Нови текстуални корпуси за моделовање српског језика

УДК 811.163.4'322.2

САЖЕТАК: Овај рад ће представити текстуалне корпусе за српски (и српскохрватски) који се могу користити за тренирање великих језичких модела, а који су јавно доступни на једном од неколико значајних веб репозиторијума. Сваки корпус ће бити класификован помоћу више метода и његове карактеристике ће бити детаљно описане. Поред тога, рад ће представити три нова корпуса: нови кровни веб-корпус за српскохрватски, нови висококвалитетни корпус заснован на докторским дисертацијама похрањеним у Националном репозиторијуму докторских дисертација са свих универзитета у Србији, и паралелни корпус превода сажетака из истог извора. Јединственост старих и нових корпуса биће оцењена путем стилометријских метода заснованих на фреквенцији, и укратко ће се дискутовати о резултатима.

КЉУЧНЕ РЕЧИ: корпуси, српски језик, језички модели, евалуација.

РАД ПРИМЉЕН: 12. април 2024.
РАД ПРИХВАЋЕН: 15. мај 2024.

Михаило Шкорић
mihailo.skoric@rgf.bg.ac.rs
ORCID: 0000-0003-4811-8692
Универзитет у Београду
Рударско-геолошки факултет
Београд, Србија

Никола Јанковић
nikolajankovickv@gmail.com
ORCID: 0000-0003-3484-4220
Универзитет у Београду
Филолошки факултет
Београд, Србија

1. Увод

С наглим повећањем доступних текстуалних података кроз феномен *Big Data* почетком двадесет првог века, убрзо се увидело да се ти подаци могу користити за изградњу корпуса за моделовање природног језика. Подаци са интернета, чија количина брзо расте, најпре су коришћени као додатак подацима из књига, који су споријe расли, али, са повећаним интересовањем за количину, полако али сигурно достигли су и надмашили удео података заснованим на књигама у разним корпусима

за тренирање језичких модела. Данас, већина јавно доступних корпуса користи податке са интернета, углавном због лабавијих ограничења ауторских права.

У контексту овог истраживања, категорисаћемо скупове података у следеће категорије на основу њиховог порекла:

O_1 Веб-корпуси – корпуси који садрже текстове прикупљене са отвореног интернета, првенствено коришћењем аутоматског сакупљања HTML страница;

O_2 Корпуси високог квалитета – корпуси који садрже текстове из уџбеника или сличних извора, односно научне публикације свих врста (научне монографије и чланци у научним часописима, радови са конференција/зборника, тезе итд.) и других објављених нелитерарних дела (владини и правни текстови који нису присутни на интернету, филозофски текстови, кулинарски рецепти итд.) прикупљени првенствено из дигитално насталих докумената или сканираних физичких копија;

O_3 Литерарни корпуси – корпуси који садрже објављене литерарна дела укључујући митологију и религиозне текстове;

O_4 Синтетички корпуси – корпуси који садрже текстове који нису са Интернета, а који су креирани машинским путем коришћењем метода генерисања природног језика или машинског превођења;

O_5 Мешовити корпуси – корпуси креирани коришћењем текстова који потичу из мешавине наведених и других извора.

Осим очигледне разлике у изворној грађи, ове категорије се разликују и по процесу креирања текстова које садрже: док су текстови из прве три групе (O_1-O_3) углавном креирани од стране људи (нико не може гарантовати да преузета HTML страница није машински генерисан текст), само друга и трећа група садрже оне текстове који су нужно уређени, тј. прочитани и исправљени пре објављивања, што захтева време, али осигурава виши квалитет. Текстови креирани од стране машина (O_4) обично се припремају најбрже од свих, али су повезани са нижим квалитетом.

Још једна важна класификација је облик корпуса. С тим у вези, корпусе класификујемо на следећи начин:

F_1 Корпуси необележеног текста – корпуси који садрже само текстове који се могу користити за претренирање великих језичких модела као што су *BERT*[1] (Devlin et al. 2018) и *GPT*[2] (Radford et al. 2018);

F_2 Обележени корпуси – корпуси где су токени, реченице или други сегменти обележени, нпр. текстови обележени према врсти речи, или реченице обележене на основу сентимента– ови корпуси се обично користе за фино подешавање модела за задатке обележавања текста;

F_3 Паралелни корпуси – корпуси где свака реченица (или други сегмент) има специфичан пар, нпр. превод на други језик или одговор на питање – ови корпуси се обично користе за тренирање модела са генеративном компонентом као што је *T5*[3] модел (Raffel et al. 2020).

У наредном одељку, рад ће се фокусирати на корпусе српског и српскохрватског језика прибављене из три значајна веб извора за корпусе и скупове података: *Hugging Face*,[4] *CLARIN virtual language repository*,[5] и *European Language Grid*.[6] Други појединачни репозиторијуми (нпр. *GitHub* пројекти) нису претраживани. Основне информације о корпусима преузетим са ових платформи биће представљене у одељку 2., новоприпремљени корпуси биће представљени у одељку 3., а експеримент везан за процену њихове јединствености биће представљен у одељку 4. Коначно, дискусија о резултатима ове процене налази се у одељку 5.

2. Доступни корпуси

Као што је већ поменуто, у овом одељку биће представљена листа корпуса која је састављена прегледом три веб извора: *Hugging Face* (HF), *CLARIN virtual language repository* (VLO) и *European Language Grid* (ELG), с тим што су језици корпуса ограничени на домен српско-хрватског макројезика, односно српски, црногорски, хрватски

1. Bidirectional Encoder Representations from Transformers
2. Generative Pre-trained Transformer
3. Text-To-Text Transfer Transformer
4. huggingface.co, највеће чвориште на Интернету за објављивање великих језичких модела.
5. www.clarin.eu дигитална инфраструктура која пружа податке, алате и сервисе за подршку истраживањима заснованим на језичким ресурсима.
6. live.european-language-grid.eu, платформа за све европске језичке технологије настале из MetaNet-a.

и босански, јер тренирање са сродним језицима може бити корисно у одређеним сценаријима, посебно за задатке који укључују вишејезично разумевање или превођење. Преузети корпуси су категорисани и по форми (примарна класификација) и по пореклу (секундарна класификација). За корпусе необележеног текста (F_1), минимални услов за укључивање био је тридесет милиона речи за веб-корпусе (O_1), и три милиона речи за остале (O_2-O_5). Пошто је већина корпуса у овој категорији веб порекла, спроведено је и кратко истраживање о дедупликацији. Када су у питању обележени (F_2) и паралелни корпуси (F_3), критеријум за величину је био постављен на 3000 реченица, јер су ови ресурси много ређи и обично се користе само за фино подешавање.

