Building a linguistic corpus infrastructure: The South-Slavic example

Nikola Ljubešić
Jožef Stefan Institute
Ljubljana, Slovenia

Corpora in Language Learning, Translation and Research
Zadar, 24 August 2023

South Slavic language group

State of LRT affairs in 2010
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	❌	❌	✅

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….

ReLDI + CLARIN.SI + CLASSLA

ReLDI - Regional Linguistic Data Initiative, grassroots project

- SNF project 2015-2017, partners from Zürich, Zagreb, Belgrade
- NGO in Belgrade 2019-

CLARIN.SI - Slovenian national node of the CLARIN ERIC infrastructure for language resources and technologies

CLASSLA - CLARIN ERIC Knowledge Centre for South Slavic languages, led by CLARIN.SI, partners are Institute for Croatian Language (IHJ) and Bulgarian CLARIN+DARIAH national consortium (CLADA-BG)

Overview

Necessary components for a potent linguistic corpus infrastructure

1. Linguistic processing pipeline
2. Training data for the linguistic processing pipeline
3. Additional data enrichment (genre, sentiment)
4. Data harvesting and encoding (web corpora, parliamentary corpora)
5. Data accessibility (repositories, concordancers)

Linguistic processing pipeline

Linguistic processing pipeline

A few examples say more than a thousand words…

CLASSLA-Stanza, a fork of the Stanza pipeline
https://pypi.org/project/classla/

Paper describing the latest 2.1 version (Terčon and Ljubešić, 2023)
https://arxiv.org/abs/2308.04255

The Scope of CLASSLA-Stanza

Language	Variety	Tok	Morph	Lemma	Depparse	NER	SRL
Slovenian	standard	✓	✓	✓	✓	✓	✓
	nonstandard	✓	✓	✓	X	✓	X
Croatian	standard	✓	✓	✓	✓	✓	X
	nonstandard	✓	✓	✓	X	✓	X
Serbian	standard	✓	✓	✓	✓	✓	X
	nonstandard	✓	✓	✓	X	✓	X
Bulgarian	standard	✓	✓	✓	✓	✓	X
	nonstandard	X	X	X	X	X	X
Macedonian	standard	✓	✓	✓	X	X	X
	nonstandard	X	X	X	X	X	X

Table 1: Table illustrating which tasks are supported by CLASSLA-Stanza for every language and variety. The abbreviations for each task are as follows: Tok - tokenization, Morph - morphosyntactic tagging, Lemma - lemmatization, Depparse - dependency parsing, NER - named entity recognition, SRL - semantic role labeling

Training data for CLASSLA-Stanza

Language	Variety	Morph	Lemma	Depparse	SRL
Slovenian	standard	1,025,639	1,025,639	267,097	209,791
	nonstandard	222,132	222,132	n/a	n/a
Croatian	standard	499,635	499,635	199,409	n/a
	nonstandard	89,855	89,855	n/a	n/a
Serbian	standard	97,673	97,673	97,673	n/a
	nonstandard	92,271	92,271	n/a	n/a
Bulgarian	standard	253,018	253,018	156,149	n/a
Macedonian	standard	153,091	153,091	n/a	n/a

Table 4: Overview table of the number of tokens annotated on every annotation layer for all training datasets used. The abbreviations for each task are as follows: Morph - morphosyntactic tagging, Lemma - lemmatization, Depparse - dependency parsing, SRL - semantic role labeling

Stanza vs. CLASSLA-Stanza

Task	Stanza	CLASSLA-Stanza	Rel. error reduction
Sentence segmentation	0.819	0.997	98%
Tokenization	0.998	0.999	50%
Lemmatization	0.974	0.992	69%
Morphosyntactic tagging - XPOS	0.951	0.983	65%
Dependency parsing LAS	0.865	0.911	34%

Table 3: Comparison of performance on the SloBENCH evaluation dataset by both pipelines

Take-home message #1

“Together we are stronger!”
https://www.degruyter.com/document/doi/10.1515/9783110767377-017/html

We increased the quality and scope of linguistic processing for South-Slavic languages by centrally performing pipeline improvements (inflectional lexicon usage, additional training data, closed class control etc.)

