CLARIN.SI

CLARIN K Centre

ReLDI

Together we are stronger: Collaborative development of resources and technologies for South Slavic languages

Nikola Ljubešić[1] and Tanja Samardžić[2]
Jožef Stefan Institute, Ljubljana, Slovenia
University of Zürich, Switzerland

Overview

CLARIN.SI

CLARIN K Centre

ReLDI

Nikola Ljubešić , Tomaž Erjavec , Maja Miličević Petrović and Tanja Samardžić
Together We Are Stronger: Bootstrapping Language Technology Infrastructure for South Slavic Languages with CLARIN.SI. The CLARIN Book. 2022.

Overview

Nikola Ljubešić , Tomaž Erjavec , Maja Miličević Petrović and Tanja Samardžić *Together We Are Stronger: Bootstrapping Language Technology Infrastructure for South Slavic Languages with CLARIN.SI*. The CLARIN Book. 2022.

The Regional Linguistic Data Initiative (ReLDI)

The CLARIN Knowledge Centre for South Slavic Languages (CLASSLA)

Overview

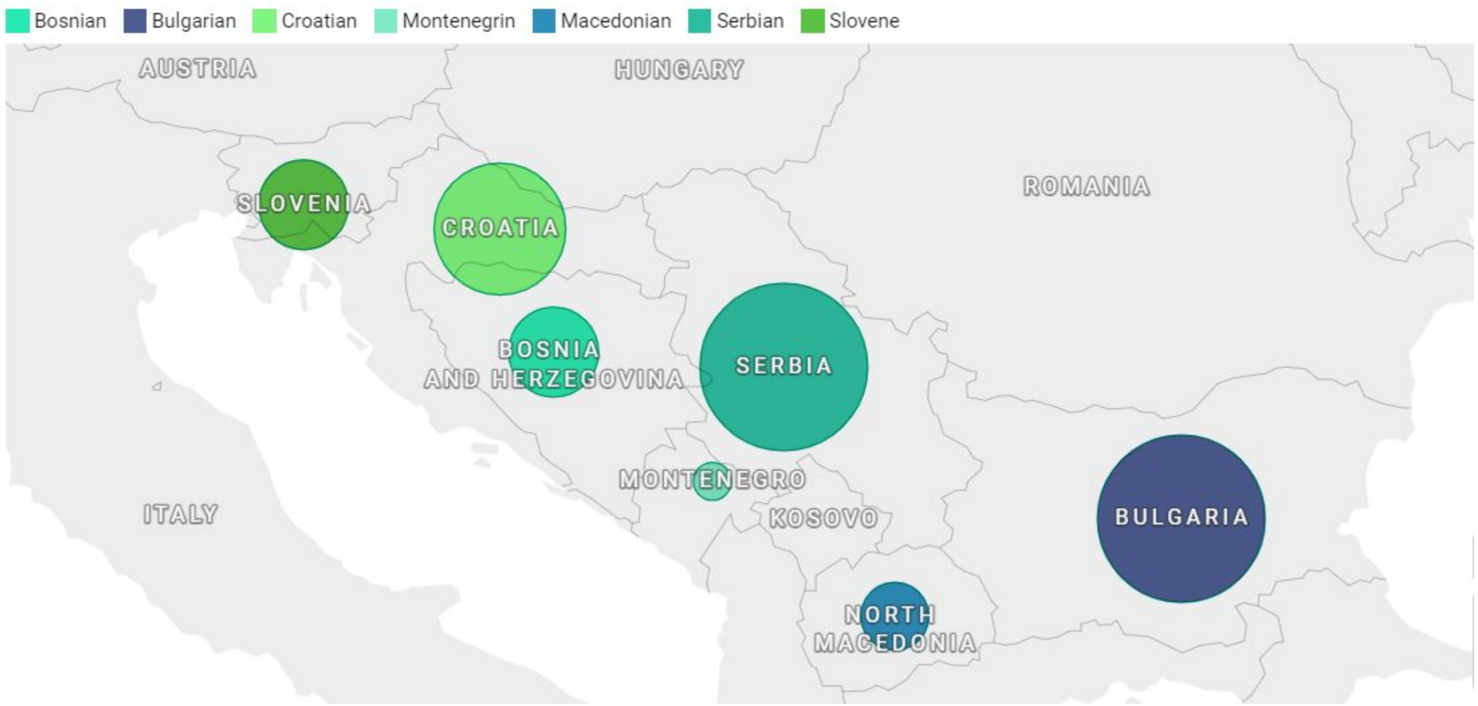
Nikola Ljubešić , Tomaž Erjavec , Maja Miličević Petrović and Tanja Samardžić
Together We Are Stronger: Bootstrapping Language Technology Infrastructure for South Slavic Languages with CLARIN.SI. The CLARIN Book. 2022.