2.1 Јавно доступни корпуси необележеног текста

Табеле 1 и 2 детаљно описују двадесет четири преузета корпуса необележеног текста. Прва табела се фокусира само на српске корпусе, док друга представља остале корпусе у оквиру српско-хрватског макројезика, укључујући неколико кровних корпуса (произведених обједињавањем и дедупликацијом неколико других корпуса).

Назив	**Јез.**	**Порекло**	**Вел.**	**Издавач**	**Чвор.**
srWaC	sr	web	493	ReLDI	VLO
cc100_sr	sr	web	711	Conneau et al.	HF
mC4-sr	sr	web	800	Google	HF
OSCAR-sr	sr	web	632	OSCAR proj.	HF
CLASSLA-sr	sr	web	752	CLASSLA	VLO
MaCoCu-sr	sr	web	2.491	Bañón et al	VLO
PDRS1.0	sr	web	602	ISJ	VLO
SrpKorNews	sr	web	468	JeRTeh	HF
SrpELTeC	sr	literary	5,3	JeRTeh	HF

Табела 1. Корпуси необележеног текста искључиво на српском језику, доступни на истраженим чвориштима скупова података. Величина је представљена у милионима речи (M)

Назив	Јез.	Порекло	Вел.	Издавач	Чвор.
meWaC	cnr	web	80	ReLDI	VLO
hrWaC	hr	web	1.250	ReLDI	VLO
bsWaC	bs	web	256	ReLDI	VLO
cc100_hr	hr	web	2.880	Conneau et al.	HF
CLASSLA-hr	hr	web	1.341	CLASSLA	VLO
CLASSLA-bs	bs	web	534	CLASSLA	VLO
hr_news	hr	web	1.433	CLASSLA	HF
MaCoCu-cnr	cnr	web	161	Bañón et al	VLO
MaCoCu-hr	hr	web	2.363	Bañón et al	VLO
MaCoCu-bs	bs	web	730	Bañón et al	VLO
riznica	hr	mixed	87	IHJJ	VLO
HPLT 1.2-sh	mixed	web	10.030	HPLT proj.	HF
BERTić-data	mixed	∼web	8.388	CLASSLA	VLO
MaCoCu-hbs	mixed	web	5.490	CLASSLA	HF
XLM-R-BERTić-data	mixed	∼web	11.539	CLASSLA	HF

Табела 2. Српско-хрватски корпуси необележеног текста доступни на истраженим чвориштима скупова података, који нису искључиво на српском језику. Величина је представљена у милионима речи (M).

Осим четири корпуса, остали корпуси садрже искључиво материјал преузет са Интернета. Један корпус је литерарни, један је мешовити, а преостала два су агрегирана и класификована као готово веб-корпуси (∼web), пошто укључују поменути мешовити корпус, *riznica* (Ćavar and Brozović Rončević 2012).

Већина ових корпуса је производ неколико засебних подухвата. Корпуси *srWac*, *meWac*, *hrWac* и *bsWac* настали су преузимањем садржаја веб страница са домена највишег нивоа за поменута четири језика (Ljubešić and Klubička 2014; Ljubešić and Erjavec 2021), што је резултирало са више од две милијарде речи. Процес је касније поновљен како би се произвели *CLASSLA-sr*, *CLASSLA-hr* и *CLASSLA-bs*, прикупивши преко 2,5 милијарди речи (Ljubešić and Lauc 2021).

Додатних шест корпуса је изведено из скупа података *Common Crawl*: *OSCAR-sr* (632M) је изведен из скупа података пројекта *OSCAR* (Suárez, Sagot, and Romary 2019), *mC4-sr* (800M) је изведен из

вишејезичног скупа података C4, који је креирала компанија Гугл (Xue et al. 2021), *HPLT 1.2-sh* (преко 10 милијарди речи) је добијен из скупа података пројекта HPLT (Aulamo et al. 2023), док су *cc100_sr* и *cc100_hr* (преко 3,5 милијарди речи) изведени из cc100 скупа података (Conneau et al. 2019). Треба напоменути да су неки од ових подухвата произвели додатне корпусе који нису укључени у ову студију јер нису испунили минимални услов од три милиона речи.

MaCoCu пројекат сакупљања веб страница (Banón et al. 2022; Kuzman and Ljubešić 2023) произвео је *MaCoCu-sr* (2,5 милијарде речи), *MaCoCu-cnr* (151M), *MaCoCu-hr* (2,2 милијарде), *MaCoCu-bs* (686M), и њихов дедупликовани агрегат, *MaCoCu-hbs* (5,5 милијарди речи).

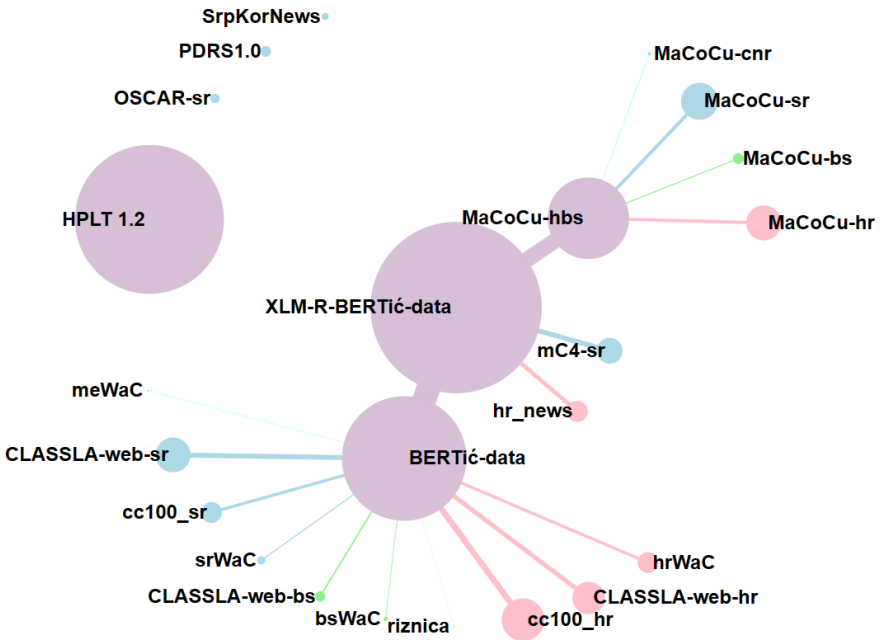
Још један значајан подухват дедупликације укључује четири *WaC* корпуса, три *CLASSLA* корпуса, *cc100_sr*, *cc100_hr* и корпус *riznica*. Тиме је произведен обједињени корпус од а 8,4 милијарди речи, објављен под именом *BERTić-data* (Ljubešić and Lauc 2021).

