
Novi CLASSLA mrežni korpusi za hrvatski i ostale južnoslavenske jezike: usporedivost, uvid u žanr, potpunost

Nikola Ljubešić
Institut Jožef Stefan, Slovenija

CLARIN.SI

- CLARIN.SI - slovenski konzorcij CLARIN ERIC infrastrukture, B (podatkovni) centar
- četiri konkordansera (bonito, crystal, crystal-log, kontekst) i podatkovni repozitorij s ~600 zapisa, 290 za slovenski, 81 za hrvatski, 74 za srpski, 31 za bugarski...
- Centri znanja unutar CLARIN.SI
 - CLASSLA - jezični resursi i tehnologije za južnoslavenske jezike
 - ELEXIS - leksikografija
 - CMC - računalno posredovana komunikacija

CLASSLA

- Centar znanja za jezične resurse i tehnologije za južnoslavenske jezike
- Partneri: CLARIN.SI, Institut za hrvatski jezik, CLADA-BG - Bugarski konzorcij za ERIC infrastrukture CLARIN i DARIAH
- helpdesk.classla@clarin.si, "FAQ" dokumenti za slovenski, hrvatski, srpski, makedonski, bugarski jezik
- CLASSLA-Stanza cjevovod za lingvističku obradu tekstova
- CLASSLA-wiki korpusi svih 7 Wikipedija na južnoslavenskim jezicima
- CLASSLA-web korpusi

Novosti u CLASSLA mrežnim korpusima

- usporedivost
 - svi su korpusi izgrađeni istom tehnologijom (alati za prikupljanje, čišćenje i obradu podataka) primjenjenom u istom vremenskom periodu

Novosti u CLASSLA mrežnim korpusima

- usporedivost
 - svi su korpusi izgrađeni istom tehnologijom (alati za prikupljanje, čišćenje i obradu podataka) primjenjenom u istom vremenskom periodu
- uvid u žanr
 - svaki tekst svakog korpusa automatski je označen informacijom o žanru

Novosti u CLASSLA mrežnim korpusima

- usporedivost
 - svi su korpusi izgrađeni istom tehnologijom (alati za prikupljanje, čišćenje i obradu podataka) primjenjenom u istom vremenskom periodu
- uvid u žanr
 - svaki tekst svakog korpusa automatski je označen informacijom o žanru
- potpunost
 - korpus za svaki nacionalni južnoslavenski jezik, uključujući i
 - crnogorski jezik - prvi opći korpus crnogorskog jezika (ili mreže?)
 - makedonski jezik - prvi opći lingvistički označeni korpus makedonskog jezika

Usporedivost

Prikupljanje i čišćenje podataka

- SpiderLing alat za prikupljanje podataka, prilagođen unutar MaCoCu projekta
- Za svaki nacionalni jezik prikupljaju se podaci s pripadajuće vršne domene (.hr)
- Prikupljanje ne staje na nacionalnim vršnim domenama, već zalazi i u tzv. generičke domene (.com, .net itd.), između 7% i 52% podataka
- Čišćenje podataka alatima chared (prepoznavanje kodiranja), jusText (uklanjanje menija, zaglavlja i sl.), onion (uklanjanje bliskih duplikata), fastText i fastSpell (razlikovanje jezika)
- Uklanjanje dokumenata samo s kratkim odlomcima te tekstom kraćim od 75 riječi (prekratak tekst za pouzdano predviđanje žanra, uz umjerenu informativnost)

Lingvističko označavanje

- CLASSLA-Stanza alat (Ljubešić i Dobrovoljc, 2019; Terčon i Ljubešić, 2023)
- Poboľšan alat Stanza kroz sljedeće intervencije
 - Rastavnik na riječi i rečenice temeljen na pravilima
 - Korištenje flektivnih leksikona
 - Korištenje više podataka za učenje nego što je dostupno u projektu UD
 - Dostupnost modela za obradu nestandardnog teksta
 - Modul za obradu mrežnih tekstova - rastavnik za standardne tekstove, modeli za obradu nestandardnih tekstova
 - Kvaliteta oznaka - morfosintaksa ~95%, lematizacija ~99%.

