Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1691–1699
Marseille, 20-25 June 2022

Cross-Level Semantic Similarity for Serbian Newswire Texts

Vuk Batanović *, Maja Miličević Petrović †
* Innovation Center of the School of Electrical Engineering, University of Belgrade,
Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia
† Department of Interpreting and Translation, University of Bologna,
Corso della Repubblica 136, 47121 Forlì, Italy
vuk.batanovic@ic.etf.bg.ac.rs, maja.milicevic2@unibo.it

Abstract

Cross-Level Semantic Similarity (CLSS) is a measure of the level of semantic overlap between texts of different lengths. Although this problem was formulated almost a decade ago, research on it has been sparse, and limited exclusively to the English language. In this paper, we present the first CLSS dataset in another language, in the form of *CLSS.news.sr* – a corpus of 1000 phrase-sentence and 1000 sentence-paragraph newswire text pairs in Serbian, manually annotated with fine-grained semantic similarity scores using a 0–4 similarity scale. We describe the methodology of data collection and annotation, and compare the resulting corpus to its preexisting counterpart in English, *SemEval CLSS*, following up with a preliminary linguistic analysis of the newly created dataset. State-of-the-art pre-trained language models are then fine-tuned and evaluated on the CLSS task in Serbian using the produced data, and their settings and results are discussed. The *CLSS.news.sr* corpus and the guidelines used in its creation are made publicly available.

Keywords: cross-level semantic similarity, semantic textual similarity, corpus annotation

1. Introduction and Related Work

Semantic similarity refers to the extent to which the meanings of two given text items are similar to each other. The level of semantic similarity is commonly expressed as a numerical score on a Likert scale. Establishing such similarity measurements is an integral part of various Natural Language Processing (NLP) tasks, such as Information Retrieval (Hliaoutakis et al., 2006), Question Answering (Risch et al., 2021), Text Summarization (Mnasri, de Chalendar, and Ferret, 2017), etc.

Semantic similarity tasks typically focus on texts of similar length, such as individual words (Rubenstein and Goodenough, 1965), word senses (Budanitsky and Hirst, 2006), or sentences (Li et al., 2006). A well-known task of this sort is *Semantic Textual Similarity* (STS) (Corley and Mihalcea, 2005; Mihalcea, Corley, and Strapparava, 2006; Islam and Inkpen, 2008), popularized via a series of *SemEval* shared tasks (Agirre et al., 2012, 2013, 2014, 2015, 2016; Cer et al., 2017). In STS, the inputs being compared are short texts, usually the length of a sentence.

By contrast, in *Cross-Level Semantic Similarity* (CLSS) the goal is to evaluate texts of different lengths, for example to semantically compare a phrase to a sentence, or a sentence to an entire paragraph. Naturally, this task has an additional level of complexity compared to its STS counterpart, since the length discrepancy all but ensures that the longer text will carry a greater amount of salient information than the shorter one. In effect, CLSS aims to measure how well the meaning of the longer text is summarized in the shorter one. As pointed out by Jurgens, Pilehvar, and Navigli (2014), who first formulated CLSS and its corresponding *SemEval* shared task, CLSS can also be considered a generalization of STS to items of different types.

There have been comparatively few previous studies of Cross-Level Semantic Similarity and its applications, and all of them have been, to the best of our knowledge, focused solely on the English language. Jurgens, Pilehvar, and Navigli (2014, 2016) conceptualized the task and provided the first annotated datasets for it, composed of four types of pairs of different text lengths: paragraph to sentence, sentence to phrase, phrase to word, and word to sense. Their data was drawn from a variety of genres, including newswire, travel, scientific, review, etc. The *SemEval* word-to-sense dataset was subsequently used by Camacho-Collados, Pilehvar, and Navigli (2016), Pilehvar and Collier (2016), and Iacobacci and Navigli (2019) for evaluating their representation techniques. Furthermore, Pilehvar, Jurgens, and Navigli (2013) and Pilehvar and Navigli (2015) proposed a graph-based approach for measuring the semantic similarity of texts regardless of their length, making it applicable to linguistic items at multiple levels, ranging from word senses to full texts. The *SemEval* sentence-to-paragraph similarity dataset was utilized by Rekabsaz et al. (2017) to evaluate their text similarity methods. Regarding related tasks, Conforti, Pilehvar, and Collier (2018a, 2018b) dealt with the problem of cross-level stance detection, where the stance target is described in the form of a complete sentence, and the text to be evaluated is a long document.

In this paper, we present *CLSS.news.sr*[1] – the first non-English annotated CLSS dataset, comprising phrase-sentence and sentence-paragraph newswire text pairs in Serbian. Previous work on semantic similarity in Serbian has been relatively limited. Batanović, Furlan, and Nikolić (2011) and Furlan, Batanović, and Nikolić (2013) presented *paraphrase.sr*, a corpus of Serbian newswire texts manually annotated with binary similarity judgments, and used it to train and evaluate several paraphrase identification approaches. Batanović, Cvetanović, and Nikolić (2018) refined and extended this dataset with fine-grained similarity scores, and the resulting *STS.news.sr* corpus[2] was utilized to compare several unsupervised and supervised models. More recently, Batanović (2020) demonstrated that multilingual pre-trained language models such as *multilingual BERT* (Devlin et al., 2019) outperform all traditional methods on this task, while Batanović (2021) showed that their counterpart for Serbian and other closely related languages – *BERTić* (Ljubešić and Lauc, 2021) – yields even better results.