И *BERTić-data* и *MaCoCu-hbs* су поново обједињени, заједно са *mC4-sr* и *hr_news* (1,4 милијарде речи) како би се креирао *XLM-R-BERTić-data* (11,5 милијарди речи), највећи објављени српскохрватски обједињени корпус. Он, међутим, не укључује *Common Crawl* скупове *HPLT 1.2-sh* и *OSCAR-sr*, нити недавно објављени корпус *PDRS1.0* (Wasserscheidt 2023) (600M), а ни корпус *SrpKorNews* (Krstev and Stanković 2023) (468M). Ово указује на могућност креирања нових, већих кровних корпуса како за српски, тако и за српскохрватски језик. Комплетан садашњи подухват агрегације корпуса је приказан на слици 1.

Једини литерарни корпус доступан на вебу који испуњава величинске услове и даље је *SrpELTeC* корпус (Krstev 2021; Stanković et al. 2022), који обухвата 120 романа и садржи 5 милиона речи. Процес објављивања није био ограничен, јер корпус садржи материјале који су прекорачили праг истека ауторских права.

2.2 Јавно доступни обележени корпуси

Испитивањем одабраних чворишта са скуповима података пронашли смо 10 обележених корпуса за српски језик (табела 3). Усредсредили смо се на корпусе који су обележени помоћу скупа ознака *Universal POS* (17 класа), при чему није било ограничења када је у питању препознавање именованих ентитета (NER) и обележавање сентимента. За разлику од корпуса необележеног текста који су се примарно налазили на HF и VLO, ови корпуси су такође у значајном броју пронађени на ELG.



Слика 1. Хијерархија агрегације јавно доступних српскохрватских корпуса необележеног текста. Плава боја представља корпусе српског језика, цијан црногорског, ружичаста хрватског, зелена босанског, а љубичаста боја представља корпусе мешовитог језичког порекла.

Корпуси српског језика обележени према врсти речи укључују четири ресурса: *SrpKor4Tagging* (Stanković et al. 2020) са 60K, *Intera* (Gavrilidou et al. 2004; Stanković et al. 2020) са 47,5K, *1984* (Krstev, Vitas, and Erjavec 2004) са 6,7K и *reldi_sr* (Miličević and Ljubešić 2016) са 5,5K реченица. Два додатна ресурса у којима су обележене и врсти речи (POS) и именовани ентитети (NER): *SETimes_sr* (Samardžić et al. 2017) и *NormTagNER-sr* (Ljubešić et al. 2023) са 7K и 4K реченица, редом, и 5 NER класа. Три додатна ресурса у којима су обележени само именовани ентитети (NER): *SrpELTeC-gold* (Todorović et al. 2021; Frontini et al. 2020), заснован на делу литерарног корпуса који обухвата 52K реченица (7 NER класа), српски део корпуса *WikiAnn* (Rahimi, Li, and Cohn 2019) са 40K реченица (3 NER класе), и српски део корпуса *polyglot_ner* (Al-Rfou et al. 2015) са пола милиона реченица и

3 NER класе. Последњи ресурс је српски део *mms* корпуса обележених сентиментом (Augustyniak et al. 2023) са 76K реченица (3 класе).

Назив	Јез.	Порекло	Анот.	Вел.	Издавач	Чвор.
SrpKor4Tagging	sr	mixed	POS	60	JeRTeH	ELG
1984	sr	literary	POS	6,7	MULTEXT-East	ELG
Intera	sr	textbook	POS	47,5	META-SHARE	ELG
reldi_sr	sr	web	POS	5,5	ReLDi	HF
SETimes_sr	sr	web	POS, NER	4	ReLDi	HF
NormTagNER-sr	sr	web	POS, NER	7	ReLDi	VLO
SrpELTeC-gold	sr	literary	NER	52	JeRTeh	ELG
wikiann-sr	sr	∼textbook	NER	40	Wikimedia	HF
polyglot_ner-sr	sr	mixed	NER	560	Al-Rfou et al	HF
mms-sr	sr	web	sentiment	76	Brand24	HF

Табела 3. Обележени корпуси искључиво српског језика доступни на истраженим чвориштима скупова података. Величина је представљена у хиљадама реченица (K).

Слична ситуација је уочена када је у питању шира језичка слика, где је пронађено још десет ресурса (табела 4), при чему се листа већином састоји од хрватских еквивалената истих, претходно поменутих обележених ресурса. Ови ресурси обухватају обележени хрватски превод романа *1984*, потом *reldi_hr*, *NormTagNER-hr*, хрватски део *WikiAnn* и *polyglot_ner* корпуса и хрватске реченице обележене према сентименту из *mms* корпуса. Такође, ту је и босански *mms*, као и и српскохрватски *WikiAnn*, који је изведен из српскохрватских чланака на Википедији. Величине ових корпуса су углавном упоредиве са њиховим српским еквивалентима.

Једини јединствени ресурс са ове листе је *hr500k* (Ljubešić et al. 2016), којије обележен са NER ознакама (5 класа) у целости (25K реченица), а само делимично са скупом ознака Universal POS (9K реченица).

2.3 Јавно доступни паралелни корпуси

Укупно тринаест паралелних ресурса је пронађено за српски језик (табела 5), укључујући пет корпуса паралелних превода, један

Назив	Јез.	Порекло	Анот.	Вел	Издавач	Чвор.
1984	hr	literary	POS	6,7	MULTEXT-East	ELG
hr500k	hr	literary	POS	9	ReLDi	HF
reldi_hr	hr	web	POS	7	ReLDi	HF
NormTagNER-hr	hr	web	POS, NER	8	ReLDi	VLO
hr500k	hr	literary	NER	25	ReLDi	HF
wikiann-hr	hr	∼textbook	NER	40	Wikimedia	HF
wikiann-sh	sh	∼textbook	NER	40	Wikimedia	HF
polyglot_ner-hr	hr	mixed	NER	630	Al-Rfou et al	HF
mms-hr	hr	web	sentiment	78	Brand24	HF
mms-bs	bs	web	sentiment	36	Brand24	HF

Табела 4. Корпуси обележеног српскохрватског текста на истраженим чвориштима скупова података, који не садрже искључиво српски језик. Величина је представљена у хиљадама реченица (K).

корпус са паровима текст-резиме, један корпус парафраза, два корпуса са паровима питање-одговор (*QA*) и четири скупа података са инструкцијама (парови текстуалних инструкција и решења).

