Info*theca* Journal of Digital Humanities

Impressum

FOR THE EDITOR:
Prof. Aleksandar Jerkov, PhD office@unilib.rs
University Library "Svetozar Marković"
Faculty of Philology, University of Belgrade

EDITOR:
Faculty of Philology, University of Belgrade
University Library "Svetozar Marković"
Serbian Academic Library Association

EDITORS IN-CHIEF:
Aleksandra Trtovac, PhD aleksandra@unilib.rs
Prof. Ranka Stanković, PhD ranka.stankovic@gmail.com

EDITOR OF ONLINE EDITION:
Jelena Gavrilović, PhD andonovski@unilib.rs
Ivana Gavrilović gavrilovic@unilib.rs

EDITORIAL BOARD:
Prof. Aleksandar Jerkov, PhD, Prof. Cvetana Krstev, PhD, Prof. Biljana Dojčinović, *PhD, Faculty of Philology, University of Belgrade*; Prof. Elisabeth Burr, PhD, *Institut für Romanistik, Universität Leipzig*; Prof. Vladan Devedžić, PhD, *Faculty of Organization Sciences, University of Belgrade*; prof. Milena Dobreva, PhD, *Faculty of Media and Knowledge Sciences, University of Malta*; Tomaž Erjavec, PhD, *Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana*; Prof. Svetla Koeva, PhD, *Institute for Bulgarian Language, Bulgarian Academy of Sciences*; Prof. Denis Maurel, PhD, prof. Agata Savary, PhD, *Université Francois Rabelais de Tours*; Prof. Ivan Obradović, PhD, *Faculty of Mining and Geology, University of Belgrade*; Prof. Gordana Pavlović Lažetić, PhD, prof. Duško Vitas, PhD, *Faculty of Mathematics, University of Belgrade*; Prof. Katerina Zdravkova, PhD, *Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje*

Belgrade, Vol. 24, No. 1(2024), February 2025

ISSN 1450-9687 (print edition)
ISSN 2217-9461 (online edition)

WEB PORTAL:
Jelena Gavrilović, PhD
Ivana Gavrilović
University Library "Svetozar Marković"

LECTORS FOR ENGLISH:
Prof. Ivan Obradović, PhD
University of Belgrade
Mina Dizdar
University Library "Svetozar Marković"

DESIGN AND PREPRESS:
Info*theca* Team

REDACTOR OF REFERENCES AND UDC:
Nataša Dakić, PhD
University Library "Svetozar Marković"

DOI REDACTOR:
Miloš Utvić, PhD
Faculty of Philology, University of Belgrade

JOURNAL REDACTION: +381 11 3370–211
Journal **Info***theca* infothecajdh@gmail.com
11000 Belgrade, Bulevar kralja Aleksandra 71

PRINTED BY:
Mamigo plus
Belgrade

Journal is published twice a year

Info*theca* Vol. 24, No. 1(2024)

Table of Contents

Scientific papers

New Language Models for Serbian
Mihailo Škorić 7

Building the dictionary of football terminology through data-driven and Ontolex model
Jelena Lazarević, Olivera Kitanović 29

Creating a Synthetic Evaluation Dataset for Serbian SentiWordNet Using Large Language Models
Saša Petalinkar 53

New Textual Corpora for Serbian Language Modeling
Mihailo Škorić, Nikola Janković 71

Professional paper

The Galaksija computer – the road to stars paved with words
Olga Mirković Maksimović 97

Review

Introduction to Digital Humanities: workshops for the implementation of 'remote readers' in research practice
Milica Rabrenović 118

New Language Models for Serbian

UDC 811.163.41'322.2
DOI 10.18485/infotheca.2024.24.1.1

ABSTRACT: The paper will briefly present the development history of transformer-based language models for the Serbian language. Several new models for text generation and vectorization, trained on the resources of the Society for Language Resources and Technologies, will also be presented. Ten selected vectorization models for Serbian, including two new ones, will be compared on four natural language processing tasks. The paper will analyze which models are the best for each selected task, how their size and the size of their training sets affects the performance on those tasks, and what is the optimal setting to train the best language models for the Serbian language.

KEYWORDS: language models, Serbian language, vectorization, natural language processing.

PAPER SUBMITTED: 27 January 2024
PAPER ACCEPTED: 21 February 2024

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

1 Introduction

The beginning of the twenty-first century, brought a sharp increase in the amount of available textual data, followed by a sharp increase in computing power, triggering a wave of research based on the idea of *deep learning* (LeCun, Bengio, and Hinton 2015). In the case of natural language processing, the research culminated in the appearance of the transformer architecture (Vaswani et al. 2017), based on the use of encoders, responsible for text analysis, and decoders, responsible for text synthesis. The first extremely popular model of this type was *BERT*[1] (Devlin et al. 2018), based exclusively on the transformer encoder (encoder-only model). This model had made a major breakthrough on multiple natural language processing tasks,

1. Bidirectional Encoder Representations from Transformers

primarily the ones based on vectorization of text (word embedding). Its variations, *RoBERTa*[2] (Liu et al. 2019) and *DeBERTa*[3] (He et al. 2020) to this day achieve state-of-the-art results for word embedding, word annotation (e.g. word type marking and named entity recognition) and classification of sentences and documents. On the other hand, the appearance of *GPT* (generative pretrained transformer) (Radford et al. 2018) and *GPT-2* (Radford et al. 2019) popularized language models based on the transformer decoder (decoder-only models), which is a group of models currently developing the fastest. Models that combine the use of encoders and decoders, such as *BART* (Lewis et al. 2020) and *T5* (Raffel et al. 2020), remain underrepresented despite their outstanding results on text transformation tasks e.g. machine translation, document summarization and style transfer.

1.1 Overview of published models for the Serbian language

Transformer-based language models made way into the Serbian language through multilingual models, firstly through *MBERT*[4] (Devlin et al. 2018), and then and through *XLM-RoBERTa*[5] (Conneau et al. 2019), for the training of which about 4 billion tokens from texts written in Serbian or another closely related language (Croatian, Bosnian) were used. The latter model was released in December 2019 in two variants, *base* (279 million parameters) and *large* (561 million parameters). Even today, as one of the largest encoder models, *XLM-RoBERTa* is being used for the processing of Serbian texts and achieves good results, especially after additional, specific training.

In early 2021, a model called *BERTić* (*classla/bcms-bertic*) (Ljubešić and Lauc 2021) was published on the platform *Huggingface*.[6] The model is based on the *ELECTRA* architecture (Clark et al. 2020) (110 million parameters), and it was trained on a corpus of over 8 billion tokens, Bosnian (800 million), Croatian (5.5 billion), Montenegrin (80 million) and Serbian (2 billion).

Later that same year, the first models were trained specifically for Serbian and published as part of a wider linguistic research for the Macedonian language (Dobreva et al. 2022). More precisely, the Serbian version of the *RoBERTa-base* model, *macedonizer/sr-roberta-base* (120 million parameters) and the Serbian version of the *GPT2-small* model, *macedonizer/sr-gpt2* (130

2. Robustly Optimized *BERT*
3. Decoding-Enhanced *BERT*
4. Multilingual *BERT*
5. Cross-lingual Language Model
6. Huggingface, the largest web hub for publishing language models.

million parameters) were published. Both of these models were trained on the Serbian Wikipedia corpus and support only the Cyrillic alphabet.

Not long after, a similar undertaking begun, during which five *RoBERTa-base* models were trained for the Serbian language (Cvejić 2022). The initial model, *Andrija/SRoBERTa*, had 120 million parameters and was trained on a small corpus of 18 million tokens known as *Leipzig* (Biemann et al. 2007), while the remaining four models each had 80 million parameters and were trained on an increasingly larger corpus. For the *Andrija/SRoBERTa-base* model, the corpus *OSCAR* (Suárez, Sagot, and Romary 2019) (220 million tokens) was added, for the *Andrija/SRoBERTa-L* model, the *srWAc* (Ljubešić and Klubička 2014) (490 million tokens) was added, for the *Andrija/SRoBERTa-XL* model, a part of the *cc100-hr* (21 billion tokens) and *cc100-sr* (5.5 billion tokens) (Wenzek et al. 2020) corpora were added, while for the model *Andrija/SRoBERTa-F* all of the mentioned corpora were used in their entirety.

By the end of 2022, three experimental generative models for Serbian were published (Škorić 2023). The control model, *procesaur/gpt2-srlat*, was again based on the *GPT2-small* architecture, had 138 million parameters and was trained on a subset of the Society for Language Resources and Technologies corpora (260 million tokens) (Krstev and Stanković 2023). The other two models, *procesaur/gpt2-srlat-sem* and *procesaur/gpt2-srlat-synt*, were created by retraining the control model using two specially prepared corpora with the aim of separately modeling the semantics and the syntax of the text. The three models were then used for the experiment of combining language models on the sentence classification task (Škorić, Utvić, and Stanković 2023).

Early next year, researchers from the University of Niš published the *JelenaTosic/SRBerta* model (75 million parameters) based on the *RoBERTa-base* architecture, trained using the *OSCAR* corpus (Suárez, Sagot, and Romary 2019). What is interesting about this model and its second version (*nemanjaPetrovic/SRBerta*, 120 million parameters), is that they were retrained before publication using texts from the law domain (Bogdanović, Kocić, and Stoimenov 2024).

Between the publication of these two models, the first question-answering model for Serbian, *aleksahet/xlm-r-squad-sr-lat* (Cvetanović and Tadić 2023), was created by adapting the RoBERTa model using the *SQuAD* (Rajpurkar, Jia, and Liang 2018) dataset translated into Serbian.

In mid-2023, two more generative models for Serbian based on the GPT architecture were released. Both were trained on the same dataset: the cor-

pora of the Society for Language Resources and Technologies (Krstev and Stanković 2023), doctoral dissertations downloaded from the NARDUS platform,[7] corpus of public discourse of the Serbian language by the Institute of Serbian Language SANU dubbed *PDRS* (Wasserscheidt 2023), and some additional publicly available corpora from the web, such as the already mentioned *srWAc* (Ljubešić and Klubička 2014) and *cc100-sr* (Wenzek et al. 2020). The total number of tokens in this dataset was about 4 billion. The larger model, *jerteh/gpt2-orao*,[8] has 800 million parameters, it is based on the *GPT2-large* architecture and is currently the largest available model pretrained for the Serbian language. The smaller model, *jerteh/gpt2-vrabac*,[9] has 136 million parameters and it is based on the *GPT2-small* architecture. Both models were trained using the computing resources of the National Platform for Artificial Intelligence of Serbia. In addition to the training corpus, the two models also share a dictionary and a tokenizer, specially equipped to pair Cyrillic and Latin characters, enabling equal support for both alphabets.

After the publication of the 800 million parameter generative model (*jerteh/gpt2-orao*), the focus slowly shifted to retraining available large models for English using Serbian language texts. Hence, two models based on the *Alpaca* (Taori et al. 2023) architecture, *datatab/alpaca-serbian-3b-base* (3 billion parameters) and *datatab/alpaca-serbian-7b-base* (7 billion parameters) were published, while the publication of another 7 billion parameter model based on the *Mistral-7b* (Jiang et al. 2023) architecture, trained on Croatian, Bosnian and Serbian texts numbering 11.5 billion tokens, was announced. The same 11.5 billion tokens corpus was used to retrain the *XLM-RoBERTa-large*. This model was published under the name *classla/xlm-r-bertic* (Ljubešić et al. 2024) and has 561 million parameters, the same number as the original *XLM* model.

Finally, the dataset which was used to train *jerteh/gpt2-orao* and *jerteh/gpt2-vrabac*, was also used to train more encoding models from scratch. The larger model, *jerteh/Jerteh-355*,[10] is based on the *RoBERTa-large* architecture and has 355 million parameters, while the smaller model, *jerteh/Jerteh-81*,[11] is based on the *RoBERTa-base* architecture and has 81 million parameters. As with the *jerteh/gpt2-orao* model, the goal was to

7. NARDUS – National Repository of Doctoral Dissertations from all Universities in Serbia.
8. jerteh/gpt2-orao
9. jerteh/gpt2-vrabac
10. jerteh/Jerteh-355
11. jerteh/Jerteh-81

train the models on the highest quality corpora. This paper will present an analysis of the performance of these two models individually, as well as in comparison with the performance of other selected models, in order to establish their place in the hierarchy of Serbian language models for text vectorization.

1.2 The Experiment

In the previous section, it was pointed out that there is a large number of multilingual models that support the processing of the Serbian language to a different extent, and that there are about twenty models that have been prepared specifically for the processing of Serbian. Published models differ from each other by several features: the family (architecture) of the model, the number of parameters, the dictionary or tokenizer on which the model is based, the corpora used for its training, the task on which the model was trained, and the training length. It should be noted that some of the information on models is missing, but also that some available information (primarily the properties of the training set) is not verifiable.

In the following sections, the paper will focus on ten selected encoding models (general type). Basic information about those models will be presented in Section 2, an experiment comparing their performance on four prepared tasks will be presented in Section 3, and the results of the experiments will be presented and discussed in Section 4. Finally, in Section 5, the process of training new models for Serbian will be proposed. This paper will not focus on generative models due to the lack of a reliable (automatic) mechanisms for measuring their performance. Encoder-decoder models specially developed for the Serbian language are yet to be published.

2 Selected encoder-based models

For the purposes of this paper, ten of the previously mentioned models (Section 1) were selected, and will be analyzed in more detail. They include four *SRoBERTa* models, which, due to the fact that they differ only in the training data set, are very suitable for this experiment. Furthermore, the oldest model, *classla/bcms-bertic* and the newest model, *classla/xlm-r-bertic*, published by the Center for South Slavic Languages *CLASSLA*, will be examined, as well as the two most popular multilingual models *xlm-roberta-base* and *xlm-roberta-large*. The last two models to be analyzed, *jerteh-81* and *jerteh-355*, trained on the resources of the Society for Language Resources

and Technologies, are presented for the first time in this paper. The basic features of the ten models are shown in Table 1.

No.	1	2	3	4	5	6	7	8	9	10
ID	Andrija/SRoBERTa-base	Andrija/SRoBERTa-L	Andrija/SRoBERTa-XL	Andrija/SRoBERTa-F	classla/bcms-bertic	classla/xlm-r-bertic	xlm-roberta-base	xlm-roberta-large	jerteh/jerteh-81	jerteh/jerteh-355
Tokenizer	SRoBERTa				bertic	XLM-R			jerteh	
Architecture	RoBERTa				ELE.	XLM-R			RoBERTa	
Model Size	80				110	561	279	561	81	355
Dataset Size	500	1000	3750	5700	8400	11500	4000*		4000	
Serbian	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Croatian			✓	✓	✓	✓	✓	✓		
Bosnian					✓	✓	✓	✓		
Montenegrin					✓		✓	✓		

Table 1. Ten selected encoder-based models for Serbian and their features: tokenizer, architecture, model size in millions of parameters and training set size in millions of tokens. The data was taken from the *HuggingFace* platform. *The size of the training set for models 7 and 8 (*xlm-roberta-base*, *xlm-roberta-large*) refers to the part of the set in Serbian, Croatian or other related language. The lower part of the table shows in which of these languages the models were trained.

From the table, as well as from the description of the models in the previous section, it is apparent that the most popular architecture is *RoBERTa* (6 out of 10 selected models), with the additional three models being based on a close-related, *XLM-RoBERTa* architecture. The one remaining model, *bcms-bertic* based on the ELECTRA architecture, is the only selected model not pre-trained on the masked language modeling task (prediction of parts of text masked behind a special label).

The size of the selected models varies from 80 (for four *SRoBERTa* models) to over 560 million parameters (for models based on *XLM-RoBERTa-large*). The size of the training set varies from 500 million for model 1 (*SRoBERTa-base*) to as much as 11.5 billion tokens for model 6 (*classla/xlm-r-bertic*), where it should be noted that this model was not trained from scratch, but rather a *xlm-roberta-large* model addapted on Croatian, Bosnian and Serbian texts. Only four out of ten models were trained exclusively on Serbian texts, namely models 1, 2, 9 and 10, i.e. the first two *SRoBERTa* models, *jerteh/jerteh-81* and *jerteh/jerteh-355*.

It is also important to note that the ten presented models use only four different dictionaries/tokenizers:

X_1 *SRoBERTa* tokenizer – the first 4 models;
X_2 *bertic* tokenizer – model 5;
X_3 *XLM-R* tokenizer – models 6 to 8;
X_4 *jerteh* tokenizer – the last 2 models (9 and 10).

3 Performance evaluation setting

Ten selected models were evaluated on four separate tasks to compare their performance:

T_1 Masked language modeling (guessing missing tokens);
T_2 Calculation of (semantic) sentence similarity;
T_3 Part-of-speech annotation;
T_4 Named entity recognition.

The first two tasks belong to the group of so-called *upstream* tasks, which use models in their basic state, while the other two tasks belong to the group of *downstream* tasks because they require the models to be fine-tuned and evaluated on a specially prepared, task-specific datasets.

3.1 Model evaluation - upstream tasks

As already mentioned, upstream tasks do not require model adaptation, so only preparation of test sets is necessary.

In order to evaluate the models on the masked language modeling task (T_1), a special data set was prepared using texts in which one random token is masked in each sentence behind a mask `<MASK>`. Four sources were used for the textual material:

Y_1 *Dečko*, Serbian translation of the novel *The Adolescent* (*Подросток*) by Dostoyevsky;
Y_2 *Mladić*, alternative translation of *Dečko*;
Y_3 Serbian translation of Jules Verne's novel *Around the World in 80 Days*;
Y_4 Croatian translation of Jules Verne's novel *Around the World in 80 Days*.

The first two sources were not used to train any of the models, while the other two have been available on the web for a long time (Vitas et al. 2008) and were therefore probably used to train most, if not all, of the models listed.

In order that no model has a particular advantage, the texts were tokenized using all four tokenizers (X_1 to X_4) and then masked. Each of the ten models had the task of unmasking each of the sixteen prepared texts (four sources tokenized and masked in four different ways). One token was masked in each sentence, and the models offered three candidates in its place. Every instance where the masked (i.e., requested) token appeared in the set of candidates provided by the model for the given sentence was counted as a successful hit, and the accuracy on this task was used for assessment of the test results.

To evaluate models on the second task, namely, calculating the similarity between sentences (T_2), triplets based on extraction of the same sentence from parallelized novels were used (Y_1 and Y_2, i.e. Y_3 and Y_4). Since the novels were parallelized at the sentence level, it was easy to create pairs of sentences with the same meaning. Each triplet was formed by adding a similar length drawn a different point of the counterpart novel, that sharing as many tokens as possible with the first sentence. Example of a triplet: To evaluate models on the second task, namely, calculating the similarity between sentences (T_2), triplets based on extraction of the same sentence from parallelized novels were used (Y_1 and Y_2, i.e. Y_3 and Y_4). Since the novels were parallelized at the sentence level, it was easy to create pairs of sentences with the same meaning. Each triplet was formed by adding a

sentence sharing as many tokens as possible with the first sentence, and of a similar length, but extracted from a different point of the counterpart novel. Example of a triplet:

1. "Zaista, ko ne bi obišao svet i za manju cenu?" (control sentence, Y_3: *Around the World in 80 Days*, Serbian)
2. "Doista, nije li i za manje od toga vrijedno izvršiti put oko svijeta?" (pair, Y_4: *Around the World in 80 Days*, Croatian)
3. "He! he! pa konačno zašto ne bi uspio?" (false pair, Y_4: *Around the World in 80 Days*, Croatian)

The models were tasked with recognizing two of the sentences in the assigned triplets that actually match (the similarity between the first and the second sentence has to be greater than between the first and the third), and the accuracy on that task was used to evaluate performance. The similarity between the sentences is calculated as the difference of the number 1 and the cosine distance of the calculated sentence vectors. To compute sentences vectors, the model first assigns vector values to each token in the sentence, and then the value of those vectors is averaged to obtain a vector representation of the sentence.

3.2 Model evaluation - downstream tasks

For the purpose of evaluating the performance of models on the remaining two planned tasks, the models were fine-tuned and tested on specially-prepared datasets. The publicly available set, *SrpKor4Tagging* (Stanković et al. 2020) (three hundred and fifty thousand tagged tokens), was used for the part-of-speech tagging task (T_3), while another publicly available set, *SrpELTeC-gold* (Šandrih Todorović et al. 2021), was used for the named entity recognition task (T_4). For both tasks, the models were fine-tuned on 90% of labeled sentences from each set and tested on the remaining 10%. As both tasks are multi-class classification problems, the F_1-score obtained during the classification of the sentences from the test set were used to access the model performance.

4 Evaluation results

The results of the first test (T_1) i.e. the average accuracy of the selected models on the task of guessing the missing tokens in sixteen prepared texts, are shown in Table 2.