Chasing the State-of-the-Art

dataset	language	variety	CLASSLA	mBERT	cseBERT	BERTić
hr500k	Croatian	standard	93.87	94.60	95.74	***95.81
reldi-hr	Croatian	non-standard	-	88.87	91.63	***92.28
SETimes.SR	Serbian	standard	95.00	95.50	**96.41**	96.31
reldi-sr	Serbian	non-standard	-	91.26	93.54	***93.90

Table 2: Average microF1 results on the morphosyntactic annotation task over five training iterations. The highest score per dataset is marked with bold. The statistical significance is tested with the two-sided t-test over the five runs between the two strongest results. Level of significance is labeled with asteriks signs (*** p<=0.001).

Chasing the State-of-the-Art

dataset	language	variety	CLASSLA	mBERT	cseBERT	BERTić
hr500k	Croatian	standard	80.13	85.67	88.98	****89.21
ReLDI-hr	Croatian	non-standard	-	76.06	81.38	****83.05
SETimes.SR	Serbian	standard	84.64	**92.41**	92.28	92.02
ReLDI-sr	Serbian	non-standard	-	81.29	82.76	***87.92

Table 3: Average F1 results on the named entity recognition task over five training iterations. The highest score per dataset is marked with bold. The statistical significance is tested with the two-sided t-test over the five runs between the two strongest results. Level of significance is labeled with asteriks signs (*** p<=0.001, **** p<=0.0001).

Take-home message #2

We work in a hastily developing research area

We knew already while working on the first version of CLASSLA-Stanza, its performance will be surpassed before it gets released

Computational linguistics does not need state-of-the-art as much as
1. ease of use, 2. stability, 3. reproducibility

Expansion of take-home #1 - we need coordination to be relevant in this quickly developing research area

Training data

Training data

(Almost) all modern tools for linguistic processing rely on machine learning and learning from manually annotated language samples (training data)

Training data for linguistic processing require multiple (never-ending?) iterations over data with improvement of the annotations, but the formalism as well

Very good example is the Universal Dependencies project
https://universaldependencies.org

We heavily coordinate development of Croatian and Serbian training data
https://github.com/reldi-data

Stable releases of the training datasets via the CLARIN.SI repository
https://www.clarin.si/repository/xmlui/

Another demo

Best way to identify limitations of the pipeline, mostly to be followed back to limitations in the training data

Croatian standard pipeline

Croatian non-standard (Internet) pipeline

Take-home message #3

The machine learning tools are as good as their training data

This is less the case with pre-trained models (large language models and similar) where raw, non-annotated data can help a lot, but without data, computers simply cannot know things

Relationship to take-home message #2 - work on data availability (raw or annotated), data do not get obsolete, while technology very much does!

The Macedonian case

By far the most under-resourced language in our language group

Still not in Universal Dependencies, so not supported by Stanza et al.

First open resources - MULTEXT-East project / initiative

- inflectional lexicon
- “1984” novel corpus, non-disambiguated tags and lemmas

Recent work of Katerina Zdravkova (University of Skopje) - “1984” disambiguation

Ongoing work with Biljana Stojanovska (University of Rijeka) in manual annotation of the SETimes.MK corpus - highly needed for training data diversification

Macedonian example (just a few days ago)

```
>>> import classla
>>> nlp = classla.Pipeline('mk') # run classla.download('mk') beforehand if necessary
>>> doc = nlp('Крсте Петков Мисирков е роден во Постол.')
>>> print(doc.to_conll())
# newpar id = 1
# sent_id = 1.1
# text = Крсте Петков Мисирков е роден во Постол.
1       Крсте   крсте   ADJ     Afpms-n Definite=Ind|Gender=Masc|Number=Sing    _       _
2       Петков  петков  NOUN    Ncmsnn  Case=Nom|Definite=Ind|Gender=Masc|Number=Sing   _
3       Мисирков        мисирков        NOUN    Ncmsnn  Case=Nom|Definite=Ind|Gender=Masc|N
4       е       сум     AUX     Vapip3s-n       Aspect=Prog|Mood=Ind|Number=Sing|Person=3|P
5       роден   роден   ADJ     Ap-ms-n Definite=Ind|Gender=Masc|Number=Sing|VerbForm=Part
6       во      во      ADP     Sps     AdpType=Prep    _       _       _       _
7       Постол  постол  NOUN    Ncmsnn  Case=Nom|Definite=Ind|Gender=Masc|Number=Sing   _
8       .       .       PUNCT   Z       _       _       _       _       _
```