Сви *QA* корпуси и корпуси са инструкцијама су синтетичког порекла и креирани су помоћу машинског превођења постојећих корпуса на енглеском језику: *m_mmlu-sr* и *m_hellaswag-sr* су, редом, преводи *mmlu* (Hendrycks et al. 2021) и *HellaSwag* (Zellers et al. 2019) уз помоћ GPT3.5; *airoboros-3.0-sr* је превод скупа података *airoboros-3.0* (Tunstall et al. 2023), а *open-orca-slim-sr*, *ultrafeedback-bin-sr* и *alpaca-cleaned-sr* су српске верзије скупова података *SlimOrca* (Lian et al. 2023), *UltraFeedback* (Cui et al. 2023) и *alpaca-cleaned* (Taori et al. 2023), редом, преведени помоћу сервиса *Google translate*. Заједно са српским делом полусинтетичког скупа *tapaco* (Scherrer 2020) (оригинални текст можда није синтетички, али су реченице упариване коришћењем једноставног алгоритма), већина доступних паралелних ресурса за српски језик су синтетички.

Када су у питању ресурси који нису синтетички, пет корпуса паралелних превода (преводи са енглеског) обухватају литерарни корпус *80 jours* (Vitas et al. 2008) са 3,7K реченица и *biblenlp-sr*, који је изведен из електронске верзије Библије са 32K реченица derived from the elec-

tronic version of the *Bible* with 31K senteces, паралелну енглеско-српску верзију *Intera* корпуса (47,5K реченица), српско-енглески паралелни подскуп *Opus* скупа података (Tiedemann 2012), који обухвата 1 милион реченица, као и паралелизовани подскуп *MaCoCu* скупа података, *MacocuParallel-sr* (Banón et al. 2022), са 2 милиона реченица. Иако су два веб корпуса много већа по величини, могуће је да ће код њих бити неопходно уклањање дупликата.

Једини скуп података са сажецима је *XL-Sum* са 15 хиљада парова текстова и резимеа (Hasan et al. 2021).

Назив	**Јез.**	**Порекло**	**Pair type**	**Вел.**	**Издавач**	**Чвор.**
80 jours	sr	literary	translation	3,7	JeRTeh	ELG
biblenlp-sr	sr	literary	translation	31	eBible	HF
Intera	sr	mixed	translation	47,5	META-SHARE	ELG
OPUS-sr	sr	web	translation	1000	Opus project	HF
MacocuParallel-sr	sr	web	translation	2000	Bañón et al	HF
XL-Sum	sr	web	summary	15	Hasan et al	ELG
tapaco-sr	sr	∼synthetic	paraphrase	8	Scherrer, Yves	HF
m_mmlu-sr	sr	synthetic	*QA*	14,5	Alexandra Inst.	HF
m_hellaswag-sr	sr	synthetic	*QA*	9,5	Alexandra Inst.	HF
airoboros-3.0-sr	sr	synthetic	instruction	46	draganjovanovich	HF
open-orca-slim-sr	sr	synthetic	instruction	515	DataTab	HF
ultrafeedback-bin-sr	sr	synthetic	instruction	63	DataTab	HF
alpaca-cleaned-sr	sr	synthetic	instruction	52	DataTab	HF

Табела 5. Корпуси искључиво српског језика доступни на истраженим чвориштима скупова података. Величина је представљена у хиљадама речи(K).

Паралелни скупови података који нису на српском представљени су у табели 6 и састоје се већином од алтернативних подскупова истих ресурса: енглеско-хрватски подскуп скупа *biblenlp*, подскупови *OPUS* скупа који садрже преводе са српскохрватског, хрватског и босанског на енглески, као и подскупови скупа *MaCoCu Parallel*, који садрже преводе

са црногорског, хрватског и босанског на енглески. Ови скупови чине већину реченица, којих има преко 5 милиона.

Два јединствена ресурса са ове листе су хрватски подскуп *mfaq* скупа података који је сачињен од 5K парова питања и одговора сакупљених са веба (De Bruyn et al. 2021) и српскохрватски подскуп *exams* скупа података који садржи 5K парова питања и одговора преузетих са стварних тестова са ученицима (Hardalov et al. 2020). Упркос релативно малој величини, ови ресурси су важни у премало заступљеној групи *QA* скупова података, где су једина алтернатива синтетички преводи.

Назив	**Јез.**	**Порекло**	**Врста пара**	**Вел.**	**Издавач**	**Чвор.**
biblenlp-hr	hr	literary	translation	31	eBible	HF
OPUS-sh	sh	web	translation	271	Opus project	HF
OPUS-hr	hr	web	translation	1000	Opus project	HF
OPUS-bs	bs	web	translation	1000	Opus project	HF
MacocuParallel-me	cnr	web	translation	218	Bañón et al	HF
MacocuParallel-hr	hr	web	translation	2000	Bañón et al	HF
MacocuParallel-bs	bs	web	translation	500	Bañón et al	HF
mfaq-hr	hr	web	*QA*	5	CLiPS	HF
exams-sh	mixed	textbook	*QA*	5	Hardalov et al	HF

Табела 6. Српскохрватски паралелни корпуси доступни на истраженим чвориштима скупова података, који нису искључиво на српском. Величина је приказана у хиљадама реченица (K).

3. Нови корпуси

У оквиру овог истраживања предвиђена су три нова корпуса. Први (*Кишобран*) је нови и већи кровни корпус необележеног текста који ће који ће обухватити све тренутно објављене веб корпусе под једним ентитетом. Израда корпуса Кишобран је описана у одељку 3.1. Други предвиђени корпус (*S.T.A.R.S.*) осмишљен је да повећа заступљеност јавно доступних извора високог квалитеета и заснован је на докторским дисертацијама преузетим са платформе НаРДуС,[7] које су претходно

7. НаРДуС – Национални репозиторијум дисертација у Србији.

коришћене за тренирање тренутно највећег језичког модела који је претрениран за српски језик, *gpt2-orao* (Škorić 2024). Процес припреме *S.T.A.R.S.* корпуса биће објашњен детаљније у одељку 3.2. Последњи корпус (*PaSaž*) обухвата поравнате преводе сажетака издвојених из ових дисертација помоћу метода представљених у одељку 3.3.

3.1 Кровни веб корпус - Кишобран

Нови српскохрватски кровни корпус, *Umbrella corp.* резултат је обједињавања двадесет постојећих корпуса који су прегледани и разматрани у овом раду. Избегавали смо коришћење постојећих кровних корпуса као материјала како бисмо обезбедили могућност касније екстракције одређеног језика, на пример српског дела корпуса, али треба напоменути да је већина коришћених корпуса већ била дедупликована како у односу на друге корпусе са листе, тако и интерно. Једини корпус са мешаним језицима на листи, *HPLT 1.2-sh* је додатно обрађен како би био подељен у корпусе који садрже документа на српском (*HPLT-sr*) и хрватском језику (*HPLT-hr*). Основа за ову врсту меке класификације је била учесталост специфичних речи (тко/ко, што/шта, увјет/услов, уопће/уопште, итд.), као и однос слова *е* и подниске *је*, указујући на ијекавски дијалекат.