No.	1	2	3	4	5	6	7	8	9	10
ID	Andrija/SRoBERTa-base	Andrija/SRoBERTa-L	Andrija/SRoBERTa-XL	Andrija/SRoBERTa-F	classla/bcms-bertic	classla/xlm-r-bertic	xlm-roberta-base	xlm-roberta-large	jerteh/jerteh-81	jerteh/jerteh-355
X_1-Y_1	0.43	0.63	0.66	0.70	/	0.43	0.46	0.51	0.70	**0.75**
X_1-Y_2	0.43	0.62	0.64	0.69	/	0.42	0.46	0.50	0.69	**0.73**
X_1-Y_3	0.37	0.56	0.59	0.63	/	0.34	0.38	0.43	0.66	**0.72**
X_1-Y_4	0.36	0.55	0.64	**0.68**	/	0.34	0.38	0.42	0.58	0.63
X_2-Y_1	0.36	0.47	0.51	0.54	/	0.47	0.50	0.55	0.57	**0.60**
X_2-Y_2	0.37	0.48	0.51	0.54	/	0.46	0.50	0.54	0.56	**0.59**
X_2-Y_3	0.31	0.41	0.45	0.48	/	0.42	0.45	0.50	0.50	**0.54**
X_2-Y_4	0.31	0.42	0.47	**0.51**	/	0.42	0.46	**0.50**	0.47	**0.51**
X_3-Y_1	0.37	0.49	0.52	0.54	/	0.48	0.50	0.55	0.57	**0.60**
X_3-Y_2	0.37	0.48	0.51	0.54	/	0.46	0.50	0.54	0.57	**0.59**
X_3-Y_3	0.30	0.41	0.44	0.47	/	0.41	0.45	0.49	0.50	**0.54**
X_3-Y_4	0.31	0.42	0.47	**0.51**	/	0.41	0.46	**0.50**	0.47	**0.50**
X_4-Y_1	0.42	0.60	0.63	0.67	/	0.43	0.47	0.51	0.73	**0.78**
X_4-Y_2	0.41	0.58	0.61	0.65	/	0.41	0.45	0.49	0.71	**0.75**
X_4-Y_3	0.35	0.53	0.55	0.60	/	0.33	0.38	0.42	0.69	**0.76**
X_4-Y_4	0.34	0.50	0.58	0.62	/	0.33	0.37	0.41	0.62	**0.66**
average	0.36	0.51	0.55	0.59	/	0.41	0.45	0.49	0.60	**0.64**

Table 2. Model accuracy on the task of guessing masked tokens (three candidates) for each of the sixteen prepared masked texts and on the average. Each masked text is marked (at the beginning of each line) with a unique label representing a combination of a tokenizer (X) and a source (Y). The best result in each row ($\pm 1\%$) is marked in bold.

The results show a clear superiority of the new model, *jerteh-355*, which achieved the best result in thirteen out of sixteen cases, and shared the best result (±1%) in two more cases. Moreover, in nine out of twelve cases, the *jerteh-355* model outperformed models on texts masked by tokenizers of those same models. The only model that managed to surpass it in two cases is *SRoBERTa-F*, which performed best in processing the source (Y_4), written in Croatian (the language included in a large percentage in its training set). However, its average accuracy is lower than the accuracy of *jerteh-81*, the other new model. Model 5 (*classla/bcms-bertic*) was not included in the evaluation on this task because that would put in a disadvantage, as unlike the others, it was not trained on the masked language modeling task.

The results of the second test, calculating sentence similarity (T_2) are shown in Table 3. The values show the model accuracy in recognizing sentences with the same/similar meaning in triplets extracted from two Serbian translations of the same novel, Y_1 and Y_2 (first row of values), from the Serbian and Croatian translations of the same novel, Y_3 and Y_4 (second row), and on the average (third row).

No.	1	2	3	4	5	6	7	8	9	10
ID	Andrija/SRoBERTa-base	Andrija/SRoBERTa-L	Andrija/SRoBERTa-XL	Andrija/SRoBERTa-F	classla/bcms-bertic	classla/xlm-r-bertic	xlm-roberta-base	xlm-roberta-large	jerteh/jerteh-81	jerteh/jerteh-355
Y_1-Y_2	0.93	**0.95**	**0.96**	**0.96**	0.92	0.76	0.90	0.87	**0.95**	**0.95**
Y_3-Y_4	0.83	0.89	**0.92**	**0.91**	0.79	0.66	0.78	0.71	0.89	0.83
average	0.88	0.92	**0.94**	**0.93**	0.85	0.71	0.84	0.79	0.92	0.89

Table 3. Performance of the selected models (accuracy) on the task of recognizing sentences with the same or similar meaning in the triplets extracted from the translations of Dostoyevsky (Y_1-Y_2), Verne (Y_3-Y_4), and their average. In each row, the best result (±1%) is marked in bold.

The results for the first set of triplets are very good for several models, where *SRoBERTa-L*, *SRoBERTa-XL*, *SRoBERTa-F*, *jerteh-81* and *jerteh-355* achieve a similar accuracy of around 95%. The *SRoBERTa-XL* model achieves the best results by a very small margin, but also the best results for the second set of triplets containing sentences in the Croatian language (92% accuracy). Therefore, it also has the best overall performance on this task. The only other model that achieves an accuracy of over 90% for the second set of triplets is *SRoBERTa-F*, which was to be expected, because it was also trained on Croatian texts.

The results achieved by the models (F_1-score) on downstream tasks T_3 (part-of-speech-tagging) and T_4 (named entity recognition) are shown in Table 4.

No.	1	2	3	4	5	6	7	8	9	10
ID	Andrija/SRoBERTa-base	Andrija/SRoBERTa-L	Andrija/SRoBERTa-XL	Andrija/SRoBERTa-F	classla/bcms-bertic	classla/xlm-r-bertic	xlm-roberta-base	xlm-roberta-large	jerteh/jerteh-81	jerteh/jerteh-355
T_3	0.974	0.980	0.982	0.982	**0.986**	**0.987**	0.984	**0.986**	0.985	**0.986**
T_4	0.908	0.922	0.929	0.935	**0.942**	**0.942**	0.933	0.935	0.928	0.928

Table 4. F_1-score achieved by the models on tasks T_3 (part-of-speech-tagging) and T_4 (named entity recognition). In each row, the best result (±0.1%) is marked in bold.

From the results shown, it is apparent that on the task T_3 (part-of-speech-tagging) nine out of ten models perform quite well (F_1-score of over 98%), where the results achieved by the four top-performing models (*classla/bcms-bertic*, *classla/xlm-r-bertic*, *xlm-roberta-large* and *jerteh/jerteh-355*) differ by less than 0.02%, indicating that the models are slowly approaching the upper limits of performance for this task.

When it comes to the results on the last task, T_4 (named entity recognition), the highest performance was achieved by models *classla/bcms-bertic* and *classla/xlm-r-bertic*, but the results across the models are not similar to those for task T_3 ($\sim 4\%$ for task T_4 compared to $\sim 1\%$ for task T_3). However, the performance gap between the best and worst performing model is still significantly smaller than the one on the upstream tasks ($\sim 28\%$ gap for task T_1).

In the following section, the achieved results will be discussed, together with the apparent reasons that led to those results, with the aim of determining the most favorable conditions for training Serbian language models in the future.

5 Discussion

The previously presented evaluation results (tables 2–4) show that there is not a single model (or group of models) that performs best in general, but rather that different models (and model groups) are better (or worse) at different tasks. In this sections results for each task will be accessed individually, with an emphasis on the relationship between the achieved performance and the size of the model, the size of its training set and the quality of that set.

5.1 Masked Language Modeling

The results for the masked language modeling task (Table 2) show a substantial advantage of the *jerteh/jerteh-355* model, with *jerteh/jerteh-81* also achieving good results as the second best-performing model on the average. The most probable cause is that these models used the same training dataset. Correlations of the average accuracy of models and the number of their parameters and the models and the size of their training sets are shown in Figure 1.

At first glance, some prominent exceptions are noticeable, primarily the models based on the *XLM-R* architecture, which achieve some of the worst results on this task. If their results are removed, new trends appear (see Figure 2). Thus, when looking only at the *RoBERTa* models, it seems that the larger the model (albeit not very convincingly) and the larger the set used to train it (very convincingly) the better the model performance.

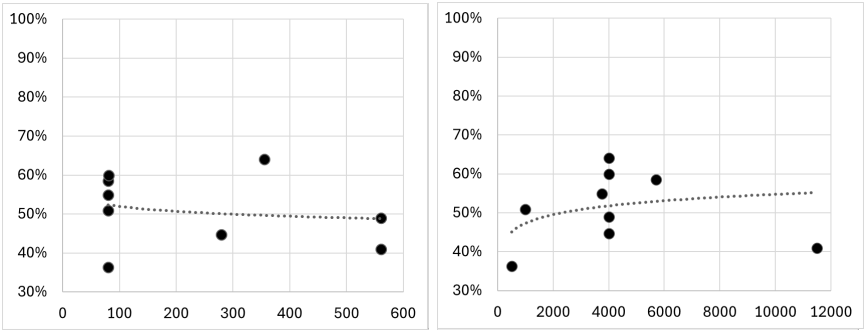


Figure 1. Correlation between each model's accuracy on the masked language modeling task with its size (left), and with the size of its training sets (right). The displayed trend curve corresponds to a logarithmic function.

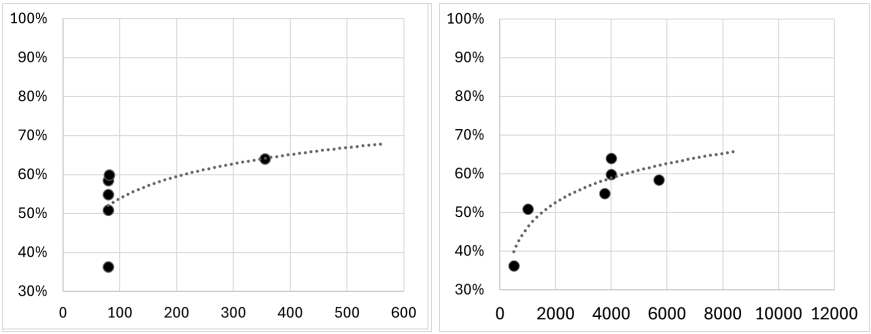


Figure 2. Correlation between each RoBERTa-based model's accuracy on the masked language modeling task with their size (left), and with the size of their training sets (right). The displayed trend curve corresponds to a logarithmic function.

5.2 Calculation of (semantic) sentence similarity

During the evaluation of the T_2 task, it was established that the best results were achieved by the models *SRoBERTa-L*, *SRoBERTa-XL*, *SRoBERTa-F*, *jerteh-81* and *jerteh- 355* when it comes to recognizing sentence pairs in Serbian, and *SRoBERTa-XL* and *SRoBERTa-F* when it comes to recognizing bilingual sentences pairs (Serbian and Croatian, Table 2). This indicates that the key for good sentence embedding is to pre-train the model for the lan-

guages being processed. When embedding Serbian sentences only, the best models are those that were previously pre-trained on a sufficiently large set of sentences in Serbian, but when embedding both Serbian and Croatian sentences, models pre-trained on both Serbian and Croatian have an advantage. On the other hand, models based on the *XLM-R* architecture and pre-trained on one hundred world languages are under-performing on this task, probably due to the large noise that the diverse training set produces.

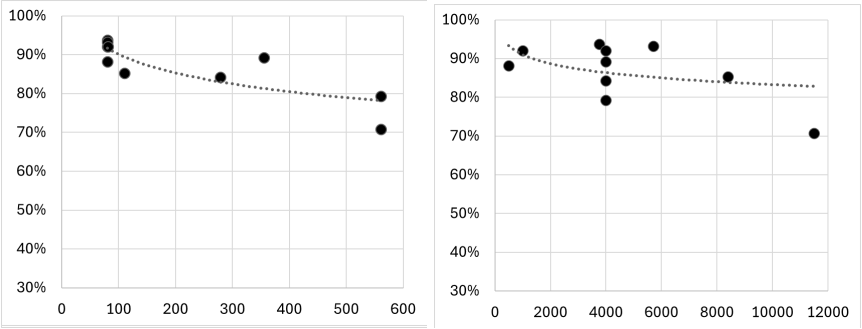


Figure 3. Correlation between each model's accuracy on the sentence embedding task with its size (left), and with the size of its training sets (right). The displayed trend curve corresponds to a logarithmic function.

Figure 3 shows the effect of the model size and the training set size on the performance of the model on this task and, interestingly, the trend lines indicate that the performance decreases with the increase of either parameter. The impact of the dataset size can be attributed (to some extent) to the previously described phenomenon affecting the *XLM-R* architecture. On the other hand, when it comes to the effect of model size, there are additional indicators that smaller models are better for this task, especially for bilingual embeddings where *jerteh/jerteh-81* outperforms *jerteh/jerteh-355*. The reason could be that the smaller model, due to its size, is less adapted to the Serbian language (underfit), but has an aadvantage in generalizing ability.

5.3 Downstream Tasks

Unlike the evaluation on the upstream tasks, the results achieved by the models on the downstream tasks are much more even. Nearly all mod-

els achieve great results for part-of-speech tagging (T_3, Table 4), including those based on the *XLM-R* architecture. Moreover, *classla/xlm-r-bertic* and *xlm-roberta-large*, which are *XLM-R*-based are two of the four models that achieve the best results on this task (the other two being *classla/bcms-bertic* and *jerteh/jerteh-355*).

What these four models have in common is that they are either the largest models or models trained on the largest datasets. A positive correlation between the performance and the size of the model, as well as between the performance and the size of the training sets can also be observed in Figure 4.

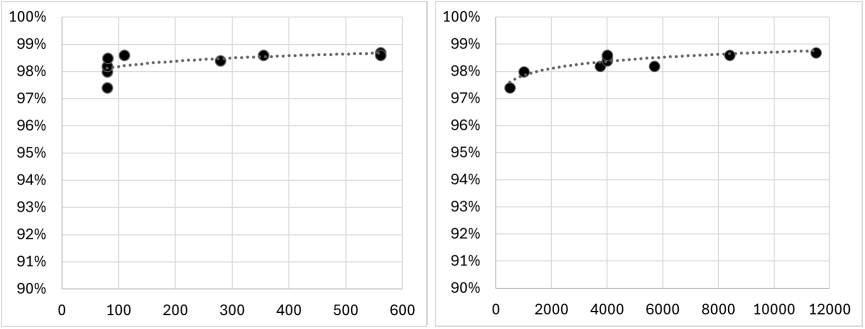


Figure 4. Correlation between each model's F_1-score on part-of-speech tagging with their size (left), and with the size of their training sets (right). The displayed trend curve corresponds to a logarithmic function.

The correlation between the size of the training set is even more obvious in the case of named entity recognition task (Figure 5). The best results on this task (T_4) were achieved by the two models with the largest training sets, with outstanding performance of *XLM-R*-based models. Model size is also shows a (slight) positive correlation with performance on this task.

5.4 Conclusion

When it comes to masked language modeling, it seems that the development of new models for Serbian is going in the right direction. The *jerteh/jerteh-355* model achieves by all means the best results, at least when it comes to working with high-quality texts, even when they are tokenized by unknown

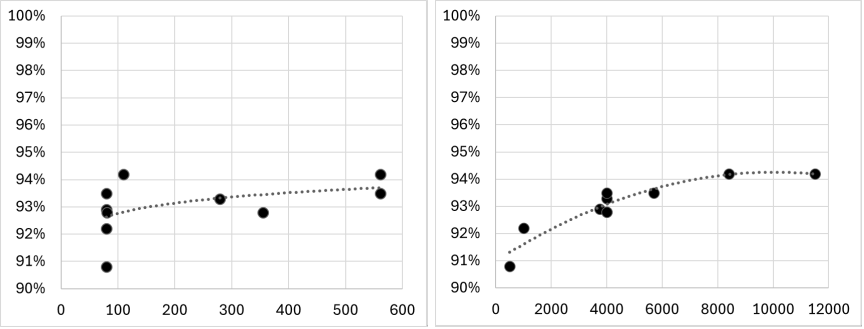


Figure 5. Correlation between each model's F_1-score on named entity recognition with their size (left), and with the size of their training sets (right). The displayed trend curve corresponds to a logarithmic function.

tokenizers (Table 1). Although the training data set size has a positive correlation with the performance of the model (fig.2), the quality of the set should not be neglected, since *jerteh/jerteh-355* and *jerteh/jerteh-81* outperform the *Andrija/SRoBERTa-F* and *Andrija/SRoBERTa-XL* models trained on larger training sets, indicating that web corpora may not always be sufficient to train quality models for this task. This is consistent with the conclusion of another recent research (Li et al. 2023). However, new research should include non-literary sources in the evaluation set, in order to obtain a more comprehensive outlook of the situation.

On the task of calculating the similarity between sentences (sentence embedding) models *Andrija/SRoBERTa-F* and *Andrija/SRoBERTa-XL* stood out, followed by *Andrija/SRoBERTa-L*, *jerteh/jerteh-81* and *jerteh/jerteh-355*, at least when it comes to embedding Serbian sentences (Table 3). What sets these models apart is that they are smaller compared to other models and trained on a larger set, so generalization seems to be the key for this task, that is, larger data sets in combination with smaller models. Also, when it comes to processing sentences of a wider linguistic spectrum (e.g. South Slavic languages), it would be necessary to include sentences from the complete spectrum in the training set or, even better, to adapt the dictionary to map a wider range of tokens, and therefore allow for the correct vectorization of these sentences. New research on this topic should also explore a more recent method of sentence vectorization, for ex-

ample, using the *sentence transformer* architecture (Reimers and Gurevych 2019).

In the case of both assessed upstream tasks, the performance achieved by the models based on the *XLM-R* architecture is significantly lower than the one of the models based on the *RoBERTa* architecture. In the case of the first task (T_1), this can be explained by their significantly larger token dictionary (which makes the selection of the appropriate token more difficult). However, such an explanation would not be adequate for the second task (T_2). On the other hand, the models based on the *XLM-R* architecture proved to be the best (by a small margin) on downstream tasks, primarily for named entity recognition (T_4). It seems that in order to successfully solve this task, it is best that the model encounters a wide variety of tokens during pre-training, while additional training on Serbian texts brings additional improvements. It seems that in order to improve the performance, it would be optimal to retrain the *XLM-RoBERTa-large* model using the largest and highest quality set of texts in the Serbian language. When it comes to part-of-speech tagging, it seems that any of the new models would be adequate to handle this task.

Acknowledgment

The most important datasets for the training of *GPT2-orao*, *GPT2-vrabac*, *jerteh-81* and *jerteh-355* models were provided by the Society for Language Resources and Technologies.[12]

Computer resources for the training of *GPT2-orao* and *GPT2-vrabac* models were provided by the National Platform for Artificial Intelligence of Serbia.

Computer resources for the training of *jerteh-81* and *jerteh-355* models were provided by the Faculty of Mining and Geology, University of Belgrade.

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

Thank you!

References

Biemann, Chris, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. "The Leipzig corpora collection-monolingual corpora of standard size." *Proceedings of corpus linguistic* 2007.

12. JeRTeh

Bogdanović, Miloš, Jelena Kocić, and Leonid Stoimenov. 2024. "SRBerta-A Transformer Language Model for Serbian Cyrillic Legal Texts." *Information* 15 (2). ISSN: 2078-2489. https://doi.org/10.3390/info15020074.

Clark, Kevin, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. "Electra: Pre-training text encoders as discriminators rather than generators." *arXiv preprint arXiv:2003.10555.*

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." *arXiv preprint arXiv:1911.02116.*

Cvejić, Andrija. 2022. "Prepoznavanje imenovanih entiteta u srpskom jeziku pomoću transformer arhitekture." *Zbornik radova Fakulteta tehničkih nauka u Novom Sadu* 37 (02): 310–315.

Cvetanović, Aleksa, and Predrag Tadić. 2023. "Synthetic Dataset Creation and Fine-Tuning of Transformer Models for Question Answering in Serbian." In *2023 31st Telecommunications Forum (TELFOR),* 1–4. IEEE.

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.*

Dobreva, Jovana, Tashko Pavlov, Kostadin Mishev, Monika Simjanoska, Stojancho Tudzarski, Dimitar Trajanov, and Ljupcho Kocarev. 2022. "MACEDONIZER-The Macedonian Transformer Language Model." In *International Conference on ICT Innovations,* 51–62. Springer.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. "DEBERTA: Decoding-Enhanced BERT with Disentangled Attention." In *International Conference on Learning Representations.*

Jiang, Albert Q, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. "Mistral 7B." *arXiv preprint arXiv:2310.06825.*

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.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep learning." *nature* 521 (7553): 436–444. https://doi.org/10.1038/nature14539.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 7871–7880.

Li, Yuanzhi, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. "Textbooks are all you need ii: phi-1.5 technical report." *arXiv preprint arXiv:2309.05463.*

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Roberta: A robustly optimized BERT pretraining approach." *arXiv preprint arXiv:1907.11692.*

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

Ljubešić, Nikola, Vit Suchomel, Peter Rupnik, Taja Kuzman, and Rik van Noord. 2024. *Language Models on a Diet: Cost-Efficient Development of Encoders for Closely-Related Languages via Additional Pretraining.* Submitted for review.

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

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. "Language models are unsupervised multitask learners."

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.

Rajpurkar, Pranav, Robin Jia, and Percy Liang. 2018. "Know what you don't know: Unanswerable questions for SQuAD." *arXiv preprint arXiv:1806.03822.*

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 3982–3992.

Šandrih Todorović, Branislava, 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.

Škorić, Mihailo. 2023. "Композитне псеудограматике засноване на паралелним језичким моделима српског језика." Докторска дисертација. PhD diss., Универзитет у Београду.

Škorić, Mihailo, Miloš Utvić, and Ranka Stanković. 2023. "Transformer-Based Composite Language Models for Text Evaluation and Classification." *Mathematics* 11 (22): 4660.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. "Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian." In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 3954–3962.

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. "Alpaca: A strong, replicable instruction-following model." *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html* 3 (6): 7.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." *Advances in neural information processing systems* 30.

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.

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

Wenzek, Guillaume, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data" [in English]. In *Proceedings of the 12th Language Resources and Evaluation Conference,* 4003–4012. Marseille, France: European Language Resources Association, May. ISBN: 979-10-95546-34-4. https://www.aclweb.org/anthology/2020.lrec-1.494.

