# Cross-Level Semantic Similarity in Newswire Texts and Software Code Comments: Insights from Serbian Data in the AVANTES Project

**Maja Miličević Petrović,[*] Vuk Batanović,[†] Radoslava Trnavac,[‡] Borko Kovačević[‡]**

[*] Department of Interpreting and Translation, University of Bologna
Corso della Repubblica 136, 47121 Forlì
maja.milicevic2@unibo.it
[†] Innovation Center of the School of Electrical Engineering, University of Belgrade
Bulevar kralja Aleksandra 73, 11120 Belgrade
vuk.batanovic@ic.etf.bg.ac.rs
[‡] Faculty of Philology, University of Belgrade
Studentski trg 3, 11000 Belgrade
radoslava.trnavac@fil.bg.ac.rs, borko.kovacevic@fil.bg.ac.rs

**Abstract**

This paper presents the Serbian datasets developed within the project *Advancing Novel Textual Similarity-based Solutions in Software Development* – AVANTES, intended for the study of Cross-Level Semantic Similarity (CLSS). CLSS measures the level of semantic overlap between texts of different lengths, and it also refers to the problem of establishing such a measure automatically. The problem was first formulated about a decade ago, but research on it has been sparse and limited to English. The AVANTES project aims to change this through the study of CLSS in Serbian, focusing on two different text domains – newswire and software code comments – and on two text length combinations – phrase-sentence and sentence-paragraph. We present and compare two newly created datasets, describing the process of their annotation with fine-grained semantic similarity scores, and outlining a preliminary linguistic analysis. We also give an overview of the ongoing detailed linguistic annotation targeted at detecting the core linguistic indicators of CLSS.

## 1. Introduction

One of the central meaning-related tasks in Natural Language Processing (NLP) is Semantic Textual Similarity (STS; Agirre et al., 2012). The goal of STS is to establish the extent to which the meanings of two short texts are similar to each other, which is typically encoded as a numerical score on a Likert scale. The similarity scores can subsequently be used in more complex tasks, such as Question Answering (Risch et al., 2021) or Text Summarisation (Mnasri et al., 2017).

In the related task of Cross-Level Semantic Similarity (CLSS) the goal is to contrast texts of non-matching size, such as a phrase and a sentence, or a sentence and a paragraph. CLSS was first formulated as a *SemEval* shared task by Jurgens et al. (2014), who saw it as a generalisation of STS to items of different lengths. Clearly, the length discrepancy brings an additional level of complexity, as longer texts tend to carry a greater amount of salient information than shorter texts, so CLSS can be understood as aiming to measure how well the meaning of the longer text is summarised in the shorter one.

Previous work on CLSS has generally been sparse and, to the best of our knowledge, focused entirely on English. In addition, there is a large discrepancy between the NLP models, which are based on linguistically opaque text properties, and linguistic analyses of semantic similarity. The main aim of this paper is to describe the first non-English annotated CLSS datasets, *CLSS.news.sr* and *CLSS.codecomments.sr*, developed within the project *Advancing Novel Textual Similarity-based Solutions in Software Development* – AVANTES. Both datasets comprise phrase-sentence and sentence-paragraph text pairs in Serbian and both are (being) manually annotated for CLSS. After providing some background, we describe the dataset creation and CLSS annotation, outline a preliminary linguistic analysis, and explain how the linguistic properties identified as relevant for recognising different similarity levels are being annotated further, with a view to improving linguistic descriptions of semantic similarity and testing linguistically informed NLP models.

## 2. Related work

Previous studies of CLSS are few. The NLP task was introduced by Jurgens et al. (2014, 2016), who provided the first annotated datasets for English, composed of text pairs of different lengths (paragraph to sentence, sentence to phrase, phrase to word, and word to sense), in genres including newswire, travel, scientific, review, and others. The initial datasets were re-used in subsequent work on developing and evaluating CLSS methods at different specific levels (e.g., Rekabsaz et al., 2017 for sentence to paragraph), or regardless of text length (e.g., Pilehvar and Navigli, 2015). Among related tasks, Conforti et al. (2018) dealt with the problem of cross-level stance detection, where the stance target is a sentence, and the text to be evaluated is a long document.