[1] http://vukbatanovic.github.io/CLSS.news.sr/

[2] http://vukbatanovic.github.io/STS.news.sr/

The remainder of this paper is structured as follows: in section 2, we present the methodology used to create and annotate the *CLSS.news.sr* corpus. In section 3, we give a statistical overview of CLSS datasets and outline a preliminary linguistic analysis of the new corpus. Section 4 focuses on the fine-tuning and evaluation of state-of-the-art pre-trained language models (Bommasani et al., 2021) for CLSS in Serbian and discusses their settings and results. Finally, section 5 contains our conclusions and pointers for future work.

2. Dataset Creation and Annotation

The *CLSS.news.sr* corpus was developed in the context of a broader project that aims to analyze the similarity between blocks of source code, written in a programming language, and the semantic similarity between their respective documentation comments, written in a natural language – English or Serbian. Code comments can be of arbitrary length, so the setup of phrase-sentence and sentence-paragraph cross-level semantic similarity naturally arises. In addition, the language used in code comments has long been known to diverge from the standard language, for instance in often being syntactically incomplete (see e.g. Zemankova and Eastman, 1980). For this reason, we decided to also explore phrase-sentence and sentence-paragraph CLSS in standard language, choosing newswire texts as its representative, since news-based STS corpora are available in both English (Agirre et al., 2012, 2013, 2014, 2015, 2016) and Serbian (Batanović, Cvetanović, and Nikolić, 2018).

In order to enable comparative analyses, it was important to establish a common methodology for dataset creation and annotation. Since the only pre-existing CLSS dataset was the *SemEval* one for English, produced by Jurgens, Pilehvar, and Navigli (2014), we decided to take their approach as a (partial) model for our work. We retained their five-point Likert similarity scale, with scores ranging from 0 to 4 (0 – unrelated, 1 – slightly related, 2 – somewhat related but not similar, 3 – somewhat similar, 4 – very similar), as well as their definitions for each score. However, their method of text pair construction, where the annotators were given a longer text and then asked to generate a shorter one with a designated similarity score in mind, would be ill-suited for the domain of source code comments, given the highly technical and often project-specific terminology encountered in them. For this reason, we chose to develop all our datasets by providing the annotators with numerous text samples of different lengths (phrases, sentences, and paragraphs), and asking them to combine these naturally occurring texts into phrase-sentence and sentence-paragraph pairs. While doing so, the annotators were asked to aim for a balanced score distribution for the pairs they construct. By creating text pairs in this manner, we have also avoided the impact of a potential paraphrasing bias that the annotators could inadvertently introduce into the dataset.

The source texts for the *CLSS.news.sr* corpus were gathered from *naslovi.net*, a news aggregator website in Serbian. The approach was the same one used in the construction of various newswire STS and paraphrasing corpora (Dolan, Quirk, and Brockett, 2004), based on exploiting the journalistic convention that the beginning sections of an article often provide a summary of its content. Since each news item can be reported on in different forms by different media outlets, cross-linking the texts of these different reports allows for the creation of text pairs with varying degrees of semantic similarity. *Naslovi.net* provides a headline and an introductory paragraph for each news report, sometimes with a subhead as well. We treated the headlines as source material for phrases, subheads as source material for sentences, and introductory paragraphs as source material for paragraphs for our corpus. However, the annotators were instructed to carefully evaluate whether an item in a certain category really was a phrase, a sentence, or a paragraph. To this end, we defined a paragraph as text containing a minimum of two sentences, where only complete sentences were to be taken into account. A sentence had to contain at least one finite verb form, whereas a phrase was not allowed to contain finite verbs (infinite forms such as infinitives and participles were allowed, as were deverbal nouns).

Since we aimed for our dataset to be comparable in size to the *SemEval* one, we set out to create 1000 phrase-sentence and 1000 sentence-paragraph pairs. In total, close to 18000 news reports, written between June and August 2021, were scraped from *naslovi.net* using the *scrapy* Python library[3], in order to provide the annotators with a sufficient quantity of raw texts for creating adequate pairs.

Score	Example
4	Veliki požar na železničkoj stanici u Londonu *A large fire at a London railway station*
	Veliki požar izbio je danas na metro stanici u centralnom delu Londona. *A large fire broke out today at an underground station in central London.*
3	Novi nacionalni praznik: Džuntint *A new national holiday: Juneteenth*
	Američki Kongres usvojio je predlog zakona prema kojem je 19. jun proglašen praznikom u znak sećanja na kraj ropstva i odlazak poslednjih robova 1865. godine u državi Teksas. *The American Congress passed a Draft law declaring 19 June a holiday to commemorate the end of slavery and the liberation of the last slaves in 1865 in the state of Texas.*
2	Veliki problem za Portugal *A major problem for Portugal*
	Loše vesti stižu za Portugal pred start Evropskog prvenstva. *Bad news arrives for Portugal just before the start of the European Championship.*
1	Svađa pred svadbu *A pre-wedding argument*
	Mirko Šijan i Bojana Rodić uskoro očekuju svoje prvo dete, a uveliko se sprema i njihova svadba. *Mirko Šijan and Bojana Rodić are expecting their first child soon, and their wedding is being prepared.*
0	Otvaranje silosa u Zrenjaninu *A silo opening in Zrenjanin*
	Maja Žeželj, voditeljka, ispričala je kako je svojevremeno jedva izvukla živu glavu. *Maja Žeželj, TV presenter, told the story of how some time ago she nearly died.*

Table 1. Guideline examples of phrase-sentence pairs for each similarity score.