Veličina konačnih korpusa

CLASSLA.web-sl	1,768,555,396 riječi
CLASSLA-web.hr	2,189,723,427 riječi
CLASSLA-web.bs	686,231,779 riječi
CLASSLA-web.cnr	150,672,975 riječi
CLASSLA-web.sr	2,342,626,265 riječi
CLASSLA-web.mk	512,349,136 riječi
CLASSLA-web.bg	3,249,563,781 riječ
sveukupno	10,899,722,759 riječi

Uvid u žanr

Kako mi automatski označavamo podatke

1. Definiramo problem i shemu za označavanje (žanr - novinski, mišljenje, informacija...)
2. Provedemo shemu na određenom skupu podataka ručnim označavanjem (stručnjak svaki tekst označi nekom od oznaka), često koristimo i više stručnjaka da osiguramo da se stručnjaci slažu među sobom (ako slaganje nije dovoljno, vraćamo se na 1)
3. Dio ručno označenih podataka izdvojimo u *skup za provjeru* automatskog označavanja (završni test koji će računalo rješavati samostalno)
4. Preostale podatke koristimo da računalo naučimo rješavati zadatak - *skup za učenje*
5. Jednom kada je računalo završilo učenje, skupom za provjeru ispitamo računalo - izrazito je važno da znamo s kojom uspješnosti računalo rješava zadatak

Kako mi automatski označavamo podatke

1. Definiramo problem i shemu za označavanje (žanr - novinski, mišljenje, informacija...)
2. Provedemo shemu na određenom skupu podataka ručnim označavanjem (stručnjak svaki tekst označi nekom od oznaka), često koristimo i više stručnjaka da osiguramo da se stručnjaci slažu među sobom (ako slaganje nije dovoljno, vraćamo se na 1)
3. Dio ručno označenih podataka izdvojimo u *skup za provjeru* automatskog označavanja (završni test koji će računalo rješavati samostalno)
4. Preostale podatke koristimo da računalo naučimo rješavati zadatak - *skup za učenje*
5. Jednom kada je računalo završilo učenje, skupom za provjeru ispitamo računalo - izrazito je važno da znamo s kojom uspješnosti računalo rješava zadatak

