New Textual Corpora for Serbian Language Modeling

UDC 811.163.4'322.2
DOI 10.18485/infotheca.2024.24.1.4

ABSTRACT: This paper will present textual corpora for Serbian (and Serbo-Croatian) that can be used for the training of large language models and that are publicly available at one of the several important online repositories of such resources. Each corpus will be classified using multiple methods and its characteristics will be described in details. Additionally, the paper will introduce three new corpora: a new umbrella web corpus of Serbo-Croatian, a new high-quality corpus based on the doctoral dissertations from all Universities in Serbia, stored within the National Repository of Doctoral Dissertations (NARDUS), and a parallel corpus of dissertation abstracts and their translations, derived from the same source. The uniqueness of both old and new corpora will be accessed via frequency-based stylometric methods, and the results will be briefly discussed.

KEYWORDS: corpora, Serbian language, language models, evaluation.

PAPER SUBMITTED: 12 April 2024
PAPER ACCEPTED: 15 May 2024

Mihailo Škorić
mihailo.skoric@rgf.bg.ac.rs
ORCID: 0000-0003-4811-8692
University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

Nikola Janković
nikolajankovickv@gmail.com
ORCID: 0000-0003-3484-4220
University of Belgrade
Faculty of Philology
Belgrade, Serbia

1 Introduction

With the rapid increase of available textual data within the *Big Data* phenomenon at the beginning of the twenty-first century, it was soon realized that these data can be used to build corpora for natural language modeling. The fast-growing web-based data was first used as an add-on to the existing slower-growing book-based data, but with an increased interest in quantity, it slowly but steadily caught up with and surpassed the latter's share in various

language model training corpora. Today, most of the publicly available corpora use web-based data, mostly due to looser copyright constraints.

In the context of this research, we will categorize the datasets into the following categories based on their origin:

O_1 Web corpora – corpora containing texts collected from the open web, primarily using automatic scraping of HTML pages;
O_2 Textbook corpora – corpora containing texts from textbooks or of similar origin, namely scientific publications of all sorts (scientific monographs, journal articles, conference/proceedings papers, theses etc.), and other published non-literary works (government and legal texts not present on the web, philosophy, culinary recipes etc.) collected primarily from the digitally created documents or physical copy scans;
O_3 Literary corpora – corpora containing published literary works including mythology and religious texts;
O_4 Synthetic corpora – corpora containing non-web texts created by machines using natural language generation or machine translation;
O_5 Mixed corpora – corpora created using texts originating from a mix of the aforementioned and other sources.

Apart from the obvious difference in source material, these classes are also differentiated by the creation process for the containing texts: while texts from the first three groups (O_1-O_3) are mostly human-created (no one can guarantee that a scrapped HTML page is not machine generated text), only the second and third group contain texts that are necessarily curated, i.e. they were read and corrected before publication, which takes time, but ensures higher quality. Texts created by machines (O_4) are usually prepared the fastest but are also associated with lesser quality.

Another important classification is corpora form. In that regard, we also classify corpora as follows:

F_1 Plain-textual corpora – corpora incorporating only plain texts which can be used to pre-train large language models such as BERT[1] (Devlin et al. 2018) and GPT[2] (Radford et al. 2018);
F_2 Annotated corpora – corpora where tokens, sentences or other segments are annotated, e.g. part-of-speech-annotated texts or sentences labeled with sentiment – these corpora are usually used to fine-tune models on downstream annotation and labeling tasks;

1. Bidirectional Encoder Representations from Transformers
2. Generative Pre-trained Transformer

F_3 Parallel corpora – corpora where each sentence (or other segment) has a specific pair, e.g. a translation to another language or an answer to a question – these are usually used for the training of models with a generative component such as the *T5*[3] model (Raffel et al. 2020).

In the following section, the paper will focus on the Serbian and Serbo-Croatian corpora procured from three important online corpora/dataset sources: Hugging Face,[4], CLARIN virtual language repository[5] and European Language Grid.[6] Other singular repositories (e.g. GitHub projects) were not searched. Basic information about corpora retrieved from these hubs will be presented in Section 2, newly-prepared corpora will be presented in Section 3, and the experiment in accessing their uniqueness will be presented in Section 4. Finally, the results of the assessment will be discussed in Section 5.

