University of Zurich[UZH]

University Research Priority Program
Language and Space

ReLDI

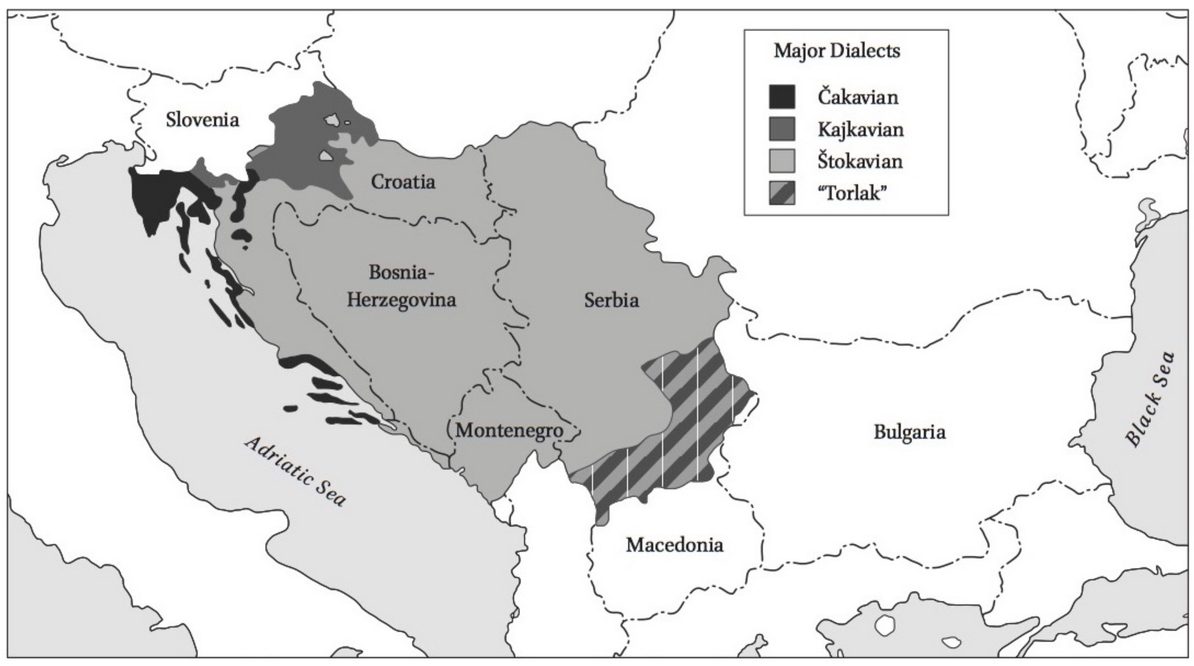
What Twitter Data Say about Bosnian, Croatian, Montenegrin and Serbian (BCMS)

Tanja Samardžić (Language and Space Lab)
Partners:
Maja Miličević Petrović (University of Belgrade) and **Nikola Ljubešić** (Institute Jozef Stefan, Ljubljana)

The work supported by the SNSF-SCOPES Institutional Partnership "Regional Linguistic Data Initiative" and URPP "Language and Space"

Background and motivation:

- One dialect, four standards
- Internationally known, heated debates on the name(s) and the distinction criteria
- Missing:
  - objective discourse
  - scientific evidence
  - systematic surveys

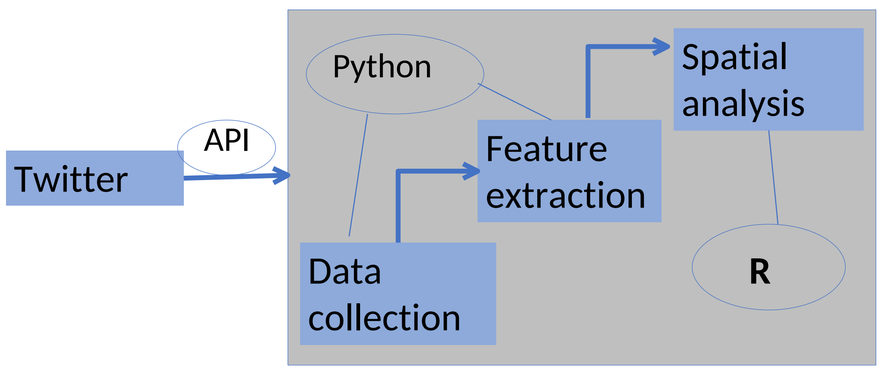


South Slavic dialects:
Štokavian -- the basis for all four standards BCMS
(Source: Alexander 2013)

