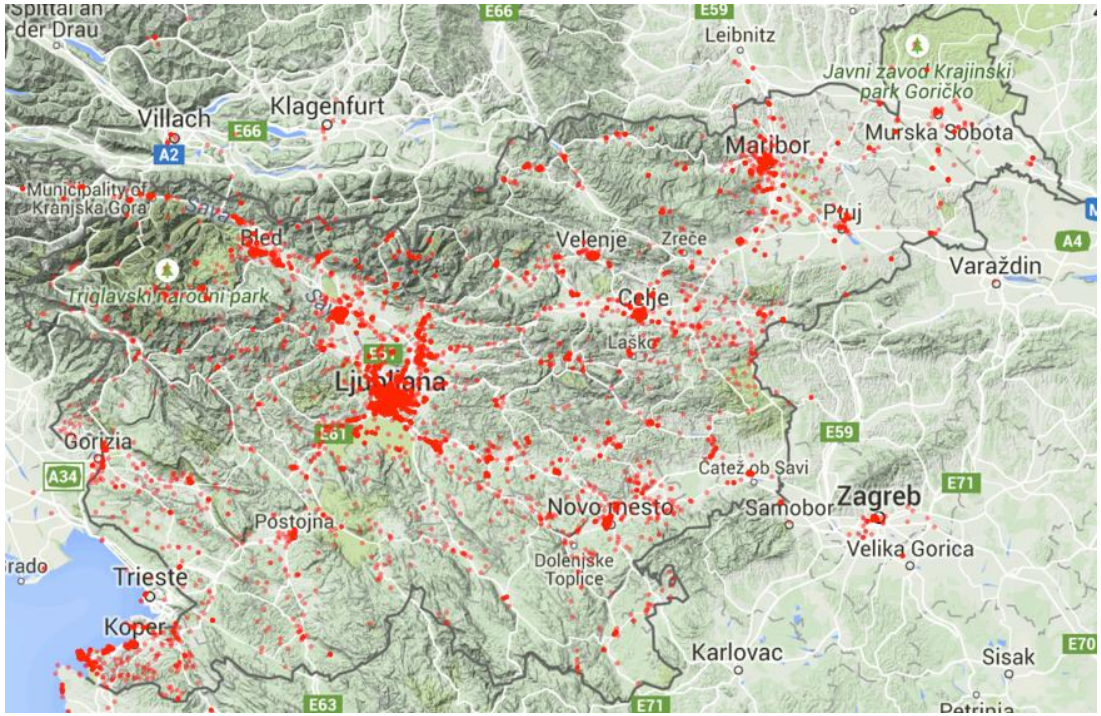
- high degree of regional language variation
- ~48 dialects (7 main dialect groups)

Motivation

- regional language variants in user-generated content
  - in Slovene UGC (Erjavec & Fišer 2013; Zwitter Vitez & Fišer 2015)
- linguistic analysis

Dataset Preparation

- 130,000 geotagged tweets from the JANES corpus of Slovene UGC (Fišer et al. 2016)



Regional Subcorpora

- 9 regional subcorpora (7 main dialectal groups + Ljubljana and Maribor)
- 90% threshold

Sampling

- Sampling criteria:
  - 500 tweets per region
  - only private users
  - non-standard tweets (L3) (Ljubešić et al. 2015)
  - all users included
  - max. 30–50 random tweets per user

- This paper: Primorska, Gorenjska, Štajerska

Typology of Non-Standard Slovene Language Elements

- manual analysis, bottom-up approach
- **7 main categories**, 105 different tags
  - **non-standard vocabulary** (ejga, čuj, nanka)
  - **reductions and ellipses** (čudno → čudn_)
  - **alternative graphemes** (ne vem → ne wem)
  - **non-standard morphology** (imate → imaste)
  - **spelling variants of frequent standard words** (jaz → jz, js, jst, jest, jast...)
  - **frequent transformations** (-aj- to -ej-, e.g. nekaj → nekej, včeraj → včerej)
  - **miscellaneous** (to je → toj)