2 Available Corpora

As already mentioned, this section will present a list of corpora compiled by reviewing three online sources: Hugging Face (HF), CLARIN virtual language repository (VLO) and European Language Grid (ELG) with corpora languages being limited to the domain of Serbo-Croatian macro language i.e., Serbian, Montenegrin, Croatian and Bosnian, as training with closely related languages can be beneficial in certain scenarios, particularly for tasks involving multilingual understanding or translation. Retrieved corpora were categorized both by form (primary classification) and origin (secondary classification). For plain-textual corpora (F_1), minimum requirement for including them was thirty million words for web corpora (O_1), and three million words for the other corpora ($O_2 - O_5$). Since most of the corpora in this category are web-originated, a brief study on deduplication was also conducted. As for the annotated (F_2) and parallel corpora (F_3), the size requirement was set to 3,000 sentences, as these resources are much scarcer and usually used for fine-tuning only.

3. Text-To-Text Transfer Transformer

4. huggingface.co the largest web hub for publishing language models.

5. www.clarin.eu, digital infrastructure offering data, tools and services to support research based on language resources.

6. live.european-language-grid.eu, platform for all European Language Technologies originated from MetaNet.

2.1 Publicly available plain-text corpora

Tables 1 and 2 detail the twenty-four retrieved plain-text corpora. The first table focuses on the Serbian corpora only, while the second presents the other corpora in the scope of the Serbo-Croatian macro language, including several *umbrella* corpora (produced via aggregation and deduplication of other corpora).

Name	Lang.	Origin	Size	Publisher	Hub
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

Table 1. Serbian only plain-text corpora available on surveyed dataset hubs. Size is given in millions of words (M).

All corpora, save four, were sourced exclusively from the web. One of them is literary, one is mixed, and the remaining two are aggregated and classified as almost web corpora (∼web), since they incorporate the aforementioned single mixed-origin corpus, *riznica* (Ćavar and Brozović Rončević 2012).

Most of these corpora are a product of several specific efforts. Corpora *srWac*, *meWac*, *hrWac* and *bsWac* originate from a single web scrape of top-level domains for the four languages (Ljubešić and Klubička 2014; Ljubešić and Erjavec 2021), resulting in over two billion words. The process was repeated again later in order to produce *CLASSLA-sr*, *CLASSLA-hr* and *CLASSLA-bs*, amassing over 2.5 billion words (Ljubešić and Lauc 2021).

The additional six corpora were derived from the *Common Crawl* dataset: *OSCAR-sr* (632M) is the OSCAR project dataset derivative (Suárez, Sagot,

Name	Lang.	Origin	Size	Publisher	Hub
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

Table 2. Serbo-Croatian plain-text corpora available on surveyed dataset hubs, which are not Serbian only. Size is given in millions of words (M).

and Romary 2019), *mC4-sr* (800M) was derived from the multilingual C4 dataset by Google (Xue et al. 2021), *HPLT 1.2-sh* (over 10 billion words) was gained from the HPLT project dataset (Aulamo et al. 2023), and finally, *cc100_sr* and *cc100_hr* (over 3.5 billion words) are the cc100 dataset derivatives (Conneau et al. 2019). It should be noted that some of these efforts produced additional corpora, which were not included in this study as they didn't meet the minimum size requirement of three million words.

MaCoCu project web crawling effort (Banón et al. 2022; Kuzman and Ljubešić 2023) produced *MaCoCu-sr* (2.5 billion words), *MaCoCu-cnr* (151M), *MaCoCu-hr* (2.2 billion words), *MaCoCu-bs* (686M), and their deduplicated aggregation, *MaCoCu-hbs* (5.5 billion words).