[3] http://scrapy.org/

Text pair construction was divided between five annotators, who were either trained linguists or had previous experience with text annotation on the closely related STS task. They were given the similarity score definitions as well as *SemEval* score examples to help them interpret each score. However, the provided examples proved insufficient to ensure high levels of annotation consistency. Hence, the annotators' outputs were calibrated by having all of them create a smaller set of five to six representative pairs for each similarity score and each length pairing. We reviewed these representative pairs and gave feedback to the annotators regarding any issues encountered. We then compiled a detailed set of examples, three per similarity score and length pairing, using the agreed upon representative pairs from all annotators. This set, together with the score definitions and general instructions, became an integral part of the final annotation guidelines for our task, which have been made available in Serbian (original) and English (translation) in the dataset repository. A subset of the examples used for the phrase-sentence pairs is shown in Table 1.

The annotators were then asked to construct additional pairings for each text length combination, but to avoid creating only pairs that should clearly be marked with a specific similarity score, so as to ensure that more difficult pairs are included in the dataset as well. In total, each annotator produced around 200 pairs per text length combination. After the merging of the pairs produced by different annotators, a few duplicate entries were replaced, to ensure that every text appeared only once in the dataset.

Finally, the 2000 cross-level text pairs were labeled with semantic similarity scores by all five annotators. This process was completed using the *STSAnno* tool (Batanović, Cvetanović, and Nikolić, 2018), which allows an annotator to explore the pairs in a corpus, to view in parallel the texts in a pair and to assign, change or erase their semantic similarity score. The final score for each pair was calculated by averaging the individual scores of all annotators.

Obtaining multiple parallel annotations and averaging them out was chosen instead of relying on an adjudicated double annotation (used for the *SemEval* dataset) in order to minimize individual annotator's biases. In addition, while Jurgens, Pilehvar, and Navigli (2014) allowed for finer-grained distinctions using multiples of 0.25, in our setup this was not necessary for achieving final fine-grained scores.

3. Dataset Analysis

In this section we first present a statistical overview of the CLSS data, and then outline a preliminary linguistic analysis of the Serbian corpus.

3.1 Statistical Overview

A basic statistical overview of the new *CLSS.news.sr* dataset, and its *SemEval* counterpart in English, is shown in Table 2. The statistics are calculated independently for phrase-sentence and sentence-paragraph pairs in both corpora. The newswire subsets of the *SemEval* data are also shown separately, since the *SemEval* dataset includes several other genres. Note also that for the *SemEval* dataset we merge the train, trial, and test portions of the data.

As can be seen from the table, *CLSS.news.sr* is comparable to *SemEval* both in the number of text pairs and in token counts. The average lengths of phrases, sentences, and paragraphs are also similar between the two datasets, particularly if we only consider the newswire portion of the *SemEval* corpus. The vocabularies, on the other hand, are larger in the Serbian dataset, which is expected given the morphological complexity of the language. Finally, the average similarity scores for the Serbian data are closer to the scale's mean value of 2 than the scores for English, especially in comparison to the *SemEval* newswire pairs. A more detailed step plot overview of the distribution of text pairs across semantic similarity scores in English and Serbian CLSS corpora is shown in Figure 1 for the phrase-sentence pairs, and in Figure 2 for the sentence-paragraph pairs. We again consider separately the entire *SemEval* CLSS corpus and its newswire subset.

The figures show that the distribution is more balanced in *CLSS.news.sr* than in the English corpora across the entire score range, including intermediate values, likely due to the averaging of multiple individual annotations in the Serbian dataset (note that Jurgens, Pilehvar, and Navigli (2014, 2016) already described their scores as evenly distributed).

Dataset	Language	Text pairs	Tokens	Average phrase length in tokens	Average sentence length in tokens	Average paragraph length in tokens	Vocabulary size	Average similarity score
CLSS.news.sr phrase-sentence	SR	1000	30K	~6	~23	/	12K	1.96
CLSS.news.sr sentence-paragraph	SR	1000	86K	/	~22	~64	27K	1.91
SemEval CLSS phrase-sentence	EN	1036	26K	~5	~19	/	8K	1.90
SemEval CLSS phrase-sentence (newswire)	EN	425	13K	~5	~24	/	4K	1.76
SemEval CLSS sentence-paragraph	EN	1034	93K	/	~19	~71	20K	1.84
SemEval CLSS sentence-paragraph (newswire)	EN	301	26K	/	~20	~66	7K	1.68

Table 2. A statistical overview of the new *CLSS.news.sr* dataset and the previous CLSS datasets in English.

