LREC-COLING 2024

CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation

Nikola Ljubešić and Taja Kuzman

Jožef Stefan Institute, University of Ljubljana, Institute of Contemporary History, Slovenia

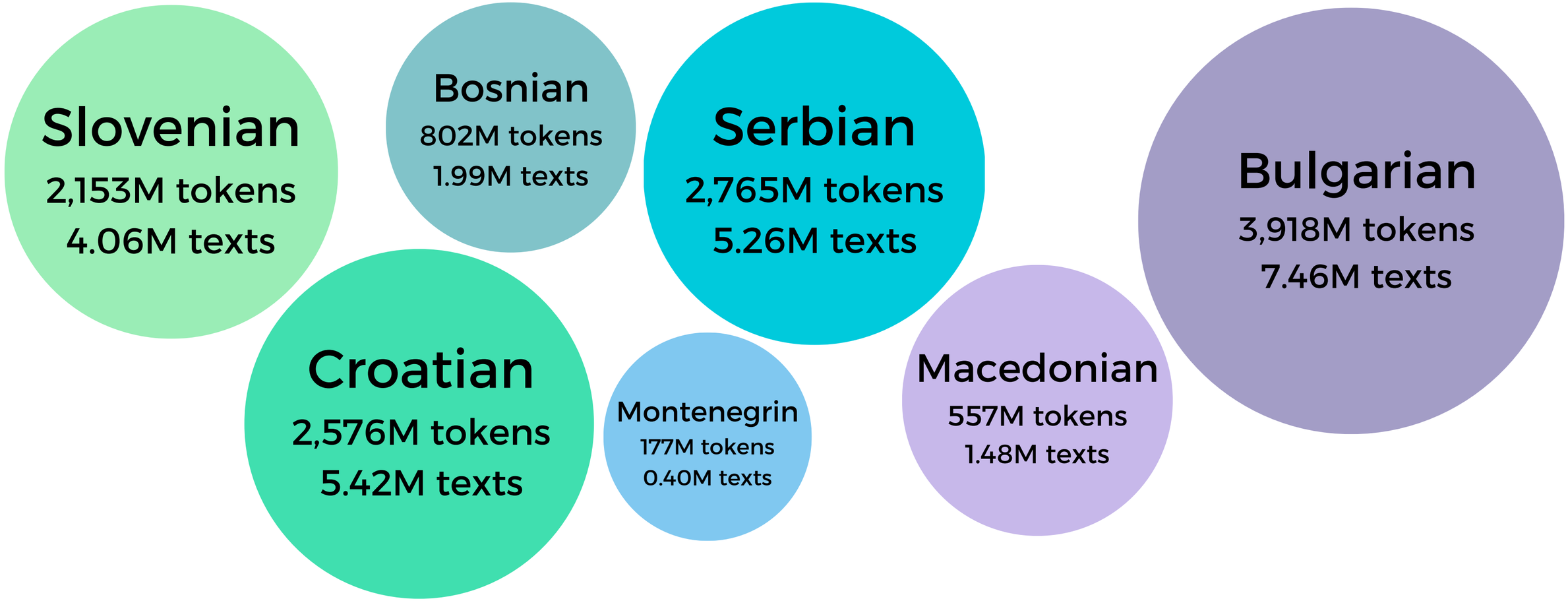
Institut "Jožef Stefan" Ljubljana, Slovenija
CLARIN.SI
CLASSLA K-centre
mezzanine
EMMA
macocu
Co-financed by the Connecting Europe Facility of the European Union

CLASSLA-web Corpora

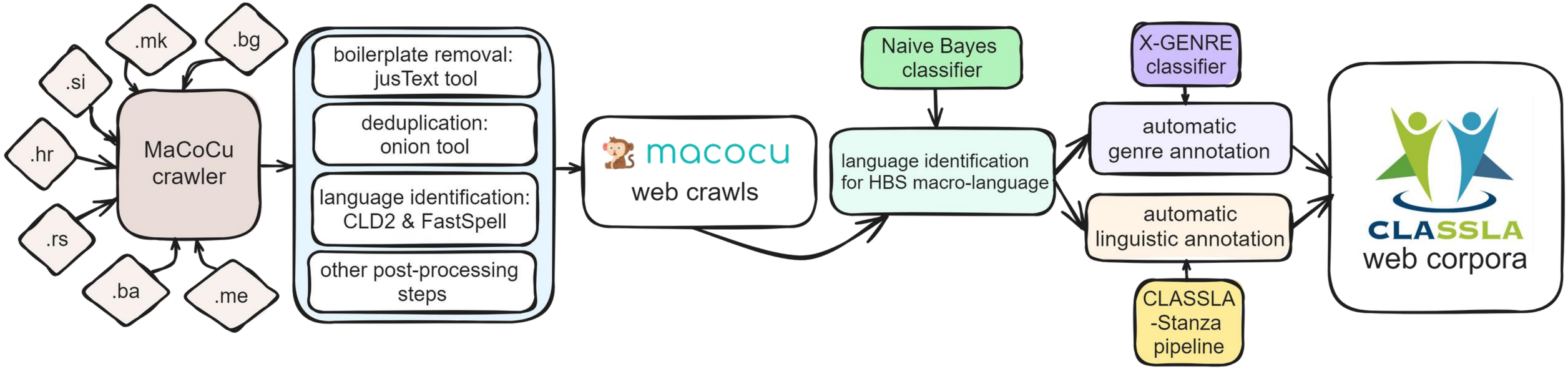
- First comparable corpus collection covering the entire language group
- The biggest general corpus for each South Slavic language
- Automatic genre and linguistic annotation