Another notable deduplication effort included four WaC corpora, and three *CLASSLA* corpora, *cc100_sr*, *cc100_hr* and *riznica* corpora. This

produced a 8.4 billion word aggregated corpus, published under the name *BERTić-data* (Ljubešić and Lauc 2021).

Both *BERTić-data* and *MaCoCu-hbs* were aggregated again, together with *mC4-sr* and *hr_news* (1.4 billion words) in order to create *XLM-R-BERTić-data* (11.5 billion words), the largest published Serbo-Croatian aggregated corpus. It, however, does not incorporate *Common Crawl* datasets *HPLT 1.2-sh* and *OSCAR-sr*, nor the recently published *PDRS1.0* (Wasserscheidt 2023) (600M) and *SrpKorNews* (Krstev and Stanković 2023) corpora (468M). This opens the possibility of creating new and bigger umbrella corpora for both Serbian and Serbo-Croatian. The full current effort in corpus aggregation is visualized in Figure 1.

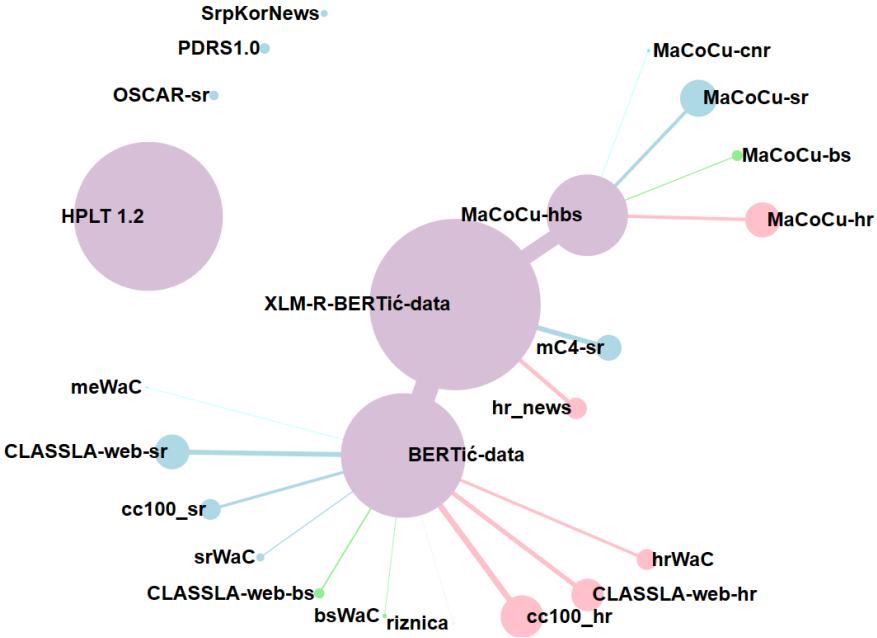

Figure 1. Aggregation hierarchy of publicly available Serbo-Croatian plain-text corpora. Blue color represents Serbian corpora, cyan Montenegrin, pink Croatian, lime Bosnian, while corpora of mixed language origin are represented in purple.

The only literary corpus of the required size available online is still the *SrpELTeC* corpus (Krstev 2021; Stanković et al. 2022), which includes 120 novels and contains 5 million words. There were no constraints for publication, as the corpus comprises materials for which the copyright has expired.

2.2 Publicly available annotated corpora

Analyzing the selected dataset hubs we found 10 annotated corpora for Serbian (Table 3). We focused on the corpora annotated using the Universal POS tag set (17 classes), while there were no constraints regarding named entity recognition (NER) and sentiment annotation. As opposed to the plain-text corpora found primarily on HF and VLO, these corpora were also found on the ELG in numbers.