Building the Dictionary of Football Terminology Through Data-Driven and Ontolex Model

UDC 811.163.41'322.2
DOI 10.18485/infotheca.2024.24.1.2

ABSTRACT: This paper investigates the innovative approach to developing a dictionary of football terminology through the application of linked data. Using traditional approaches of dictionary creation includes sequentially written entries and manual analysis of corpora which can be very time consuming. Our approach uses the techniques of computational linguistics and automatization in order to increase the speed of dictionary creation process. The Digital Dictionary of Football Terminology which we have developed is meant for both human and software application use, giving users a fast approach. Furthermore, this gives the possibility of constant updating as well as the integration with different digital resources. This paper focuses on the Serbian Language section of the trilingual Serbian-Spanish Dictionary of Football Terminology which is developed within the doctoral dissertation of Jelena Lazarević.

KEYWORDS: dictionary, dictionary of football terminology, corpus, lexical units, computational linguistics.

PAPER SUBMITTED: 10 June 2024
PAPER ACCEPTED: 27 June 2024

Jelena Lazarević
jelazarevic1@gmail.com
ORCID: 0009-0004-4481-2729
PhD student,
University of Belgrade
Faculty of Philology
Belgrade, Serbia

Olivera Kitanović
olivera.kitanovic@rgf.bg.ac.rs
ORCID: 0000-0002-7571-2729
University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

1 Introduction

The integration of language resources represents one of the principal topics in contemporary research, relating language technologies. There is an increase in the implementation of the data-driven approach for its improved possibilities of data processing possibilities.

A data-driven approach is a methodology leaning on the analysis and interpretation of data as a basis of decision-making and problem-solving. Instead of using intuition or experience through previous theoretical knowledge, a data-driven approach uses concrete, objective and measurable information gathered from the data. This approach is very popular in disciplines such as linguistics, medicine and various information technologies (Kitanović et al. 2021).

This paper will feature researching the possibility of streamlining and improving the traditional approach of creating dictionaries through the example of developing a bilingual dictionary of football terminology.

The traditional approach of creating dictionary entries, one by one, most often in an alphabetic order relies on the analysis within the corpora or excerpted examples with the consultation of referent dictionaries for each individual entry. Using computational linguistics technologies makes this process quicker than the aforementioned practices. Moreover, the entries are not processed sequentially. Instead, the data-driven approach is implemented, led by data and parallel processing of a greater amount of data at once.

Bergh and Ohlander (Bergh and Ohlander 2019) had noticed that XX century features a rise in the use of football terminology in general speech, while some terms had become common in the speech of fans. The language of football is in constant change, adapting itself to events in and around the game. Due to its importance and a great media presence, football, as the People's game, influences general English dictionaries to include football terms as part of the General language.

Digital dictionaries have vast advantages in comparison to traditional, printed dictionaries. According to Fuentes-Olivera and Tarp (Fuertes-Olivera and Tarp 2014), the ideal solution for contemporary dictionaries is in fact their online edition. The first and key advantage of an online dictionary is the speed of search which enables fast word and phrase search, while printed dictionaries need to be manually revisited. The second advantage is the possibility of constant updating of the entries, unlike printed dictionaries which can expire in their relevance within a matter of years.

Online dictionaries also allow the adaptation of fonts, letter size and other parameters to fulfill user needs, while containing additional information, such as detailed etymology and pronunciation. The availability of mobile applications and desktop platforms always enables users access to the dictionary. It is very important to add the possibility of search within a single word entry, to quickly find meaning or information on a certain term while reading.

Putting dictionaries onto digital language learning platforms enables progress following through making word lists, notes, flash cards and other helpful tools that streamline the language learning process. Integrating the dictionaries with the text corpora is very significant because users are allowed to learn words and expressions based on concrete examples of their use. Unlike the global practice, Serbia still features greater commonality levels of traditional dictionary use due to a feeling of authenticity or a simple habit. Currently, digital dictionaries have become a necessary tool for language learners and those who speak multiple languages. At this point, we need to mention that any digital dictionary can be printed.

The Dictionary of Football Terminology featured in this paper is a digital dictionary intended for both human and machine use, regarding software applications. It represents a part of a broader research and the compilation of the first Digital Dictionary of Football in the Serbian language. The entries are translated to Spanish, within the doctoral dissertation of Jelena Lazarević under the title: "Language Characteristics of the New Media Discourse on Football: a Contrastive Analysis of the Serbian and Spanish Language Corpora". The emphasis stays on the connection between computational linguistics and the implementation of technologies of semantical networks in order to fulfill the needs of individual users, as well as software tools.

The Dictionary is intended for the use of individuals working in sport: athletes, coaches, referees, professors, students, sports journalists as well as anyone wishing to master the precise interpretation of footballing terms. This paper particularly focuses on the Serbian section of the bilingual Serbian/Spanish Dictionary of Football Terms.

In Section 2 of this paper we will view the two ways of excerpting information necessary for the creation of the Dictionary: automatic excerption from the corpus *srFudKo*, term candidates, frequencies and examples of use (Krstev et al. 2015; Ivanović et al. 2022), as well as the extraction from the preexisting dictionaries and glossaries available online (Kitanović et al. 2021).

2 Data preparation for the Dictionary compilation

The corpus *srFudKo* (Lazarević et al. 2023) was created from media articles about football in the Serbian language, compiled from five online news sites: *B92, Blic, Mondo, Politika* and *Sport Klub*. The articles were automatically

downloaded through various techniques, which was followed by deduplication of articles, eliminating articles shorter than 3000 characters, sentences in other languages and tables that contained only numerical results. After filtering the articles and text debriding the corpus *srFudKo* was created; it contains 10,100,553 tokens, out of which a total of 8,618,426 represents words, while the rest is interpunction. The corpus content was then annotated by word type and lemmatized though taggers for the Serbian language SrpKor4Tagging-TreeTagger (Stanković et al. 2020; Stanković, Škorić, and Šandrih Todorović 2022).

The corpus *srFudKo* was then used for information excerption. The initial result of the automatic excerption was a list of word candidates, followed by frequency of use. In the first phase, single part terms were extracted, evaluated, and tagged, which was followed by the multiple compound terms (Krstev et al. 2015), most frequent syntactical patterns based on the morphological dictionaries of the Serbian language and local grammars (Krstev 2008). Aside from single part terms, their keyness[1] was also calculated, comparing frequencies in *srFudKo*, with the Contemporary Serbian Language Corpus *SrpKor2013* (Утвић 2011). In terms of compound terms, more complex associative measures were used, such as *T-Score* and *CValue* (Vu, Aw, and Zhang 2008). The details of this stage are described in the paper: "Football terminology: compilation and transformation into *OntoLex-Lemon* resource"(Lazarević et al. 2023).

In terms of information extraction from the preexisting dictionaries and glossaries available online, various files of various types and levels of detail were downloaded. Instead of consulting a dictionary for each individual entry, crossing, harmonizing and processing of all sources was planned to streamline and accelerate the creating of the first contemporary dictionary of football, rich in information. Data integration relied on multiple sources, but the main roles belong to: 1) The four language dictionary based on Serbian, English, French and Spanish terms lists (Mihajlović 2003) and 2) *Kicktionary*[2] (Schmidt 2009), semantically based on the frames of *FrameNet*[3] (Fillmore et al. 2003).

1. https://www.sketchengine.eu/documentation/simple-maths/

2. http://www.kicktionary.de/ the multilingual lexicon that covers terms and their definitions related to football matches, tactics, players, and equipment, with the aim of facilitating learning and understanding of football terminology through visualizations and explanations.

3. *FrameNet* is a lexical database that documents semantic frames and how they connect words to their meanings in different contexts.

Frame Semantics is the theory of meaning focused on conteptual structures – frames, which represent the background necessary for understand words and expressions in various contexts. These frames include associated elements and relations activated while using a certain term, allowing a better understanding of their meaning and context of use. Frames and lexical units in English were crossed with the English side of the QUAD dictionary in order to enrich the Serbian side of the dictionary with semantic frames. Considering that no language of Football Terminology in Serbian had descriptive entry definitions and given the lack of machine translating for the sake of a faster compilation of the Dictionary, we used Glossaries of football terminology in English and Spanish with semi-automatic processing. Aside from *Kictionary*, which also contains certain definitions, dictionaries available at the moment of data processing were also used. We need to add that some of these dictionaries have become unavailable on the previously known addresses:

- The Field[4] Glossary with 900 of the most used football terms is organized through illustrated scenarios. The base language is Portuguese, while containing equivalents in English and Spanish, alongside examples of use. Homonymy and Polysemy was implemented through separated dictionary entries and its structure was inspired through the concepts of Semantic Frames
- The UEFA[5] Football Glossary we have used features a total of 8,086 entries: 1,893 in German, 2,306 in English and 1,887 in French.
- Football Glossary *UsingEnglish*[6] contains different domains and has only 74 entries for frequent terms.
- The NINE language dictionary[7] contains 77 terms with translated equivalents in English, German, French, Spanish, Italian, Portuguese, Polish, Russian and Ukrainian.

The Dictionary of Football Terms *FudLe* was created based on the doctoral dissertation of Jelena Lazarević and contains now a total of 2,648 entries with translation equivalents in Serbian and English that were enriched by

4. previously available on http://dicionariofield.com.br/, see https://www.facebook.com/dicionariofield

5. previously available on https://www.uefa.com/insideuefa/dictionary/

6. https://www.usingenglish.com/glossary/football-vocabulary/

7. https://www.uefa.com/MultimediaFiles/Download/competitions/General/01/80/75/52/1807552_DOWNLOAD.pdf

the word types and frequencies of the *srFudKo* corpus. A total of 93 entries feature an added definition.

Within the Serbian partition, there are: 990 single part term, 996 two-part term, 369 featuring three components, 200 featuring four components, while 123 contain five and more components. Connecting with the semantic frames of the dictionary *Kicktionary* was performed on 142 lexical units. After the extraction of the terminological phrases from the *srFudKo* corpus, manual evaluation and tagging domain markers (whether the term refers to sport or distinctively to football), a list of 1,943 terms was extracted that was also given a code of their syntactic group, frequency, and association measure.

3 Compilation and transformation of the Dictionary into the *Ontolex* Model

The use of the ontology lexicon *OntoLex-Lemon*,[8] short for Ontolex (*LExicon Model for ONtologies*) (McCrae et al. 2017) is on the rise in terms of the implementation of lexical resources as a collection of connected data available online. Determinants, whether they pertain to either one word term or multi-word term from the football domain, alongside the results of extraction from the corpus *srFudKo* (frequencies and examples) were enriched with information downloaded and processed from the dictionaries and glossaries, that were then represented by using the *OntoLex* model.

Aside from the example, dictionary entries point out the word type, of each individual entry and potentially their grammatical attributes. Additional information, such as the definition, meaning, conceptual relations regarding other entries within other resources such as knowledge bases, ontology or lexical bases helps users to better understand the term they are searching for. The context of use through various examples extracted from the corpus, alongside the frequencies are there to complete the terminological image.

The first uses of Lemon, as well as the *OntoLex-Lemon* in Serbian, are connected to the creation of the lexical base Leximirka[9] (Stanković et al. 2018) and the translation of the systems of Serbian electronic dictionaries (Krstev 2008) into a relational base (Рујевић 2022).

8. https://www.w3.org/2016/05/ontolex/ and https://jogracia.github.io/ontolex-lexicog/

9. https://leximirka.jerteh.rs/

This paper features the *OntoLex* model for modeling the RDF[10] (eng. Resource Description Framework), which means representing linguistic descriptions of lexical units from the generated dictionary with the amplification of the lexical layer in order to enable their connections and a machine-provided understanding online. The *OntoLex* modules are introduced through the examples of implementation, beginning with the basic *OntoLex* module which defines lexical units, word forms and lexical meanings. The *vartrans*[11] module contains two important classes: Translation and TranslationSet, where the translation represents a liaison between two lexical meanings, connected to the lexical units in various languages. The *LexInfo*[12] type ontology, type value and property value are used as a catalogue of data categories, such as: grammatical gender, number, word type, etc.

Translation categories are represented by pairing with an external catalogue *OEG Translation Categories.*[13] Other frequently used dictionaries, such as: *Dublin Core, DCMI Metadata Terms*,[14] are used for metadata on sources, authorship, version, or license data. For modeling, recommendations are given in the following guidelines: Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries (v2.0)[15] (Río Gracia et al. 2023) (Río Gracia et al. 2023), as well as: "Guidelines and best practices for LLOD."

According to the aforementioned recommendations, a bilingual dictionary is organized so that individual languages and separated RDF charts are connected through an additional chart which contains connections of translational equivalents. The generated dictionary is formed by four charts, i.e. collections of data:

- Source lexicon, in this case of the Serbian language,
- Target lexicons, in this case Spanish,
- Compilation of equivalents, also known as *TranslationSet.*

The approach of using resources prepared in the previous phases, shown on the left and bottom parts of the image, as well as models used for transforming them into the RDF form, depicted on the right are illustrated in Figure 1. The output is *FudLe* a bilingual dictionary: in Serbian and Spanish, containing entries enriched with various additional information.

10. https://www.w3.org/RDF/
11. https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans
12. https://lexinfo.net/, https://github.com/ontolex/lexinfo
13. http://purl.org/net/translation-categories
14. http://purl.org/dc/elements/1.1/
15. https://www.w3.org/community/bpmlod/wiki/Guidelines_and_best_practices_for_LLOD

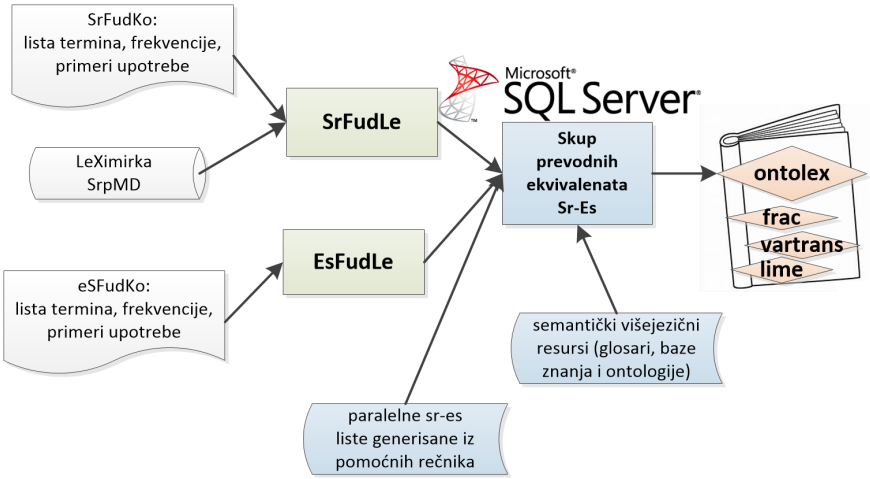


Figure 1. Graphic representation of language data integration from various resources into the *FudLe* Dictionary.

The *OntoLex-FrAC* module (Chiarcos et al. 2020) was used to include information on frequencies, examples in context and collocations extracted from the *srFudKo* corpus into the *FudLe* dictionary. Automatically extracted terms were then manually evaluated, tagged and marked, depending on whether they belonged to the sporting domain or a specialized footballing domain. They were then completed with inflective forms which were further transformed into the *OntoLex* model. The terms were then connected to the examples of use which were chosen by using the GDEX[16] (Good Dictionary Examples) algorithm (Kilgarriff et al. 2008).

The morphological dictionary of compound terms was made by using the tool *LeXimir* (Stanković et al. 2011), and then transformed through the application which follows *OntoLex* specifications and published examples (Chiarcos et al. 2022).

Grammatical information, morpho-syntactic properties of word types were shown in concordance with *LexInfo* vocabulary. The following sections show the use of the basic *OntoLex* Module, as well as the frequency module *OntoLex-FrAC*, examples of use and other corpus-based information (Chiarcos et al. 2022).

16. https://www.sketchengine.eu/documentation/manual-for-gdex/

3.1 The core of the *FudLe*

The corpus *srFudKo* was compiled in Latin script, which means that the proveniente dictionary *FudLe* was also written by using Latin script. The example of the term: *football match* was shown in previous paper (Lazarević et al. 2023) while this paper features the example of the term: yellow card through the *OntoLex* model, which was amplified with additional information on the meaning and the relations towards other resources and languages.

A yellow card is defined as a warning that a referee is giving a player due to their unsportsmanlike conduct. By using the yellow card, the referee is sanctioning a foul in cases where the severity of the conduct hasn't yet merited a red card and immediate expulsion, such as a fight or an intentional foul meant to injure the opponent. This term also exists in the Serbian Morphological Dictionary SrpMD of compound words (Krstev and Vitas 2009) in the form of DELAC script (Savary, Krstev, and Vitas 2007). The source script is the following:

```
žuti(žut.A8:adms1g) karton(karton.N1:ms1q),
NC_AXN+DOM=Sport+Comp+Conc
```

The final transducer NC_AXN generates inflective forms for electronic morphological dictionaries of compound words, where NC marks a noun collocation, and AXN shows a collocation in the form: Adjective-Noun, with the adjective concurring with the noun's grammatical number, gender, case, and animacy. For the components of a compound the transducer provides information on their lemmas, in this case *yellow* and *card*, their corresponding inflection transducers (A8 and N1) and values of the grammatical properties of forms in which they occur (adms1g и ms1q). The grammatical properties are as follows: a-positive grade, d-determined (long) form, m-male grammatical gender, s-singular, 1-nominative case, g-the information on animacy is not important, q-inanimate noun (objects). Most of the grammatical properties are easily mirrored into the ontology *LexInfo*, only the information that the animacy is of no importance is not foreseen. Taking into account the fact that the collocation *yellow card* can be found in various terminological dictionaries, by using the *OntoLex* model, this lexicalized collocation can be tagged with its particular meaning.

A part of the data base for the system *LeXimirka*, implemented within the MS SQL Server (Stanković et al. 2018; Pујевић 2022), is shown in Figure 2 which illustrates tables: *LexicalEntry* and flective forms (marked as *Forms*), as well as compound (marked as Components).

One of the roles of the *LeXimirka* system is to generate flective forms and their grammatical information for single and multi-word terms in the *FudLe* Dictionary. Grammatical information is connected to the inflective forms through Data Categories and their values (*DatCatValues*). The system is equipped with metadata that refer back to the connected information between data categories in Serbian morphological dictionaries and the *LexInfo Dictionary*. A single lexical unit may have various meanings, saved in the table *LexicalSense* which compiles individual categories through the function *SenseProperties* from the *DataCategories* catalogue.

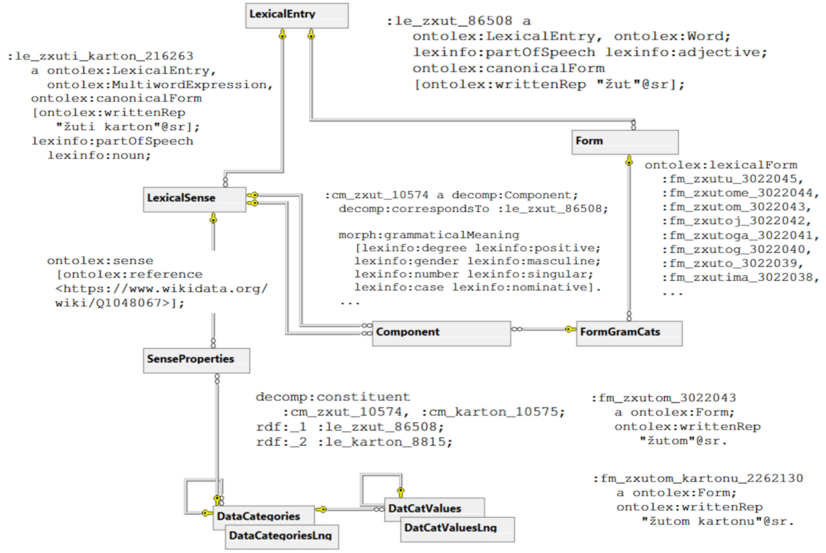


Figure 2. MS SQL Server Data Base diagram with the connected tables and the example of data exported into *OntoLex*.

The script in Figure 2 depicts the collocation *yellow card* with the title of the lexical unit *le_ zxuti_ karton_ 216263*. It consists of the prefix *le*, derived from *LexicalEntry*, the term itself, in this case *yellow card*, and the primary key 216263 in the *LexicalEntry* table within the entire *LeXimirka* data base.

Regarding individual components, the prefix is chosen in a similar way cm, as to mark entries coming from the Component table, with the prefix

fm marking entries from the Form table. The specific letters are changed based on the Aurora coding (Vitas 1979), created by professor Duško Vitas to provide equality in meaning ASCII coding, used for names TITLES whole literals. This means that written forms are coded by using the coding scheme UTF-8.