In Serbian, previous work on semantic similarity has been relatively limited. Batanović et al. (2011) and Furlan et al. (2013) introduced *paraphrase.sr*, a corpus of Serbian newswire texts manually annotated with binary similarity judgments; they also used it to train and evaluate several paraphrase identification approaches. Batanović et al. (2018) extended this dataset with fine-grained similarity scores, using the resulting *STS.news.sr* corpus to compare several automatic models. Finally, Batanović (2020) showed that multilingual pre-trained models such as *multilingual BERT* (Devlin et al., 2019) outperform all traditional methods, while Batanović (2021) obtained even better results using BERT's counterpart for Serbian and other closely related languages, *BERTić* (Ljubešić and Lauc, 2021).

In terms of linguistic analysis, semantic similarity is not systematically defined and described, and the contributing phenomena tend to be explored in isolation from each other

(e.g., synonymy in lexical semantics, diathesis alternations in morphosyntax). A somewhat more integrated approach is found with regard to the neighbouring notion of *paraphrase*, intended as a relation of (near-)equivalence of meaning between phrases and/or sentences (Mel'čuk, 2012: 46), i.e. as an instance of high semantic similarity (albeit a non-symmetrical one). According to Milićević (2007), paraphrases can be of different types based on the nature of information that underlies equivalence (linguistic vs. extra-linguistic), the level of linguistic representation involved (morphology, lexicon, semantics, syntax), and the depth of relation. A detailed typology of changes involved in paraphrase has been proposed by Vila Rigat (2013) and Vila et al. (2014) in view of the NLP task of automatic paraphrase detection. This typology combines several criteria and multiple levels of granularity into a taxonomy that will be presented in more detail in Section 4.2, as the basis for our linguistic analysis of CLSS.

## 3. Datasets and CLSS annotation

The corpora of phrase-sentence and sentence-paragraph text pairs presented in this paper are developed within the AVANTES project. The aim of this project is to support the analysis of correspondences between blocks of source code, written in a programming language, with an analysis of the level of semantic similarity between their respective documentation comments, written in a natural language (English or Serbian), with the goal of detecting code similarity and clones. A CLSS setup is highly appropriate for the textual similarity task due to arbitrary comment length, which can range from single words to phrases, sentences and entire paragraphs. Since the language used in comments is known to diverge from the standard language, for instance in being syntactically incomplete (Zemankova and Eastman, 1980), we add to our study setup CLSS in standard language, choosing newswire texts as its representative.

In the context of the project, comparative analyses are planned both between text domains and between languages. For this reason, it was important to establish a common methodology for the creation and annotation of datasets. Since the only pre-existing CLSS dataset was the *SemEval* one for English, we adopted the approach of Jurgens et al. (2014) as a (partial) model for our work. We retained their five-point similarity scale, with scores ranging from 0 to 4, as well as their definitions for each score: 0 – unrelated, 1 – slightly related, 2 – somewhat related but not similar, 3 – somewhat similar, 4 – very similar. However, we altered the method of text pair construction. Namely, while Jurgens et al. (2014) provided annotators with a longer text and asked them to generate a shorter one with a designated similarity score in mind, we pre-prepared numerous text samples of different lengths (phrases, sentences, and paragraphs), and asked the annotators to combine these texts into phrase-sentence and sentence-paragraph pairs, aiming for a balanced score distribution for the pairs they construct. The main motivation for this choice was that the generation of texts by annotators would have been very difficult to implement in the domain of source code comments, given the highly technical and often project-specific terminology encountered in them. At the same time, our approach prevented a potential paraphrasing bias that the annotators could inadvertently introduce.

### 3.1. *CLSS.news.sr*

The initial texts for the *CLSS.news.sr* dataset were obtained from the Serbian news aggregator website *naslovi.net*. This website provides a headline and an introductory paragraph for each news report; a subhead is frequently included too. 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, exploiting the journalistic convention that the beginning sections in an article commonly provide a summary of its content; our approach was the same one used in the construction of multiple other newswire STS and paraphrasing corpora (Dolan et al., 2004). Since news item are commonly reported differently by different media outlets, cross-linking the texts of different reports allowed for the creation of text pairs with varying degrees of semantic similarity. Close to 18,000 news reports, published between June and August 2021, were scraped using the *scrapy* Python library,[1] to ensure the annotators had a sufficient quantity of raw text available for creating adequate pairs. To ensure comparability with the *SemEval* dataset, our target dataset size was 1,000 phrase-sentence and 1,000 sentence-paragraph pairs.

The construction of the 2,000 text pairs was divided between five annotators, who were either trained linguists or had previous experience with text annotation for the closely related STS task. Even though they received text samples pre-classified based on length, they were instructed to evaluate whether an item in a certain category really was a phrase, a sentence, or a paragraph, and were allowed to change the categorisation. Paragraphs were defined 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 (non-finite forms such as infinitives and participles were allowed, as were deverbal nouns).