POS-annotated corpora for Serbian include four resources: *SrpKor4Tagging* (Stanković et al. 2020) with 60K sentences, *Intera* (Gavrilidou et al. 2004; Stanković et al. 2020) with 47.5K, *1984* (Krstev, Vitas, and Erjavec 2004) with 6.7K, and *reldi_sr* (Miličević and Ljubešić 2016) with 5.5K sentences. Two additional resources were annotated by both POS and NER tags: *SETimes_sr* (Samardžić et al. 2017) and *NormTagNER-sr* (Ljubešić et al. 2023) with 7K and 4K sentences respectively and 5 NER classes. Three additional resources were annotated via NER only: *SrpELTeC-gold* (Todorović et al. 2021; Frontini et al. 2020) based on a part of a literary corpora comprising 52K sentences (7 NER classes), Serbian part of the *WikiAnn* corpus (Rahimi, Li, and Cohn 2019) with 40K sentences (3 NER classes), and Serbian part of the *polyglot_ner* corpus (Al-Rfou et al. 2015) with over half of a million sentences and 3 NER classes. The one remaining resource is the Serbian part of the *mms* sentiment-annotated corpus (Augustyniak et al. 2023) with 76K sentences (3 sentiment classes).

A similar situation was observed regarding the other Serbo-Croatian languages, where another ten resources had been retrieved (Table 4), with the list mostly consisting of the Croatian counterparts of the same, previously mentioned annotated resources. These include annotated Croatian translation of the *1984*, *reldi_hr*, *NormTagNER-hr*, Croatian part of the *WikiAnn* and *polyglot_ner* corpora and the Croatian sentiment-annotated sentences from the *mms* corpora. Additionally, there are the Bosnian *mms* and Serbo-Croatian *WikiAnn* extracted from the Serbo-Croatian Wikipedia articles. The sizes of these corpora are mostly comparable to those of their Serbian counterparts.

Name	Lang.	Origin	Annotat.	Size	Publisher	Hub
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

Table 3. Serbian only annotated corpora available on surveyed dataset hubs. Size is given in thousands of sentences (K).

The only unique resource from this list is *hr500k* (Ljubešić et al. 2016), which was fully annotated (25k sentences) with NER tags (5 classes), and only partially with Universal POS tag set (9k sentences).

2.3 Publicly available parallel corpora

A total of thirteen parallel resources were found for Serbian (Table 5), including five corpora of parallel translations, one corpus with text-summary pairs, one corpus of paraphrases, two corpora of question-answer pairs (QA) and four instruction datasets (textual instruction and solution pairs).

All QA and instruction corpora found are synthetic in origin, created using machine translation of existing QA and instruction corpora for English: *m_mmlu-sr* and *m_hellaswag-sr* are translations of datasets *mmlu* (Hendrycks et al. 2021) and *HellaSwag* (Zellers et al. 2019) using GPT3.5, translated respectively; *airoboros-3.0-sr* is the translation of the *airoboros-3.0* dataset (Tunstall et al. 2023), and *open-orca-slim-sr*, *ultrafeedback-bin-sr* and *alpaca-cleaned-sr* are Serbian versions of the *SlimOrca* (Lian et al. 2023), *UltraFeedback* (Cui et al. 2023) and *alpaca-cleaned* (Taori et al. 2023) datasets, respectively, translated using the *Google translate* service. Together with the Serbian part of the semi-synthetic

Name	Lang.	Origin	Annotat.	Size	Publisher	Hub
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

Table 4. Serbo-Croatian annotated corpora available on analyzed dataset hubs, which are not Serbian only. Size is given in thousands of sentences (K).

tapaco set (Scherrer 2020) (original text may not be synthetic, but the sentences were paired using a simple algorithm), most of the available parallel resources for Serbian are synthetic.

As for the non-synthetic resources, five parallel translation corpora (English translations) include literary corpora *80 jours* (Vitas et al. 2008) with 3.7K sentences, and *biblenlp-sr* derived from the electronic version of the *Bible* with 31K senteces, parallel English-Serbian version of *Intera* (47.5K sentences), Serbian-English parallel subset of the *Opus* dataset (Tiedemann 2012) with 1 million sentences, and the parallelized subset of the *MaCoCu* set, *MacocuParallel-sr* (Banón et al. 2022) with 2 million sentences. While the two web-originated corpora are much bigger in size, it might not be necessary to deduplicated them.

The only summary dataset is *XL-Sum* with 15K summarized text pairs (Hasan et al. 2021).