The Regional Linguistic Data Initiative (ReLDI)

The CLARIN Knowledge Centre for South Slavic Languages (CLASSLA)

The story of a language group with heavily lacking LRTs and a decade in which a group of younger researchers managed to turn the tide through… collaboration

South Slavic language group



Language data infrastructure

Two main options for language data infrastructure development

- **Top-down** - organisations (bodies, institutes, states) that care for a language X ensure significant amount of knowledge
  - Publishers
  - Media
  - Libraries
  - Archives
  - The Web
- **Bottom-up** - lack of such (successful) organisations - researchers, enthusiasts have to organise among themselves
  - The Web

State of LRT affairs in 2010 freely available for research

CLARIN.SI CLARIN K Centre ReLDI

	Croatian	Serbian
POS tagger and lemmatizer	❌	❌
POS and lemma training data	❌	✅
Corpus in concordancer	✅	✅
>500M corpus in concordancer	❌	❌
Corpus for offline research	❌	❌
Inflectional lexicon	❌	❌

State of LRT affairs in 2010 freely available for research

CLARIN.SI

CLARIN K Centre

ReLDI

	Croatian	Serbian	Slovenian
POS tagger and lemmatizer	❌	❌	✅
POS and lemma training data	❌	✅	✅
Corpus in concordancer	✅	✅	✅
>500M corpus in concordancer	❌	❌	✅
Corpus for offline research	❌	❌	✅
Inflectional lexicon	❌	❌	✅