The annotators were provided with the similarity score definitions and *SemEval* examples to help them interpret each score. Since these examples proved insufficient to ensure high annotation consistency, the outputs were calibrated by having all annotators create a smaller set of five to six representative pairs for each similarity score and each length pairing. These pairs were reviewed by project researchers and feedback was provided regarding any issues encountered. The following step was the compilation of a detailed set of examples, three per similarity score and length pairing, using the agreed upon representative pairs from all annotators. This set, the score definitions and general instructions became an integral part of the final annotation guidelines for our task, available in the dataset repository in Serbian (original) and English (translation).[2] A subset of examples is shown in Table 1.

The annotators were subsequently asked to construct a total of 200 pairs for each text length combination, trying to include both pairs clearly corresponding to a specific score, and less clear-cut ones. The resulting 2,000 cross-level text pairs were labelled with semantic similarity scores by all five annotators, using the *STSAnno* tool (Batanović et al., 2018). The final score for each pair was calculated by averaging the scores of all individual annotators. Obtaining multiple parallel annotations and

[1] http://scrapy.org/

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

averaging them out was chosen instead of relying on an adjudicated double annotation (used for the *SemEval* dataset) in order to minimise individual annotator's biases. In addition, while Jurgens et al. (2014) allowed finer-grained score distinctions using multiples of 0.25, in our setup with five annotators this was not necessary.

| Score | Examples |
|---|---|
| 4 | Veliki požar na železničkoj stanici u Londonu<br>*A large fire at a London railway station* |
| | Veliki požar izbio je danas na metro stanici u centralnom delu Londona.<br>*A large fire broke out today at an underground station in central London.* |
| 3 | Novi nacionalni praznik: Džuntint<br>*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.<br>*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<br>*A major problem for Portugal* |
| | Loše vesti stižu za Portugal pred start Evropskog prvenstva.<br>*Bad news arrives for Portugal just before the start of the European Championship.* |
| 1 | Svađa pred svadbu<br>*A pre-wedding argument* |
| | Mirko Šijan i Bojana Rodić uskoro očekuju svoje prvo dete, a uveliko se sprema i njihova svadba.<br>*Mirko Šijan and Bojana Rodić are expecting their first child soon, and their wedding is being prepared.* |
| 0 | Otvaranje silosa u Zrenjaninu<br>*A silo opening in Zrenjanin* |
| | Maja Žeželj, voditeljka, ispričala je kako je svojevremeno jedva izvukla živu glavu.<br>*Maja Žeželj, TV presenter, told the story of how some time ago she nearly died.* |

Table 1: Guideline examples of phrase-sentence pairs in the newswire dataset for each similarity score.

The final *CLSS.news.sr* dataset comprises 30 thousand tokens in the phrase-sentence subset, and 86 thousand tokens in the sentence-paragraph subset. The average sentence length is ~22 tokens in the sentence-paragraph pairs and ~23 tokens in the phrase-sentence ones. The average phrase length is ~6 tokens, while the average paragraph length is ~64 tokens. The average similarity scores are close to the scale's mean value of 2: 1.91 in the sentence-paragraph subset, and 1.96 in the phrase-sentence subset. The distribution of different scores is fairly uniform, especially for the phrase-sentence pairs; the peaks include a marked one around 0, and a less evident one around 3. The annotation (self-)agreement levels are very high. For the phrase-sentence subset, the average binary agreement between each annotator and the mean of other annotators' scores yields a Krippendorff's alpha coefficient of $\alpha = 0.929$, while the Pearson and the Spearman correlation coefficients are equal, $r = \rho = 0.938$. In the case of sentence-paragraph pairs these values are $\alpha = 0.922$, $r = 0.937$ and $\rho = 0.934$. More details and a comparison with the English *SemEval* dataset are reported in Batanović and Miličević Petrović (2022).

### 3.2. *CLSS.codecomments.sr*

A particularly innovative part of the work conducted in the AVANTES project is the creation of a corpus of software code comments, to be made publicly available for download and use in testing NLP models once the annotation of semantic similarity is completed. The sources that the code comment dataset was drawn from include public repositories such as GitHub, student projects, coursework and teaching materials from various computing courses at the School of Electrical Engineering of the University of Belgrade and other academic institutions in Serbia, as well as software projects developed at the Computing Center of the School of Electrical Engineering. In order to prevent our work from being focused on the specificities of a single programming language or programming paradigm, we opted to collect comments from eight programming languages: C, C++, C#, Java, JavaScript/TypeScript, MATLAB, Python, and SQL.