Корпуси су додатно пречишћени и дедупликовани. Дедупликација садржаја корпуса је извршена на свим документима помоћу *onion* (Pomikálek 2011) алата за обраду корпуса који врши расплинуту дедупликацију и лоцира дупликате докумената на основу н-торки дуплираних кроз читав корпус. За овај експеримент користили смо претрагу дупликата шесторки и елиминисали све документе изнад прага дуплицирања од 75%.

Након ове дедупликације укупан број речи је смањен за трећину, са 28 милијарди речи на 18,6 милијарди. Са овим укупним бројем, Кишобран је постао највећи доступни кровни корпус необележеног текста уједичком опсегу, са додатним побољшањима (као што су додатно уклањање артефакта и корекција текста) планираним у будућности. .

Целокупна листа коришћених корпуса, њихове величине у милионима речи пре и после дедупликације и чишћења, као и њихов укупни удео у финалној верзији корпуса *Umbrella corp.* представљени су у табели 7.

Име	Јез.	Вел. пре	Вел. после	Удео
srWaC	Serbian	493	307	1,65%
meWaC	Montenegrin	80	41	0,22%
hrWaC	Croatian	1.250	935	5,01%
bsWaC	Bosnian	256	194	1,04%
OSCAR-sr	Serbian	632	410	2,20%
cc100_hr	Croatian	2.880	2,561	13,73%
cc100_sr	Serbian	711	659	3,53%
CLASSLA-sr	Serbian	752	240	1,29%
CLASSLA-hr	Croatian	1.341	160	0,86%
CLASSLA-bs	Bosnian	534	105	0,56%
hr_news	Croatian	1.433	1.426	7,65%
mC4-sr	Serbian	800	782	4,19%
MaCoCu-sr	Serbian	2.491	2.152	11,54%
MaCoCu-cnr	Montenegrin	161	152	0,82%
MaCoCu-hr	Croatian	2.363	2.355	12,63%
MaCoCu-bs	Bosnian	730	700	3,75%
riznica	Croatian	87	69	0,37%
PDRS1.0	Serbian	602	506	2,71%
SrpKorNews	Serbian	469	469	2,51%
HPLT-sr	Serbian	∼5.015	2.562	9,95%
HPLT-hr	Croatian	∼5.015	1.856	13,74%
Total		28,095	**18.641**	100%

Табела 7. Листа корпуса коришћених за израду новог кровни корпуса, њихове величине пре и после чишћења и дедупликације и њихов удео у финалном корпусу. Величине су представљене у милионима речи (M).

3.2 Скуп теза и академских радова на српском – *S.T.A.R.S.*

У домену корпуса високог квалитета (O_2), докторске тезе представљају веома пожељан тип ресурса због неколико фактора. Најпре, високи академски стандарди у погледу методологије, података и језичког квалитета које докторска теза мора да задовољи, као и процес рецензије кроз који пролази, обезбеђују поуздан висок

квалитет података из овог типа извора. Затим, пошто се од докторске дисертације очекује да представља јединствен научни допринос који ће додатно унапредити релевантну научну област, корпус докторских дисертација заузима јединствену позицију када су у питању тематска разноликост у различитим областима знања и актуелност описаних тема. Коначно, велики број докторских дисертација у отвореном приступу у НаРДуС репозиторијуму, чињеница да су одређени одељци докторских дисертација стандардизовани, као и доступност структурираних метаподатака за сваку дисертацију на НаРДуС-у, чине овај ресурс идеалним кандидатом за јединствен, велики и висококвалитетни научни корпус на српском.

НаРДуС је осмишљен у оквиру структурног ТЕМПУС пројекта 5440932013, RODOS.[8] На основу измена и допуна Закона о Високом образовању[9] („Службени гласник РС“, бр. 99/2014, ступивши на снагу дана 19. 9. 2014), све високообразовне институције имају обавезу да учине докторске дисертације доступним јавности пре њихове одбране, док универзитети имају обавезу да обезбеде дигитални репозиторијум са отвореним приступом одбрањеним дисертацијама. У складу са наведеним, платформа НаРДуС је у функцији од септембра 2015. године и заснована је на софтверу *DSpace*, који подржава *OAI-PMH* протокол за пренос метаподатака (Verbić, Suvakov, and Luzanin 2017). У тренутку писања овог рада, на платформи НаРДуС се налазе 13.289 докторских дисертација.

Као први корак у креирању корпуса, комплетна листа дисертација је преузета са мапе сајта платформе НаРДуС,[10] уз помоћ Пајтон модула *Requests*, поштујући спецификације странице *robots.txt* на овом сајту. За преузимање комплетних детаљних метаподатака сваке дисертације, укључујући линкове за преузимање, коришћени су Пајтон и библиотека *Selenium*. Метаподаци за сваку дисертацију су обогаћени са четири додатна поља:

fulltext_url – У случајевима када је за једну дисертацију било доступно више линкова за преузимање (у које је често био укључен извештај комисије), сви PDF документи са линкова су преузети, број страница и линија је аутоматски добијен коришћењем модула

8. rodos.edu.rs Restructuring of Doctoral Studies in Serbia (RODOS)
9. www.pravno-informacioni-sistem.rs/SlGlasnikPortal/reg/viewAct/5f688c8e-4798-470f-9adf-5bc0739c72aa
10. nardus.mpn.gov.rs/htmlmap

PyMuPDF у Пајтону, а одговарајући линк је одабран и сачуван у овом пољу;

need_ocr – Покушано је аутоматско преузимање текста помоћу модула *PyMuPDF* за првих 10 страница сваке дисертације, да би се проценило да ли су документи са одабраних URL-ова дигитално настали или захтевају даље кораке оптичког препознавања знакова (OCR) и ручну ревизију. Резултати ове процедуре су сачувани у овом пољу у виду буловских вредности;

srpski – Ово поље приказује да ли су дисертације написане на српском језику (вредност *yes*) или на неком другом језику (вредност *no*), а језик сваког од тектова је одређиван испитивањем поља *dc.language* и *dc.language.iso*;

ARR – Ово поље приказује да ли су дисертације објављене под заштићеном лиценцом – *all rights reserved* (ARR), што је утврђено прегледом поља *dc.rights.license* у метаподацима.

Ова четири поља су коришћена за филтрирање одговарајућих дисертација-кандидата за ову верзију корпуса. Изабрали смо оне где је било могуће добити се одговарајући PDF могао добити помоћу поља *fulltext_url*, који су написани на српском (*srpski=yes*), који не захтевају OCR (*ocr=no*), и који нису објављени под лиценцом *all rights reserved*. Ово филтрирање је урађено да би се издвојили само текстови на српском језику и како би се осигурала усклађеност са лиценцама ауторских права.

Након примене овог критеријума, остало је 11.624 дисертације, што је представљало око 87,5% свих дисертација у НаРДуС-у.