State of LRT affairs in 2010 freely available for research

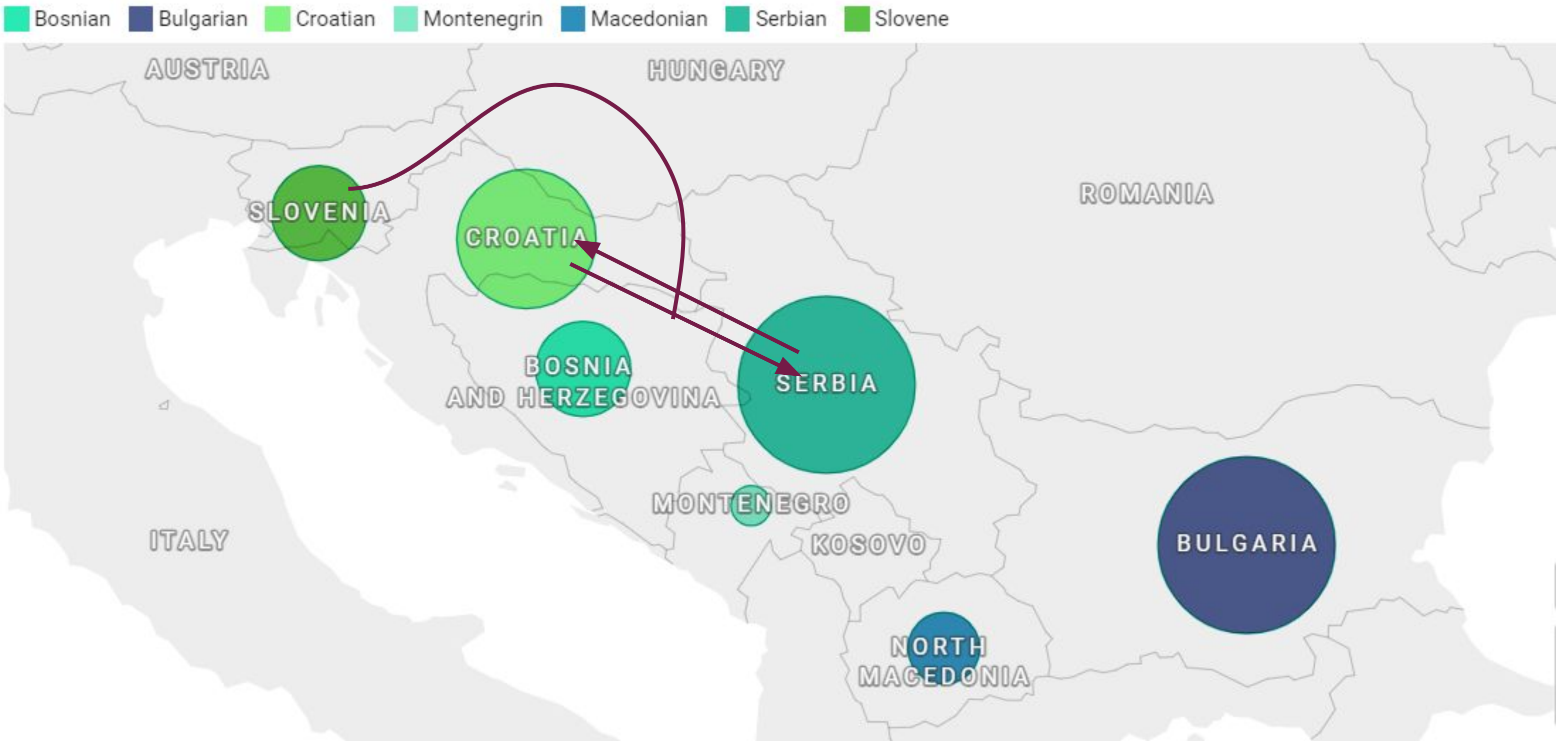
CLARIN.SI

CLARIN K Centre

ReLDI

	bottom-up		top-down
	Croatian	Serbian	Slovenian
POS tagger and lemmatizer	❌	❌	✅
POS and lemma training data	❌	✅	✅
Corpus in concordancer	✅	✅	✅
>500M corpus in concordancer	❌	❌	✅
Corpus for offline research	❌	❌	✅
Inflectional lexicon	❌	❌	✅

ReLDI institutional partnership 2015-2018



ReLDI institutional partnership 2015-2018

CLARIN.SI CLARIN K Centre ReLDI

- Funded by the Swiss National Science Foundation (SNSF)
- Awarded to the universities of:
  - Zurich
  - Belgrade (Serbia)
  - Zagreb (Croatia)
- Quickly connected with Slovenia
  - CLARIN.SI (infrastructure)
  - JANES (resources and research)
- Attached institutions and individuals from
  - Montenegro
  - Bosnia and Herzegovina
  - Macedonia