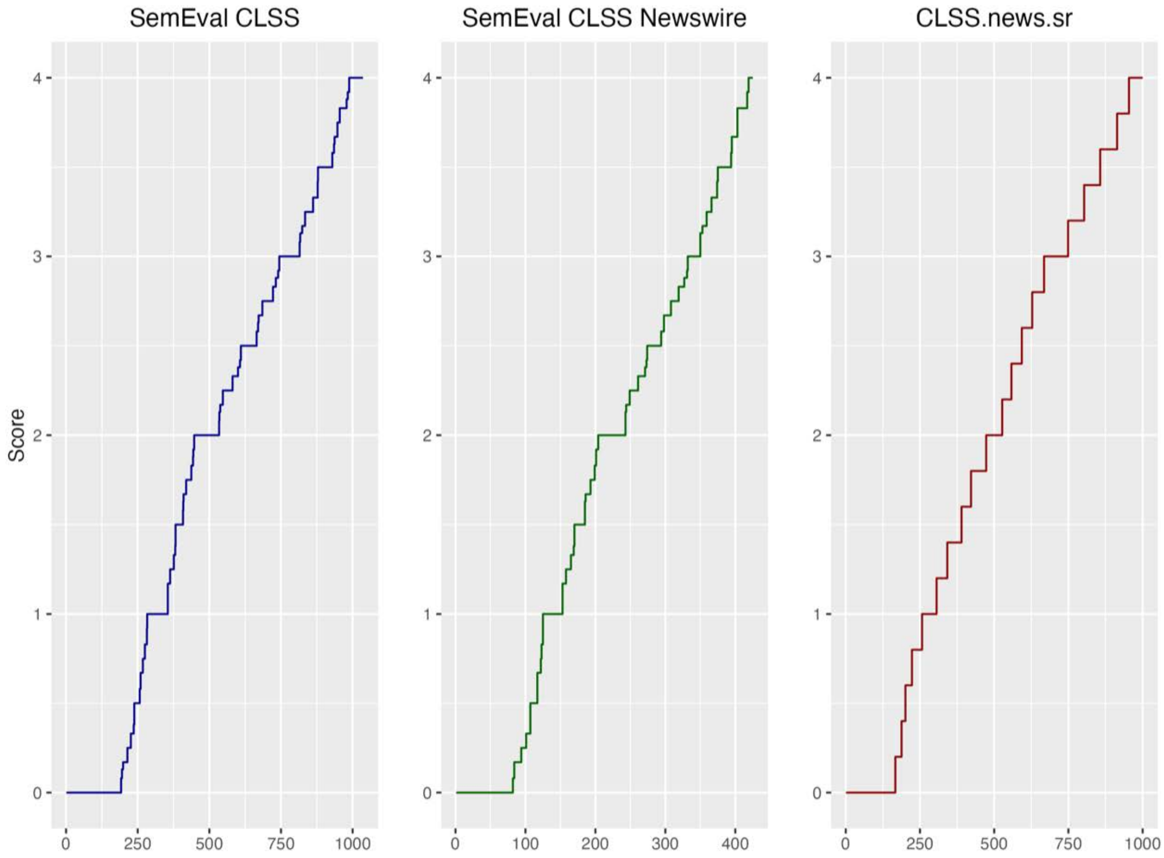

Figure 1. A step plot of the distribution of phrase-sentence pairs across averaged similarity scores in CLSS corpora.

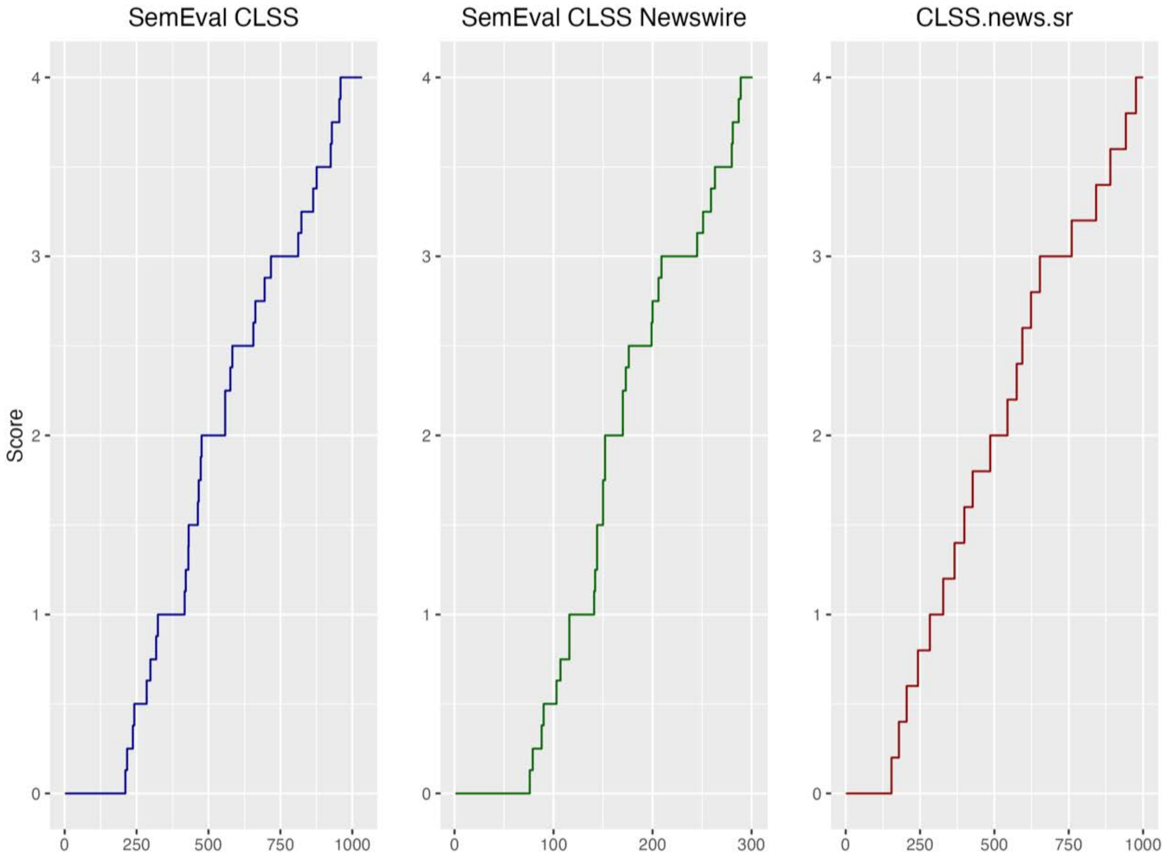

Figure 2. A step plot of the distribution of sentence-paragraph pairs across averaged similarity scores in CLSS corpora.