Variables:

e : je
rdrop
k : h
h : noh
sto : sta
dali : jeli
s : sa
mnogo : puno
ko : tko
ko : tko
long : short
da : inf
synth : nosynth
adjg
ira : isaova
trebam : treba
ica : ika

Twitter as a source of linguistic data



TweetGeo

Montenegro Census 2011
(Source: Wikipedia)

Serbian	265,895	42.88
Montenegrin	229,251	36.97
Bosnian	33,077	5.33
Albanian	132,671	5.27
Serbo-Croatian	12,559	2.03
Roma	5,169	0.83
Bosniak	3,662	0.59
Croatian	2,791	0.45
Russian	1,026	0.17
Serbo-Montenegrin	618	0.10
Macedonian	529	0.09
Montenegrin-Serbian	369	0.06
Hungarian	225	0.04
Croatian-Serbian	224	0.04
English	185	0.03
German	129	0.02
Slovene	107	0.02
Romanian	101	0.02
mother tongue	3.318	0.54
regional languages	458	0.07
without declaration	24.748	3.99

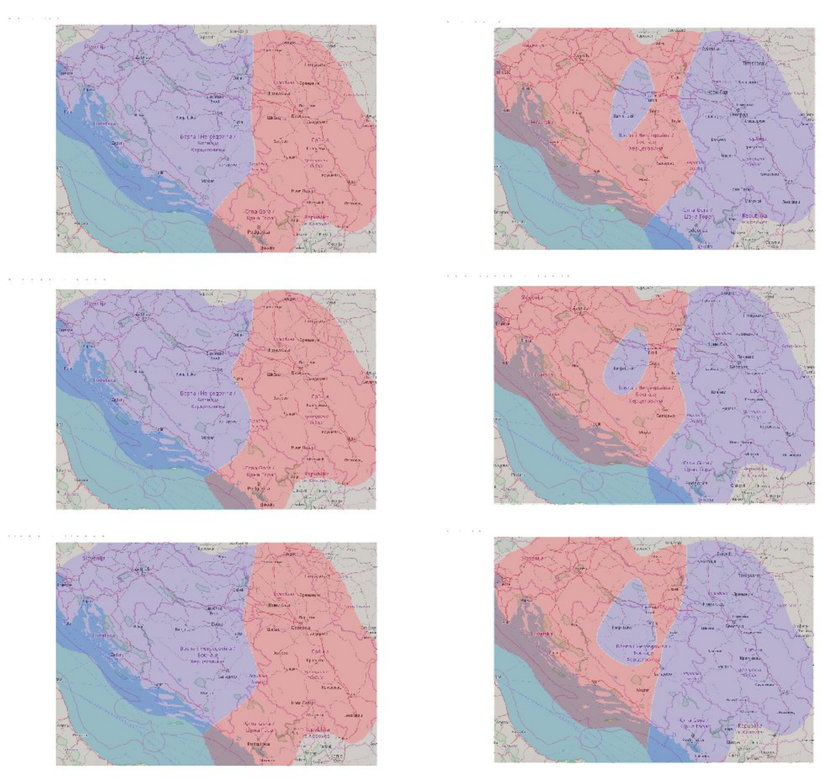
Example variable e : je

e:
Kako to misliš **devojka** si, a nikad nisi zajebala obrve? (RS)
'What do you mean you're a girl and you've never fucked up your eyebrows?'

je:
Pobise mi se neke **djevojke** ispod prozora, sto je ovo majko mila (ME) 'Some girls just got into a physical fight under my window, where is this world going'