(nadzirano) strojno učenje

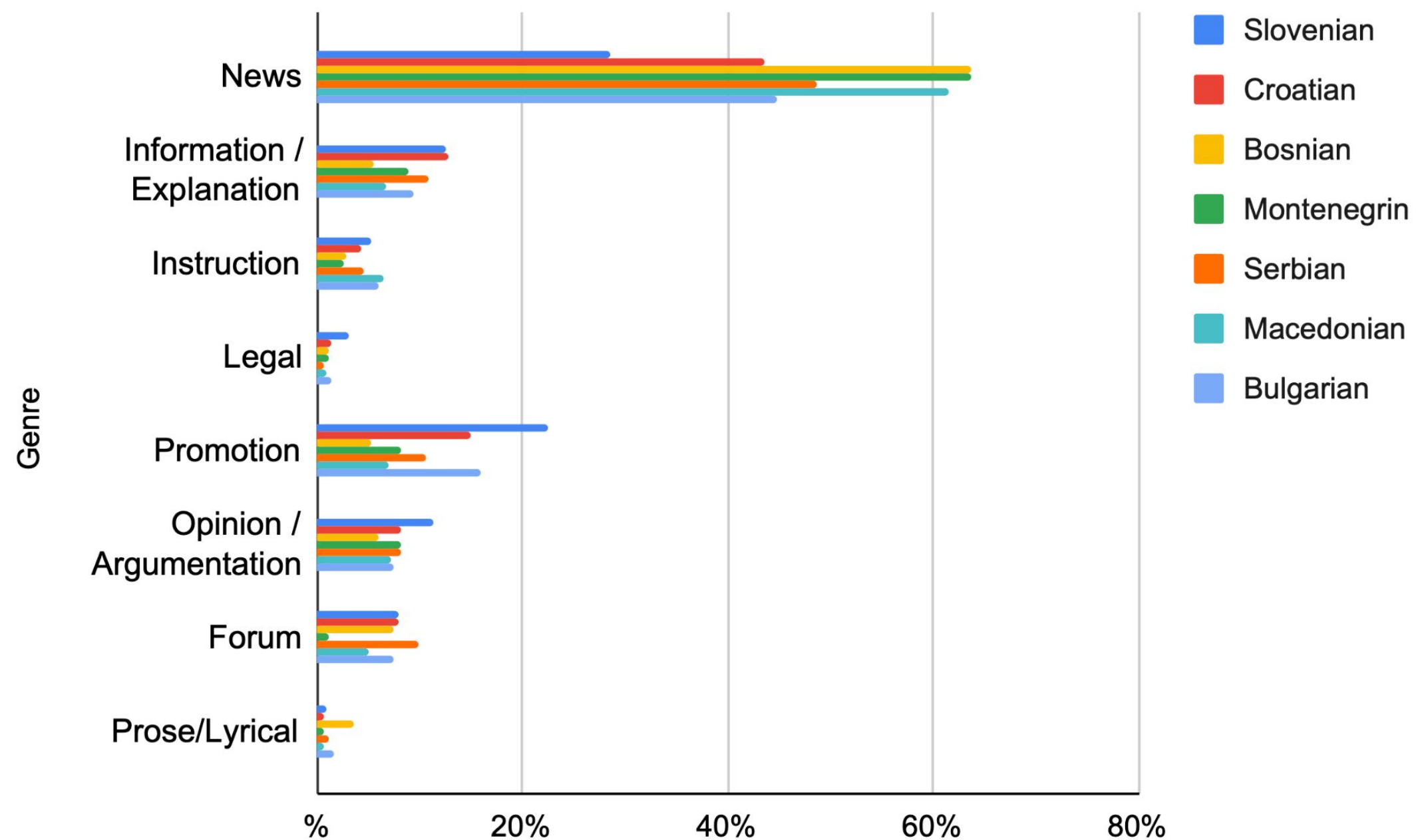
Višejezični *transformer* modeli

- Transformer modeli - osnova nedavnog napretka u području umjetne inteligencije
- Modeli se uče na velikoj količini sirovih podataka, uče zavisnosti u podacima
- XLM-R model učen na 2.5TB tekstovnih podataka iz projekta CommonCrawl, istovremeno se model uči na podacima na oko 100 jezika
- Prilagođavanje modela (*finetuning*) - prilagodba sirovog modela nekom zadatku, prilagodba se radi kroz par tisuća? primjera gdje je problem riješen (*skup za učenje*)
- Višejezični modeli prilagođeni nekom zadatku na jeziku A iznenađujuće? dobro taj zadatak rješavaju i na jeziku B, posebno ako jezici A i B nisu tipološki pretjerano udaljeni