While it is evident that all datasets contain a large portion of scores equal to zero, this effect is least pronounced in *CLSS.news.sr*. In fact, the English corpora display groupings around all five round values (the 4 being the least obvious one), and to some extent .5 values, whereas in the Serbian corpus peaks appear chiefly around the score values 0 and, to a lesser extent, 3. Apart from these two points of concentration, the overall phrase-sentence pair distribution in the Serbian dataset is fairly uniform. The sentence-paragraph pairs exhibit a more irregular score distribution in both languages, with a more pronounced peak around the score value 3 in all datasets. Unlike the phrase-sentence corpora, where both the entire *SemEval* dataset and its subset share a similar distribution, the sentence-paragraph newswire subset exhibits a visibly more irregular distribution than the entire *SemEval* corpus, with very few pairs with score values between 1 and 2.

Annotation consistency was measured using the widely applicable Krippendorff's alpha coefficient (Krippendorff, 2004; Artstein and Poesio, 2008), the Pearson correlation coefficient r and the Spearman correlation coefficient ρ. For calculating the alpha score values, we utilized the Krippendorff Python library[4]. Inter-annotator agreements are shown in Table 3 for the phrase-sentence pairs, and Table 4 for the sentence-paragraph pairs.

These figures are very high, and are consistently over the 0.8 alpha coefficient value, proposed by Krippendorff as the threshold for an agreement to be considered reliable. It is also evident that the performance of all annotators is on a similar level, with only minor differences between them. The differences between the phrase-sentence and the sentence-paragraph agreement levels are also very low.

Similarly high values are found for self-agreement scores, shown in Table 5 for all annotators on both the phrase-sentence and the sentence-paragraph pairs. These scores were calculated on the initial ~200 pair sets created and labeled by each annotator, by comparing their initial scores to the ones assigned when annotating the entire corpus.

Annotator				r / ρ						
				#1	#2	#3	#4	#5	Mean of other annotators' scores	
									Per-annotator	Average
α	Binary	#1		/	0.905 / 0.908	0.907 / 0.909	0.906 / 0.911	0.900 / 0.900	0.939 / 0.941	0.938 / 0.938
		#2		0.899	/	0.904 / 0.908	0.902 / 0.908	0.898 / 0.900	0.937 / 0.940	
		#3		0.906	0.901	/	0.899 / 0.901	0.901 / 0.899	0.937 / 0.936	
		#4		0.906	0.893	0.896	/	0.914 / 0.911	0.940 / 0.938	
		#5		0.892	0.892	0.897	0.900	/	0.938 / 0.935	
		Mean of other annotators' scores	Per-annotator	0.936	0.925	0.932	0.925	0.928	/	
			Average	0.929						
	Global	0.898								

Table 3. Inter-annotator agreement scores on the *CLSS.news.sr* phrase-sentence pairs.

Annotator				r / ρ						
				#1	#2	#3	#4	#5	Mean of other annotators' scores	
									Per-annotator	Average
α	Binary	#1		/	0.899 / 0.908	0.905 / 0.909	0.898 / 0.911	0.914 / 0.900	0.940 / 0.938	0.937 / 0.934
		#2		0.885	/	0.892 / 0.908	0.894 / 0.908	0.902 / 0.900	0.930 / 0.933	
		#3		0.905	0.880	/	0.903 / 0.901	0.896 / 0.899	0.934 / 0.931	
		#4		0.894	0.888	0.900	/	0.911 / 0.911	0.937 / 0.929	
		#5		0.914	0.890	0.901	0.908	/	0.942 / 0.940	
		Mean of other annotators' scores	Per-annotator	0.931	0.908	0.924	0.916	0.932	/	
			Average	0.922						
	Global	0.897								

Table 4. Inter-annotator agreement scores on the *CLSS.news.sr* sentence-paragraph pairs.

[4] http://github.com/pln-fing-udelar/fast-krippendorff

Annotator	Phrase-sentence pairs			Sentence-paragraph pairs		
	r	*ρ*	*α*	*r*	*ρ*	*α*
#1	0.945	0.945	0.942	0.952	0.953	0.953
#2	0.915	0.916	0.912	0.889	0.890	0.885
#3	0.921	0.923	0.914	0.914	0.914	0.909
#4	0.925	0.923	0.916	0.928	0.934	0.933
#5	0.946	0.945	0.940	0.934	0.936	0.928
Average	0.930	0.930	0.925	0.923	0.925	0.921

Table 5. Annotator self-agreement scores on the *CLSS.news.sr* dataset.

3.2 Preliminary Linguistic Analysis

A preliminary qualitative linguistic analysis was performed on a random sample of ten pairs per score (taking into account only pairs that received the same score by all five annotators), for both phrase-sentence and sentence-paragraph pairs. For both types of comparisons, the pairs unanimously marked 4 are characterized by the occurrence of the same personal name(s) and/or number(s), in addition to shared common lexical words. It is often the case that the personal name forms are not identical (e.g., they are different case forms of the same noun, as in *Kragujevcu*.LOC – *Kragujevca*.GEN ‘Kragujevac’, or an adjective and a noun, as in *vlasotinačkom*.ADJ-*Vlasotincu*.N ‘(of) Vlasotince’), but are clearly relatable on morphological grounds. The shared numbers tend to be large and either quite specific or used in a collocation (e.g., *100.620*; *3.000 dinara* ‘3000 dinars’). The overlaps in the nominal and the verbal domains of general vocabulary are also often based on morphologically related rather than identical forms (e.g., *novozaraženih* ‘newly infected’ – *novih slučajeva zaraze* ‘new cases of infection’, *stiglo*.PAST.PART – *stići*.INF ‘arrive’). Synonyms are also present, but mostly within different collocations based on the same term (e.g., *toplotni talas* – *talas vrućina* ‘heat wave’). Overall, almost all lexical words from the smaller unit are also present in the larger unit, which also contains additional words or sentences that describe the situation more extensively, but without adding new topics; the lexis from the shorter item is distributed over the entire longer item and what is added are details about these elements (e.g, *u Londonu* ‘in London’ vs. *u centralnom delu Londona* ‘in central London’). The score 3 items are distinguished by similar properties in terms of shared lexis and especially personal names, but with a presence of entirely new information in the longer item, and/or partly different information in the two components of the pair, which is reflected in a lower overall vocabulary overlap. In both score 4 and score 3 items in the phrase-sentence comparisons, the head noun of the phrase typically appears as the subject or the object of the sentence predicate. The predicate is typically the same in sentence-paragraph pairs (with additional predicates in the paragraph item).