We can see that the lexical unit *:le_zxuti_karton_216263* belongs to the class *ontolex:LexicalEntry* and that it is a compound component *ontolex:MultiwordExpression*, which is firstly stated in its canonic form, as a lemma *ontolex:canonicalForm*. This is followed by the written form of the lemma, with the added data that in this case is a noun. In terms of meaning, two external resources are used. The first is Wikidata, as a knowledge base, which contains the entry *yellow card* marked with a code *Q1048067*, with the complete URL address as follows: https://www.wikidata.org/wiki/Q1048067. It contains translational equivalents in 38 languages, which means that aside from lexicalization in Serbian, there is also a Spanish translation: *tarjeta amarilla*.

```
:le_zxuti_karton_216263 a ontolex:LexicalEntry,
   ontolex:MultiwordExpression,   ontolex:canonicalForm
   [ontolex:writtenRep "žuti karton"@sr];
   lexinfo:partOfSpeech lexinfo:noun;
   ontolex:denotes <https://www.wikidata.org/wiki/Q1048067>.

  decomp:constituent :cm_zxuti_10574, :cm_karton_10575;
    rdf:_1 :le_zxut_86508;
    rdf:_2 :le_karton_8815.

# komponente kanonskog oblika
:cm_zxuti_10574 a decomp:Component;
  decomp:correspondsTo :le_zxut_86508;

  morph:grammaticalMeaning
    [lexinfo:degree lexinfo:positive;
    lexinfo:gender lexinfo:masculine;
    lexinfo:number lexinfo:singular;
    lexinfo:case lexinfo:nominative].
  ...
```

The corresponding section in Spanish looks as follows:

```
:le_tarjeta_amarilla_10001 a ontolex:LexicalEntry,
   ontolex:MultiwordExpression,   ontolex:canonicalForm
   [ontolex:writtenRep "tarjeta amarilla"@es];
   lexinfo:partOfSpeech lexinfo:noun;
   ontolex:denotes <https://www.wikidata.org/wiki/Q1048067>;

  decomp:constituent :cm_tarjeta_1, :cm_amarilla_2;
    rdf:_1 :le_tarjeta_1;
    rdf:_2 :le_amarillo_2.
```

3.2 Defining meaning in the *FudLe* Dictionary

The class *ontolex:LexicalSense* defines the meaning of a lexical unit when it is interpreted in relation to a pertaining ontological element. The lexical meaning represents a reification of the lexical units being paired with an ontological whole to which it relates. The link between a lexical unit and its ontological entity over the object: *ontolex:LexicalSense* implies that a lexical unit may be used for pairing a given ontological entity.

The idea of a more precise connecting of the lexicon *FrameNet* and the *OntoLex* model has already been shown (Robin, Kulkarni, and Buitelaar 2023). When defining lexical units, it has been shown that they can be used for marking a determined ontological predicate by using the property *ontolex:denotes*. This means that a lexical unit is marking an object class or an ontological element. Also, it is possible to present an evocation of a certain mental concept from the given lexical unit, which doesn't pertain to the class of formal interpretation in a given model.

The class Lexical Concept (*ontolex:LexicalConcept*) represents a mental abstraction, concept or thought unit that can be lexicalized in a way that the class *ontolex:LexicalConcept* represents a subclass of the element *skos:Concept*. The definitions are added to the lexical concepts as a glossa by using property *skos:definition*.

Researching the creation of the semantic football dictionary, while focusing on the football domain corpus, Wu and Li (Wu and Li 2017) have used Word2Vec (Mikolov et al. 2013) to train a vector model for words within the corpus. Combining a statistical algorithm TF-IDF and word vectors, they have processed concepts characteristic for the football domain.

For the football term list, they have calculated the distance from the words within a vector model and separated three of the most proximal terms with the highest cosines value as synonyms for the headword to build a semantic dictionary. They have shown that preexisting semantic dictionaries such as WordNet don't correspond to the needs of certain domains, since the semantic granularity of the words is lower.

It has been necessary to create a semantic dictionary that connects the synonyms to a term, its subordinate and supraordinate terms and definitions in various languages. This is how a semantic dictionary is being created which implements semantic queries of articles from the football domain.

We have created our individual identifiers of concepts based on the gathered resources. The UEFA dictionary terms and definitions was taken as a model for defining the concepts. In the *yellow card* example we see that *:ct_ yellow_ card_ 1* starts with *ct* (abbreviation for Concept), followed by

the term "yellow card" in English, ending in an identifier (for the sake of unambiguity), using an underscore "_" instead of a space.

```
:conceptLexicon a ontolex:ConceptSet .
:le_zxuti_karton_216263 ontolex:evokes :ct_yellow_card_1.
:le_tarjeta_amarilla_10001 ontolex:evokes :ct_yellow_card_1.

:ct_yellow_card_1 a ontolex:LexicalConcept ;
  ontolex:LexicalisedSense :sn_zxuti_karton_216263, :sn_tarjeta_amarilla_10001;
  ontolex:isConceptOf  <https://www.wikidata.org/wiki/Q1048067> ;
  skos:definition "Disciplinska sankcija koju sudija izriče tokom utakmice
    protiv igrača zbog kršenja Pravila igre, posebno zbog lošeg ponašanja u
    sportu ili odlaganja ponovnog početka igre."@sr;
  skos:definición "Sanción disciplinaria impuesta durante un partido por el
    árbitro a un jugador por infracción de las Reglas de Juego, en particular
    por mala conducta deportiva o por retrasar la reanudación del juego."@es;

ontolex:isEvokedBy :le_zxuti_karton_216263, sn_tarjeta_amarilla_10001;
  skos:inScheme :FudLe .
```

Considering that a glossa can be connected to *ontolex:LexicalSense*, which then represents the specificities of use through the property *ontolex:usage*, the interpretation of a lexical unit in relation to the meaning defined within a given ontology is often adapted to the conditions of use or pragmatic implications. Especially due to the registry, connotations, or variations in meaning. These facets can be specified through the use of properties which enables gathering information on the conditions of use and pragmatic implications under which the lexical entry is being referred to a lexical meaning.

The aforementioned conditions of use are not being introduced as a replacement for the formal meaning, but as a complement which better describes a lexical unit. In the case of the *FudLe*, definitions compiled from various sources were extracted from the corpora or directly written in this way.

```
:sn_zxuti_karton_216263 a ontolex:LexicalSense ;
  ontolex:reference <http://www.kicktionary.de/LUs/Sanction/LU_1240.html>;
  ontolex:isLexicalizedSenseOf :ct_yellow_card_1 ;
  ontolex:isSenseOf :le_zxuti_karton_216263 ;
  ontolex:usage [ rdf:value "Žuti karton koji sudija vadi iz svog džepa da bi pokazao
    igraču koji je uradio nešto prilično loše, kao što je loš faul ili igranje rukom,
    ali ne tako loš kao prekršaj crvenog kartona kao što je tuča. Žuti karton često
    ide zajedno sa drugom kaznom kao što je slobodan udarac. Dva žuta kartona na jednom
    meču znače crveni karton, a dva žuta u određenom periodu suspenziju."@sr ].

:sn_tarjeta_amarilla_10001 a ontolex:LexicalSense ;
  ontolex:reference <http://www.kicktionary.de/LUs/Sanction/LU_1240.html> ;
  ontolex:isLexicalizedSenseOf :ct_yellow_card_1 ;
  ontolex:isSenseOf :le_tarjeta_amarilla_10001 ;
  ontolex:usage [ rdf:value "La tarjeta amarilla se muestra a un jugador por acciones
```

```
como faltas fuertes o tocar el balón con la mano. Puede ir acompañada de sanciones
como tiros libres. Dos tarjetas amarillas en un partido significan una roja, y dos
en un período llevan a suspensión."@es ].
```

3.3 The implementation of the frequencies and their examples in the FudLe dictionary

The documentation for the *OntoLex* module and information on frequencies, examples of use within a given context and other corpora information known as "Frequency, Attestations and Corpus data" (FrAC) are available at the following link: https://github.com/ontolex/frequency-attestation-corpus-information (Chiarcos et al. 2022; Chiarcos et al. 2020).

The *FrAC* module is directed towards the enrichment of dictionaries and other linguistical resources which contain lexicographic data through a model for the representation of statistic data extracted from the corpora: information on frequency and appearance, collocations, markers from lexical resources into corpora and other text collections – the confirmation of their use within the corpora, tagging of corpora and other sources of lexical information, lemmatization according to the dictionary and distributional semantics, meaning vectors of collocations, word, meaning and term embedding, The module treats cases of use within lexicography based on corpus, corporal linguistics, and the processing of natural language in combination with basic and other *OntoLex* modules.

The auxiliary class *:SrFudKo* has been defined in order to provide us with easier handling and shorter annotation. Indications of the frequencies used for *srFudKo* are published on "noSketch" (Kilgarriff et al. 2014) which is maintained by the Society for language resources and technologies – JePTeX.[17]

An auxiliary class *:SrFudKo_lemma_frek* was introduced for the overall frequency of all forms of a concrete term, with a reminder that a term can contain one or various components. The motive for introducing the said class was compacted coding, because it diminishes the number of code segment repetitions. Frequency comparison with the Referent and General Corpus of the Serbian Language СрпКор2021 (Krstev and Stanković 2023; Škorić and Janković 2024) was completed through the noting of frequencies within that corpus by using the object *:SrpKor2021*.

```
# korpus fudbala - srpski
```

17. JePTeX

```
:SrFudKo a owl:Class;
  rdfs:subClassOf  [a owl:Restriction;
    owl:onProperty frac:observedIn ;
  owl:hasValue <https://noske.jerteh.rs/#dashboard?corpname=SrFudKo>] .

:SrFudKo_lemma_freq rdfs:subClassOf
    frac:Frequency, :SrFudKo, [a owl:Restriction;
      owl:onProperty dct:description; owl:hasValue "lemma frequency"].

# opšti korpus savremenog srpskog jezika
:SrpKor2021 a owl:Class;
  rdfs:subClassOf  [a owl:Restriction;
    owl:onProperty frac:observedIn ;
  owl:hasValue <https://noske.jerteh.rs/#dashboard?corpname=SrpKor2021>] .

:SrpKor2021_lemma_freq rdfs:subClassOf
    frac:Frequency, :SrpKor2021, [a owl:Restriction;
      owl:onProperty dct:description; owl:hasValue "lemma frequency"].

# korpus fudbala - španski
:EsFudKo a owl:Class;
  rdfs:subClassOf  [a owl:Restriction;
    owl:onProperty frac:observedIn ;
  owl:hasValue <https://noske.jerteh.rs/#dashboard?corpname=EsFudKo>] .

:EsFudKo_lemma_freq rdfs:subClassOf
    frac:Frequency, :EsFudKo, [a owl:Restriction;
      owl:onProperty dct:description; owl:hasValue "lemma frequency"].
```

Shown below is an illustration of the lemmas in the Corpus of Football and the General Corpus of the Serbian Language. One of the factors for determining the terms, also known as *Termhood* is the notion of *Keyness*, which is calculated based on the frequencies in these two corpora. There are two ways of representing frequencies:

```
# frekvencije na nivou leme  (prvi način)
:le_zxut_86508 frac:frequency
    [a :SrFudKo_token_freq; rdf:value "2313" ] .

:le_zxut_86508 frac:frequency
    [a :SrpKor2021_token_freq; rdf:value "23377" ] .

# frekvencije na nivou leme  (drugi način)
:le_zxut_86508 frac:frequency
   :freq_SrFudKo_le_zxut_86508,
   :freq_SrpKor2021_le_zxut_86508.

:freq_SrFudKo_le_zxut_86508
   a :SrFudKo_token_freq; rdf:value "2313".

:freq_SrpKor2021_le_zxut_86508
   a :SrpKor2021_token_freq;
   rdf:value "23377".

# frekvencije na nivou leme za višečlanu reč
:le_zxuti_karton_216263 frac:frequency
```

```
    :freq_SrFudKo_le_zxuti_karton_216263,
    :freq_SrpKor2021_le_zxuti_karton_216263.
```

For the Spanish language, we use the example of frequencies on the *esFudKo* corpus. For the CQL (Corpus Query Language) the query [lemma = "amarillo"], we get a total of 1517 repetitions, out of which the following frequencies are shown in individual forms: amarilla (839), amarillas (244), amarillo (221), amarillos (85), Amarillas (55), Amarilla (41), AMARILLAS (21), Amarillo (10) и AMARILLA (1). Within the FudLe dictionary, frequencies are shown comprised, regardless of caps, which means that amarillas contains 244+55+21=320.

```
# frekvencije na nivou leme (prvi način) španski
:le_amarillo_2 frac:frequency
    [a :EsFudKo_token_freq; rdf:value "1517" ] .

# frekvencije na nivou leme  (drugi način)
:le_amarillo_2 frac:frequency :freq_ESFudKo_le_amarillo_2.

:freq_ErFudKo_le_amarillo_2
   a :EsFudKo_token_freq;
   rdf:value "1517".
```

Absolute frequencies are reached using CQL expressions, while relative frequencies are calculated by a million words – dividing the entire number of words in the corpus, multiplied by a million. In terms of frequencies of complex elements in Serbian and Spanish, one of the possible solutions is shown as follows (Lazarević et al. 2023):

```
SrFudKo_mwe_freq rdfs:subClassOf frac:Frequency, :SrFudKo,
   [owl:Restriction;
     owl:onProperty dct:description;
     owl:hasValue "mwe frequency"].

:EsFudKo_mwe_freq rdfs:subClassOf
  frac:Frequency, :EsFudKo,
   [owl:Restriction;
     owl:onProperty dct:description;
     owl:hasValue "mwe frequency"].
```

The example that follows will show the frequencies of the lemma: “yellow card” (Sr. žuti karton) and the same meaning in Serbian. Using the example of textitsrFudKo, a clarification is given, as well as the queries used to reach their frequencies.

Lemma frequencies are reached through the query “[lemma="žut"][lemma="karton"]”, which yields also the inflected forms: “žuti kartoni” (124), “žuta kartona” (95), “žutih kartona”, “žutog kartona” (84), “žute kartone” (62), “žutim kartonom” (51), “žuti kartoni” (15)...

```
# mwe frekvencija lema

:le_zxuti_karton_216263
  frac:frequency [a :SrFudKo_mwe_freq;  rdf:value "1996"];
  frac:frequency [a :SrpKor2021_mwe_freq;  rdf:value "3192"];
  frac:head :le_karton_8815  .
```

The Spanish Language example is given through the CQL query of the lemma "Tarjeta", meaning "Card": [lemma = "tarjeta"][lemma = "amarillo"], offering 312 hits: "tarjeta amarilla" (214), "tarjetas amarillas" (75), "Tarjetas amarillas" (12), "Tarjeta amarilla" (10), "tarjeta amarillas" (1). When giving frequencies, we count the insensitivity to upper and lowercase.

```
# mwe lemma frequency za primer na španskom

:fm_tarjeta_amarilla_10001
  frac:frequency [a :EsFudKo_mwe_freq;  rdf:value "224"];
  frac:head :le_tarjeta_1  .

:le_tarjetas_amarillas_10001
  frac:frequency [a :EsFudKo_mwe_freq;  rdf:value "87"];
  frac:head :le_tarjeta_1  .

:le_tarjeta_amarillas_10001
  frac:frequency [a :EsFudKo_mwe_freq;  rdf:value "1"];
  frac:head :le_tarjeta_1  .
```

It is possible to introduce an individual object for frequencies, used to then connect the data:

```
# frekvencije za španski
:le_tarjeta_amarilla_10001 frac:frequency :freq_le_tarjeta_amarilla_10001.
```

4 Translational equivalents in the *FudLe* dictionary

As previously mentioned, the Dictionary of Football Terminology FudLe comprises of various components, among which we show the source, Serbian section of the Lexicon, the key section in Spanish as well as the Translation Set.

This typology fits more naturally into the base scheme of *OntoLex* and the *Vartans* model. As a result, two separate lexicons in Serbian and Spanish language are published as connected data sets, alongside the translation set which connects them. The publication of the dictionary for other languages is also possible according to the same scheme.

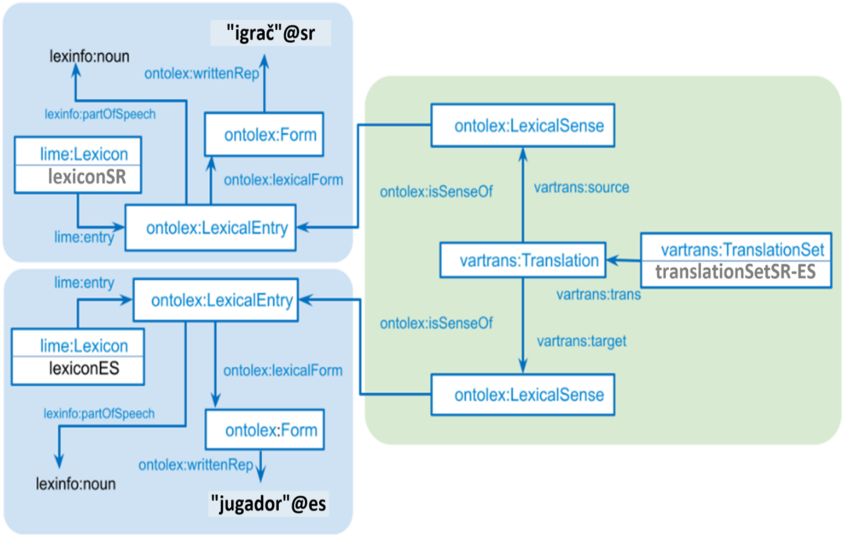


Figure 3. Translation modeling using *OntoLex* and vartrans.

Figure 3 illustrates the presentation scheme used for the translation of the lexical unit "Player" from Serbian into Spanish: "Jugador" within the context of the basic space of *OntoLex* and *Vartans*.

The lexical unit (*ontolex:LexicalEntry*) as well as its attributed properties are used to add lexical information, while *vartrans:Translation* class connects them via lexical meaning *ontolex:LexicalSense*. For example, the property of written form (*ontolex:writtenRep*) connects the subject (*ontolex:Form*) "Player" in Serbian or "Jugador" in Spanish.

There are simpler possibilities in connecting lexical units without defining their circumstantial meaning. In those cases, we use the predicate *vartrans:translatableAs*. However, translational equivalents are established between specific meanings of lexical units and the use of the class *vartrans:Translation* as a conduit between their two meanings.

The following URI identifier was chosen for publishing the dictionary: http://domain/type/concept/reference, with type declaring the type of identified resource. For example: 'id' or 'item' for palpable objects, or 'doc' for the documents describing them; 'def' is used for concepts, while 'set' is used for data sets or elements specific for the context such as: 'authority'

or 'dcterms' (Archer, Goedertier, and Loutas 2012). A similar approach was used for publishing the Serbian-German dictionary SrpNemLex (Andonovski 2023).

In the case of *FudLe*, three data sets were projected:

- The Source data set in Serbian: https://llod.jerteh.rs/id/fudle/lexiconSR
- Key Language data set in Spanish: https://llod.jerteh.rs/id/fudle/lexiconES
- Translational equivalent data set (Spanish-Serbian): https://llod.jerteh.rs/id/fudle/tranSetSR-ES

The example that follows will show the units "igrač"@sr and "jugador"@es with object names adjusted towards easier understanding of the created connections between translational equivalents on the level of word meaning.

```
# :lexiconSR a lime:Lexicon .
...
:lexiconSR lime:entry :lexiconSR/igracy-n-sr .
:lexiconSR/igracy-n-sr a ontolex:LexicalEntry ;
  ontolex:lexicalForm :lexiconSR/igracy-n-sr-form ;
  lexinfo:partOfSpeech lexinfo:noun .
:lexiconSR/igracy-n-sr-form a ontolex:Form ;
  ontolex:writtenRep "igrač"@sr .
:lexiconES a lime:Lexicon  .
...

:lexiconES lime:entry :lexiconES/jugador-n-es .
:lexiconES/jugador-n-es a ontolex:LexicalEntry ;
  ontolex:lexicalForm :lexiconES/jugador-n-es-form ;
  lexinfo:partOfSpeech lexinfo:noun .
:lexiconES/jugador-n-es-form a ontolex:Form ;
  ontolex:writtenRep "jugador"@es .
...

:tranSetSR-ES a vartrans:TranslationSet   ;
...
:tranSetSR-ES vartrans:trans
  :tranSetSR-ES/igracy_jugador-n-sr-sense-jugador_igracy-n-es-sense-trans.
:tranSetSR-ES/igracy_jugador-n-sr-sense a ontolex:LexicalSense ;
  ontolex:isSenseOf :lexiconSR/igracy-n-sr .
:tranSetSR-ES/jugador_igracy-n-es-sense a ontolex:LexicalSense ;
  ontolex:isSenseOf :lexiconES/jugador-n-es .
:tranSetSR-ES/igracy_jugador-n-sr-sense-jugador_igracy-n-es-sense-trans
  a vartrans:Translation ;
  vartrans:source :tranSetSR-ES/igracy_jugador-n-sr-sense ;
  vartrans:target :tranSetSR-ES/jugador_igracy-n-es-sense .
```

The module FrAC enables illustrating collocations and embeddings (Chiarcos et al. 2022; Chiarcos et al. 2020), and various vector representations.

5 Conclusion

The development of dictionaries through the methodology of connected data has shown significant advantages in comparison to the traditional approach. Process automatization, use of corpora and integration of various digital resources has allowed a faster compilation of dictionaries. Digital dictionaries offer various advantages, such as: high-speed search, constant updates and user-friendly adaptivity, making them still irreplaceable. The application of techniques provided by computational linguistics and semantical networks has provided the possibility of creating a dictionary which can be used by humans, as well as machines. We expect this approach to be implemented towards creating dictionaries in other fields, which only further enriches the area of lexicography and language technology.

Acknowledgment

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

References

Andonovski, Jelena. 2023. "SrpNemKor and Linked Open Data." *Infotheca - Journal for Digital Humanities* 23 (1): 33–60. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2023.23.1.2.

Archer, Ph., S. Goedertier, and N. Loutas. 2012. *Study on Persistent URIs, with Identification of Best Practices and Recommendations on the Topic for the MSs and the EC.* Technical report. European Commission.

Bergh, Gunnar, and Sölve Ohlander. 2019. "A hundred years of football English: A dictionary study on the relationship of a special language to general language."

Chiarcos, Christian, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022. "Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC." In *Proceedings of the 29th International Conference on Computational Linguistics,* 4018–4027.