На крају, цео текст из сваке од ових дисертација је издвојен уз помоћ модула *PyMuPDF* и сачуван у облику текстуалних датотека, које су коришћене за изградњу корпуса. Само линије које су биле део пасуса текста су задржане у састављању финалног корпуса, што је резултирало корпусом са преко 560 милиона речи.

3.3 Корпус паралелних сажетака – *ПаСаж*

С обзиром на то да су паралелни језички ресурси, посебно за мање распрострањене језике (као што је српски), релативно ретки, паралелни корпуси су мањи, али и мање бројни у односу на једнојезичне корпусе. Пошто су сажеци српских докторских дисертација написани на српском и енглеском језику, они представљају вредан велики и квалитетан ресурс за креирање паралелног српско-енглеског корпуса

научног језика. Међутим, наслов, формат и локација сажетака унутар документа значајно варирају у зависности од научне институције. У овој студији прво је извршена ручна анализа великог броја дисертација из различитих институција, након чега су документи подељени у две групе; у првој су биле дисертације чији су се сажеци налазили у самом тексту (било на почетку или крају документа), а у другој дисертације чији су сажеци били представљени у оквиру табеле у одељку *Кључна документацијска информација*. Од 11.624 дисертације, 2.594 су другој групи, која је користила табеле. Из докумената су извучене три врсте података (како на српском, тако и на енглеском): сажеци, кључне речи и научна област дисертације. Пошто су кључне речи већ присутне у метаподацима, њихово извлачење је рађено само за прву групу докумената, где су постојала два скупа кључних речи (због могућих разлика између две верзије). За другу групу су извучени само сажеци и научна област дисертације.

С обзиром на то да метаподаци за већину дисертација у НаРДуС-у садрже делимичне сажетке на оба језика, Пајтон скрипта је коришћена за налажење почетка оба сажетка (првих 6 речи) у првој групи докумената. Уколико дисертација није имала делимичан сажетак у метаподацима или је претрага била неуспешна, налажење почетка сажетка је покушано помоћу регуларних израза који су одговарали могућем наслову одељка сажетка. Пошто су информације у овом одељку увек биле представљене следећим редом: 1) сажетак, 2) кључне речи, 3) научна област (за оба језика), примењен је приступ који је користио регуларне изразе за идентификовање краја сваког сегмента и почетка наредног, након чега су са овим информацијама ажурирани постојећи метаподаци дисертације.

За другу групу, у којој су жељени одељци били присутни у делу *Кључна документацијска информација*, исти приступ је примењен, уз коришћење другачијих регуларних израза за сегменте табела.

Након ових корака, извучено је 7.687 паралелних сажетака, а метаподаци ових дисертација су ажурирани сажецима и научном облашћу дисертације. Где је било могуће, кључне речи из текста дисертација су такође извучене и сачуване у новом пољу у метаподацима (*keywords_from_text*), у случају да се вредности разликују од оних већ постојећих у метаподацима добијеним са НаРДуС-а.

4. Евалуација

Евалуација података у оквиру овог рада фокусира се само на корпусе необележеног текста на српском језику, а сама евалуација се врши кроз грубу процену јединствености сваког корпуса, односно процене колико се он разликује од осталих. Да бисмо проценили аспект јединствености, одлучили смо да спроведемо евалуацију засновану на учесталости речи, инспирисану постојећим експериментом (Rayson and Garside 2000). За сваки корпус-кандидат, извод од милион речи је коришћен за компилацију фреквенција речи, а затим је извучено 1000 најчешћих речи из сваког извода корпуса. Ове најчешће речи из сваког од десет извода корпуса коришћене су за изградњу листе карактеристика (јединственог скупа речи), што је резултирало са укупно 3.257 карактеристика из десет корпуса. Релативне фреквенције (по милион речи) сваке речи из скупа карактеристика у сваком корпусу коришћене су за попуњавање вектора карактеристика (ако је реч једна од одабраних 1000 за тај корпус, у супротном је релативној фреквенцији додељена вредност 0). Када су вектори карактеристика попуњени, израчунате су косинусне сличности (на десети) између сваког пара како би се генерисала матрица сличности корпуса. Корпус који има најмању релативну сличност са осталима сматра се најјединственијим. Израчуната матрица сличности корпуса приказана је у табели 8.

Из представљених резултата је евидентно да је најјединственији корпус према учесталости речи *SrpELTeC*, са просечном сличношћу од 0,40, посебно у поређењу са укупном просечном сличношћу од 0,71. Такође је најудаљенији од сваког појединачног корпуса у матрици сличности.

Нови корпус, *S.T.A.R.S.*, је други по удаљености од свих осталих корпуса појединачно и у просеку, што га чини другим најјединственијим корпусом (просечна сличност од 0,57). Оно што је посебно занимљиво јесте да су два корпуса са најмањом међусобном сличношћу управо *S.T.A.R.S.* и *SrpELTeC*, са сличношћу од само 0,22, што их ставља на супротне крајеве спектра, док се веб корпуси групишу у средини, што је посебно видљиво на дводимензионалној репрезентацији матрице сличности. (слика 2).

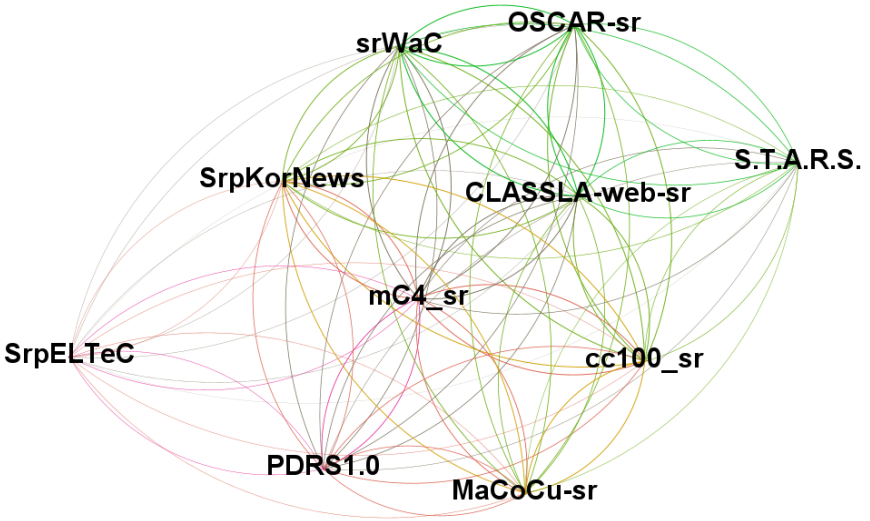
	srWaC	cc100_sr	mC4_sr	OSCAR-sr	CLASSLA-sr	MaCoCu-sr	PDRS1.0	SrpKorNews	SrpELTeC	S.T.A.R.S.
srWaC		0,93	0,88	0,95	0,98	0,79	0,72	0,87	0,36	0,71
cc100_sr	0,93		0,91	0,91	0,90	0,92	0,75	0,93	0,44	0,62
mC4_sr	0,88	0,91		0,84	0,87	0,84	0,78	0,88	0,50	0,61
OSCAR-sr	0,95	0,91	0,84		0,94	0,78	0,70	0,82	0,37	0,69
CLASSLA-sr	0,98	0,90	0,87	0,94		0,75	0,71	0,85	0,34	0,70
MaCoCu-sr	0,79	0,92	0,84	0,78	0,75		0,71	0,89	0,48	0,52
PDRS1,0	0,72	0,75	0,78	0,70	0,71	0,71		0,73	0,47	0,47
SrpKorNews	0,87	0,93	0,88	0,82	0,85	0,89	0,73		0,42	0,56
SrpELTeC	0,36	0,44	0,50	0,37	0,34	0,48	0,47	0,42		0,22
S,T,A,R,S,	0,71	0,62	0,61	0,69	0,70	0,52	0,47	0,56	0,22	
Average	0,80	0,81	0,79	0,78	0,78	0,74	0,67	0,77	0,40	0,57