Non-Serbian parallel datasets are presented in Table 6 and consist mostly of alternative subsets of the same resources: English-Croatian subset of *biblenlp*, subsets of *OPUS* dataset containing translations from Serbo-Croatian, Croatian and Bosnian to English, and subsets of the *MaCoCu Parallel* dataset containing translations from Montenegrin,

Name	Lang.	Origin	Pair type	Size	Publisher	Hub
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

Table 5. Serbian only parallel corpora available on analyzed dataset hubs. Size is given in thousands of pairs (K).

Croatian and Bosnian into English. They account for most of the sentences, over 5 million.

Two unique resources from the list are the Croatian subset of the *mfaq* dataset comprising 5k web-scraped QA pairs (De Bruyn et al. 2021) and the Serbo-Croatian subset of the *exams* dataset containing 5k QA pairs taken from actual student tests (Hardalov et al. 2020). Despite the relatively small size, these resources are important in the heavily underrepresented group of QA datasets, the only alternative being synthetic translations.

3 New Corpora

A total of three new corpora are envisioned in the scope of this research. The first (*Umbrella corp*) is the new and larger aggregated (umbrella) plain-text corpus that will encompass the entirety of currently published web corpora under a single entity. The creation of *Umbrella corp* is elaborated in Section 3.1. The second envisioned corpus (*S.T.A.R.S.*) is designed to

Name	Lang.	Origin	Pair type	Size	Publisher	Hub
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

Table 6. Serbo-Croatian parallel corpora available on analyzed dataset hubs, which are not Serbian only. Size is given in thousands of pairs (K).

boost the representation of openly available textbook-quality sources and it is based on the doctoral dissertations downloaded from the NaRDuS platform,[7] which were previously used to train currently largest language model pre-trained for Serbian, *gpt2-orao* (Škorić 2024). The preparation process for *S.T.A.R.S.* will be explained in more detail in Section 3.2. The last corpus (*PaSaž*) represents aligned translations of abstracts extracted from the dissertations using methods presented in Section 3.3.

3.1 Aggregated web corpus - *Umbrella corp*

New umbrella Serbo-Croatian web corpus, *Umbrella corp.* is an aggregation of the twenty existing corpora analyzed and discussed in this paper. We avoided using existing aggregated corpora as material to ensure the possibility of extraction of a specific language later on, e.g. the Serbian part of the corpora, but it should be noted that the majority of the used corpora had already been deduplicated both against other corpora from the list and internally. The only mixed-language corpora on the list, *HPLT 1.2-sh* was additionally processed to be split into corpora containing documents in Serbian (*HPLT-sr*) and Croatian i.e. non Serbian (*HPLT-hr*). The basis for

7. NaRDuS – National Repository of Doctoral Dissertations from all Universities in Serbia.

this fuzzy classification was the frequency of specific words (*tko/ko*, *što/šta*, *uvjet/uslov*, *uopće/uopšte* etc.) and the ratio of the character *e* and substring *je*, which indicates the Ijekavian dialect.

The corpora was additionally cleaned and deduplicated. Deduplication of the corpora content was performed on all documents using the *onion* (Pomikálek 2011) corpus processing tool which performs fuzzy deduplication, and locates document duplicates on the bases of n-grams duplicated across the whole corpus. For this experiment we used *6-gram* duplicates search and eliminated all documents exceeding the 75% *n*-gram duplication threshold.

After this deduplication the the total number of words was reduced by a third, from 28 billions to 18.6 billions. With this total number of words, *Umbrella corp* became the biggest available umbrella plain-text corpus in the language scope, with additional improvements (e.g. additional boilerplate removal and text correction) planned in the future.

The full list of corpora used, their sizes in millions of words before and after deduplication and cleaning, as well as their total share in the final version of the *Umbrella corp.* is presented in Table 7.