Chiarcos, Christian, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck, and John Philip McCrae. 2020. "Modelling frequency and attestations for ontolex-lemon." In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography,* 1–9.

Fillmore, Charles J, Miriam RL Petruck, Josef Ruppenhofer, and Abby Wright. 2003. "FrameNet in action: The case of attaching." *International journal of lexicography* 16 (3): 297–332.

Fuertes-Olivera, Pedro A, and Sven Tarp. 2014. *Theory and practice of specialised online dictionaries: Lexicography versus terminography.* Vol. 146. Walter de Gruyter GmbH & Co KG.

Ivanović, Tanja, Ranka Stanković, Branislava Šandrih Todorović, and Cvetana Krstev. 2022. "Corpus-based bilingual terminology extraction in the power engineering domain." *Terminology* 28 (2): 228–263.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. "The Sketch Engine: ten years on." *Lexicography* 1 (1): 7–36.

Kilgarriff, Adam, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ. 2008. "GDEX: Automatically finding good dictionary examples in a corpus." In *Proceedings of the XIII EURALEX international congress,* 1:425–432. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra ...

Kitanović, Olivera, Ranka Stanković, Aleksandra Tomašević, Mihailo Škorić, Ivan Babić, and Ljiljana Kolonja. 2021. "A data driven approach for raw material terminology." *Applied Sciences* 11 (7): 2892.

Krstev, Cvetana. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries.* Faculty of Philology of the University of Belgrade.

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, Ranka Stanković, Ivan Obradović, and Biljana Lazić. 2015. "Terminology Acquisition and Description Using Lexical Resources and Local Grammars." In *Proceedings of the 11th Conference on Terminology and Artificial Intelligence.* Granada, Spain: LexiCon (Universidad de Granada).

Krstev, Cvetana, and Duško Vitas. 2009. "An Effective Methode for Developing a Comprehensive Morphological E-Dicitonary of Compounds." In *Proceedings of the 28th Conference on Lexis and Grammar,* edited by B. Lamiroy, E. Laporte, and T. Kyriakopoulou, 204–212. 29th September - 3rd October 2009, in Arena Romanistica. Bergen: University of Bergen, Department of Foreign Languages.

Lazarević, Jelena, Ranka Stanković, Mihailo Škorić, and Biljana Rujević. 2023. "Football terminology: compilation and transformation into OntoLex-Lemon resource." In *Proceedings of LDK 2023 – 4th Conference on Language, Data and Knowledge.* 12–15 September, Vienna, Austria. Lisabon: NOVA FCSH - CLUNL. https://doi.org/10.34619/srmk-injj.

McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. "The Ontolex-Lemon model: Development and applications." In *Proceedings of eLex 2017 conference,* 19–21.

Mihajlović, Aleksandar. 2003. *Fudbalski rečnik: srpsko-englesko-francusko-španski.* Beograd : A. Mihajlović, 2003 (Beograd : Solidarnost).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems* 26.

Río Gracia, Jorge del, et al. 2023. *Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries (v2.0) Draft Community Group Report.* Technical report. 03 October 2023. October. https://bpmlod.github.io/Bilingual_Dictionaries_Report/.

Robin, Cécile, Atharva Kulkarni, and Paul Buitelaar. 2023. "Identifying FrameNet Lexical Semantic Structures for Knowledge Graph Extraction from Financial Customer Interactions." In *Proceedings of the 12th Global Wordnet Conference,* edited by German Rigau, Francis Bond, and Alexandre Rademaker, 91–100. University of the Basque Country, Donostia - San Sebastian, Basque Country: Global Wordnet Association, January. https://aclanthology.org/2023.gwc-1.11/.

Savary, Agata, Cvetana Krstev, and Duško Vitas. 2007. "Inflectional non compositionality and variation of compounds in French, Polish and Serbian, and their automatic processing." In *Bulag — Bulletin de Linguistique Appliquée et Générale: Les langues slaves et le français : approches formelles dans les études contrastives,* edited by Aleksandra Dziadkiewicz and Izabella Thomas, 73–94. 32. Besançon: Presses Universitaires de Franche Comté.

Schmidt, Thomas. 2009. "The Kicktionary–A multilingual lexical resource of football language." *Multilingual FrameNets in computational lexicography. Methods and applications,* 101–134.

Škorić, Mihailo, and Nikola Janković. 2024. "New Textual Corpora for Serbian Language Modeling." Faculty of Philology, University of Belgrade, *Infotheca* 24 (1).

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic Dictionaries - from File System to lemon Based Lexical Database." In *Proceedings of the 11th International Conference on Language Resources and Evaluation – W23 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science (LDL-2018).* LREC 2018, May 7–12, 2018. Miyazaki, Japan: ELRA.

Stanković, Ranka, Ivan Obradović, Cvetana Krstev, and Duško Vitas. 2011. "Production of morphological dictionaries of multi-word units using a multipurpose tool." In *Proceedings of the Computational Linguistics-Applications Conference,* 77–84.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian." In *Proceedings of the 12th Language Resources and Evaluation Conference,* 3954–3962. Marseille, France: ELRA.

Stanković, Ranka, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. "Parallel bidirectionally pretrained taggers as feature generators." *Applied Sciences* 12 (10): 5028.

Vitas, Duško. 1979. "Prikaz jednog sistema za automatsku obradu teksta." *Zbornik radova XIV jugoslovenskog medunarodnog simpozijuma o obradi podataka Informatica* 79:7–101.

Vu, Thuy, Aiti Aw, and Min Zhang. 2008. "Term extraction through unithood and termhood unification." In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II.*

Wu, Jiguang, and Ying Li. 2017. "Research on construction of semantic dictionary in the football field." In *Proceedings of the 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA),* 303–306. IEEE Xplore Full-Text PDF available. IEEE.

Рујевић, Биљана. 2022. "Речници у дигиталном добу – информатичка подршка за српски језик." Докторска дисертација, Универзитет у Београду, Филолошки факултет.

Утвић, Милош. 2011. "Анотација Корпуса савременог српског језика." *ИНФОтека* 12, no. 2 (December): 39–51.

Creating a Synthetic Evaluation Dataset for the Serbian SentiWordNet

UDC 811.163.41'322.2
DOI 10.18485/infotheca.2024.24.1.3

ABSTRACT: This study presents the creation of a synthetic evaluation dataset for the Serbian SentiWordNet using Large Language Models (LLMs), specifically focusing on the Mistral model. Addressing the scarcity of the sentiment analysis resources for Serbian, this research aims to bridge this gap by generating a dataset to evaluate and enhance sentiment analysis tools for Serbian. Sentiment polarity values from the English SentiWordNet were automatically mapped to Serbian WordNet via the Inter-Lingual Index (ILI). To refine these values for better alignment with the Serbian language, a new evaluation dataset was created. Initially, 500 synsets from the Serbian WordNet were selected based on their alignment with the *senti-pol-sr* lexicon and with the mapped values from SentiWordNet. These synsets underwent sentiment polarity classification using the Mistral model. A balanced subset of 75 synsets was then randomly extracted. It was further refined for sentiment gradation, and manually reviewed. The findings demonstrate a high model reliability, with approximately 93% of responses meeting the established acceptability criteria.

KEYWORDS: SentiWordNet, synthetic dataset, Large Language Models, Serbian, sentiment analysis

PAPER SUBMITTED: 23 May 2024
PAPER ACCEPTED: 24 June 2024

Saša Petalinkar
sasa5linkar@gmail.com
ORCID: 0009-0007-9664-3594
University of Belgrade
Multidisciplinary PhD Studies
Intelligent Systems
Belgrade, Serbia

1 Introduction

Sentiment Analysis is the process of computationally determining the emotional tone behind the text to understand the attitudes, opinions, and emotions expressed by it. One of the two main methods for Sentiment Analysis is based on sentiment lexicons. Sentiment lexicons are specialized dictionaries that associate words and phrases with sentiment values, facilitating the automated analysis of emotions in text (Liu 2010).

One problem noted with sentiment lexicons is that some words can have multiple meanings, and different senses of the word can convey different sentiments. The solution to this problem is to assign sentiment value by sense rather than by word. A prominent example of such a lexicon is SentiWordNet (SWN), which extends the Princeton WordNet (PWN) dictionary by assigning sentiment scores to each synset (a set of cognitive synonyms) to reflect the collective emotional tone of literals comprising synsets. This process involves selecting a small number of synsets with high polarity values and then expanding this set by adding synsets through a semi-automated method that identifies relationships within WordNet that either preserve or reverse polarity. The resulting expanded set is then used to train classifiers based on the definitions of these synsets (Baccianella, Esuli, and Sebastiani 2010).

Wordnets for many languages are mutually interconnected through the Inter-Lingual Index (ILI) — a synset in a wordnet of a particular language is assigned a value from ILI representing an abstract concept, which is assigned to synsets in other languages that lexicalize the same concept. Both EuroWordNet (Vossen 2004) and BalkaNet (Tufis, Cristea, and Stamou 2004) connect to PWN in this way, and thus to SWN. EuroWordNet includes wordnets for eight languages, while BalkanNet adds five more (Bulgarian, Greek, Romanian, Serbian, and Turkish) (Krstev et al. 2004; Krstev, Koeva, and Vitas 2006; Stanković et al. 2018). It is trivial to map SWN to any wordnet that has ILI, which is the case for all wordnets belonging to the Open Multilingual Wordnet (OMW) project (Bond and Paik 2012), which currently spans 200 languages.

Research has demonstrated that the sentiment values from the SWN are applicable to non-English languages (Denecke 2008). Serbian WordNet (SrpWN) contains sentiment values obtained from SWN by direct mapping of synsets using ILI (Mladenović, Mitrović, and Krstev 2014). It has already been used in the creation of a hybrid framework for sentiment analysis in Serbian (Mladenović et al. 2015).

Our hypothesis is that the lexicon can be improved by replacing mapped sentiment values with values that better represent the Serbian language. Similar improvements have been made for Turkish (Dehkharghani et al. 2016), Arabic (Alhazmi and Black 2013), Vietnamese (Vu and Park 2014), Odia (Mohanty, Kannan, and Mamidi 2017), Indonesian (Wijayanti and Arisal 2021), and Hindi (Bakliwal, Arora, and Varma 2012). However, an evaluation dataset, a subset of synsets from SrpWN already annotated with sentiment polarity, is needed to evaluate these improvements for Serbian.

For SWN, an evaluation dataset exists: Micro-WNO(Cerini et al. 2007) is a manually labeled subset of synsets from PWN, available online.[1] Creating a comparable evaluation dataset for the Serbian language manually would necessitate a significant effort, involving either a small number of expert annotators or a larger group of less skilled annotators. Having in mind that such experts are not at our disposal, an alternative approach becomes imperative.

Synthetic evaluation datasets are artificially created collections of data designed to test and validate computational models, particularly in domains where real-world data may be scarce, biased, or too sensitive to use. These datasets are generated through algorithms or simulations that aim to mimic the statistical properties of real data, allowing researchers to conduct robust evaluations under controlled conditions (Lu et al. 2024).

The emergence of Large Language Models (LLMs) has enabled the creation of much better synthetic datasets. It has been shown that for purposes of nature language processing (NLP) tasks, including sentiment analysis, LLMs can successfully annotate with only a few examples (Amatriain 2024; Brown et al. 2020).

In a zero-shot learning scenario model is making predictions or annotations without having seen any explicit examples of the task during training. This capability is particularly useful for sentiment analysis in languages or contexts where annotated data are scarce, as it allows the LLM to apply its pre-existing knowledge to new, unseen tasks (Amatriain 2024; Brown et al. 2020).

Few-shot learning, on the other hand, provides the model with a small number of examples from which it can learn to perform a task. This approach is especially advantageous for refining the model's understanding and increasing its accuracy in specific applications, such as distinguishing nuanced sentiment expressions (Brown et al. 2020).

1. https://github.com/aesuli/Sentiwordnet/blob/master/data/Micro-WNop-WN3.txt

The utilization of Prompts and Responses with LLMs enables these learning paradigms to be applied effectively. By crafting prompts that guide the model towards the desired output, researchers can leverage the LLM's capabilities for sentiment analysis. This involves providing a prompt that clearly states the task, such as identifying the sentiment of a given text, and then allowing the model to generate a response based on its training and the context provided by the prompt. Such an approach helped harness the power of LLMs for detailed sentiment analysis, offering a flexible and efficient method for analyzing sentiment across diverse datasets (Amatriain 2024).

This raises the question of whether LLMs could be employed not just to create an evaluation dataset, but to annotate the entire SrpWN with sentiment polarity values. The decision to focus on creating a small evaluation dataset stems from the prohibitive computational expense associated with annotating the entire network.

The proposed solution entails the creation of a synthetic dataset comprising synsets annotated by an LLM. The primary motivation behind this approach is to enable modification of the existing sentiment values in SrpWN obtained by mapping from SWN, particularly targeting synsets containing words that would be identified as having a specific polarity in a sentiment lexicon derived from a Serbian corpus but were mapped as purely objective in sentiment.

To achieve a balanced sample suitable for effective evaluation, especially in the later application of machine learning models for sentiment classification, a set of 500 synsets was randomly selected and processed using the *Mistral* model. This initial processing aimed to categorize the synsets into positive, negative, and objective sentiment groups. Subsequently, an equal number of synsets from each sentiment category were chosen for finer gradation.

The results underwent a manual annotation process, where each synset, along with the values returned by the LLM, was assessed and assigned a simple 'pass' or 'fail' grade based on their alignment with the expected sentiment annotations.

Originally, the methodology was designed to incorporate a few-shot learning approach for fine sentiment gradation, utilizing examples from synsets not selected for the primary dataset—specifically, those synsets that exhibited self-evident sentiment alignment across both Serbian and English SWN. However, the preliminary results obtained through the zero-shot approach were found to be sufficiently satisfactory, rendering the few-shot component

unnecessary. Consequently, the study proceeded exclusively with the zero-shot learning paradigm, where the *Mistral* model was applied without prior examples specific to the task of sentiment analysis.

Manual annotation of the outputs confirmed the validity of this streamlined approach. The zero-shot methodology demonstrated a remarkable success rate of over ninety percent, affirmatively showing that even without the inclusion of few-shot learning and the additional context it provides, the LLM could effectively discern and classify sentiment within the selected Serbian synsets.

The main LLM used in this research is *Mistral 7B — Instruct*[2], a fine-tuned variant of the 7-billion-parameter *Mistral 7B model* designed for superior performance and efficiency. *Mistral 7B* outperforms the *Llama 2 13B model* on various benchmarks and even surpasses *Llama 1 34B* in reasoning, mathematics, and code generation (Jiang et al. 2023). The *Mistral 7B — Instruct* model used in this research also outperforms the *Llama 2 13B — Chat* model on both human and automated benchmarks (Jiang et al. 2023). Released under the Apache 2.0 license, the model offers a flexible tool for researchers, with the ability to run locally without cost (Jiang et al. 2023).

This paper is organized into several key sections to comprehensively discuss the research and findings. In Section 2, the methodology is discussed in detail, including the selection of synset definitions for processing, the specific prompts used, and the application of these prompts. Additionally, there is a brief mention of attempts to use other Large Language Models (LLMs) for comparison. Section 3 presents the results, describing the accuracy of the model with examples of synsets evaluated as both correct and incorrect, as well as outputs that were not in the proper format. Finally, Section 4 concludes the paper with a summary of the findings and a discussion of potential future work that could extend this research.

2 Methodology

The *senti-pol-sr* is a polarity lexicon for the Serbian language, annotated at the literal level rather than at the sense level (Stanković et al. 2022). It includes literals that exhibit clear polarity, categorized as either positive or negative, and do not contain words considered to be objective.

In our research, the lexicon was employed to select a sample suitable for annotation by LLM. This was achieved by identifying all synsets from the

2. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

SrpWN containing literals (words) present in the *senti-pol-sr* lexicon but having a neutral sentiment value (0,0) as mapped from SWN.

Four primary reasons were detected that produced discrepancies between the sentiment values in Serbian expressed in the *senti-pol-sr* lexicon and those derived from SWN. Firstly, while a word may convey a polarizing sentiment, the actual sense it is used in may not. For example, synset `ENG30-06828389-n` (def. "a star-shaped character * used in printing") which in SrpWN contains the word "zvezda" (Eng. "star") in other synsets represents different meanings that can have positive connotations. Secondly, the sentiment values in SWN, generated through machine learning methods, may be incorrect. For example, synset `ENG30-02585489-v` (def. "cause to suffer") which is obviously very negative was assigned a neutral sentiment value. Thirdly, some synsets from SrpWN do not have corresponding synsets in PWN, such as `BILI-00000941` (referring to the specific kind of mourning) and were assigned polarity (0,0) by default.

The last and most interesting is the possibility that while a sense is considered objective in English while it carries a sentiment in Serbian. No such synsets have been found yet.

The initial analysis identified 2,956 from 25,320 synsets within the SrpWN (Mladenović, Stanković, and Krstev 2017) that contained literals annotated with clear polarity in the senti-pol-sr lexicon while having polarity value (0,0) in the SrpWN; 1,511 of them exhibited positive sentiment and 1,445 negative. Given the substantial volume, processing all these synsets with LLM was deemed impractical. Consequently, a random sample of 500 synsets was selected for further research.

For the sake of creating a balanced set of samples, definitions or glosses of synsets from this random sample were processed using the *LangChain Python library*, a powerful tool for creating, experimenting with, and analysing language models and agents (Chase 2022). The *LangChain Python* library acts as an interface for multiple large language model modules. This project uses wrappers designed for the Hugging Face Hub and the Transformer libraries. The chain used in this research was simple, containing just one prompt template with one input variable – the definition of the synset, and the *Mistral 7B – Instruct* model downloaded from Hugging Face Hub. The model was executed on the local machine. The *LangChain* language chain allows prompt templates with variables, which are marked by curly brackets, to be invoked with values of those variables, sent to the LLM module, which is returning the response.

Using the appropriate prompt as shown in Figure 1, the sample was divided into those marked as positive, negative, objective, and those not properly marked. There were 290 objective, 102 negative, 33 positive, and 75 erroneously marked synsets. To harness the full capabilities of the LLM for sentiment analysis, the prompt was meticulously designed to encapsulate three essential elements: role-playing, clear instructions, and expected outcomes.

Kao ekspert za analizu sentimenta, analizirajte sledeći tekst na srpskom jeziku i odredite njegov sentiment.

Sentiment treba da bude striktno klasifikovan kao "pozitivan", "negativan", ili "objektivan". Nijedan drugi odgovor neće biti prihvaćen.

Tekst: {text}

Sentiment:

Figure 1. Prompt template for determining polarity.[3]

1. Role Playing – Instructing LLM as an Expert: The prompt initiates with a role-playing scenario, instructing the LLM to assume the role of an expert in sentiment analysis (Figure 1). This approach was adopted to prime the model's response generation towards a more analytical and focused examination of the text, drawing on its extensive pre-trained knowledge and understanding of sentiment analysis nuances.
2. Clear Instructions: The essence of effective communication with an LLM lies in the clarity of instructions. The prompt explicitly details the task at hand, guiding the LLM to identify and analyze the sentiment expressed

3. Translation to English of the prompt for determining polarity:
"As an expert in sentiment analysis, analyze the following text in Serbian and determine its sentiment.
The sentiment should be strictly classified as 'positive', 'negative', or 'objective'. No other answers will be accepted.
Text: {text}
Sentiment: "

in a given piece of a text (synset definition). This clarity ensures that the model's analytical capabilities are directed towards accurately assessing sentiment, minimizing ambiguity in its responses.

3. Expected Outcomes: To further refine the model's output, the prompt delineates the expected outcomes of the analysis. It specifies the desired format of the response, whether it be a sentiment classification (positive, negative, neutral) or a more nuanced sentiment rating.

Kao ekspert za analizu sentimenta, analizirajte sledeći tekst na srpskom jeziku i odredite da li ima pozitivan sentiment.

Sentiment treba da bude striktno klasifikovan kao "nije pozitivan", "slabo pozitivan", "umereno pozitivan", "veoma pozitivan", ili "ekstremno pozitivan". Nijedan drugi odgovor neće biti prihvaćen.

Nijedan drugi odgovor neće biti prihvaćen.

Tekst: {text}

Pozitivan sentiment:

Figure 2. Prompt template for fine-grained classification of positive sentiment.[4]

The prompt used for classification was refined through experimental testing on a smaller subset of synset definitions. Comparative analysis of prompt instruction texts in English and Serbian revealed that the Serbian language prompt instruction yielded more accurate results. To ensure comprehensive

4. Translation to English of the prompt for fine-grained classification of positive sentiment:
"As an expert in sentiment analysis, analyze the following text in Serbian and determine if it has a positive sentiment.
The sentiment should be strictly classified as "not positive", "slightly positive", "moderately positive", "very positive", or "extremely positive". No other answers will be accepted.
No other answers will be accepted.
Text: {text}
Positive sentiment: "

coverage of potential responses, the number of tokens in the output was set to five.