ReLDI institutional partnership
Main outcomes

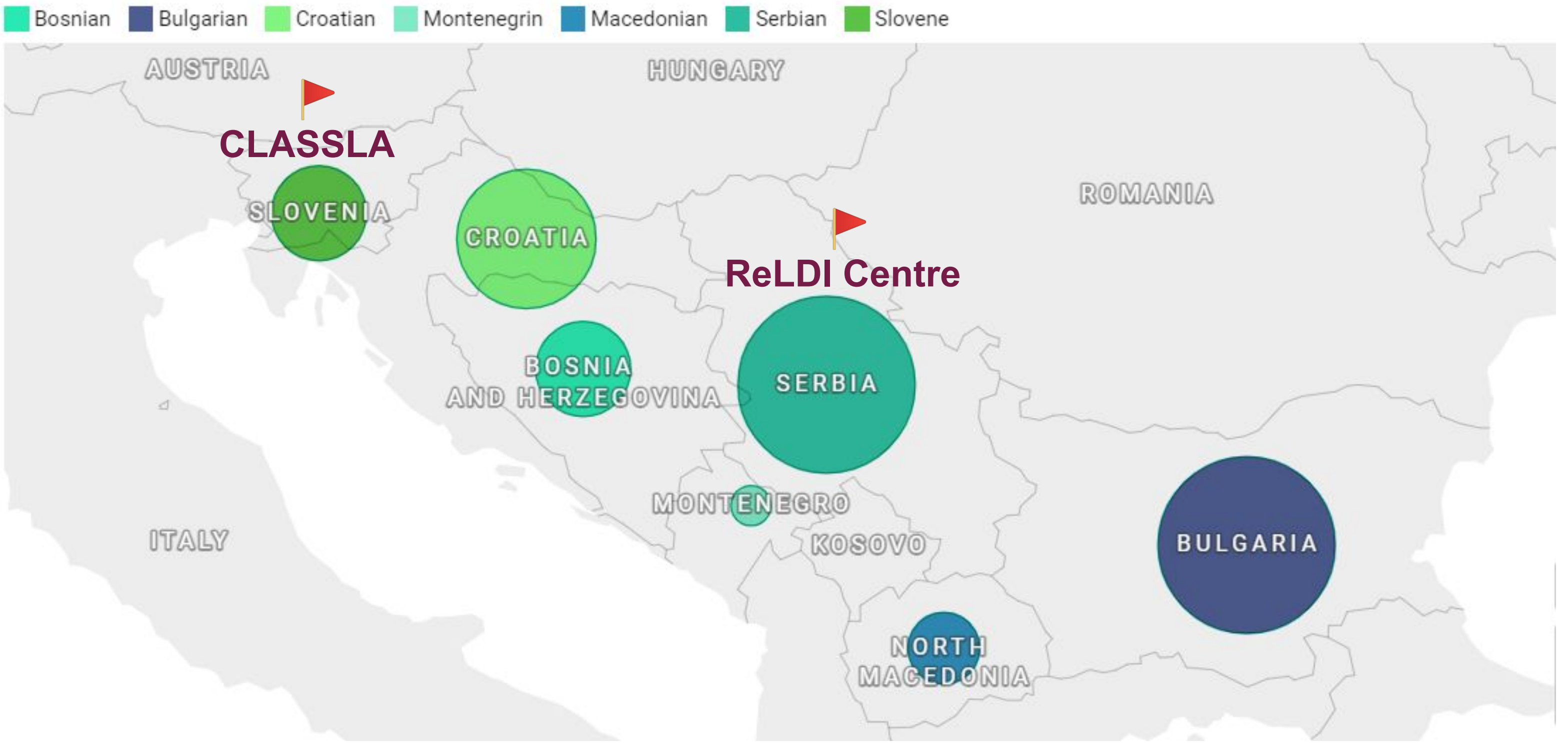
CLARIN.SI CLARIN K Centre ReLDI

- **Data** sets
  - SETimes.SR (new)
  - ReLDI-NormTagNER-hr, ReLDI-NormTagNER-hr (new)
  - hr500k (improved)
  - hrLex, srLex (improved)
- **NLP**: integrated and standardised multilingual processing
  - reldi-tokeniser (supports Slovene, Croatian, Serbian, Macedonian and Bulgarian)
  - reldi-tagger (supports Slovene, Croatian and Serbian, now obsolete)
- **Research** in linguistics: Spatial analysis of Twitter data, accommodation
- **Education**: 6 seminars in empirical methods (in linguistics)

Since 2018

CLARIN K Centre

ReLDI



Scaling the ReLDI success story to all South-Slavic

CLASSLA - CLARIN Knowledge Centre for South Slavic Languages

Envisioned by CLARIN.SI, CLADA-BG joined at its mere beginning, last year expanded with the Institute for Croatian Language and Linguistics