3.2 Set of theses and academic research in Serbian - *S.T.A.R.S.*

In the domain of textbook corpora (O_2), doctoral theses represent a highly desirable resource type due to several factors. Firstly, the high academic standards regarding the methodology, data, and linguistic quality that a doctoral thesis needs to satisfy, as well as the peer review process that it undergoes, ensure a reliable high quality of the data from this source type. Secondly, since a doctoral dissertation is expected to be a unique scientific contribution that would advance the relevant science field further, a corpus of doctoral dissertations is uniquely positioned in terms of the thematic variety across various domains of knowledge and the timeliness of the topics described. Finally, the large number of open-access doctoral dissertations in the NaRDuS repository, the fact that certain sections of doctoral dissertations are standardized, as well as the availability of structured metadata for each dissertation on NaRDuS, make this resource the ideal candidate for a unique, large, and high-quality scientific corpus in Serbian.

NaRDuS was envisioned within the structural TEMPUS project 5440932013, RODOS.[8] Pursuant to the amendments to the Law on Higher

8. rodos.edu.rs Restructuring of Doctoral Studies in Serbia (RODOS)

Name	Language	Size before	Size after	Share
srWaC	Serbian	493	307	1.65%
meWaC	Montenegrin	80	41	0.22%
hrWaC	Croatian	1250	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	1341	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%

Table 7. List of corpora used to build the new umbrella corpus, their sizes before and after cleaning and deduplication and their share in the final corpus. Sizes are given in millions of words (M).

Education[9] (Amendments published in the Official Gazette, no. 99/2014, entered into force on 19 September 2014), all higher education institutions

9. Official Gazette/Law on Higher Education

have an obligation to make PhD dissertations available to the public prior to their defense, while universities have an obligation to provide an open access digital repository of the defended dissertations. In accordance with the above, the NaRDuS platform has been in operation since September 2015. It is based on the DSpace software, which supports the OAI-PMH protocol for metadata transfer (Verbić, Suvakov, and Luzanin 2017). At the moment of writing this article, there were 13,289 doctoral dissertations on NaRDuS.

As the first step in creating the corpus, the full list of dissertations was obtained from the sitemap of the NaRDuS website,[10], with Python's *Requests* module, complying with the specifications on the site's robots.txt page. *Selenium* and Python were used to retrieve the full detailed metadata of each dissertation, including download links. Metadata for each dissertation was also enriched with four additional fields:

fulltext_url – In the cases where multiple download links were provided for one dissertation (which often included the committee report), all of the PDF documents from the links were downloaded, their page count and number of lines were automatically retrieved using the *PyMuPDF* module in Python, and the appropriate link was selected and stored in this field;

need_ocr – Automatic text retrieval using *PyMuPDF* for the first 10 pages was attempted to evaluate whether the documents from the selected URLs are digital-born or they require further steps of optical character recognition (OCR) and manual revision, and the results of the procedure were stored in this field in Boolean form;

srpski – This field indicates whether the dissertations were written in Serbian (*yes* value) or another language (*no* value), and the language of each text was determined through examination of both *dc.language* and *dc.language.iso* fields;

ARR – This field indicates whether the dissertations were published under the copyright license – *all rights reserved* (ARR), which was determined by examining the *dc.rights.license* metadata field.

These four fields were used to filter out appropriate candidate-dissertation for this rendition of the corpus. We selected the ones where the respective PDF could be obtained through the *fulltext_url*, which were written in Serbian (*srpski=yes*), do not need OCR (*need_ocr=no*),

10. NaRDuS website

and were not published under the *all rights reserved* licence (*ARR=no*). This filtering was done to extract only texts in Serbian and to ensure compliance with the copyright licenses.

After applying this criterion, 11,624 dissertations remained, which is around 87.5% of all of the dissertations in the NARDUS repository.

Finally, the full text from each of these dissertations was extracted using *PyMuPDF*, saved as a TXT file, and used to build the corpus. Only the lines that are part of textual paragraphs were preserved during the compilation of the final corpus file, resulting in an corpus of over 560M words.