Kao ekspert za analizu sentimenta, analizirajte sledeći tekst na srpskom jeziku i odredite da li ima negativan sentiment.
Sentiment treba da bude striktno klasifikovan kao "nije negativan", "slabo negativan", "umereno negativan", "veoma negativan", ili "ekstremno negativan". Nijedan drugi odgovor neće biti prihvaćen.
Nijedan drugi odgovor neće biti prihvaćen.
Tekst: {text}
Negativan sentiment:

Figure 3. Prompt template for fine-grained classification of negative sentiment[5]

Further analysis revealed that duplicates exist within the set, due to some synsets containing multiple literals from *senti-pol-sr*, such as `ENG30-01375831-a` 'celebrated, famed, far-famed, famous, illustrious, notable, noted, renowned: widely known and esteemed' ('slavan, poznat, čuven, značajan, renomiran, priznat, proslavljen: opšte prihvaćen i uvažen'). Notably, all but two literals ('renomiran', 'priznat') from this synset appear in the lexicon. After removing duplicates, there were 279 objective, 97 negative, and 27 positive synsets. At this point, there was no reason to count

5. Translation to English of the prompt for fine-grained classification of negative sentiment:
"As an expert in sentiment analysis, analyze the following text in Serbian and determine if it has a negative sentiment.
The sentiment should be strictly classified as "not negative", "slightly negative", "moderately negative", "very negative", or "extremely negative". No other answers will be accepted.
No other answers will be accepted.
Text: {text}
Negative sentiment: "

improperly marked synsets since removing duplicates from them serves no practical purpose.

To facilitate more detailed sentiment analysis, a random subset comprising 25 synsets from each sentiment category was selected, forming what is referred to as the "balanced sample."

The balanced sample was processed through two prompt templates for greater sensitivity, one for positive and one for negative sentiment (Figures 2 and 3). These prompt templates were designed experimentally, through iterative testing of various options and making incremental adjustments to the text, using small sets of synset definitions from SrpWN. For example, repeating the sentence "No other answers will be accepted" ("Nijedan drugi odgovor neće biti prihvaćen") twice significantly reduced the number of out-of-bounds responses.

A singular prompt for determining both values was not chosen to maintain consistency with the structure of SWN, where a synset can have both positive (POS) and negative (NEG) values above zero, as long as their sum is less than or equal to one.

During this phase, other LMM models were tested on a small sample – *GPT2-ORAO* (Škorić 2024), *Llama-2-7b* (Touvron et al. 2023), *alpaca-serbian-7b-base* and *WizardLM-1.0-Uncensored-Llama2-13b*. The idea was to compare differing results, simulating annotation by multiple annotators. However, the results led to their exclusion from further work due to too many improperly marked synsets and a lack of finer grading. To accommodate the anticipated length of the output strings, the number of output tokens was increased to nine.

For each definition in the balanced set, two values were assigned in that way. The response options were:

- For the positive sentiment:
  - "nije pozitivan" (not positive)
  - "slabo pozitivan" (slightly positive)
  - "umereno pozitivan" (moderately positive)
  - "veoma pozitivan" (very positive)
  - "ekstremno pozitivan" (extremely positive)
- For the negative sentiment:
  - "nije negativan" (not negative)
  - "slabo negativan" (slightly negative)
  - "umereno negativan" (moderately negative)
  - "veoma negativan" (very negative)
  - "ekstremno negativan" (extremely negative)

The resulting dataset is stored in the form of a comma-separated values (CSV) file.[6] Columns in that file are as follows:

ILI: Inter-Lingual Index, which connects the synset to its equivalents in other languages' wordnets.
Definition: The gloss of the synset, providing its meaning or explanation.
Lemma_names: The lemmas (base forms) of the words contained within the synset.
Sentiment_SWN: The sentiment value mapped from SWN, indicating the original sentiment score in the English version.
Sentiment_lexicon: The sentiment value derived from a sentiment lexicon *senti-pol-sr* created for the Serbian language, reflecting local sentiment nuances.
Sentiment_sa: The initial classification by the LLM, categorizing the sentiment as positive, negative, or neutral.
Sentiment_sa_positive: Fine-grained sentiment classification for positive sentiment, indicating the degree of positivity ranging from "not positive" to "extremely positive."
Sentiment_sa_negative: Fine-grained sentiment classification for negative sentiment, indicating the degree of negativity ranging from "not negative" to "extremely negative".

Given the small sample size of 75, the author could conduct a manual evaluation on the entire dataset. This process was made more efficient by using the *Data Wrangler*[7] extension in *Visual Studio Code*, which is a data visualization and cleaning utility that meshes well with *VS Code* and *VS Code Jupyter Notebooks*. It offers an interactive interface for data review and analysis, delivers informative statistics and visual presentations, and can streamline data cleaning by automatically producing Pandas script for data modification. The tool was employed for a meticulous examination comparing synset definitions with their respective detailed sentiment feedback.

3 Results

The output summary generated by the LLM using the fine-grained prompt templates is presented in Tables 1 and 2.

6. https://github.com/sasa5linkar/SWN-synth-eval-set/blob/main/balanced_sample2.csv
7. microsoft/vscode-data-wrangler (github.com)

Row Labels	**Count of sentiment_sa_negative**
ekstremno negativan	1
Nije negativan	60
Umereno negativan	1
umjereno negativan	2
veoma negativan	11
Grand Total	75

Table 1. Negative sentiment.

Row Labels	**Count of sentiment_sa_positive**
Nije (Ova iz)	1
Nije pozitivan	39
Nijedan (O)	6
Nijedan Tekst	1
Učimo se držati	1
Umereno pozitivan	3
Umereno pozitivan The	2
Umjereno pozitivan	6
veoma pozitivan	12
Veoma pozitivan (O)	1
Veoma pozitivan (R)	2
Veoma pozitivan Ov	1
Grand Total	75

Table 2. Positive sentiment.

As shown in Tables 1 and 2, the output of the LLM was not constrained to values offered in the prompt template, yet it remained within easily correctable limits. The output generated from the negative template prompt closely matched the suggested output, with only minor variations (different pronunciation). In contrast, the output from the positive template prompt included one nonsensical response, "Učimo se držati," (Engl. We learn to hold on,) and some responses that could be loosely interpreted as not positive. The subsequent manual examination, performed by one person, classified all of these responses as non-positive. Additionally, minor typos, such as extra letters, incorrect articles, or casing errors, were disregarded during manual

examinations. The manual examination revealed that the output did not classify any synset either as slightly positive or as slightly negative.

For content, there were two synsets clearly incorrectly marked, and three that were questionable.

Cleary incorrect are:

- `BILI-00000941`, synset with the definition in Serbian "Glasno izražavati žalost, zapevati, tužiti." referring to the specific kind of mourning, characteristic of Serbia (and Balkans), involving loud wailing and sombre singing. It was marked as purely objective (neither positive nor negative), white it obviously conveys a negative sentiment, moreover, a strongly negative. It does not have a corresponding English synset
- `ENG30-04525038-n` 'velvet: a silky densely piled fabric with a plain back' ('baršun, somot, pliš, kadifa: Svilenkasta, gusta, čupava tkanina, ravna sa poleđina.'), which was marked as mildly positive. As a type of textile, it should be objective.

Synsets that are considered questionable are:

- `ENG30-01215137-v` 'collar, nail, apprehend, arrest, pick up, nab, cop: take into custody' ('uhapsiti, zatvoriti, skembati: Staviti u pritvor, kao osumnjičene kriminalce; o policiji.') was marked as mildly negative while the evaluator considered it as strongly negative.
- `ENG30-00309647-n` 'expedition: a journey organized for a particular purpose' ('ekspedicija: Putovanje organizovano sa specijalnim ciljem.') was marked as very positive while the evaluator considered it neutral.
- `ENG30-03135152-n` 'Cross: a representation of the structure on which Jesus was crucified; used as an emblem of Christianity or in heraldry' ('krst: krst kao znamenje hrišćanstva') was marked very positive while the evaluator considered it neutral.

From the majority of synsets accurately labelled throughout our sentiment analysis process to we present here a selection of illustrative examples. These examples should also demonstrate the criteria used in evaluation when considering what is the correct labelling in the classifications to positive, negative, and neutral sentiments. Some examples are:

- `ENG30-01220336-n:` Synset associated with actions like 'kleveta' (defamation), 'klevetanje' (slander), and 'omalovažavanje' (disparagement), described as "oštar napad na čiju ličnost ili dobro ime" (a harsh attack on someone's personality or good name). This synset's sentiment was graded

as "Nije pozitivan" (Not positive), reflecting the absence of positive sentiment, and more precisely, "Ekstremno negativan" (Extremely negative), accurately capturing the negative connotations of the described actions.

- `ENG30-06828389-n`: This synset refers to "karakter nalik na zvezdu (*) koji se koristi u štampanim tekstovima" (a character resembling a star (*) used in printed texts), with lemmas including 'asterisk', 'zvezdica' (little star), and 'zvezda' (star). The sentiment for this synset was correctly classified as "Nije pozitivan" (Not positive) and "Nije negativan" (Not negative), indicating its objective nature without inherent positive or negative sentiment.
- `ENG30-10407310-n`: Pertains to "onaj koji voli i brani svoju zemlju" (one who loves and defends her/his country), with lemmas including 'domoljub', 'patriota', and 'rodoljub' (patriot). This synset's sentiment was finely graded as "Veoma pozitivan" (Very positive) and "Nije negativan" (Not negative), effectively capturing the positive connotations associated with patriotism.

4 Conclusion

The analysis of the output generated by the LLM, specifically the Mistral model, indicates that 70 out of 75 responses met the acceptability criteria defined for this study. This outcome, representing a substantial majority of the dataset, underscores the Mistral model's reliability and efficacy in performing sentiment classification for the Serbian language. With an approximate success rate of 93.3%, it can be confidently concluded that the dataset is both workable and of sufficient quality for the intended task of evaluating proposed corrections of sentiment polarity scores within the SrpWN.

The failure of other models tested could be due to prompts being optimized for the Mistral model. Devising prompts for each model individually may allow the use of multiple models for better quality.

A significant area for future investigation involves the augmentation of prompt templates with specific examples, transitioning the current zero-shot learning approach to a few-shot learning paradigm. This modification is anticipated to refine the model's capacity for nuanced sentiment gradation, offering a more precise and contextually aware analysis. It is particularly recommended that this enhancement be applied selectively to the fine sentiment gradation templates, where the potential for increased accuracy and sensitivity to linguistic subtleties is most pronounced (Brown et al. 2020).

Further, the exploration of advanced prompting methodologies, such as the "chain-of-thought" approach, warrants attention. This technique, by facilitating a more structured and logical progression of thought within the model's processing, may significantly improve the model's ability to extract relevant data from responses. The potential of such advanced prompting strategies to overcome the inherent challenges in accurately identifying and classifying sentiment expressions within complex linguistic contexts presents a promising avenue for research (Amatriain 2024).

This study serves as a "proof-of-concept" for utilizing the *Mistral* model to generate additional synthetic datasets for Serbian, particularly in areas where resources are scarce or challenging to acquire from natural, not synthesized, corpora. The successful application of the *Mistral* model for sentiment analysis underscores its potential as a valuable tool in addressing these challenges.

References

Alhazmi, Samah, and William Black. 2013. "Arabic SentiWordNet in Relation to SentiWordNet 3.0." *John McNaught International Journal of Computational Linguistics (IJCL),* no. 4, 1.

Amatriain, Xavier. 2024. *Prompt Design and Engineering: Introduction and Advanced Methods.* ArXiv:2401.14423 [cs], January. Accessed February 7, 2024. https://doi.org/10.48550/arXiv.2401.14423. http://arxiv.org/abs/2401.14423.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. "SENTIWORDNET 3.0: An Enhanced Lexical Resource." In *7th LREC,* 2200–2204. http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf.

Bakliwal, Akshat, Piyush Arora, and Vasudeva Varma. 2012. "Hindi Subjective Lexicon: A Lexical Resource for Hindi Adjective Polarity Classification." In *International Conference on Language Resources and Evaluation.* https://api.semanticscholar.org/CorpusID:1657748.

Bond, Francis, and Kyonghee Paik. 2012. "A Survey of Wordnets and Their Licenses." In *Proceedings of the 6th Global WordNet Conference (GWC 2012),* 64–71. Matsue, Japan.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. *Language Models are Few-Shot Learners.* ArXiv:2005.14165 [cs], July. Accessed February 8, 2024. https://doi.org/10.48550/arXiv.2005.14165. http://arxiv.org/abs/2005.14165.

Cerini, S., V. Compagnoni, A. Demontis, Maicol Formentelli, and C. Gandini. 2007. "Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining." *Language Resources and Linguistic Theory* (January): 200–210.

Chase, Harrison. 2022. *LangChain,* October. https://github.com/langchain-ai/langchain.

Dehkharghani, Rahim, Y. Saygin, B. Yanikoglu, and Kemal Oflazer. 2016. "SentiTurkNet: a Turkish polarity lexicon for sentiment analysis." *Language Resources and Evaluation* 50:667–685. https://doi.org/10.1007/s10579-015-9307-6.

Denecke, Kerstin. 2008. "Using SentiWordNet for multilingual sentiment analysis." In *International Conference on Data Engineering,* 507–512. https://doi.org/https://doi.org/10.1109/ICDEW.2008.4498370.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, et al. 2023. *Mistral 7B.* ArXiv:2310.06825 [cs], October. Accessed August 1, 2024. https://doi.org/10.48550/arXiv.2310.06825. http://arxiv.org/abs/2310.06825.

Krstev, Cvetana, Svetla Koeva, and Duško Vitas. 2006. "Towards the Global Wordnet." In *Conference Abstracts of the First International Conference of Digital Humanities Organisations (ADHO) Digital Humanities 2006,* 114–117. Paris-Sorbonne, July. http://poincare.matf.bg.ac.rs/~cvetana/biblio/DigHum06-KKV.pdf.

Krstev, Cvetana, Gordana Pavlović-Lažetić, Duško Vitas, and Ivan Obradović. 2004. "Using Textual and Lexical Resources in Developing Serbian Wordnet." Publisher: Romanian Academy, Publishing House of the Romanian Academy, *Romanian Journal of Information Science and Technology* 7 (1-2): 147–161. http://poincare.matf.bg.ac.rs/~cvetana/biblio/RJIST-MATF.pdf.

Liu, Bing. 2010. "Sentiment Analysis and Subjectivity." In *Handbook of Natural Language Processing, Second Edition,* edited by Fred J. Damerau and Nitin Indurkhya, 629–661. Chapman & Hall/CRC.

Lu, Yingzhou, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. 2024. *Machine Learning for Synthetic Data Generation: A Review.* ArXiv:2302.04062 [cs] version: 6, June. Accessed August 1, 2024. https://doi.org/10.48550/arXiv.2302.04062. http://arxiv.org/abs/2302.04062.

Mladenović, Miljana, Jelena Mitrović, and Cvetana Krstev. 2014. "Developing and Maintaining a WordNet: Procedures and Tools." In *Proceedings of the Seventh Global WordNet Conference 2014,* edited by Heili Orav, Christiane Fellbaum, and Piek Vossen, 55–62. Tartu, Estonia: University of Tartu, January. ISBN: 978-9949-32-492-7.

Mladenović, Miljana, Jelena Mitrović, Cvetana Krstev, and Duško Vitas. 2015. "Hybrid sentiment analysis framework for a morphologically rich language." *Journal of Intelligent Information Systems* 46 (3): 599–620. Accessed January 1, 2023. https://doi.org/10.1007/s10844-015-0372-5.

Mladenović, Miljana, Ranka Stanković, and Cvetana Krstev. 2017. "A WordNet Ontology in Improving Searches of Digital Dialect Dictionary." In *New Trends in Databases and Information Systems: ADBIS 2017 Short Papers and Workshops - SW4CH (Semantic Web for Cultural Heritage),* 767:373–383. Lecture Notes in Computer Science. Nicosia, Cyprus: Springer International Publishing, September.

Mohanty, Gaurav, Abishek Kannan, and Radhika Mamidi. 2017. "Building a SentiWordNet for Odia." In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis,* edited by Alexandra Balahur, Saif M. Mohammad, and Erik van der Goot, 143–148. Copenhagen, Denmark: Association for Computational Linguistics, September. https://doi.org/10.18653/v1/W17-5219. https://aclanthology.org/W17-5219/.

Škorić, Mihailo. 2024. "New Language Models for Serbian." Publisher: Zajednica biblioteka univerziteta u Srbiji, Beograd, *Infotheca - Journal for Digital Humanities* 24 (1). https://doi.org/https://doi.org/10.48550/arXiv.2402.14379. https://arxiv.org/abs/2402.14379.

Stanković, Ranka, Miloš Košprdić, Milica Ikonić Nešić, and Tijana Radović. 2022. "Sentiment Analysis of Serbian Old Novels." In *2nd Workshop on Sentiment Analysis and Linguistic Linked Data (SALLD)*, 31–38. Marseille: ELRA, June. Accessed August 21, 2023. https://aclanthology.org/2022.salld-1.6.

Stanković, Ranka, Miljana Mladenović, Ivan Obradović, Marko Vitas, and Cvetana Krstev. 2018. "Resource-based WordNet Augmentation and Enrichment." In *Proceedings of the Third International Conference on Computational Linguistics in Bulgaria (CLIB 2018)*, 104–114. Sofia, Bulgaria: Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences, May. https://aclanthology.org/2018.clib-1.14/.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models.* ArXiv:2307.09288 [cs], July. Accessed August 1, 2024. https://doi.org/10.48550/arXiv.2307.09288. http://arxiv.org/abs/2307.09288.

Tufis, D., D. Cristea, and S. Stamou. 2004. "BalkaNet: Aims, Methods, Results and Perspectives. A General Overview." Publisher: Institute for Artificial Intelligence, Romanian Academy, Bucharest, Romania, *Romanian Journal of Information Science and Technology* 7 (1-2): 9–43.

Vossen, Piek. 2004. "EuroWordNet: A Multilingual Database of Autonomous and Language-Specific WordNets Connected via an Inter-Lingual Index." _Eprint: https://academic.oup.com/ijl/article-pdf/17/2/161/2912694/ech041.pdf, *International Journal of Lexicography* 17, no. 2 (June): 161–173. ISSN: 0950-3846. https://doi.org/10.1093/ijl/17.2.161. https://doi.org/10.1093/ijl/17.2.161.

Vu, Xuan-Son, and Seong-Bae Park. 2014. "Construction of Vietnamese SentiWordNet by using Vietnamese Dictionary." *ArXiv* abs/1412.8010.

Wijayanti, Rini, and Andria Arisal. 2021. "Automatic Indonesian Sentiment Lexicon Curation with Sentiment Valence Tuning for Social Media Sentiment Analysis." *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 20:1–16. https://doi.org/10.1145/3425632.

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.

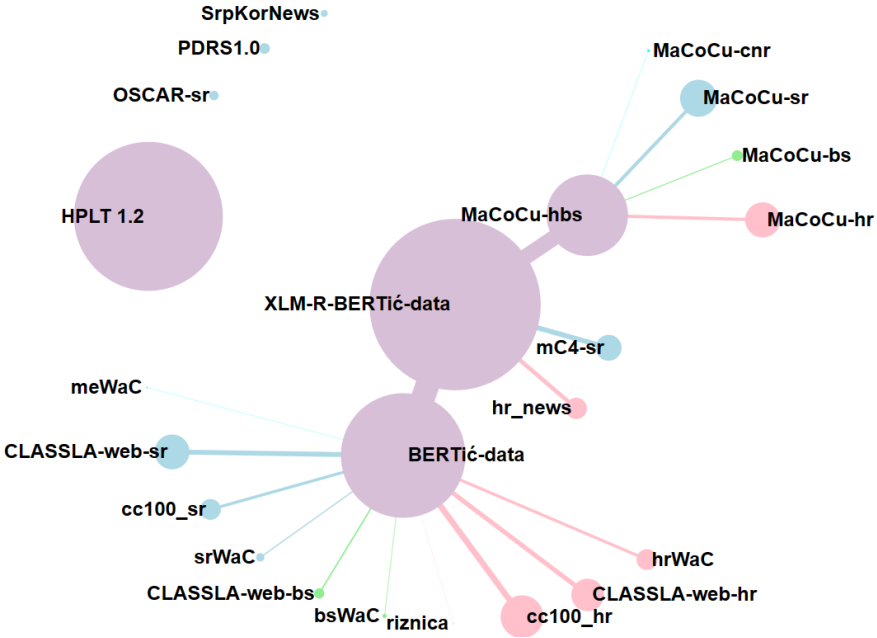


Figure 1. Aggregation hierarchy of publicly available Serbo-Croatian plain-text corpora. Blue color represents Serbian corpora, cyan Montenegrin, pink Croatian, lime Bosnian, while corpora of mixed language origin are represented in purple.

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:

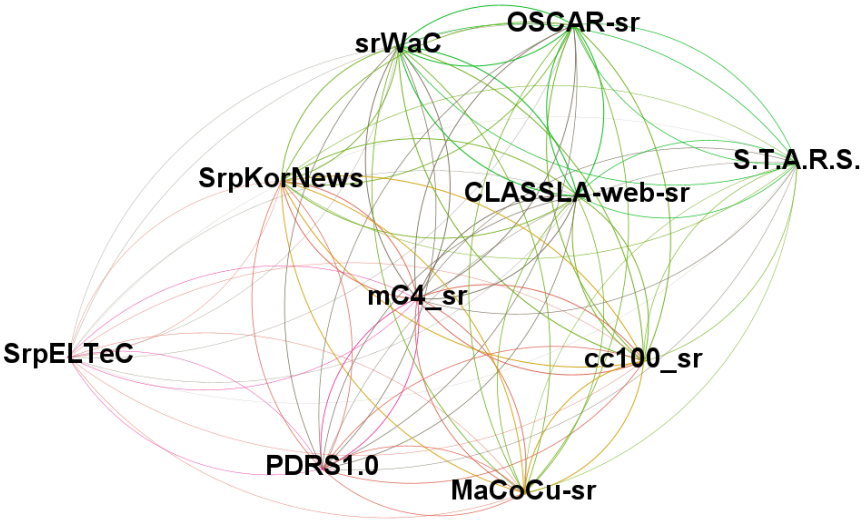


Figure 2. Corpus uniqueness visualized through a two-dimensional network graph representation of the calculated corpus-similarity matrix, where the edges represent distances between each corpus and the colors represent the clustering spectrum.