Among the less similar pairs, those marked 2 tend to be somewhat mixed, as they either contain different personal names and similar common vocabulary, or vice versa. The predicate in the sentence is typically not related to the head noun in the phrase. The pairs marked 1 and 0 contain barely any overlapping personal names. Score 1 items do share some common lexical words, but synonyms and terms from the same semantic field appear to be more frequent than identical or morphologically closely related words (e.g., *pljuskovi* ‘showers’ – *kiša* ‘rain’, or *povreda* ‘injury’ – *bolovi* ‘pains’); a common situation is also for the relatedness of lexical items in the pair to be based on real world knowledge rather than on linguistic information (e.g., *žreb za Ligu konferencija* ‘Conference League draw’ – *fudbaleri* ‘football players’; *vakcinacija* ‘vaccination’ – *virus korona* ‘corona virus’). Items marked 0 typically do not share any lexical words at all.

4. Model Evaluation

Previous work on the closely related task of Semantic Textual Similarity in Serbian (Batanović, 2020, 2021) demonstrated a significant performance superiority of fine-tuning massive pre-trained language models, also known as foundation models (Bommasani et al., 2021), over the previous approaches. Due to this, we limit our explorations to two representative language models. The first one is *multilingual BERT* (Devlin et al., 2019), a multilingual extension of the original *BERT* neural architecture, pre-trained on 104 different languages. The second model is *BERTić* (Ljubešić and Lauc, 2021), based on the computationally more efficient *Electra* model (Clark et al., 2020), pre-trained on over 8 billion tokens of text in Bosnian, Croatian, Montenegrin and Serbian, all closely related languages. We use the implementations provided in the *HuggingFace Transformers* library (Wolf et al., 2019), which we interface with using the *SimpleTransformers* library[5]. We do not perform any pre-processing of the corpus texts, since previous STS research on Serbian has shown that applying such techniques proves detrimental for the utilized neural models (Batanović, 2020). Both models we consider retain text casing.

The evaluation was performed using 10-fold cross-validation with sorted stratification. The performance metrics used are the Pearson correlation coefficient *r* and the Spearman correlation coefficient *ρ*, calculated between the model outputs and the averaged annotated similarity scores, which we consider the gold standard. For both models, we report the figures obtained by averaging five runs of the model with different initial seed values. As a baseline we use the word overlap technique employed in the *SemEval* STS shared tasks (Agirre et al., 2012, 2013, 2014, 2015, 2016; Cer et al., 2017), where texts are lowercased and tokenized using white space and then represented as binarized bag-of-words vectors in the multidimensional token space. The similarity of such vectors is expressed via cosine similarity.

For both the phrase-sentence and the sentence-paragraph similarity task, we explore the performance effect of model fine-tuning length, ranging from one to five epochs. We also consider the impact of enlarging the training set in each

[5] http://simpletransformers.ai/

Model	Additional training data	Fine-tuning epochs	Correlation coefficient	
			Pearson *r*	Spearman *ρ*
Word overlap baseline	/	/	0.6361	0.6430
Multilingual BERT	/	1	0.8756	0.8736
		3	0.9005	0.8970
		5	0.9010	0.8990
	CLSS.news.sr sentence-paragraph pairs	1	0.8902	0.8893
		3	0.9106	0.9073
		5	0.9100	0.9060
	STS.news.sr sentence pairs	1	0.8830	0.8851
		3	0.8988	0.8962
		5	0.9000	0.8974
BERTić	/	1	0.9193	0.9239
		3	0.9441	0.9403
		5	0.9483	0.9439
	CLSS.news.sr sentence-paragraph pairs	1	0.9272	0.9277
		3	0.9501	0.9467
		5	**0.9524**	**0.9486**
	STS.news.sr sentence pairs	1	0.9231	0.9236
		3	0.9446	0.9409
		5	0.9479	0.9442

Table 6. Model results on the phrase-sentence similarity task.

Model	Additional training data	Fine-tuning epochs	Correlation coefficient	
			Pearson *r*	Spearman *ρ*
Word overlap baseline	/	/	0.6458	0.6833
Multilingual BERT	/	1	0.9048	0.8941
		3	0.9250	0.9106
		5	0.9265	0.9126
	CLSS.news.sr phrase-sentence pairs	1	0.9187	0.9056
		3	0.9324	0.9186
		5	0.9322	0.9198
	STS.news.sr sentence pairs	1	0.9110	0.9004
		3	0.9192	0.9062
		5	0.9261	0.9132
BERTić	/	1	0.9077	0.9000
		3	0.9394	0.9255
		5	0.9465	0.9334
	CLSS.news.sr phrase-sentence pairs	1	0.9225	0.9135
		3	0.9451	0.9333
		5	**0.9485**	**0.9368**
	STS.news.sr sentence pairs	1	0.9111	0.9008
		3	0.9374	0.9260
		5	0.9405	0.9292

Table 7. Model results on the sentence-paragraph similarity task.

cross-validation fold with additional data, since in both tasks we only have one thousand samples to work with. To this end, we experiment with including sentence-paragraph pairs in the training set for the phrase-sentence similarity task, and vice versa. Similarly, we examine the inclusion of sentence pairs from the *STS.news.sr* corpus in the same manner, for both CLSS tasks.