X-GENRE model za označavanje žanra

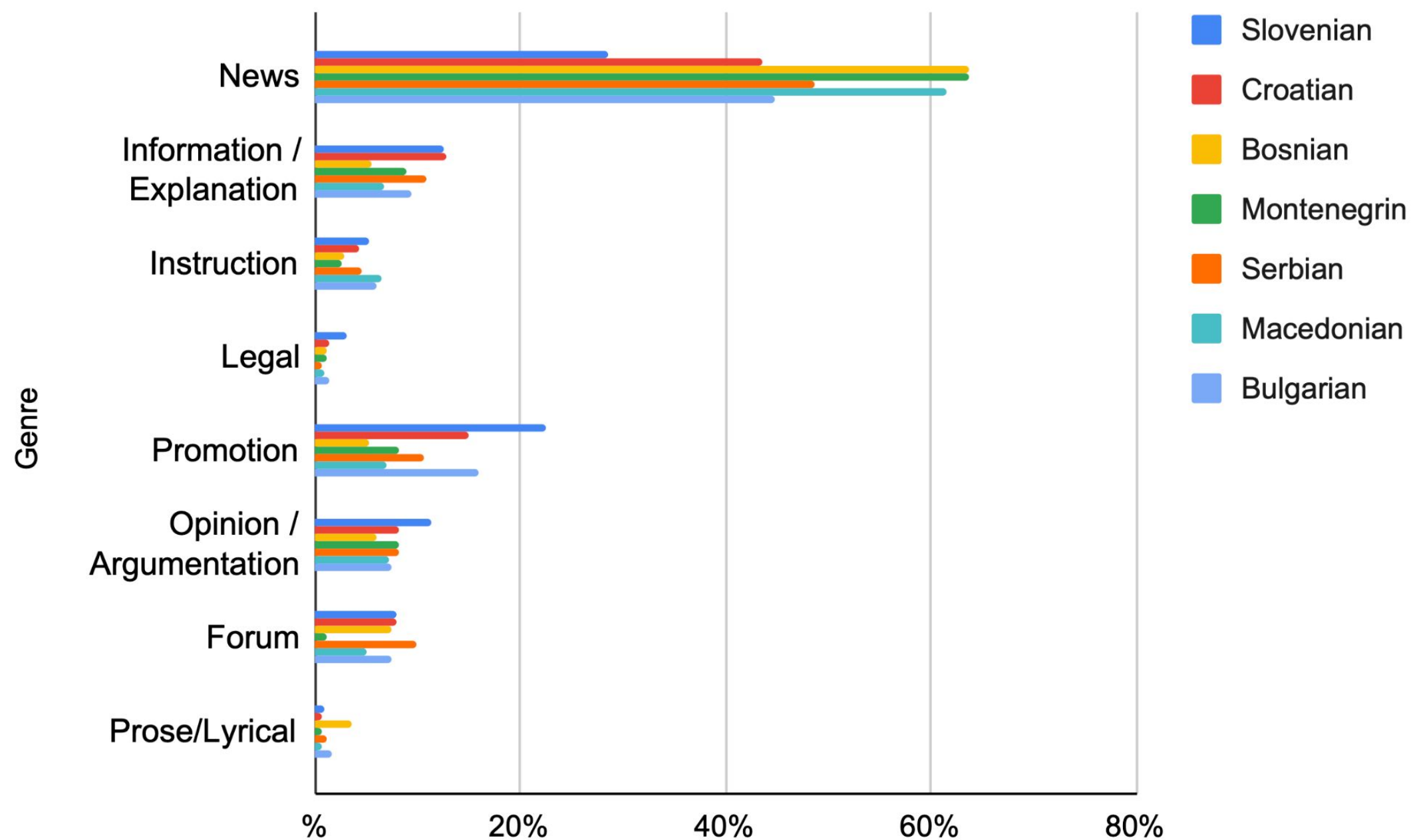
- Skup slovenskih podataka GINCO (Kuzman et al. 2022) proširen skupom engleskih podataka CORE (Egbert et al., 2015) i engleskih podataka FTD (Sharoff, 2018)
- Sva tri skupa podataka prilagođeni X-GENRE shemi (Kuzman et al. 2023)
- XLM-R model prilagođen podacima provjeren na sličnom *skupu za provjeru* pokazuje vrlo dobre rezultate (80% rezultata točno) te umjereno dobre rezultata na teškom, udaljenom *skupu za provjeru* (68% rezultata točno)
- Izrazito dobri rezultati s obzirom na rezultate prije transformer modela (30-60%)
- Razlog uspjeha - transformer modeli rade s numeričkim prikazom teksta koji uključuje i leksičke i sintaktičke (i mnoge druge) značajke teksta!