- 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.*

The *Galaksija* Computer — The Road to Stars Paved with Words

UDC 050:004GALAKSIJA"1983/1984"
DOI 10.18485/infotheca.2024.24.1.5

ABSTRACT: The research is conceived as a content analysis of the popular science journal "Galaksija" (issues published during 1983 and early 1984). The aim is to determine, at the content level, how computers, robots, and the concept of artificial intelligence (AI) are portrayed in the materials published by this journal. Additionally, the study explores to what extent such writings contributed to the success of the *Galaksija* computer, the first Yugoslav computer intended for mass use. Through the analysis, it becomes clear how an active approach to problem-solving and the encouragement of creative work and thinking can be promoted, especially in moments when the state imposes objective constraints.
KEYWORDS: galaksija, computer, artificial intelligence, content analysis.

PAPER SUBMITTED: 27 January 2024
PAPER ACCEPTED: 15 March 2024

Olga Mirković Maksimović
olga.mirkovic@gmail.com
ORCID: 0000-0003-4255-9200
University of Belgrade
Multidisciplinary PhD Studies
History and Philosophy of
Natural Sciences and Technology
Belgrade, Serbia

1 The *Galaksija* computer – what's all the fuss about?

Galaksija was the first microcomputer in the territory of SFR Yugoslavia accessible to all the enthusiasts in the period when computers were intensively developed and brought closer to end users, both in the world and in the areas of the former SFRY. It is important to understand that the import of computers into the former Yugoslavia was not allowed in 1983, thus computer and technology enthusiasts couldn't import the object of their admiration legally. To protect domestic production from foreign competition, strict import rules were introduced in the SFRY, which made the development of computers in the country different from the development in the Western world (restrictions did not apply only to the import of computers). The government tried to control the influx of foreign home computers by

imposing import restrictions on price and memory size. As the passion was stronger than the restrictions, the people were resourceful – they continued to import computers piece by piece and assemble them without violating the applicable rules and regulations. They also got by with the help of friends and relatives who worked abroad and could legally obtain all the parts. The pressures of the state only increased the creativity and inventiveness of computer enthusiasts and encouraged them to self-organize in the desire to finally own a computer suitable for home use. The best illustration of the popularization of microcomputers can be found in the 129th issue of "Galaksija" in the text "Microcomputer close up", which says that: "Millions of people of various professions, students at all levels, poets and housewives, workers and officials, medical doctors and lawyers have already come in contact with a computer".

Several companies tried to make microcomputers similar to home computers from the 1980s, such as "Ivo Lola Ribar" Institute – *Lola 8*, "EI Niš" – *Pecom* 32 and 64, "PEL Varaždin" – *Galeb* and *Orao*, "Ivel" – *Ultra* and *Z3*, etc. (Le Parisien 2000). Their failure or positioning outside the home computer market was influenced by many factors:

- they were too expensive for individual buyers (especially compared to the popular foreign computers ZX Spectrum, Commodore 64, etc.),
- the small number of entertainment and other programs made them unattractive to contemporary computer enthusiasts,
- it was not possible to buy them in stores.

Domestic computers of this generation were therefore mainly used in state institutions (which were not allowed to buy imported computers).

Galaksija was the only computer that managed to find its place in the sun and win the hearts of home computer enthusiasts in SFRY. Its creator was Voja Antonić. In an interview he gave for the "Startit" portal, Voja talked about when and how he came up with the idea to make the *Galaksija* computer:

"(the idea) ... I came up with it in Montenegro while on vacation. For me, beach vacations are boring because there is nothing to do. So I grabbed a notebook and started to make some of my projects and sketches on the beach. That's where I first came up with the idea of how to make a video unit simply and economically, which was the most complicated thing for computers at that time. The price of a computer has always depended on the complexity of the video unit. I came up with the idea of how to make it simpler, to make a computer that is bad and incredibly slow because it spends four-fifths of

its time generating an image, but which was still a computer. I realized that such a computer worked and could be built very simply, so I thought, why not offer it to someone else? Then I decided to publish it somewhere as a "do it yourself" project. The "Galaksija" magazine had a special issue about computers, which was supposed to be called "Computers in Your Home". I reached out to Dejan Ristanović, the man who was supposed to write it, and told him my idea, to which he replied: 'Of course. It would be wonderful if, for the first issue, there was some self-made thing' and that's how we started." (Startit 2017)

In the same interview, we also learned about their expectations and the number of people who would want to assemble a *Galaksija* computer, which ranged from 50 to 500. The number of purchase orders that the magazine "Galaksija" got for the assembly of *Galaksija* computers was 8000 and exceeded even the wildest hopes.

In an interview for "Yugopapir", another witness - Dejan Ristanović, editor of the special issue of "Galaksija" called "Computers in Your Home" testified about the success of the action. He talked about how they made a mistake in estimating the circulation needed for "Computers in Your Home":

"We printed 30,000 copies of the first issue of 'Computers in Your Home'. They sold out in two weeks, so we had to reprint 20,000 copies which sold out as well, so in the end a reprint of the first issue was made in another 25,000 copies. The second issue was printed immediately in 50,000 copies, and as it sold out, I think another 25,000 copies will be printed these days. The third issue is being made in 50,000 copies. If we had estimated things a little better and printed 100,000 copies from the first issue, everything would have been sold out again, because when there were no more 'Computers' on the newsstands, people exchanged and photocopied them". (Yugopapir 2014)

Ristanović also writes on his personal blog about how the magazine "Galaksija" and the eponymous computer came together: "In order for this kind of action to succeed, it took several months to prepare and keep an eye on the texts in "Galaksija", and my text in which self-assembly was presented was published in September 1983. It was accompanied by a preliminary purchase order, and the response was incredible: over a thousand readers of "Galaksija" expressed their desire to build their first computer! Such a large response is probably due to the fact that the import of computers (as well as any other goods more expensive than 1.500 dinars at the time) was prohibited and thousands of people interested in computers saw no other way to expand their knowledge. The characteristics of the *Galaksija* computer were not very good, but it was usable, balanced, and, most importantly,

attainable – you just had to gather the courage and pick up the soldering iron!" (Ristanović)

The technical characteristics of the *Galaksija* computer (Figure 1) were as follows:

- Processor: ZiLOG Z80A at 3.072 MHz
- *Galaksija* ROM 'A' (ROM 'A' or '1') – 4 kB EPROM (2732 EPROM) with a simple operating system, and Galaksija BASIC code for the BASIC program-interpreter;
- Galaksija ROM 'B' (ROM 'B' or '2') – 4 kB (optional, same as 2732 EPROM), provides additional BASIC commands, machine language (assembler), machine code monitor and more;
- Galaksija ROM for characters – 2 kB (2716 EPROM) includes definitions of symbols (letters and characters);
- RAM memory: from 2 to 6 kB of static memory (6116 RAM) in the basic version, possible expansion up to 54 kilobytes;
- Text: 32 × 16 letters or characters, black and white
- Pseudographics: 2 × 3 matrix, total 64 × 48 dots;
- Sound: None, but a tape recorder interface could be used for this purpose, as on the Spektrum (ZX Spectrum);
- Data storage: cassette tape, recording at a speed of 280 bits per second;
- Input and output ports: 44-pin edge connector with Z80 bus, cassette recorder (DIN connector), black and white video signal with PAL synchronization (DIN connector), and UHF antenna output (RCA connector).

The main goal of the *Galaksija* computer was for people to assemble it themselves (since they could not buy a computer) and then use it and play games. The technical characteristics were such that people could, with a soldering iron and a little patience, assemble their first computer on their own. How popular self-assembly of computers was is also evident in the article from the 130th issue of the magazine "Galaksija", where the text "Club in Novi Sad" featured fifteen-year-old Ivan Zindović, who presented a self-made microcomputer at the forum of the Center for the Study of Social and Technological Progress, in the presence of about a hundred high school students. He managed to assemble a computer whose value was 6,000 dinars.

All this led to the *Galaksija* being the only computer in the homes of many in 1980s, which also enabled the popularization of the use of computers for home purposes.



Figure 1. *Galaksija* computer. (Source: https://www.limundo.com/kupovina/Racunari-i-oprema/Desktop-racunari/Retro-racunari-i-oprema/Racunar-GALAKSIJA/70103817)

2 The magic of the "Galaksija" magazine

"Galaksija" was a popular science magazine that was published once a month. Each issue had 66 pages. Its price from January to June 1983 was 45 dinars, and 50 dinars after that. In that period, the editor-in-chief was Gavrilo Vučković, and the editorial board consisted of:

- Esad Jakupović, deputy editor-in-chief;
- Tanasije Gavranović, assistant editor-in-chief;
- Aleksandar Milinković and Jova Regasek, editors.

"Galaksija" began to be published on March 1, 1972, as a "Journal of Aviation, Astronautics and Future Research", edition of the newspaper publishing company "Duga". In November 1972 (issue 8), the magazine was repositioned and the new header read "Journal for the Popularization of Science", and remained so until the end of the period we are covering in the research. If we look at the list of people who wrote for the "Galaksija" magazine, we can see that it tried its best to distance itself from pseudoscience and take an objective attitude towards topics that were of interest. This is evidenced by

the publishing council of the magazine, which consisted of, among others, PhDs, six of them, as well as two University professors.

In that period, topics that were interesting and talked about a lot in the media were the Apollo mission, nuclear power plants, computers, etc. The Cold War was fought with science and this caused a very wide interest of the population in topics from these fields.

The maximum circulation of "Galaksija" was approximately 85,000 copies; it started with 50,000 copies (Planeta 2021), although there were times when it dropped to 30,000–35,000 copies. The concept was such that each issue had its lead topic, which was presented on the front page and covered in detail in the magazine through several texts. Other topics were not permanent and differed from one issue to another. The "Games, Hobbies, DIY" section was always there and usually occupied the last six pages of the magazine.

The "Galaksija" magazine was characterized by the great loyalty of its fans, which we can see in the letters of the readers, but also in the way in which the editors engaged in direct communication with them. The editors were so open that they even recommended foreign magazines that dealt with the same topic - even though they were their competition. This is directly evidenced in issue 132 in the text "Foreign magazines", which says: "We would recommend two: 'MC microcomputer' and 'Micro and personal computer'. The subscription is about 30,000 TRL lira, and the addresses to write to are..."

3 Methodological concept and research instruments

When something needs to be done for the first time, when a novelty is introduced either at the level of society or a smaller community - change is initiated by communication: how is the novelty presented, what is the message that is being sent and in what way, and what are the important questions that need to be answered. As the introduction of computers into the everyday life of the people in Yugoslavia was one of the very important tasks for the further development of industry, the economy, and thus the entire community, it is extremely important to analyze the communication that preceded the great success of the first domestic computer intended for wide use (*Galaksija*), which users could assemble themselves. It is important to understand the messages, their intensity and frequency, their quantity and quality – especially if we want to repeat the success of introducing novelties like this in the future. The quality of writing about a particular problem, as

Stjepan Gredelj states in "On the other side of the mirror" (Гредељ 1986), includes: the number of texts in the sample, their size, and communicative value. The first two indicators are immediately visible and easily measurable. Communication value is a derived indicator that is the result of an assessment of the importance attached to the text considering its position in the structure of the paper. The most important are the texts that occupy the main pages of the newspaper, the front page, the first, and the second, while the texts on the other pages are less important.

Content analysis is as a research procedure that strives to build a systematic experiential record of social communication (Manić 2017). Content analysis in sociology). Content analysis mainly deals with the analysis of secondary material – the one that was not created for the research, but rather independently. It was the chosen method because it allows for both quantitative and qualitative analysis of the material – in this case, the analysis of the content of the "Galaksija" magazine, which was then the most important scientific magazine in the Yugoslav market. One of the few that, at that time, systematically covered topics in the field of computer development and technology. The success of the magazine justifies its choice as one of the sources of material.

The sources used to study the contents of the magazine "Galaksija" about the eponymous Yugoslav computer were:

- existing historical and scientific theoretical literature on Yugoslav society during 1983 and 1984;
- statements and testimonies of people who participated in the creation of the *Galaksija* computer and wrote texts for the magazine "Galaksija" at that time (Voja Antonić, Dejan Ristanović, Stanko Popović);
- methodological literature on the method of content analysis;
- selected editions and texts of the magazine "Galaksija" on which content analysis was applied.

All covered copies of the magazine were found and purchased through the Kupindo platform as collector's editions, given that digital versions of the issue could not be obtained. As the copies were acquired in a rather poor condition: the sheets were damp and exposed to smoke for a long time, and given the type of binding, any disassembly would have completely destroyed the magazine copies. Due to the lack of professional equipment and knowledge in the field of paper restoration, I decided to apply the original method of content analysis, which consists of reading texts and manually counting terms and selected phrases.

The sample for content analysis includes all "Galaksija" magazines that were published in 1983 from issue 129 (January 1983) to the first two issues of 1984: 141 and 142 (except for issue 136, which could not be found on the market, and access to libraries was difficult due to restrictions caused by the global coronavirus pandemic). The time frame was defined in accordance with the subject of the research, that is, the creation of the *Galaksija* computer, which was the focus of the research. As the original plan for the release of the *Galaksija* computer was the last quarter of 1983, the goal was to track the way the topic was introduced to the target group, the intensity of writing before the action was initiated, the change in intensity of writing during the action, the connotation given to the topic, as well as the impact of writing on the sales results. Therefore, the period from January 1983 to February 1984 was chosen as ideal.

According to Dejan Ristanović, one of the authors of the "Galaksija" magazine, for him the "Galaksija" project began on June 24, 1983, when Zoran Vasiljević contacted him with the idea of taking a look at a small home computer that could be of interest to the readers of this magazine. Thus, the period before the June issue of the magazine "Galaksija" can essentially be viewed in one way, as the creation of texts that the authors organically dedicated to a certain topic. On the other hand, the period after June, when the magazine became a part of the joint project in a different way – it was more focused on action and propagation of certain ideas and actions with a pre-defined sales goal. Dejan Ristanović wrote about it on his blog: "... I wondered how willing the magazine would be to 'get involved' in something that is not its main interest, but it turned out that the interest was huge: Jovan Regasek, then the editor of the section in which I wrote about computers, and Gavrilo Vučković, the long-time editor-in-chief of 'Galaksija', immediately noticed the importance of this idea and showed a great willingness to fully commit themselves to its implementation. With one important change: the project will not be published in 'Galaksija', but in a special edition of 'Computers in Your Home', which was to be published at the end of 1983". (Ristanović)

The analysis included two basic units: topics and symbols. The analysis of topics determined the most important subjects of writing that dominated in the selected period and their relation to the phenomena and events to which they are dedicated. Also, whether something preceded the action dedicated to assembling the computer, happened in parallel with it, or after the action itself. This procedure required the interpretation of the material, and the created assessments were checked in other data sources – primarily, in

the scientific literature related to that period and topic, as well as in the testimonies of the actors themselves.

The analysis of symbols attempted to determine the terms used on the pages of the magazine "Galaksija" during 1983 and at the beginning of 1984. In the value and motivational impact of means of communication, certain words are carriers of a specific message and even the entire topic. The frequency of use of these terms and their orientation (whether it is positive, negative, or neutral) were measured. In addition, placing words in a certain context is essential when it comes to influencing changes in certain processes - in this case, ordering and self-assembling *Galaksija* computers.

Before the categorial system for the analysis was prepared, preliminary research was carried out on a smaller sample of texts from the observed period, followed by grouping of categories by logical connection. We encounter two types of categories:

1. content categories – considering what was said;
2. categories of forms of presentation of content – considering how something is said.

Content categories examine the subject and direction of messages. The subject answers the question of what the message is about. Direction refers to the symbols and determines whether they are positively disposed toward the subject, neutral, or negatively disposed.

The research distinguished the texts within the section "Games, Hobbies, DIY", which was always located at the back of the monthly issue (the last six pages), the texts of the topic that was presented on the cover of the magazine, as well as the texts found in the edition related to a topic, but not related to the section or the topic of the issue. The quantity of writing about the selected problems comprises the number of writings included in the sample, their size, and communication value. The selected topics are as follows:

- man and machine, robot – as one of the topics, due to the illustrative fusion of man and machine into a robot;
- artificial intelligence – as a topic that was seriously discussed both in the world and in our country, and which sparked various controversies;
- pocket, home computers, computers as one topic – because they are often used as synonyms;
- microcomputers as a separate segment due to the different use of the term, primarily in the context of business and professional computer use.

When it comes to the senders, those to whom the message is intended, and the effect of the message itself, indirect data were extracted, checked, and supplemented with data found in reports and analyses available from that period. The senders of the message are, first of all, the authors of the texts and the editorial staff of the magazine "Galaksija". The recipients of the message are the readers of this magazine, who were interested in computers and innovations on the market and did not want to be excluded from world trends. The answer to the question of "how" lies in the magazine "Galaksija", at that time one of the most popular science magazines that dealt with topics in the field of technology and computers. Texts aimed at dispelling myths about computers themselves, promoting self-assembly of computers, as well as the "Do It Yourself" section actively promoted the topic.

4 Quantity of writings on selected issues

The number of articles included in the sample and their volume are directly perceivable as the number and size of the texts, while the communication value is a derived indicator, the result of an assessment of the importance attached to the text considering its position in the structure of the paper. The highest weighting factor was assigned to the text promoted on the front page, which was the topic of the issue with the largest number of pages devoted to it. The texts on other pages were less important. There are a total of five weighting factors shown in Table 1. If one of the selected topics ("man and machine, robot", "artificial intelligence", "pocket, home computers, computers", "microcomputers") belongs to one of the categories, it is assigned a corresponding weighting factor:

Section	weight
Topic of the issue	5
Second page	4
Third page	3
Fourth page	2
Other pages	1

Table 1. Weighting of texts.

The sample included 13 issues. Each issue of "Galaksija" had 66 pages. The number of pages devoted to the "Games, Hobbies and DIY" section is

on average six (ranging between five and seven) – so 9% of the magazine is devoted to topics of interest. In addition to this section, texts devoted to the topics of computers, robots, and artificial intelligence appear in different segments in seven issues, which make up 53% of the covered issues. In those seven issues, a total of 12 texts are devoted to selected topics: man and machine, robot, artificial intelligence, pocket computers, home computers, computers, and microcomputers. The representation can be seen in Table 2.

Issue No.	No.of pages Games, Hobby	weight	No.of for computer assembling	weight	Additional texts dedicated to topic	No.of pages additional texts	weight	Total No. of pages dedicated to topic
129	5	5	1.5	1.5	0	0	0	1.5
130	5	5	1.5	1.5	0	0	0	1.5
131	6	6	2	2	1			2
132	6	6	1.5	1.5	1	1.5	1.5	3
133	7	7	2	2	2	4	4	6
134	6	6	0.5	0.5	1	1.5	1.5	2
135	6	6	1	1	2	9	45	10
137	6	6	1.5	1.5	0	0	0	1.5
138	6	6	2.5	2.5	4	10	50	12.5
139	6	6	2	2	0	0	0	2
140	7	7	0	0	0	0	0	0
141	6	6	2	2	1	1	1	3
142	6	6	2	2	0	0	0	2

Table 2. Weights for pages dedicated to the topics.

We notice that there is no issue in which at least one of the topics (man and machine, robot, artificial intelligence, pocket, home computers, computers, and microcomputers) is not covered. In the graph (Figure 2), we can see an uneven distribution of the total number of pages dedicated to topics and a clear downward trend at the beginning of 1984, after the release of a special edition dedicated to computer assembly. The peak was reached in the October issue 138, where the topic was the computer, so it is not surprising that the maximum attention was paid to the chosen topic. The graph also clearly shows the editorial interest in the topic and the increase in representation even before learning about the existence of the *Galaksija* computer and the plan of action.

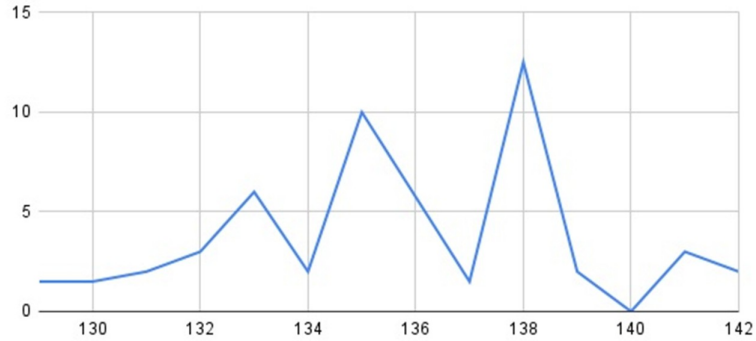


Figure 2. Total number of pages dedicated to topics per issue of "Galaksija".

When it comes to the cover, there are a total of three issues (23%) where one of these three elements is present (the cover contains the word *computer*, the cover contains the term *artificial intelligence*, and the cover contains a graphic representation of a computer). In issue 138, the topic is the computer, so this word is present in the title itself, as well as in the graphic illustration on the front page. The connotation of themes is never negative: it is either positive or neutral. Table 3 shows the percentage representation of topics on the cover of "Galaksija".

Description	%
The cover includes the word *computer*	7.69%
The cover includes the term *artificial intelligence*	7.69%
The cover includes a graphic representation of a computer	23.08%

Table 3. Percentage representation of topics on the cover of "Galaksija".