3.3 Parallel abstracts corpus – *PaSaž*

Due to the fact that parallel language resources, especially for less widely spoken languages (like Serbian), are relatively rare, parallel corpora are smaller and fewer in number compared to their monolingual counterparts. Since abstracts of the Serbian doctoral dissertations are written both in Serbian and English, they represent a valuable large, high-quality resource for the creation of a parallel Serbian-English corpus of scientific language. However, the title, format and location of the abstracts within the document varies significantly, depending on the respective scientific institution. In this study, a manual analysis of a large number of dissertations from different institutions was first performed, after which the documents were separated into two groups; in the first were the dissertations whose abstracts appeared in the text itself (either at the beginning or the end of the document), and in the second, dissertations whose abstracts were presented within a table in the *Key Word Documentation* section. Out of the 11,624 dissertations, 2,594 were in the second, table-based group. Three types of data were extracted from the documents (in both Serbian and English): the abstracts, keywords, and the scientific field of the dissertation. Since the keywords are already present in the metadata, their extraction was only done for the first group of documents, where there were two sets of keywords (in case of possible differences between the two versions). For the second group, only abstracts and the scientific field of the dissertation were extracted.

Since the metadata for the majority of dissertations in NaRDuS contain partial abstracts in both languages, Python script was used to search for the beginning of each abstract (first 6 words) in the first group of documents. If the dissertation did not have a partial abstract in the metadata, or the search was unsuccessful, finding the start of the abstract was attempted using regular expressions which matched the heading string of the abstract

section. Since the information in this section always appears in the order 1) abstract, 2) keywords, 3) scientific field (for both languages), an approach using regular expressions was used to identify the end of each section and the beginning of the next, after which the existing metadata for the dissertation was updated with this information.

For the second group, in which the desired sections were present in the *Key Word Documentation* section, the same approach was used, with different regular expressions for the table sections.

After these steps, 7,687 parallel abstracts were extracted, and the metadata of these dissertations was updated with the abstracts and the scientific field of the dissertation. Where possible, the keywords from the dissertation texts were also extracted and stored in a new field in the metadata (*keywords_from_text*), in case the values differ from the ones already existing in the metadata obtained from NaRDuS.

4 Evaluation

Dataset evaluation in the scope of this paper focuses on the plain-text Serbian corpora only, and the evaluation is performed by rough assessment of the uniqueness of each corpus, i.e., how much it differs from the others. In order to assess the uniqueness aspect, we decided to perform a word-frequency-based evaluation, inspired by an existing experiment (Rayson and Garside 2000). For each candidate corpus, a million-words excerpt was used to compile word frequencies, and the 1000 most frequent words from each corpus excerpt were extracted. The most frequent words from each of the ten corpora excerpts were used to build a features list (a unique set of words), resulting in a total of 3,257 features from ten corpora. The relative frequencies (per million words) of each feature word from each corpus were used to populate feature vectors, if the word was one of the chosen 1000 for that corpus. If it was not, the relative frequency value was set to 0. Once the feature vectors were populated, we calculated cosine similarities (to the power of ten) between each pair of vectors to generate a corpora similarity matrix. The corpus having the lowest relative similarity with the other corpora is regarded as the most unique. The corpus similarity matrix is presented in Table 8.

The presented results show that the most unique corpus according to word frequencies is *SrpELTeC*, with an average similarity of 0.40, especially when compared to the total average similarity of 0.71. It is also placed furthest from every other corpus in the embedding matrix.

	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

Table 8. Word frequency-based similarity matrix of Serbian plain-text corpora available on surveyed dataset hubs. The values represent cosine similarities to the power of ten, obtained by comparing feature (word) frequency vectors.

The new corpus, *S.T.A.R.S.*, is the second furthest corpus both from each of the other corpora and on the average, which makes it the second most unique one (average similarity of 0.57). What is especially interesting is that the two corpora that have the lowest mutual similarity are actually *S.T.A.R.S.* and *SrpELTeC* with a similarity of only 0.22, making them the two opposite ends of the spectrum. The web corpora are clustered in the middle, which is clearly visible on the two-dimensional representation of the embedding matrix (Figure 2).