Main results since 2019

- FAQ for Slovenian, Croatian, Serbian, Macedonian, Bulgarian
- The CLASSLA-Stanza linguistic processing pipeline for all languages
- CLASSLA Wikipedia corpora - first annotated corpus of Macedonian!
- CLASSLA web corpora - the largest corpora for all 7 languages!
- Croatian, Bosnian, Serbian parliamentary proceedings in ParlaMint
- Network of researchers intertwined with ReLDI

CLASSLA-Stanza

CLARIN.SI CLARIN K Centre ReLDI

Main differences to Stanza

- Rule-based tokenizer
- More XPOS/UPOS/FEATS training data beyond UD
- Seq2seq lemmatizer dependent on XPOS, not UPOS
- Inflectional lexicon
- Closed classes of words (propagated from tokenizer or lexicon upstream)
- NER for most languages, SRL for Slovenian
- Additional languages and modalities (tokenizer tagger lemmatizer for Macedonian, non-standard written for Slovenian Croatian Serbian, spoken for Slovenian)

CLASSLA-Stanza

SloBENCH official comparison to Stanza v1.5.0 (F1 given CoNLL18-eval)

	Stanza v1.5.0	CLASSLA-Stanza v2.0
Sentence	0.819	0.997
Token	0.998	0.999
Lemma	0.974	0.992
XPOS	0.951	0.983
LAS	0.865	0.911

CLASSLA-Stanza

SloBENCH official comparison to Stanza v1.5.0 (F1 given CoNLL18-eval)

	Stanza v1.5.0	CLASSLA-Stanza v2.0	
Sentence	0.819	98%	0.997
Token	0.998	50%	0.999
Lemma	0.974	69%	0.992
XPOS	0.951	65%	0.983
LAS	0.865	34%	0.911

relative error reduction

Web data!
An important ingredient of our success

- slWaC, hrWaC in 2011
- hrWaC2, bsWaC, srWaC in 2014
- CLASSLA web corpora coming to you in 2023
  - Slovenian 1.9 billion words
  - Croatian 2.4 billion words
  - Bosnian 0.7 billion words
  - Montenegrin 0.2 billion words
  - Serbian 2.5 billion words
  - Macedonian 0.5 billion words
  - Bulgarian 3.5 billion words
- Data are currently being annotated with the latest CLASSLA-Stanza models - continuously improving the linguistic training data for all languages!

SIGWAC

ACL Special Interest Group in Web As a Corpus

Recently took over the lead with three colleagues

First goal - to build a community around SIGWAC that is ready to deal with the challenges in using web data - lifeline for so many less-resourced languages!

https://www.sigwac.org.uk

Join the SIG / mailing list!

http://devel.sslmit.unibo.it/mailman/listinfo/sigwac

ReLDI Centre Belgrade since 2018

CLARIN.SI

CLARIN K Centre

ReLDI

- Collaboration with **CLASSLA** and **CLARIN.SI**
  - Further improvement of hr500k and SETimes.SR (mostly UD)
  - Speech data (50h published, another 50h in progress)
  - COPA-SR, COPA-MK
  - Network events, trainings
- Collaboration with **ZRC SAZU** (Slovenia)
  - MetaLangNEWS-Sr, MetaLangNEWS-Bs, MetaLangNEWS-Me
  - MetaLangNEWS-COMMENTS-Sr, MetaLangNEWS-COMMENTS-Bs, MetaLangNEWS-COMMENTS-Me
- Collaboration with **companies**: Serbian NLP community
- **Network** of institutional collaborations (Serbia, Croatia, Slovenia, Bosnia, Montenegro)

Serbian NLP initiative

COMtext.SR

CLARIN.SI

CLARIN K Centre

ReLDI

The name likely changing to **Serbian.ai**

Funded by **companies**! – service contracts, sponsorships, social responsibility

To create common technology basis for industry

Priorities

- Improve text search in Serbian
- Develop general semantic processing (e.g. text similarity)
- Develop training materials (tutorials) for upskilling