Distribucija žanrova kroz CLASSLA-web



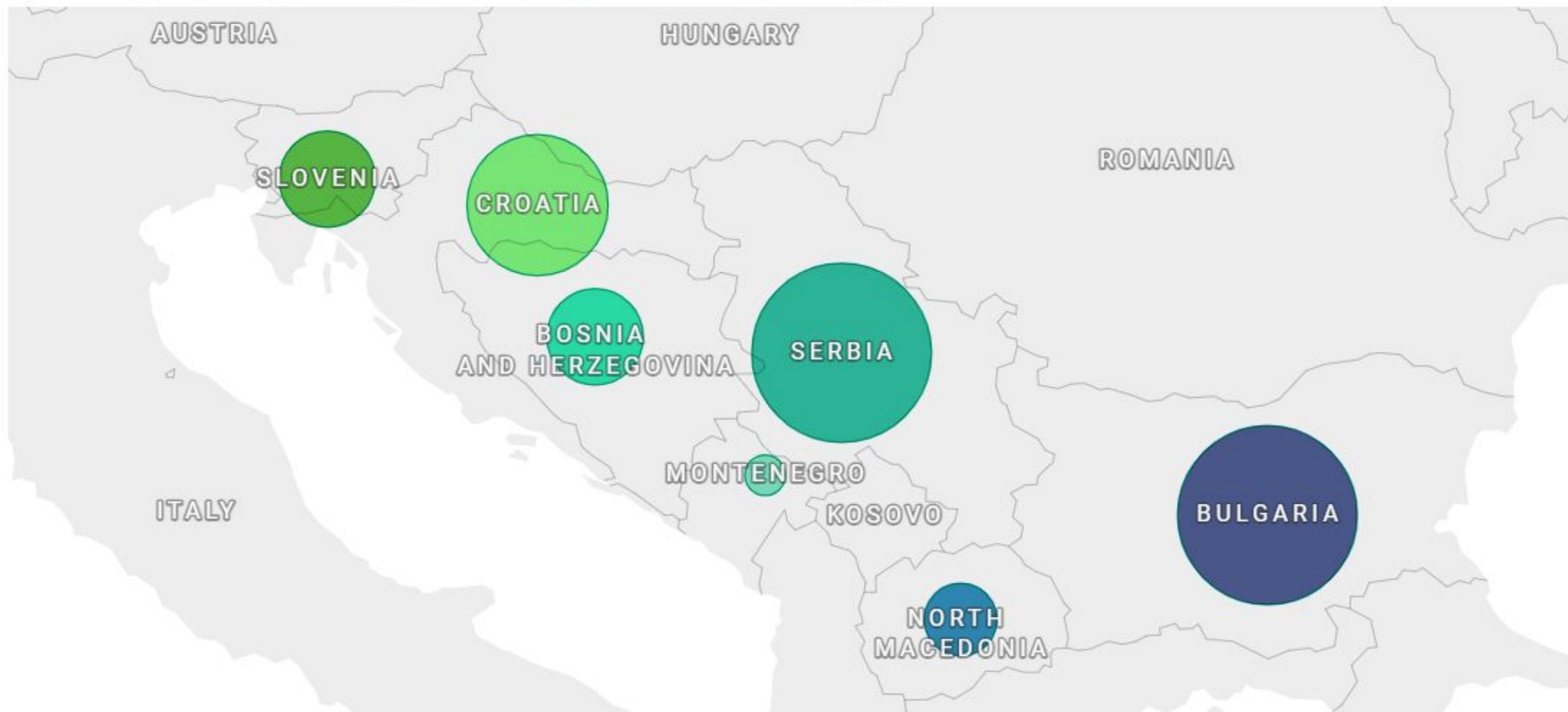
Distribucija žanrova kroz CLASSLA-web

- Poprilično podudaranje između jezika, no ne savršeno
- Novinski tekstovi variraju od 27% u slovenskom do 66% u crnogorskom i bosanskom
- Korelacija između BDP-a po stanovniku (PPP) i količine novinskih tekstova -0.9!!!, promotivnih tekstova 0.94!!!



Potpunost

■ Bosnian ■ Bulgarian ■ Croatian ■ Montenegrin ■ Macedonian ■ Serbian ■ Slovene



Dijeljenje podataka u pojedine korpusu

- Podjela podataka između hrvatskog, bosanskog, crnogorskog i srpskog korpusa
- Vršna domena je glavni kriterij za razlikovanje - tekst pisan ekavskom varijantom unutar `.hr` domene priključen je CLASSLA-web.hr korpusu
- Generičke domene (`.com`, `.net` itd.) su razvrstane metodom strojnog učenja, jedinstvena odluka za sve dokumente unutar neke domene
- Planirana aktivnost - identifikacija specifičnih varijabli unutar svakog teksta, podjela na četiri jezika ovako i onako je pretjerano pojednostavljenje problema
- 16 varijabli iz Ljubešić et al. (2018) "Borders and Boundaries in Bosnian, Croatian, Montenegrin, Serbian: Twitter Data to the Rescue". *Journal of Linguistic Geography*.

Lingvistička obrada makedonskog jezika

- Makedonski jezik do nedavno nije imao alat za morfosintaktičko označavanje i lematizaciju
- Naša prva inačica je također bila manjkava (“Крсте Петков Мисирков е роден...” bio je označen kao ADJ NOUN NOUN...), posljednja inačica značajno je bolja uključivanjem i novinskih podataka u skup za učenje povrh Orwellove “1984.”
- Od prije nekoliko dana makedonski je uključen i u UD kroz uzorak od 1000 rečenica
- CLASSLA-web.mk je prvi opći, lingvistički označeni korpus makedonskoga (2023!)
- Nije sve tako crno za “zanemarene” jezike - mjereći performanse velikih jezičnih modela na različitim jezicima, pad na makedonskom je minimalan!