A total of 47 texts were covered in the research. Each magazine was read following the method of content analysis, all the phrases and words that were covered were counted, and the data compiled into tables. The size of the texts was also covered, as it is a specific indicator of the publicity given to a certain issue and topic at a specific time. "Galaksija", as a popular science magazine, is specific because it was not intended for one-time use (unlike daily editions of newspapers), but its copies were kept, re-read, and

in certain segments served as textbooks and manuals. This makes the value of each text greater – the impact of that text is more far-reaching, it is not just a temporary entertainment and pastime. The topic of computer assembly is always present. On average, one and a half pages in each issue are dedicated to the topic. The maximum number of pages devoted to the topic is 2.5, and only one issue did not have a text devoted to that topic.

Galaksija issue	man and machine, robot	AI, intelligent machines	home computers, computers etc.	micro computer	computer democratization
129	0	1	16	5	0
130	0	0	13	4	0
131	0	4	40	2	1
132	19	1	34	2	0
133	0	2	133	1	0
134	0	0	30	0	0
135	184	8	60	1	0
137	1	1	43	1	0
138	5	14	129	32	0
139	0	0	30	0	0
140	0	0	25	0	0
141	5	0	53	11	0
142	0	0	53	2	0
total	214	31	659	61	1
average	16.46	2.38	50.69	4.69	0.08

Table 4. Representation of the topic in the covered editions of the "Galaksija" magazine.

Table 4 clearly shows a stable trend and the presence of texts related to computer assembly. On the other hand, the presence among the texts outside that section is very unstable and with large peaks, and it is clear that part of the systematic efforts of the editors was to push the selected topic and draw attention to it. It can be noticed that in the first months of 1983, the selected topics did not appear outside the Hobby section, but from the summer that year, the intensity has changed considerably, as shown in graph (Figure 3).

Keywords and phrases that appear most often within the analyzed texts:

1. *Pocket computer, home computer*, and *computer* (used as synonyms) – a total of 659 times, an average of 50 times per issue.
2. *Man* and *machine*, *robot* – a total of 214 times, an average of 16 times per issue.
3. *Microcomputer* – 61 times, an average of four times per issue. The term microcomputer is used more for business use of computers and is not a usual term when discussing home and personal use of computers – therefore it is chosen as a keyword in the analysis.
4. *Artificial intelligence* – 31 times, an average of 2.38 times per issue.
5. *Computer democratization* – only once, in issue 131, when the computer fan clubs were announced. Before that, number 130 introduces the concept of clubs as an excellent way of organizing computer enthusiasts. Given the period in which the magazine was created, it is clear that the word democratization was not popular and, regardless of its completely benign use in this context, it was never used by the editorial staff later even though these topics were discussed.

Through the keyword analysis itself, it becomes clear that the emphasis was placed on personal computer use and that the choice of words included terms and phrases close to the readers, which do not create repulsion. Preference was also given to the relationship between man and machine, as well as robots, rather than artificial intelligence itself, even though that branch of science was developing very intensively at that time. This statement is illustrated by the following fact: in issue 129, when defining the CPU, it is stated that: "The central processor is the heart and brain of the computer, the part of the machine that thinks". Graph (Figure 3) shows the distribution of keywords by issue and clearly shows the influence of the topic of the issue on the representation and frequency of keywords. It can also be seen that from the summer until the publication of the special issue dedicated to the assembly of *Galaksija* computers, keywords were intensively covered, indicating that the editorial team was preparing the ground for action.

5 The sky is the limit: the wildest dream – 1000 purchase orders; 8000 arrived

In the SFRY, in 1983 and 1984, the import of goods with a value of 1,500 dinars and more was regulated. As Tanjug reported on July 22, in the first half of the year, the Yugoslav trade deficit with the West was more than

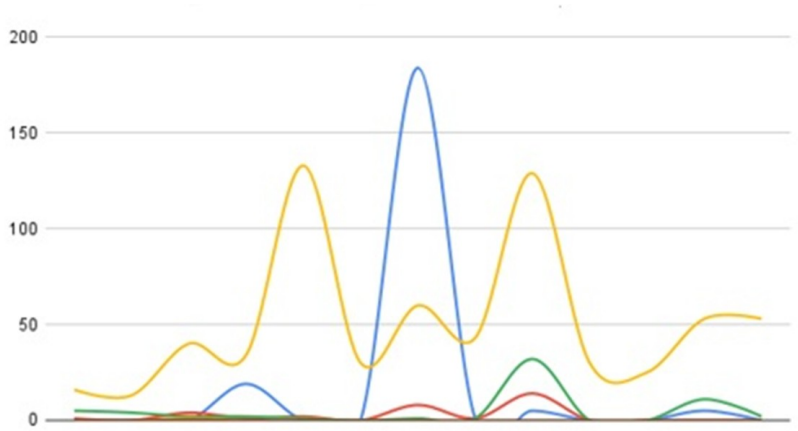


Figure 3. Distribution of keywords and phrases: blue line: *man and machine*, *robot*; red line: *AI*, *intelligent machines*; yellow line: *home computers*, *computers*, etc.; green line: *microcomputer*; orange line: *computer democratization*.

halved (from 2.12 to 0.99 billion dollars) thanks to reduced imports and increased exports. In addition, the Constitutional Court objected to the Federal Assembly because of the law requiring a deposit of 5,000 dinars for travel abroad (introduced in October 1982), which made frequent trips abroad very difficult. Due to customs measures, the import of complete computers was impossible. Yugoslavia officially went bankrupt in 1983, although it never announced this to its citizens (but others did), and stopped paying all its obligations to foreign countries, which resulted in large shortages. On the other hand, the domestic industry failed to make computers for home use quickly enough and at reasonable prices. Many people were interested, personal computers were a "hot topic" among technology enthusiasts, who used all kinds of means to get at least the components to assemble their computer and start programming or enjoy playing games. The editorial board of the magazine, primarily the three people most creditable for following the topic, Stanko Popović, Dejan Ristanović, and Voja Antonić, each of them through his network of contacts, his activity, conversations with the readers, monitoring of local events, and the questions they received, noticed an increase in interest in computers and decided to respond to the increase in demand.

When reading the texts, one notices very open communication with the readers based on questions and answers, published in the section intended for this purpose. It can also be seen that the editorial team is aware of local events in the cities, especially the clubs of computer enthusiasts, and gladly visits them and writes about them. This helps them build networks of contact and continuously receive information from all parts of the country - constantly checking the pulse of readers and events. This is exactly how editors and journalists learned about the topics that interest the readership and came up with a way to respond to them. In addition, by reading the contents of the "Galaksija" magazine, it is very clear that journalists followed all foreign editions, translated articles, and adapted them for the local market. In his interview for "Yugopapir", talking about foreign magazines on the topic of computers that could be obtained in Serbia, Dejan Ristanović says: "The real quality that these foreign magazines offer us is the outstanding presentation of new types of computers and peripherals, but that is still too little for the price of the paper and postage, which is paid in foreign currency". (Yugopapir 2014)

According to the data of the Statistical Office, the average monthly salary in Serbia was 15,161 dinars (Republički zavod za statistiku 2006), and the special edition of "Galaksija", which dealt with computer assembly, cost 200 dinars. This means that about 1.3% of the average salary had to be set aside for that special issue of "Galaksija". In 1983, the price of one kilogram of apples was 35 dinars, and a kilogram of chocolate cost 257 dinars (Milićević 2018) – therefore, based on the data of "Makroekonomija", we could assume that the price of "Galaksija" could cover the daily expenses of an average family in Serbia.

As Ristanović himself points out, "it had its purpose at a time when computers needed to be popularized and when there were not enough programs in the country...". (Yugopapir 2014)

At the very beginning, the most important thing was the realization that self-assembly was possible, and that is what preceded the writings of "Galaksija". All integrated circuits could be legally imported by mail from abroad, and the acquisition of printed circuit boards, keyboards, and boxes could be organized in Yugoslavia. The plan was to program the EPROM memories, because the initial assumption was a maximum of 100 requests, the bravest expected 1000 – however, 8000 requests arrived.

In the words of Ristanović: "In order for this kind of action to succeed, it took several months to prepare and pursue it in the texts in 'Galaksija', so in September 1983 my text which presented the self-assembly was published.

It was accompanied by a preliminary purchase order, and the response was incredible: over a thousand readers of 'Galaksija' expressed their desire to build their first computer! (...) When an action goes well, the people who participate in it are very motivated to continue with their effort: Voja Antonić improved the hardware and software of the computer countless times in the following months, and 'Galaksija' tried to find the most suitable way to organize the supply of kits for self-assembly, relying on the 'small economy', which was just emerging in those years. Thus, we reached an agreement that 'Mipro' and 'Elektronika' from Buje, in cooperation with the Institute for Electronics and Vacuum Technology, would deliver printed circuit boards, keyboards, and masks, that 'Mikrotehnika' from Graz would send chips, and that 'Galaksija' would collect purchase orders and program the EPROM."

That is exactly what "Galaksija" did. It covered the topic very well and created a demand by choosing great authors – excellent experts on the topic, familiar with the trends that existed in the world, who wrote in a way that is close to the readership. The insistence on self-assembly, "do-it-yourself" and active participation in the assembly of computers, which "Galaksija" dealt with for years, paid off, and readers were convinced that they could pick up a soldering iron and assemble their own computer and start doing something useful on it. Excellent instructions, graphic displays, positive connotation of the computer were the wind in the sails of the whole project. Here are some examples:

- In issue 131, in the text entitled "Communication with a computer", there are six illustrations of the process of communication with a computer. As can be seen in Figure 4, in each illustration the computer is smiling and happily communicates, and the cooperation with the human is represented as teamwork.
- In issue 135, the text with the creative title *The Varaždin 'seagull' has taken off*, announced the first Yugoslav microcomputer intended for personal and home use – that is, its entry into mass production. The price of the computer is said to be 89,035 dinars and it is stated that it is quite high for what is offered. The text is accompanied by a photo of a computer with the explanation: "In terms of external appearance, the 'seagull' can be compared to many foreign computers".
- In issue 139, under the title *"More and more powerful"*, graphics of the "Galaksija" computer was shown. The first message written on the screen read: "You too can make a *Galaksija* computer" – no better motivation was needed.



Figure 4. Communication with a computer.

Problems arose in the realization of the project itself – logistics and procurement were the first to break down. "First, there were no fiberglass epoxy laminated sheets, then there were no keys, then there was not enough money for production (payment was made by cash on delivery), then there were no chips... Murphy's laws were very much in action: chip prices were down for years, only to suddenly increase at the beginning of 1984. Static RAMs became more expensive right at that time, there were problems in obtaining 4 kilobyte EPROMs, the Z-80 became cheaper, but Galaksija's Z-80A became more expensive. EPROM programming was a little slow, 'Mikrotehnika' was late by several months, MIPRO and 'Elektronika' by almost a whole year, and all these problems made the editorial phone of 'Galaksija' constantly ring. Thousands of phone calls were the main cause

of both the *Computer 2* delay and the slow progress of *Galaksija* computer development". (Ristanović)

The only way to get the job done involved a lot of direct work and contact with the customers, as well as a lot of patience to solve the procurement and logistics issues. At the beginning of 1984, radio came to the rescue, and Zoran Modli, Dejan and Voja joined forces. Every Friday a new episode aired, and the topics were popular and interesting for many listeners – mostly games and demos: *Castle*, *Jumping Jack*, *Mastermind*, *Evolution*, *Galactic War*, *Hammurabi*, *Biorhythm*.

The editors and creators of *Galaksija* computers were not satisfied with the creation of software by end users. Out of thousands of purchase orders, only five or six software solutions worth publishing have arrived. The editorial office concluded that creating a program was a much more demanding job than assembling a computer and that it was important to provide creators with commercial compensation for their work and effort.

6 Conclusion

Nowadays, it is difficult to imagine a situation in which you are unable to go to a store and buy anything you need: whether it is a computer, a mobile phone, or a software solution. You do not even have to go anywhere – all it takes is a "click" and you can make all purchases online, and then pick it up the next moment or have it delivered to your home address. The presented case of the first local computer, Galaksija, intended for home use and assembly, shows that even in moments of great limitations and restrictions, it is possible to find a solution if you use your own knowledge and skills, i.e. if you take an active stance towards the solution of the problem. The role assumed in that case is active, and not passive – consumer role. In this case, man is an active creator, and the media (in this case, the popular science magazine "Galaksija") can contribute to the initiation of action, encourage people to take a specific action and give them enough faith and inspiration to succeed in their intention. The goal of writing was to move people to action, without the initial desire for commercial success – which, in the end, was nevertheless achieved. In an interview he gave to the "Startit" portal when asked why he gave *Galaksija* to people for free, Voja Antonić explains: "For the same reason I did everything else. It was my goal. Later, when the first issue of the magazine was published, a company from Istria appeared that distributed components, imported them from Austria, and sold them together with programmed EPROMs. They did well financially. When I finally calculated how

much they sold, it turned out that they earned more than a million marks at the time. That, in theory, should have upset me, but it didn't. It was their goal, I lacked the greed to do something like that". (Startit 2017)

The relationship with the readers of "Galaksija", which took place through dialogue (at that time asynchronous as it required correspondence via postcards or telegrams), indicates the existence of a discourse on the topic of assembling a computer, choosing the best options in a certain price range, researching potential solutions and encouraging creativity.

This type of communication with magazine editors has almost completely disappeared today. As comments are left via the Internet and various platforms, their creators do not use enough time and energy to articulate their views and thoughts and create a constructive dialogue. On the other hand, the editors are far from devoting time and energy to the views of the readers, their questions, or dialogue with them. Moderation of comments takes place at a completely different level and within other teams, most often those in charge of social networks. The issue of posting and not posting comments and impressions of readers is the same as it was in 1983 – unfortunately, we will never know how much remained unpublished and why.

Based on the testimony of the actors, as well as the research data, it is clear that the *Galaksija* computer would never have achieved success without the systemic support of the "Galaksija" magazine. The project is a good example of how to encourage an active attitude towards problem-solving, as well as creative work and thinking at times when there are objective restrictions imposed by the state.

References

Le Parisien. 2000. *Dictionnaire Sensagent.* Retrieved from Dictionnaire Sensagent. http : / / dictionnaire . sensagent . leparisien . fr / List % 20of % 20computer % 20systems % 20from % 20the % 20Socialist % 20Federal % 20Republic%20of%20Yugoslavia/en-en/.

Manić, Ž. 2017. *Analiza sadržaja u sociologiji.* Beograd: Čigoja štampa.

Milićević, D. 2018. *Makroekonomija.* Accessed on March 18, 2018. Retrieved from Makroekonomija. http://www.makroekonomija.org/0-dragovan-milicevic/sfrj-%E2%80%93-iluzija-ili-stvarnost-boljeg-zivota/.

Planeta. 2021. *Planeta.rs.* Accessed on April 04, 2021. Retrieved from Planeta.rs. http://www.planeta.rs/54/04_temabroja1.htm.

Republički zavod za statistiku. 2006. *Zarade u Republici Srbiji 1965–2005.* Technical report. Retrieved from Republički zavod za statistiku. https://pod2.stat.gov.rs/ObjavljenePublikacije/G2006/Pdf/G20066004.pdf.

Ristanović, D. *Dejan Ristanovic.* N.d. Retrieved from dejanristanovic.com. https://www.dejanristanovic.com/galaks.htm.

Startit. 2017. *Startit.* Accessed on July 15, 2017. Retrieved from Startit.rs. https://startit.rs/voja-antonic-danas-su-kompjuteri-bela-tehnika-a-nekada-su-bili-cudo-prirode/.

Yugopapir. 2014. *Yugopapir.com.* Accessed in August 2014. Retrieved from Yugopapir. http://www.yugopapir.com/2014/08/dejan-ristanovic-intervju-sa-mladim.html.

Гредељ, Стјепан. 1986. *С ону страну огледала. Истраживање промена модела комуникације у југословенском друштву на основу анализе садржаја писања листова Борбе и Политике у периоду од 1945. до 1975. год.* 232 стр. Београд: Истраживачко-издавачки центар ССО Србије.

Introduction to Digital Humanities: Workshops on the Implementation of Distant Reading in Research Practice Tršić, December 4–8, 2023

Milica Rabrenović
milica.rabrenovic@isj.sanu.ac.rs
ORCID: 0000-0002-4441-4884
Institute for Serbian Language SASA
Belgrade, Serbia

PAPER SUBMITTED: 17 April 2024
PAPER ACCEPTED: 29 May 2024

1 Introduction

The University Library "Svetozar Marković" and the Society for Language Resources and Technologies – JeRTeh organized the first seminar, "Introduction to Digital Humanities: Workshops on the Implementation of Distant Reading in Research Practice", and it was held in Tršić from December 4 to December 8, 2023. The seminar was conducted through collaboration between its authors and the Ministry of Science, Technological Development and Innovation of the Republic of Serbia. The seminar participants were undergraduate, master's and doctoral students. The goal of the seminar was to introduce students of philological disciplines and humanities to digital humanities methods, as well as to prepare and process texts for the application of distant reading techniques. The seminar included a lecture component and a practical component, i.e., workshops.

On the first day of the seminar, two lectures were held, followed by the application of the theoretical segment in the afternoon. The first lecture, "Introduction to Digital Humanities", was delivered by PhD Vasilije Milnović. At the beginning of his presentation, PhD Milnović reflected on the results of their previous projects, introduced the participants to the COST Action CA16204: *Distant Reading for European Literary History* (2018–2022), and highlighted the importance of digital humanities. He discussed the significance of developing language technologies, with a particular focus on the so-called "small languages" (e.g., Icelandic, Estonian, or Hebrew, where substantial investments are made in the field) and their potential applications

in lexicography, the teaching of Serbian language and literature, translation, and other areas.

The lecture by PhD professor Cvetana Krstev was dedicated to the *SrpELTeC* corpus. PhD professor Krstev introduced the participants to the process of creating this corpus and the criteria that had to be met for Serbian novels from the given period to become a part of the corpus (equal representation of male and female authors; balanced coverage of the selected period (1840–1920); variety in terms of length; inclusion of both well-known and lesser-known works; diversity of authors). She discussed the challenges that occurred during the formation of the Serbian subcollection and the process of digitizing the first editions (or the oldest available ones). The digitization process includes the following steps: scanning, optical character recognition (OCR), correction, basic annotation, metadata entry and automatic advanced annotation.

The seminar participants were introduced to the process of how text correction and annotation are performed, and the way metadata is entered. On the first day of the workshop, they were given a text for correction following the annotation guidelines. In the following days, they entered metadata and worked on annotating linguistic entities in the text.

On the second day, PhD Aleksandra Trtovac, in her lecture "The University Library 'Svetozar Marković' and Distant Reading", discussed the connection between librarianship and digital humanities, as well as their role in the COST Action. PhD Trtovac dedicated her second lecture to the *Transkribus project* and digitization using mobile phones. During the workshop that day, participants created accounts on Transkribus[1] and were introduced to the *DocScan* app, which enables the uploading of handwritten works to generate text transcripts. A particularly useful segment was the introduction to advanced search in the COBISS[2] system. In addition to the selective searching, the focus was also placed on the command searching. One of the valuable pieces of information shared was that the *UNPAYWALL* section contains open-access scientific articles. Furthermore, all names and surnames of a single author (useful when dealing with pseudonyms or when a female author's surname has changed due to marriage) can be found in the *CONOR.SR* section.

On the third day, PhD professor Krstev introduced participants to the concept of named entities and the methods for annotating them in a text. PhD professor Ranka Stanković's lecture was built upon the previous one,

1. readcoop, visited 28. 6. 2024.
2. COBISS.SR, visited 28. 6. 2024.

focusing on the processing of digital texts. Workshop participants worked on the automatic annotation of text segments with named entities (places, people, organizations, professions, etc.). Notes on more advanced types of annotation were also provided, followed by the review and evaluation of automatically annotated data.

On the same day, during the workshop, participants were introduced to Wikidata entries about Serbian novels and entered basic information about the text they received on the first day using the tools *OpenRefine* and *QuickStatements*. Additionally, participants were introduced to data querying using the *SPARQL* language.

The fourth day began with a lecture by PhD professor Duško Vitas, who discussed the process of creating a culinary corpus. PhD professor Ranka Stanković continued the topic by presenting the corpora available on the NOSKE platform of the JeRTeh society.[3] Participants had the opportunity to learn the basics of the *CQL* language and how to query the *srpELTeC* corpus. The workshop focused on command searches, starting with simpler queries and progressing to more complex grammatical patterns and queries involving named entity tags.

On the final day of the seminar, participants attended the lecture "Initiative for the Development of Open NLP/NLU Resources and Tools for the Serbian Language" by Slobodan Marković. The lecture highlighted the challenges of using artificial intelligence (AI) and the possibilities and issues related to processing the Serbian language with AI. The problem lies in the fact that models have not been fundamentally adapted, their application is not practical for business solutions, and their expressive capabilities are limited. The solution is to have as much training text as possible. Ideally, this data should be publicly available, under a permissive license, and cover both Serbian dialects. The goal of the initiative for open NLP resources is to have better search results and improved text comprehension.

This seminar provided comprehensive knowledge about the methods of digital humanities and the significance of this scientific field. It has allowed for better insights into future research, especially when it comes to annotating resources and working with textual corpora, which is where the greatest application of what was learned during the seminar lies. What should be emphasized as the most important outcome of these lectures and workshops is the opportunity to connect with colleagues and develop potential collaborations on future projects and research. This is precisely the main reason why, as a participant, I would highly recommend the seminar "Introduction to

3. noske.jerteh.rs, visited 28. 6. 2024.

Digital Humanities: Workshops on the Implementation of Distant Reading in Research Practice".

Acknowledgment

This work was financed by the Ministry of Education, Science, and Technological Development of the Republic of Serbia under Contract No. 451-03-136/2025-03/200174, concluded with the Institute for Serbian Language, SASA.