How

- build small, high-quality data sets representing different topic domains
- test available models on the priority tasks to have a baseline
- write didactic documentation for using NLP models
- trace progress

The brave new world of transformer models

All South Slavic languages were covered in mBERT, also later in XLM-R

Following the CamemBERT / BERTje / AlBERTo / PhoBERT trend, we developed **BERTić** (Ljubešić and Lauc, 2021) for Croatian, Serbian, Bosnian, Montenegrin (8,4 billion tokens of web-crawled data)

UPOS/XPOS/NER are not tasks challenging enough

The brave new world of transformer models

All South Slavic languages were covered in mBERT, also later in XLM-R

Following the CamemBERT / BERTje / AlBERTo / PhoBERT trend, we developed **BERTić** (Ljubešić and Lauc, 2021) for Croatian, Serbian, Bosnian, Montenegrin (8,4 billion tokens of web-crawled data)

UPOS/XPOS/NER are not tasks challenging enough

COPA - Choice of Plausible Alternatives dataset, mBERT 54%, BERTić 66%

> Premise: The man broke his toe. What was the CAUSE of this?
> Alternative 1: He got a hole in his sock.
> Alternative 2: He dropped a hammer on his foot.

The brave new world of transformer models

All South Slavic languages were covered in mBERT, also later in XLM-R

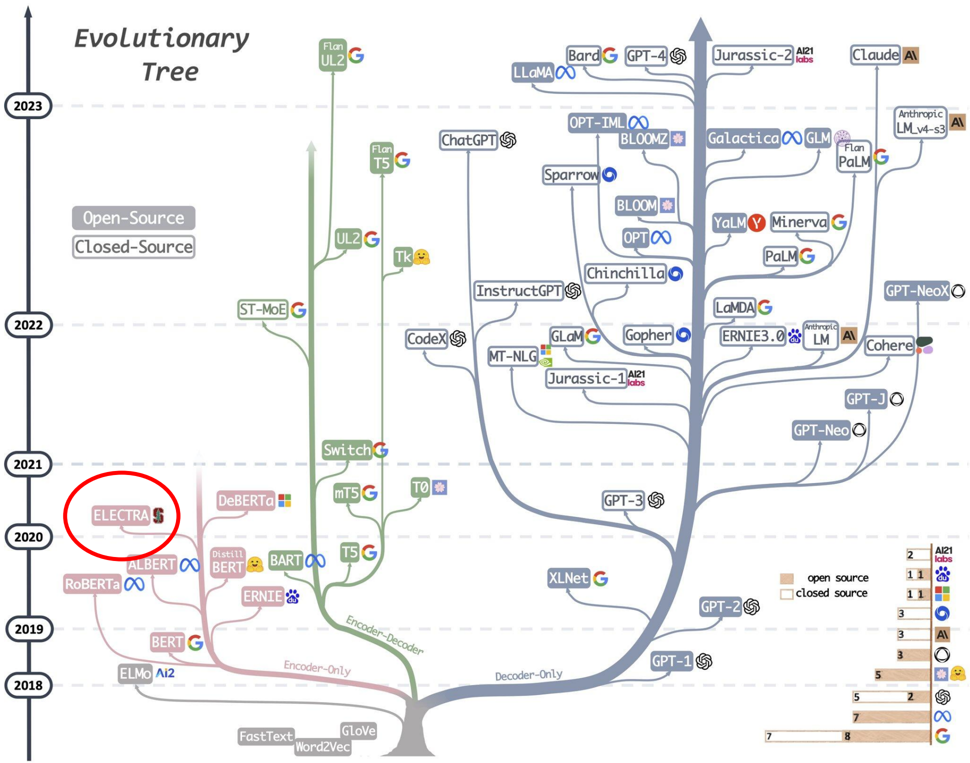
Following the CamemBERT / BERTje / AlBERTo / PhoBERT trend, we developed **BERTić** (Ljubešić and Lauc, 2021) for Croatian, Serbian, Bosnian, Montenegrin (8,4 billion tokens of web-crawled data)

UPOS/XPOS/NER are not tasks challenging enough

COPA - Choice of Plausible Alternatives dataset, mBERT 54%, BERTić 66%

Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.