Web Corpora Sizes

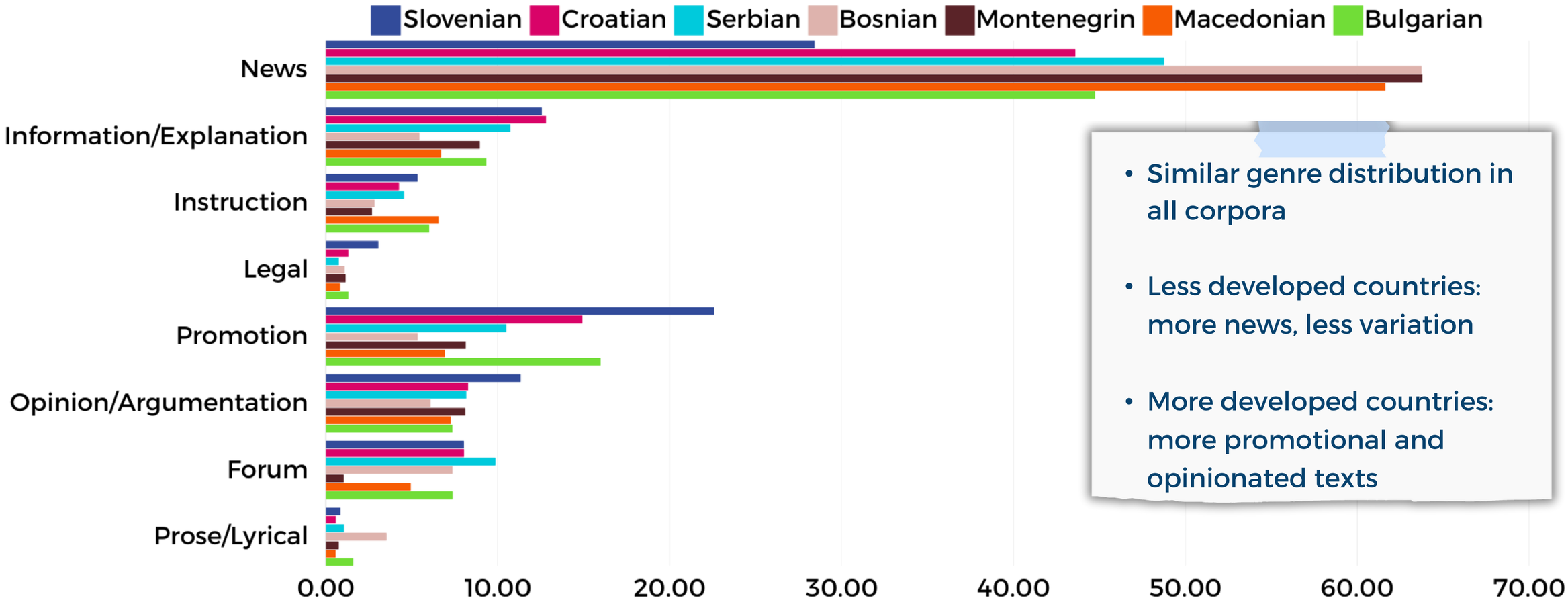
⇒ 13 billion tokens ⇒ 26 million texts



Pipeline


.mk
.bg
.si
.hr
.rs
.ba
.me
MaCoCu crawler
boilerplate removal: jusText tool
deduplication: onion tool
language identification: CLD2 & FastSpell
other post-processing steps
macocu web crawls
Naive Bayes classifier
language identification for HBS macro-language
X-GENRE classifier
automatic genre annotation
automatic linguistic annotation
CLASSLA -Stanza pipeline
CLASSLA web corpora

Genres



- Similar genre distribution in all corpora
- Less developed countries: more news, less variation
- More developed countries: more promotional and opinionated texts

Plans for the Future

- Analyses of genre classifier's multilingual performance
- Series of workshops for wider linguistic community: CLASSLA-Express, visiting 5 South Slavic countries
- Iterative crawling to expand and update corpora: CLARIN.SI crawling infrastructure

Thank you