5 Conclusion

This paper provides an overview of textual corpora for Serbian (and Serbo-Croatian) language publicly available (on Hugging Face, European Language Grid and CLARIN VLO) in the following categories:

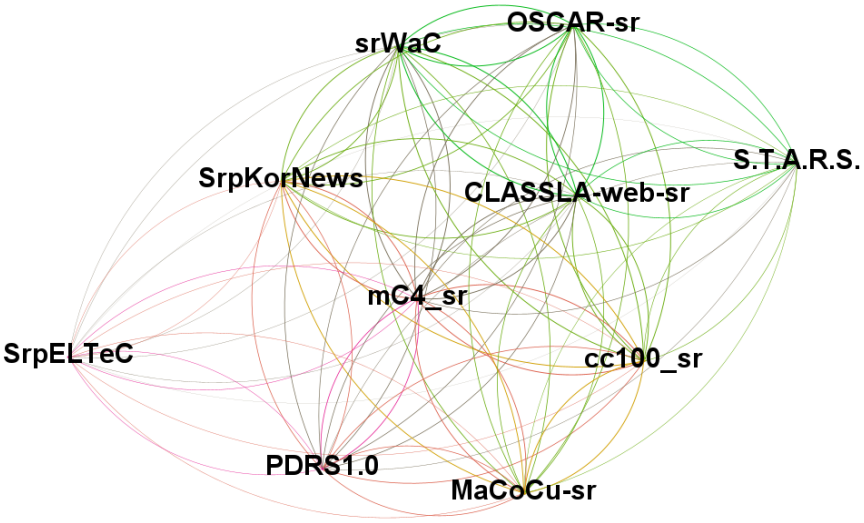

Figure 2. Corpus uniqueness visualized through a two-dimensional network graph representation of the calculated corpus-similarity matrix, where the edges represent distances between each corpus and the colors represent the clustering spectrum.

- Plain text corpora – mostly web-originated, with an exception of the literary *SrpELTeC* corpus;
- Annotated corpora – various origins, primarily annotated with POS and NER and to a lesser extent with sentiment;
- Parallel corpora – mostly web-originated (*MacocuParallel*, *OPUS*), as well as literary translations and synthetic QA and instruction sets, one web summary corpus, one semi-synthetic paraphrase corpus and two smaller curated QA sets.

We found that the vast majority of the available plain-text corpora originate from the web, and that several deduplication efforts exist, but none of them currently encompasses all the corpora. We also found that the textbook corpora were highly underrepresented, with not a single corpus available matching the size criterion of three million words minimum.

To deal with the scarcity of textbook-quality texts we proposed and compiled two new highly curated corpora based on publicly-available doctoral dissertations in Serbian: The plain text corpus dubbed *S.T.A.R.S.* (Set of theses and academic research in Serbian) and Parallel English-Serbian

corpus of abstracts dubbed *PaSaž*. The former contains 11,624 documents and features over 560 million words, and the latter contains 7,678 parallel abstracts, and additional 2,805 partial abstracts, which amounts to around eight million words, and represents a huge contribution to the publicly available textbook-quality corpora for Serbian language modeling.

We also performed a word-frequency-based evaluation that indicated the uniqueness of the dissertations corpus in the current paradigm, showing that it has a relatively low similarity to other currently available corpora, especially the literary corpus *SrpELTeC*, with the web corpora (having high mutual similarity) placed in the middle (Figure 2).

Additionally, the paper introduces a new umbrella web corpus for Serbian and Serbo-Croatian dubbed *Umbrella corp.* which accumulates all the currently available plain text web corpora in the language scope.

The future work in this area should include further procurement of textbook and literary texts that can be published, and the creation of procedures for the enhancement of the existing resources, particularly web texts.

Acknowledgment

Computer resources necessary for the deduplication of the *Umbrella corp.* were provided by the National Platform for Artificial Intelligence of Serbia.

This research was supported by the Science Fund of the Republic of Serbia, #7276, *Text Embeddings – Serbian Language Applications – TESLA*.

References

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.*