Табела 8. Матрица сличности српских корпуса необележеног текста доступних на испитиваним чвориштима скупова података, заснована на учесталости речи. Вредности представљају косинусне сличности подигнуте на десети степен, добијене поређењем вектора учесталости речи из скупова карактеристика.

5. Закључак

Овај рад пружа преглед текстуалних корпуса за српски (и српскохрватски) језик који су јавно доступни (на чвориштима *Hugging Face, European Language Grid* и *CLARIN VLO*) у следећим категоријама:

– Корпуси необележеног текста – углавном веб порекла, са изузетком књижевног корпуса *SrpELTeC*;
– Обележени корпуси – различитог порекла, примарно обележени са POS и NER информацијама;
– Паралелни корпуси – углавном веб порекла (*MacocuParallel*, *OPUS*), књижевни преводи и синтетички *QA* сетови и сетови инструкција, један веб корпус резимеа, један полусинтетички корпус парафраза и два мања уређена *QA* скупа.



Слика 2. Јединственост корпуса визуализована кроз дводимензионалну репрезентацију матрице сличности корпуса у виду графа мреже, где ивице представљају удаљености између сваког корпуса, а боје представљају спектар груписања.

Утврдили смо да велика већина доступних корпуса необележеног текста потиче са веба, и да постоји више напора њихове дедупликације, али ниједан тренутно не обухвата све корпусе. Такође смо утврдили да су корпуси високог квалитета веома слабо заступљени, без ијединог корпуса који испуњава критеријум величине од најмање три милиона речи.

Да бисмо превазишли недостатак текстова високог квалитета, предлажемо и састављамо два нова, високо уређена корпуса на основу јавно доступних докторских дисертација на српском: корпус необележеног текста назван *S.T.A.R.S.* (Set of theses and academic research in Serbian) и и паралелни енглеско-српски корпус сажетака назван *PaSaž*. Први садржи 11.624 документа и преко 560 милиона речи, а други садржи 7.678 паралелних сажетака и додатних 2.805 делимичних сажетака, што чини око осам милиона речи, и представља велики допринос јавно доступним корпусима високог квалитета за моделовање српског језика.

Такође смо спровели евалуацију засновану на фреквенцији речи која указује на јединственост дисертација у тренутној парадигми, где је показано да имају релативно ниску сличност са другим доступним корпусима, нарочито са књижевним корпусом *SrpELTeC*, док се веб корпуси (који имају високу међусобну сличност) налазе у средини (слика 2).

Поред тога, рад представља нови кровни веб корпус за српски и српскохрватски језик назван *Umbrella corp.*, који обједињује све тренутно доступне веб корпусе необележеног текста у планираном језичком обиму.

Будући рад у овој области треба да укључи даље прибављање висококвалитетних и књижевних текстова који се могу објавити, као и креирање процедура за унапређење постојећих ресурса, посебно веб текстова.

Захвалница

Рачунарски ресурси неопходни за дедупликацију корпуса *Umbrella corp.* обезбеђени су од стране Националне платформе за вештачку интелигенцију Србије.

Ово истраживање је подржано од стране Фонда за науку Републике Србије, #7276, *Text Embeddings – Serbian Language Applications – TESLA*.

Литература

Augustyniak, Łukasz, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, and Tomasz Kajdanowicz. 2023. *Massively Multilingual Corpus of Sentiment Datasets and Multi-faceted Sentiment Classification Benchmark.* arXiv: 2306.07902 [cs.CL].

Aulamo, Mikko, Nikolay Bogoychev, Shaoxiong Ji, Graeme Nail, Gema Ramírez-Sánchez, Jörg Tiedemann, Jelmer Van Der Linde, and Jaume Zaragoza. 2023. "HPLT: High Performance Language Technologies." In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation,* 517–518.

Banón, Marta, Miquel Espla-Gomis, Mikel L Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, et al. 2022. "MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages." In *23rd Annual Conference of the European Association for Machine Translation, EAMT 2022,* 303–304. European Association for Machine Translation.

Ćavar, Damir, and Dunja Brozović Rončević. 2012. "Riznica: the Croatian language corpus." *Prace filologiczne* 63:51–65.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Unsupervised cross-lingual representation learning at scale."

Cui, Ganqu, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. *UltraFeedback: Boosting Language Models with High-quality Feedback.* arXiv: 2310.01377 [cs.CL].

De Bruyn, Maxime, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2021. *MFAQ: a Multilingual FAQ Dataset.* arXiv: 2109.12870 [cs.CL].

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805.*

Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. "Named entity recognition for distant reading in ELTeC." In *CLARIN Annual Conference 2020.*

Gavrilidou, Maria, Penny Labropoulou, Elina Desipri, Voula Giouli, Vasilis Antonopoulos, and Stelios Piperidis. 2004. "Building parallel corpora for eContent professionals." In *Proceedings of the Workshop on Multilingual Linguistic Resources,* 90–93.

Hardalov, Momchil, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. "EXAMS: A Multi-subject High School Examinations Dataset for Cross-lingual and Multilingual Question Answering." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* edited by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, 5427–5444. Online: Association for Computational Linguistics, November. https://doi.org/10.18653/v1/2020.emnlp-main.438. https://aclanthology.org/2020.emnlp-main.438.

Hasan, Tahmid, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages." In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021,* 4693–4703. Online: Association for Computational Linguistics, August. https://aclanthology.org/2021.findings-acl.413.