The maximum sequence length for both *multilingual BERT* and *BERTić* is kept on the *SimpleTransformers*' default of 128 tokens for the phrase-sentence similarity task, since all paired phrases and sentences possess fewer than 128 tokens, for both of the pre-trained models' tokenizers. On the other hand, for the sentence-paragraph similarity task, as well as for the extension of phrase-sentence training data with sentence-paragraph pairs, the maximum sequence length is increased to 256, since all of the paired sentences and paragraphs are shorter than this. All other model hyperparameters, except for the number of fine-tuning epochs, are kept at their default settings.

The evaluation results on the phrase-sentence similarity task are shown in Table 6. Table 7 contains the results for the sentence-paragraph similarity task.

It is evident that the *multilingual BERT* model achieves higher scores for both correlation coefficients on sentence-paragraph similarity than on the phrase-sentence task, while the performance of *BERTić* exhibits the opposite trend. However, in both settings the *BERTić* model outperforms its multilingual counterpart, with the difference being more pronounced on the phrase-sentence similarity task. The superiority of the *BERTić* model is in line with the results previously reported on the Serbian STS corpus (Batanović, 2021). Naturally, the performance of both models improves as the fine-tuning is extended with additional epochs. The difference in results between three and five epochs is usually more noticeable with the *BERTić* model, but even in its case the performance gain is quite limited. Nevertheless, the benefit of increasing the fine-tuning length is typically most evident when no additional training data is used.

The impact of adding additional training data pairs is consistently positive when those pairs come from the *CLSS.news.sr* corpus. On the other hand, when the *STS.news.sr* corpus is used in the same manner, the effects are clearly positive only for *multilingual BERT*, when its fine-tuning is limited to a single epoch. If longer fine-tuning lengths are employed, or if the *BERTić* model is used instead, adding *STS.news.sr* pairs has a negligible effect at best, and can in many cases actually hurt the performance. This is probably due to the significant topic divergence between the STS and CLSS corpora, since *STS.news.sr* contains newswire texts that are a decade old, whereas *CLSS.news.sr* is made up of recent news reports.

Even without additional training data, *BERTić* reaches human performance on both sections of the *CLSS.news.sr* corpus. When fine-tuned for five epochs, this model outperforms the average inter-annotator agreement levels, both in terms of the Pearson and the Spearman correlation coefficient.

5. Conclusion

In this paper, we have presented *CLSS.news.sr*, the first Cross-Level Semantic Similarity corpus in a language other than English, and the methodology used to construct and annotate the data. We have compared this newly created

Serbian dataset to a similar one that already exists for English – *SemEval* CLSS – and its newswire subset (Jurgens, Pilehvar, and Navigli, 2014), showing that our fine-grained similarity annotation is even more balanced across the range of score values. A preliminary linguistic analysis was also conducted on a sample of pairs selected evenly among the similarity scores. Finally, we have evaluated a couple of pre-trained language models which support Serbian on the newly created corpus, showing that the best performances are obtained with *BERTić* (Ljubešić and Lauc, 2021).

Our planned next steps are to conduct a more extensive linguistic analysis and to examine the impact of linguistic traits on model performances. Another goal is to compare the results to those obtained for source code comments, and to develop a model that can handle both types of text. Finally, we intend to examine cross-lingual setups of the CLSS task, both in the newswire and the source code comment domain.

6. Acknowledgements

The research presented in this paper was conducted as part of the project *Advancing Novel Textual Similarity-based Solutions in Software Development – AVANTES*, supported by the Science Fund of the Republic of Serbia, grant no. 6526093, AI-AVANTES, within the “Program for Development of Projects in the field of Artificial Intelligence”. The authors would like to thank Bojan Jakovljević, Lazar Milić, Marija Lazarević, Ognjen Krešić, and Vanja Miljković for annotating the *CLSS.news.sr* corpus.

7. Bibliographical References

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., … Wiebe, J. (2014). SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, Association for Computational Linguistics, pp. 81–91.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., … Wiebe, J. (2015). SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the Ninth International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, USA, Association for Computational Linguistics, pp. 252–263.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., … Wiebe, J. (2016). SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, USA, Association for Computational Linguistics, pp. 497–511.

Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, Montreal, Canada, Association for Computational Linguistics, pp. 385–393.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. (2013). *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, Atlanta, Georgia, USA, Association for Computational Linguistics, pp. 32–43.

Artstein, R., and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4), pp. 555–596.

Batanović, V. (2020). A methodology for solving semantic tasks in the processing of short texts written in natural languages with limited resources. *PhD dissertation*, University of Belgrade - School of Electrical Engineering.

Batanović, V. (2021). Semantic Similarity and Sentiment Analysis of Short Texts in Serbian. In *Proceedings of the 29th Telecommunications forum (TELFOR 2021)*, Belgrade, Serbia, IEEE.

Batanović, V., Cvetanović, M., and Nikolić, B. (2018). Fine-grained Semantic Textual Similarity for Serbian. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, European Language Resources Association (ELRA), pp. 1370–1378.

Batanović, V., Furlan, B., and Nikolić, B. (2011). A Software System for Determining the Semantic Similarity of Short Texts in Serbian. In *Proceedings of the 19th Telecommunications forum (TELFOR 2011)*, Belgrade, Serbia, IEEE, pp. 1249–1252.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2021). On the Opportunities and Risks of Foundation Models. arXiv: 2108.07258

Budanitsky, A., and Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1), pp. 13–47.

Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2016). NASARI: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240 pp. 36–64.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval 2017)*, Vancouver, Canada, Association for Computational Linguistics, pp. 1–14.

Clark, K., Luong, M.-T., Le, Q., and Manning, C. D. (2020). Electra: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.

Conforti, C., Pilehvar, M. T., and Collier, N. (2018a). Modeling the Fake News Challenge as a Cross-Level Stance Detection Task. In *Proceedings of the 2nd International Workshop on Rumours and Deception in Social Media (RDSM 2018)*, Turin, Italy.

Conforti, C., Pilehvar, M. T., and Collier, N. (2018b). Towards Automatic Fake News Detection: Cross-Level Stance Detection in News Articles. In *Proceedings ofthe First Workshop on Fact Extraction and VERification (FEVER)*, Brussels, Belgium, Association for Computational Linguistics, pp. 40–49.

Corley, C., and Mihalcea, R. (2005). Measuring the Semantic Similarity of Texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment (EMSEE 2005)*, Ann Arbor, Michigan, USA, Association for Computational Linguistics, pp. 13–18.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Minneapolis, Minnesota, USA, Association for Computational Linguistics, pp. 4171–4186.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, Association for Computational Linguistics, pp. 350–356.

Furlan, B., Batanović, V., and Nikolić, B. (2013). Semantic similarity of short texts in languages with a deficient natural language processing support. *Decision Support Systems*, 55(3), pp. 710–719.

Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E. G. M., and Milios, E. (2006). Information Retrieval by Semantic Similarity. *International Journal on Semantic Web and Information Systems*, 2(3), pp. 55–73.

Iacobacci, I., and Navigli, R. (2019). LSTMEmbed: Learning Word and Sense Representations from a Large Semantically Annotated Corpus with Long Short-Term Memories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy, Association for Computational Linguistics, pp. 1685–1695.

Islam, A., and Inkpen, D. (2008). Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity. *ACM Transactions on Knowledge Discovery from Data*, 2(2), Article No. 10.

Jurgens, D., Pilehvar, M. T., and Navigli, R. (2014). SemEval-2014 Task 3: Cross-Level Semantic Similarity. In *Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, Association for Computational Linguistics, pp. 17–26.

Jurgens, D., Pilehvar, M. T., and Navigli, R. (2016). Cross Level Semantic Similarity: An Evaluation Framework for Universal Measures of Similarity. *Language Resources and Evaluation*, 50(1), pp. 5–33.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*, Sage, Beverly Hills, California, USA.

Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., and Crockett, K. (2006). Sentence Similarity Based on Semantic Nets and Corpus Statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8), pp. 1138–1150.

Ljubešić, N., and Lauc, D. (2021). BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2021)*, Kiev, Ukraine, Association for Computational Linguistics, pp. 37–42.

Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006)*, Boston, Massachusetts, USA, AAAI Press, pp. 775–780.

Mnasri, M., de Chalendar, G., and Ferret, O. (2017). Taking into account Inter-sentence Similarity for Update Summarization. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, Taipei, Taiwan, Association for Computational Linguistics, pp. 204–209.

Pilehvar, M. T., and Collier, N. (2016). De-conflated semantic representations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, Austin, Texas, USA, Association for Computational Linguistics, pp. 1680–1690.

Pilehvar, M. T., Jurgens, D., and Navigli, R. (2013). Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria, Association for Computational Linguistics, pp. 1341–1351.

Pilehvar, M. T., and Navigli, R. (2015). From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228 pp. 95–128.

Rekabsaz, N., Bierig, R., Lupu, M., and Hanbury, A. (2017). Toward Optimized Multimodal Concept Indexing. In *Transactions on Computational Collective Intelligence XXVI*, Springer International Publishing, pp. 144–161.

Risch, J., Möller, T., Gutsch, J., and Pietsch, M. (2021). Semantic Answer Similarity for Evaluating Question Answering Models. In *Proceedings of the Third Workshop on Machine Reading for Question Answering*, Punta Cana, Dominican Republic, Association for Computational Linguistics, pp. 149–157.

Rubenstein, H., and Goodenough, J. B. (1965). Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10), pp. 627–633.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv: 1910.03771

Zemankova, M., and Eastman, C. M. (1980). Comparative lexical analysis of FORTRAN code, code comments and English text. In *Proceedings of the 18th annual Southeast regional conference*, Tallahassee, Florida, USA, Association for Computing Machinery, pp. 193–197.

8. Language Resource References

Batanović, V., Furlan B., and Nikolić B. (2011). The Serbian Paraphrase Corpus (*paraphrase.sr*), distributed online: http://vukbatanovic.github.io/paraphrase.sr/, 1.0, ISLRN 192-200-046-033-9.

Batanović, V., Cvetanović M., and Nikolić B. (2018). The Semantic Textual Similarity News Corpus (*STS.news.sr*), distributed online: http://vukbatanovic.github.io/STS.news.sr/, 1.0, ISLRN 146-979-597-345-4.

Jurgens, D., Pilehvar, M. T., and Navigli, R. (2014). Paragraph-to-Sentence (*SemEval* 2014 Task 3: Cross-Level Semantic Similarity), distributed online: https://alt.qcri.org/semeval2014/task3/index.php?id=data-and-tools, 1.0.

Jurgens, D., Pilehvar, M. T., and Navigli, R. (2014). Sentence-to-Phrase (*SemEval* 2014 Task 3: Cross-Level Semantic Similarity), distributed online: https://alt.qcri.org/semeval2014/task3/index.php?id=data-and-tools, 1.0.