Veliki jezični modeli

- Ne bi ih trebali preskočiti niti u jednom predavanju o računalnom istraživanju jezika
- Korišteni i u pripremi CLASSLA mrežnih korpusa (XLM-R transformer modeli),
CLASSLA mrežni korpusi korišteni za učenje jezičnih modela
- Nisu spremni za upotrebu u istraživanju jezika (*model kao korpus*), no zasigurno su dovoljno važni i obećavajući da postanu predmet našeg istraživanja
- Modeli drastično napreduju - kako to mjerimo - benčmarcima / mjerilima
- COPA-HR mjerilo za kauzalno promišljanje - BERTić 66%, GPT-3.5 85%, GPT-4 97%
- COPA-HR-CKM (u pripremi) prijevod na žminjsko narječje GPT-3.5 59%, GPT-4 78%

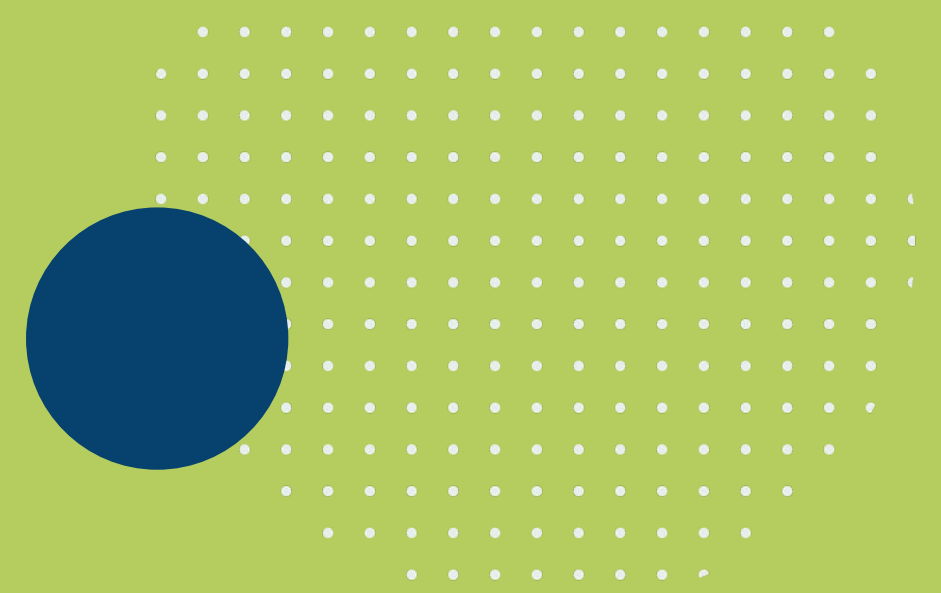
(uzorak)

Budući koraci

- Prikupljanje povratnih informacija od naših korisnika kroz helpdesk.classla@clarin.si
- Prva inačica temeljena je na podacima prikupljanima 2022., upravo se pripremamo ponovno pokrenuti prikupljanje (jedno za zapadne južnoslavenske, jedno za istočne južnoslavenske jezike) zahvaljujući postavljenoj infrastrukturi unutar CLARIN.SI
- Uz označavanje žanra, vjerojatno trebamo i druge automatski pridružene metapodatke (temu / područje?)
- Tekstovni podaci su izvrstan izvor znanja o jeziku, no nisu ni izbliza izdašni kao što su govorni podaci - tražimo način da dođemo do 1. velikih, 2. jeftinih, 3. korisnih govornih korpusa, govorne tehnologije su (uvelike) spremne!

Najzaslužnije osobe

- Taja Kuzman
- Peter Rupnik
- Luka Terčon
- Vit Suchomel
- Tomaž Erjavec
- Petra Bago (hrvatski)
- Virna Karlič (hrvatski, makedonski)
- Marija Runić (bosanski)
- Biljana Stojanoska (makedonski)
- Katerina Zdravkova (makedonski)



HVALA NA PAŽNJI!