Hendrycks, Dan, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. "Aligning AI With Shared Human Values." *Proceedings of the International Conference on Learning Representations (ICLR).*

Krstev, Cvetana. 2021. "The Serbian Part of the ELTeC Collection Through the Magnifying Glass of Metadata." *Infotheca - Journal for Digital Humanities* 21 (2): 26–42. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.2. https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2021.21.2.2_en.

Krstev, Cvetana, and Ranka Stanković. 2023. "Language Report Serbian." In *European Language Equality: A Strategic Agenda for Digital Language Equality,* edited by Georg Rehm and Andy Way, 203–206. Cham: Springer International Publishing. ISBN: 978-3-031-28819-7. https://doi.org/10.1007/978-3-031-28819-7_32.

Krstev, Cvetana, Duško Vitas, and Tomaž Erjavec. 2004. "MULTEXT-East resources for Serbian." In *Zbornik 7. mednarodne multikonference Informacijska druzba IS 2004 Jezikovne tehnologije 9-15 Oktober 2004, Ljubljana, Slovenija, 2004.* Erjavec, Tomaž and Zganec Gros, Jerneja.

Kuzman, Taja, and Nikola Ljubešić. 2023. *Montenegrin web corpus meWaC 1.0.* CLASSLA-web: Bigger and better web corpora for Croatian, Serbian and Slovenian. https://www.clarin.si/info/k-centre/classla-web-bigger-and-better-web-corpora-for-croatian-serbian-and-slovenian-on-clarin-si-concordancers/.

Lian, Wing, Guan Wang, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. *SlimOrca: An Open Dataset of GPT-4 Augmented FLAN Reasoning Traces, with Verification.* https://https://huggingface.co/Open-Orca/SlimOrca.

Ljubešić, Nikola, and Tomaž Erjavec. 2021. *Montenegrin web corpus meWaC 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1429.

Ljubešić, Nikola, Tomaž Erjavec, Vuk Batanović, Maja Miličević, and Tanja Samardžić. 2023. *Serbian Twitter training corpus ReLDI-NormTagNER-sr 3.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1794.

Ljubešić, Nikola, and Filip Klubička. 2014. "{bs, hr, sr} wac-web corpora of Bosnian, Croatian and Serbian." In *Proceedings of the 9th web as corpus workshop (WaC-9),* 29–35.

Ljubešić, Nikola, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. "New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016),* edited by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, et al. Portorož, Slovenia: European Language Resources Association (ELRA), May. ISBN: 978-2-9517408-9-1.

Ljubešić, Nikola, and Davor Lauc. 2021. "BERTić–The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian." *arXiv preprint arXiv:2104.09243.*

Miličević, Maja, and Nikola Ljubešić. 2016. "Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets." *Slovenščina 2.0: empirical, applied and interdisciplinary research* 4, no. 2 (September): 156–188. https://doi.org/10.4312/slo2.0.2016.2.156-188. https://revije.ff.uni-lj.si/slovenscina2/article/view/7007.

Pomikálek, Jan. 2011. "Removing boilerplate and duplicate content from web corpora." PhD diss., Masaryk university, Faculty of informatics, Brno, Czech Republic.

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. "Improving language understanding by generative pre-training."

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21 (1): 5485–5551.

Rahimi, Afshin, Yuan Li, and Trevor Cohn. 2019. "Massively Multilingual Transfer for NER." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 151–164. Florence, Italy: Association for Computational Linguistics, July. https://www.aclweb.org/anthology/P19-1015.

Rayson, Paul, and Roger Garside. 2000. "Comparing corpora using frequency profiling." In *The workshop on comparing corpora,* 1–6.

Al-Rfou, Rami, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. "Polyglot-NER: Massive Multilingual Named Entity Recognition." *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30- May 2, 2015* (April).

Samardžić, Tanja, Mirjana Starović, Željko Agić, and Nikola Ljubešić. 2017. "Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages." In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing,* 39–44. Valencia, Spain: Association for Computational Linguistics, April. https://doi.org/10.18653/v1/W17-1407. https://aclanthology.org/W17-1407.

Scherrer, Yves. 2020. *TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages.* V. 1.0, March. https://doi.org/10.5281/zenodo.3707949. https://doi.org/10.5281/zenodo.3707949.

Škorić, Mihailo. 2024. "Novi jezički modeli za srpski jezik." Accepted for publishing, *Infoteka* 24 (1).

Stanković, Ranka, Cvetana Krstev, Branislava Šandrih Todorović, Duško Vitas, Mihailo Škorić, and Milica Ikonić Nešić. 2022. "Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 3337–3345.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Skoric. 2020. "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian" [in English]. In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, et al., 3954–3962. Marseille, France: European Language Resources Association, May. ISBN: 979-10-95546-34-4. https://aclanthology.org/2020.lrec-1.487.

Suárez, Pedro Javier Ortiz, Benoît Sagot, and Laurent Romary. 2019. "Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures." In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7).* Leibniz-Institut für Deutsche Sprache.

Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. *Stanford Alpaca: An Instruction-following LLaMA model.* https://github.com/tatsu-lab/stanford_alpaca.

Tiedemann, Jörg. 2012. "Parallel Data, Tools and Interfaces in OPUS." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12),* edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 2214–2218. Istanbul, Turkey: European Language Resources Association (ELRA), May. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Todorović, Branislava Šandrih, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. "Serbian ner&beyond: The archaic and the modern intertwinned." In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021),* 1252–1260.

Tunstall, Lewis, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, et al. 2023. *Zephyr: Direct Distillation of LM Alignment.* arXiv: 2310.16944 [cs.LG].

Verbić, Srđan, Milovan Suvakov, and Zorana Luzanin. 2017. "NaRDuS – JAVNOST DOKTORSKIH STUDIJA Transparentnost i otvorenost podataka kroz stvaranje i razvoj nacionalnog repozitorijuma doktorskih disertacija u Srbiji (NaRDuS)." June.

Vitas, Duško, Svetla Koeva, Cvetana Krstev, and Ivan Obradović. 2008. "Tour du monde through the dictionaries." In *Actes du 27eme Colloque International sur le Lexique et la Gammaire,* 249–256.

Vitas, Duško, Ljubomir Popović, Cvetana Krstev, Ivan Obradović, Gordana Pavlović-Lažetić, and Mladen Stanojević. 2012. *Српски језик у дигиталном добу – The Serbian Language in the Digital Age.* META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Available online at http://www.meta-net.eu/whitepapers. Springer. ISBN: 978-3-642-30754-6.

Wasserscheidt, Philipp. 2023. *Serbian Web Corpus PDRS 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1752.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* edited by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, 483–498. Online: Association for Computational Linguistics, June. https://doi.org/10.18653/v1/2021.naacl-main.41. https://aclanthology.org/2021.naacl-main.41.

Zellers, Rowan, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. "HellaSwag: Can a Machine Really Finish Your Sentence?" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.*