BERTić is nice, but…



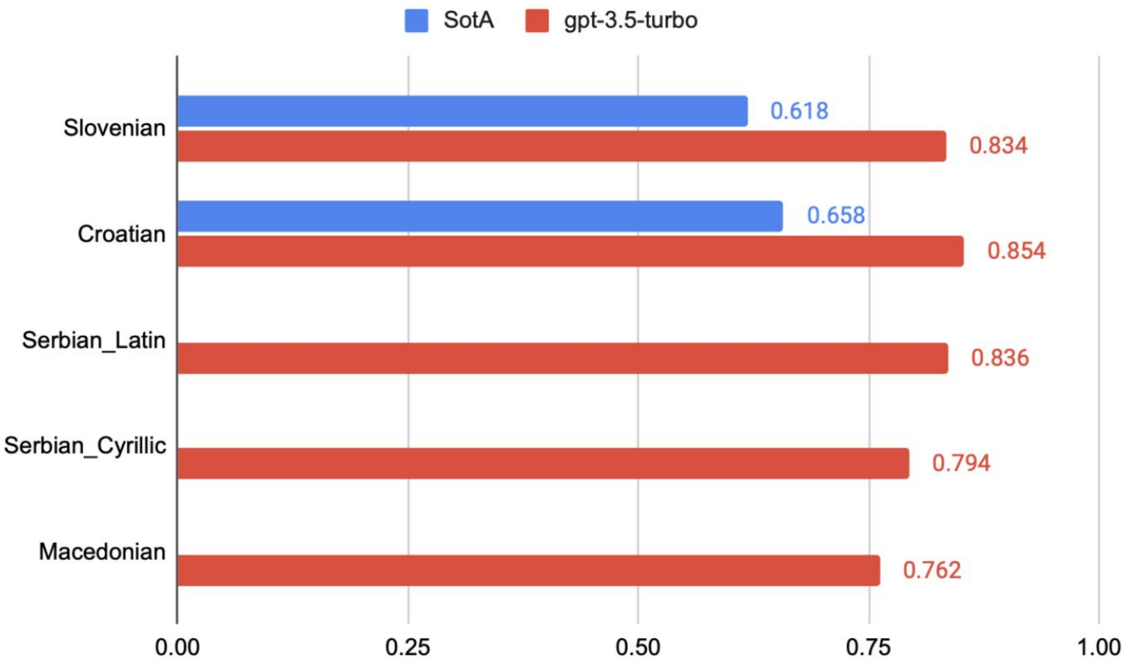
South Slavic COPAs on GPT3.5

Querying gpt-3.5-turbo through the OpenAI API

Prompt:

```
Given the premise "Предметот беше
спакуван во обвивка со меурчиња.",
and that we are looking for the
cause of this premise, which
hypothesis seems more plausible?
Hypothesis 1: "Беше кршлив.".
Hypothesis 2: "Беше мал.".
Please answer only with "1" or "2".
```