Annotation

[token/phrase]{tag 1}{tag 2}{...}

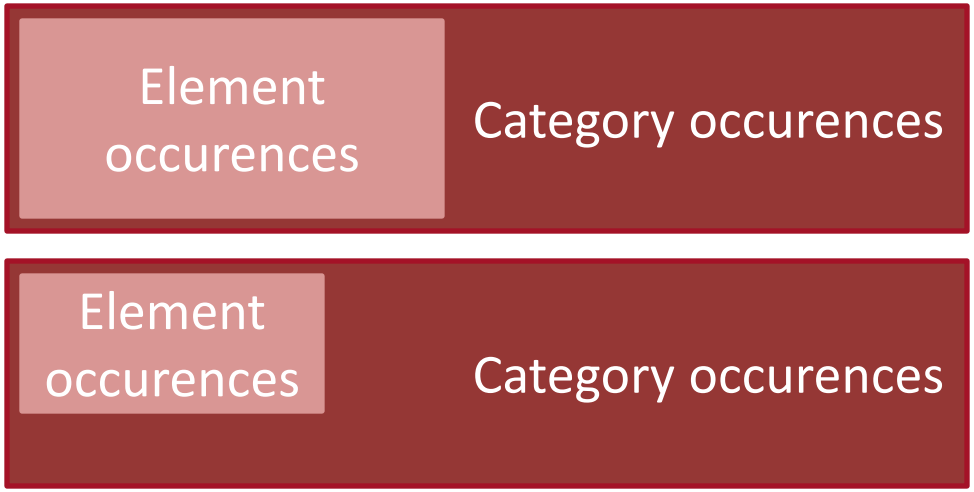
@RoganMatevz Ne ta vikend ne, [nasledn]{Rj.nj}{RksPme.i} pa [mogoč]{RksR.e} ... boš ti [nasledn]{Rj.nj}{RksPme.i} vikend doma? :) Rose me čaka ne? :D

(Preliminary) Annotation Results

Category	Primorska	Gorenjska	Štajerska
Non-standard vocabulary	394	347	371
Spelling variants of frequent standard words	233	**322**	183
Alternative graphemes	40	**54**	34
Reductions and ellipses	588	**1122**	648
Non-standard morphology	90	99	67
Frequent transformations	120	**181**	68
Miscellaneous	39	59	24
Total	1504	**2184**	1395

Measures of Regional Specificity and Dispersion

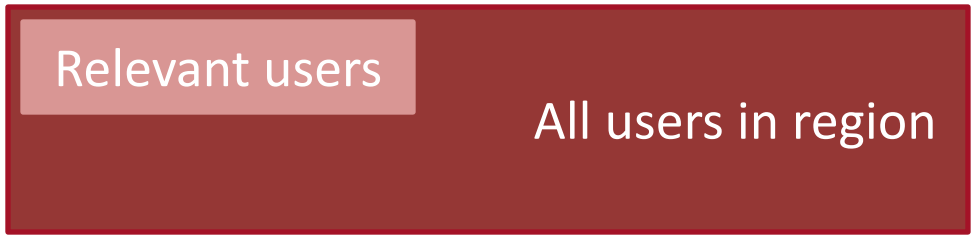
- relative frequency (f_R)
  - element occurences / category occurences
  - higher f_R → occurs more often within region
  - lower f_R → occurs less often within region



Measures of Regional Specificity and Dispersion

- user ratio (*u*)
  - percentage of users in region using the element
  - higher *u* → more widespread in region

  - lower *u* → less widespread in region



Measures of Regional Specificity and Dispersion

- type/token ratio (*t*)
  - penalises elements occurring often, but with a limited number of types
  - Example: kmalu ('soon') → kmal_ (final u ellipsis)

Measures of Regional Specificity and Dispersion

- annotation ratio (*a*)
  - tags element is used with / all tags in category
  - penalises elements occurring only with a limited number of tags (e.g. only with adverbs)
  - higher *a* → occurs in more word positions and with more PoS categories
  - lower *a* → fewer positions, fewer PoS categories

Measures of Regional Specificity and Dispersion

- coefficient of regional dispersion (δ_R)
  - summary of all measures
  - higher δ_R → more widespread and regionally specific

$$\delta_R = f_R \times u \times t \times a \times 100$$

Example: Final -i ellipsis

	Gorenjska	Štajerska	Primorska
f_R	0.52	0.60	0.47
u	**0.75**	**0.42**	0.54
t	**0.61**	0.53	0.67
a	**0.43**	**0.41**	0.24
δ_R	**10.25**	5.41	**4.08**

Conclusion

- dataset and typology for an analysis of Slovene regional language variants on Twitter
- measures of regional specificity and dispersion
- Future work:
  - annotation of all regions
  - finalisation of the typology
  - comparison with existing dialectological studies
  - comparison with statistical tests

Thank you.