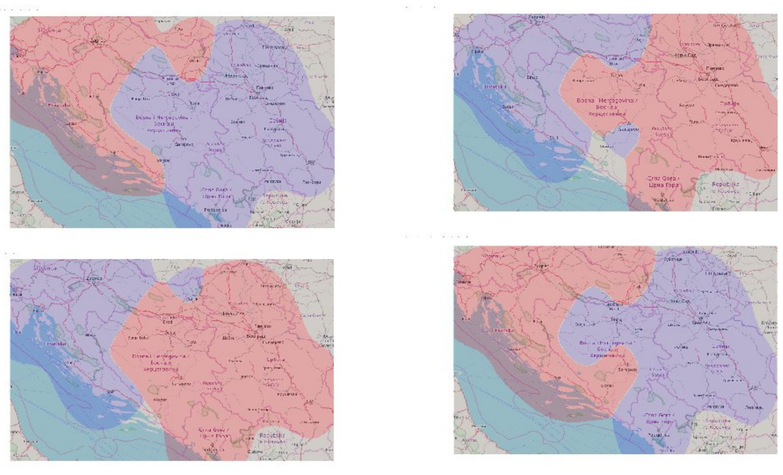
Borders and Boundaries

Kernel Density Estimation

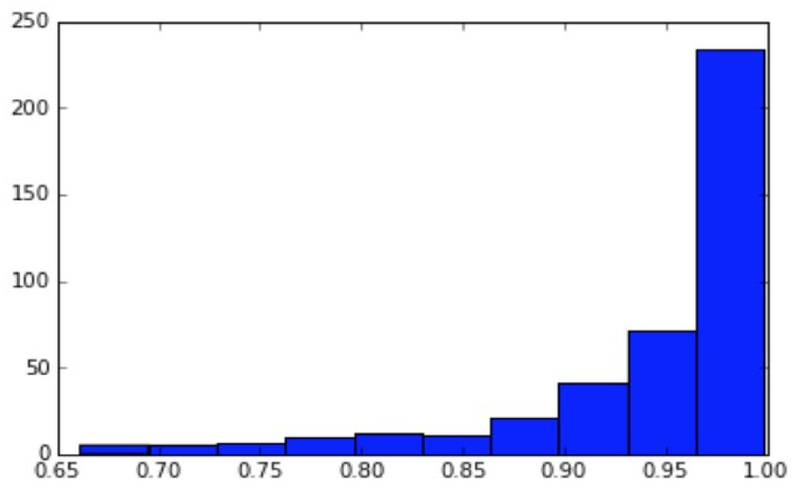


Pattern 1:
Croatia, Bosnia
vs.
Montenegro, Serbia

Pattern 2:
Croatia
vs.
Bosnia,
Montenegro,
Serbia



Accommodation in Serbian Users



Proportion of tweets by mobile users posted in Serbia
X-axis: proportion of tweets in Serbia
Y-axis: the number of users to which the proportion applies

Findings

- Significant difference between mobile and non-mobile users in 8 variables (and overall)
- No evidence of accommodation

References:

Ljubešić, N., M. Miličević Petrović, and **T. Samardžić** (2019). "Borders and boundaries in Bosnian, Croatian, Montenegrin and Serbian: Twitter data to the rescue". *Journal of Linguistic Geography* 6(2), 100-124.

Ljubešić, N., M. Miličević Petrović, and **T. Samardžić** (2019). "Language accommodation on Twitter: The case of Serbian". *Slavistična revija* 67(1), 87-106. (In Croatian)

Ljubešić, N., **T. Samardžić,** and **C. Derungs** (2016). "TweetGeo --- A tool for collecting, processing and analysing geo-encoded linguistic data". In *Proceedings of the 26th International Conference on Computational Linguistics (COLING2016)*. Osaka, Japan.

Alexander, R. (2013). "Language and identity: The fate of Serbo-Croatian". In Roumen Daskalov and Tchavdar Marinov (eds.), *Entangled histories of the Balkans. Volume 1: National ideologies and language policies*, 341–417. Leiden & Boston: Brill.