SotA was fine-tuned on train, gpt-3.5-turbo is zero-shot



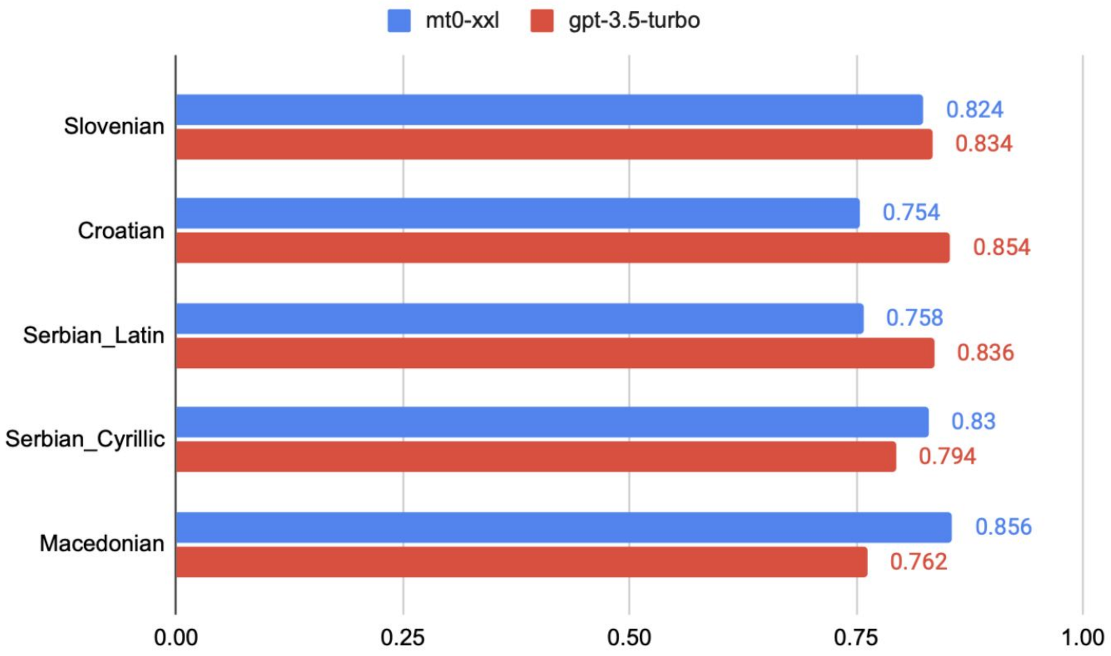
There are already open models with similar performance!

mt0-xxl

mt5-xxl of 16 billion parameters fine-tuned on 13 tasks in 46 languages (no Slavic!)

Big Science Workshop

Muennighoff et al. 2022: Crosslingual Generalization through Miltitask Finetuning



The next frontier - speech

What BERT was for text, wav2vec2 is for speech

ParlaMint going ParlaSpeech (1816 hours of Croatian, other to follow)

Slovenian is in rather good shape

- GOS corpus of 300 hours with manual transcription and rich metadata
- CLASSLA-Stanza linguistic pipeline to process transcripts
- MEZZANINE national project (500k EUR / year)
- Great opportunity to pull it off once again :-)

How to proceed in light of recent developments…

… and the expectation of further developments!

Raw data

CLASSLA web corpora - recurring craws on a bi-annual basis

Speech collections on a similar scale?

Joint modelling projects

We have donated our raw text data to OpenGPT-X (DFKI et al. training a GPT model for European languages) and HPLT (EU project on making HPC centres NLP-ready)

Evaluation benchmarks

COPA is still relevant

Ideas for additional ones?

State of LRT affairs in 2010 freely available for research

CLARIN.SI

CLARIN K Centre

ReLDI

	Croatian	Serbian	Slovenian
POS tagger and lemmatizer	❌	❌	✅
POS and lemma training data	❌	✅	✅
Corpus in concordancer	✅	✅	✅
>500M corpus in concordancer	❌	❌	✅
Corpus for offline research	❌	❌	✅
Inflectional lexicon	❌	❌	✅

State of LRT affairs in 2023 freely available for research

	Croatian	Serbian	Slovenian
POS tagger and lemmatizer	✅	✅	✅
POS and lemma training data	✅	✅	✅
Corpus in concordancer	✅	✅	✅
>500M corpus in concordancer	✅	✅	✅
Corpus for offline research	✅	✅	✅
Inflectional lexicon	✅	✅	✅

+ gigacorpora, language models, automatic speech recognition, language understanding benchmarks….

To conclude…

The power of togetherness!

Does it make sense to aim for a pan-Slavic ReLDI / CLASSLA?

Is SIGSLAV more than enough?

Any similar initiatives / good practices?

Let’s collaborate!