Macedonian example (now, with SETimes.MK added)

```
>>> import classla
>>> nlp = classla.Pipeline('mk') # run classla.download('mk') beforehand if necessary
>>> doc = nlp('Крсте Петков Мисирков е роден во Постол.')
>>> print(doc.to_conll())
# newpar id = 1
# sent_id = 1.1
# text = Крсте Петков Мисирков е роден во Постол.
1       Крсте   Крсте   PROPN   Npmsnn  Case=Nom|Definite=Ind|Gender=Masc|Number=Sing   _
2       Петков  Петков  PROPN   Npmsnn  Case=Nom|Definite=Ind|Gender=Masc|Number=Sing   _
3       Мисирков        Мисирков        PROPN   Npmsnn  Case=Nom|Definite=Ind|Gender=Masc|N
4       е       сум     AUX     Vapip3s-n       Aspect=Prog|Mood=Ind|Number=Sing|Person=3|P
5       роден   роден   ADJ     Ap-ms-n Definite=Ind|Gender=Masc|Number=Sing|VerbForm=Part
6       во      во      ADP     Sps     AdpType=Prep    _       _       _       _
7       Постол  Постол  PROPN   Npmsnn  Case=Nom|Definite=Ind|Gender=Masc|Number=Sing   _
8       .       .       PUNCT   Z       _       _       _       _       _
```

Take-home message #4

With a small amount of well-thought-through data we can make a difference for less-resourced languages / varieties

Relationship to take-home #1 - together we are stronger, we could not have done this without Katerina and Biljana, as they could not have done this without us

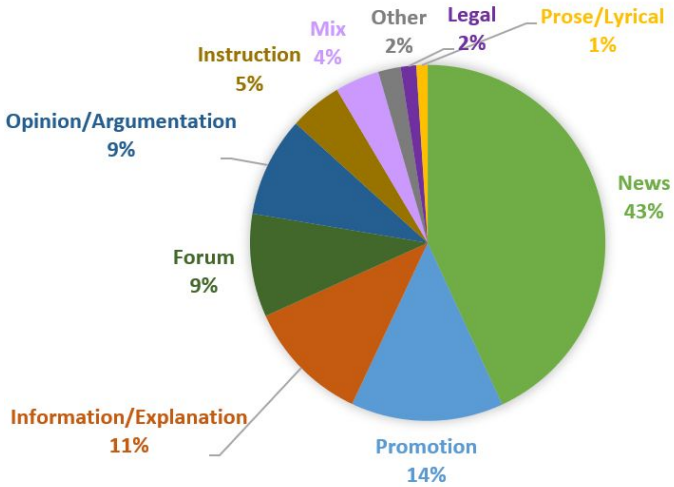
Data enrichment

Genre Identification

Used to be a very hard machine learning task as it requires both lexical and syntactic features for discrimination between genres

Joint work with Taja Kuzman (I am just supervising, Taja is doing the heavy lifting)

Training data in Slovenian and English, developing BERT-like multilingual models that enable high-quality predictions in unseen languages



Genre distribution in the CLASSLA web corpora

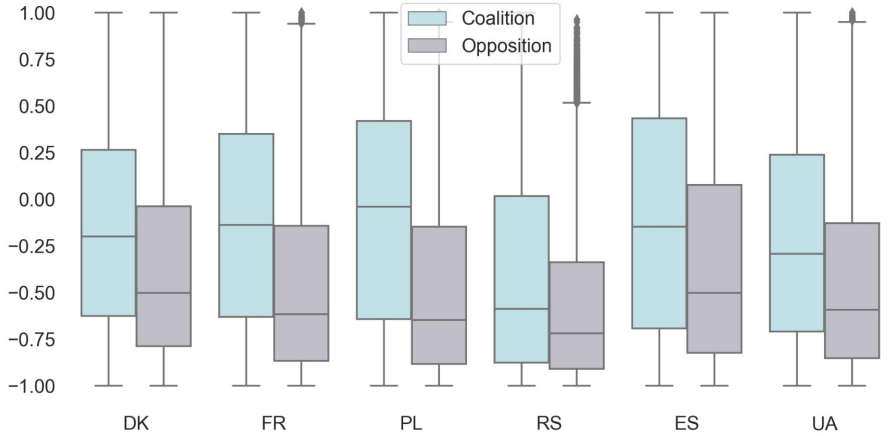
Sentiment identification in parliamentary corpora

Joint work with Michal Mochtak and Peter Rupnik

Training data in Bosnian / Croatian / Serbian, Czech, English, Slovak, Slovenian

Multilingual BERT-like models trained ensuring high-quality predictions

Ongoing work with Bojan Evkoski, Igor Mozetič, Petra Kralj Novak on differences in sentiment of speeches between coalition and opposition



Take-home message #5

Modern pre-trained language models enable data enrichment across languages with a few hundred/thousand training examples in one / a few languages

With the VERY large models few-shot (just a few training examples) or zero-shot (no training data at all) becomes a possibility, but this technology is not safe enough to be applied for wide-range data enrichment and downstream analysis at this point

Data harvesting and encoding

Web corpora

Cheap way to obtain very large general-purpose corpora of decent quality

2011 hrWaC and slWaC - largest Croatian and Slovenian corpora (cited by 139)

2014 hrWaC2, bsWaC, srWaC (cited by 192)

Reference Slovenian corpus GigaFida (publications in 2012, 2016, 2020, cited by 45), FidaPlus (publication in 2007, cited by 11), Fida (publications in 1998, cited by 30)

Croatian National Corpus (publications in 2002 and 2009, cited by 137)

Serbian reference corpus SrpKor (publications in 2011 and 2014, cited by 81)

CLASSLA web corpora

Developed an infrastructure for continuous development of web corpora inside the macocu.eu project, hosted in the CLARIN.SI infrastructure

First version of CLASSLA web corpora

CLASSLA-web.bs	0.7B
CLASSLA-web.bg	3.5B
CLASSLA-web.cnr	0.2B
CLASSLA-web.hr	2.3B
CLASSLA-web.mk	0.5B
CLASSLA-web.sr	2.4B
CLASSLA-web.sl	1.9B

CLASSLA web corpora

Slovenian, Croatian and Serbian corpora were already released in a pilot version inside CLARIN.SI concordancers - collecting feedback from the research community!

https://www.clarin.si/info/k-centre/classla-web-bigger-and-better-web-corpora-for-croatian-serbian-and-slovenian-on-clarin-si-concordancers/

Take-home message #6

While building the Croatian web corpus, why not building also the first general-purpose corpus of Macedonian? (Or Albanian?) With a fraction of required capacity we can develop language resources for additional languages.

Web is a great source of linguististic / extralinguistic knowledge.

The textual modality has already been exploited heavily.

What about the spoken modality? Are we ready for cheap and large spoken corpora, to be built for multiple languages for a fraction of the price?

Parliamentary corpora

The ParlaMint CLARIN ERIC project https://www.clarin.eu/parlamint

Textual version http://hdl.handle.net/11356/1486

Linguistically annotated version http://hdl.handle.net/11356/1488

Machine-translated version into English http://hdl.handle.net/11356/1810

ParlaSpeech - task inside ParlaMint - generating cheap speech+text collections, for training automatic speech recognition, but also for potential linguistic use

ParlaSpeech-HR - 1,815 hours of Croatian speech+text, currently building with Danijel Koržinek, Matyáš Kopp and Per Erik Solberg similar datasets for Bosnian, Czech, Norwegian, Polish, Serbian

Take-home message #7

We do not need coordination only in the bottom-up / grassroots cases

No need to develop resources (e.g., “national corpora”) for each language separately

The synergistic effect can be very strong also for well-resourced languages

Spoken corpora for many languages?

Data accessibility

Two mainstream ways to share corpus data

Repository - necessary to share data by the FAIR (Findable, Accessible, Interoperable, Reusable) principles

https://www.clarin.si/repository/xmlui/

Concordancer - main tool for non-technical people to be able to search through, or summarise large collections of data

https://www.clarin.si/info/concordances/

Take-home message #8

It all does not matter much if you do not share your data and technology with others

It is great to have research on high-quality data published, it is even better to have your research published + enable others to perform additional research, further enrich your data, use your technology on new data etc. etc.

Last, but by far not least - without a stable research infrastructure it is very hard to share your data and technology with others - special shout-out to the back-bones of CLARIN.SI Tomaž Erjavec and Simon Krek

Conclusion - take-home messages

1. Together we are stronger - we need to coordinate to get the most from our efforts!
2. Moving target - technology changes very quickly, this is an opportunity, but a curse as well - pick your battles wisely!
3. Data rule - work on your raw and annotated data, they do not get stale as technology does!
4. Small matters as well - for less-resourced languages / varieties / domains you can make a difference already with small datasets
5. Pre-trained models - enable cross-lingual enrichment, even with little (or no) training data
6. Scale to more languages - while developing / researching for language X, why not work / coordinate on a series of additional languages / varieties
7. Rich are poor as well - all the above does not hold for less-resourced-languages, but also for large well-resourced languages, we need coordination, collaboration, synergy
8. Sharing is caring - it all does not matter much if you haven’t shared your data / technology with other researchers

Thank you!

Discussion