We focused on manually pre-selecting only those code comments that describe the functionality of particular sections of code, ranging from individual code lines, to methods and functions, to classes and entire modules. To do so, we relied on a newly designed taxonomy for differentiating between types of code comments (Kostić et al., 2022), which includes the following code comment categories: Code, Functional-Inline, Functional-Method, Functional-Module, General, IDE, Notice and ToDo. The initial data collection and pre-selection were performed by master's degree students at the School of Electrical Engineering of the University of Belgrade, as part of their course project for the Natural Language Processing course. In total, after all duplicate entries were removed, 9,395 code comments belonging to the Functional categories were identified. These include 6,455 Functional-Inline comments, which describe the functionality of individual code lines or code passages, 1,829 Functional-Method comments, which address the functionality of functions and class methods, and 1,111 Functional-Module comments, which are related to the functionality of entire code modules and classes.

In order to construct text pairs, the comments were first roughly divided into candidates for phrases, sentences, and paragraphs on the basis of a set of heuristics. Using whitespace tokenisation, we treated all texts with up to six tokens as candidates for phrases. All texts containing more than six tokens, but limited to a single sentence, were treated as candidates for sentences, while those with more than one sentence were considered paragraph candidates. The number of sentences was determined using a regular expression that treated question marks, exclamation marks, and periods outside of URLs and decimal numbers as sentence boundaries. Using this procedure, the text set was divided into 4,880 phrase candidates, 3,592 sentence candidates, and 923 paragraph candidates.

Due to the high domain specificity of code comments, we entrusted the creation of CLSS pairs to two experienced

programmers. They used the provided candidate texts to form the pairs, but were instructed to carefully evaluate whether each sample truly belonged to its automatically assigned length grouping. Such an evaluation was necessary because complete standard sentences and paragraphs were rarely encountered in the data. Instead, we found that despite having a sentence-like function in the comment, many texts are not true sentences in the linguistic sense – they do not follow any punctuation rules and they lack a predicate, or possess it only implicitly (e.g., *@author Tim 2* or *Naziv komponente* 'Component name' within a paragraph item). Similarly, paragraphs in the code comment domain are often separated into units not via standard punctuation, but rather by using visual boundaries, such as moving to a new line in the source file, or (repeatedly) using special characters (e.g., * or ###). Limiting our text selection to a rigid definition of sentences and paragraphs would thus not only have reduced the size of the dataset, but it would also have led to the exclusion of numerous domain-specific phenomena, significantly impacting our linguistic analyses of code comments. We therefore decided to count as paragraphs texts consisting of at least two clearly identifiable units, even if those units were not true sentences. Similarly, we expanded the sentence set with texts containing an implicit predicate, as well as with those containing subordinate clauses without a main clause (e.g., relative clauses such as: *Metode koje se odnose na simulaciju procesa* 'Methods that refer to process simulation').

| Score | Examples |
|---|---|
| 4 | Računanje površine pravougaonika<br>*Calculating the area of a rectangle* |
| | Površina pravougaonika po formuli je a * b<br>*The area of a rectangle according to the formula is a * b* |
| 3 | POMOCNA FUNKCIJA<br>*AUXILIARY FUNCTION* |
| | Fajl koji pruza pomocne funkcije<br>*A file that provides auxiliary functions* |
| 2 | ubrzano kretanje<br>*accelerated movement* |
| | Zelimo da se ogranicimo od mogucnosti da se ubrzano krece.<br>*We want to limit the possibility of accelerated movement.* |
| 1 | Update dokumenta<br>*Document update* |
| | Ovaj program formira html dokument<br>*This program forms an html document* |
| 0 | izracunavanje faktorijela<br>*calculating the factorial* |
| | Azurira rotaciju kamere preko pomeraja misa<br>*Updates the camera rotation via mouse movement* |

Table 2: Guideline examples of phrase-sentence pairs in the code comment dataset for each similarity score.

This allowed us to construct a code comment dataset of the same size as *CLSS.news.sr*. The *CLSS.codecomments.sr* dataset therefore includes 1,000 phrase-sentence pairs, comprising 14 thousand tokens, and 1,000 sentence-paragraph pairs, comprising 39 thousand tokens. The average sentence length is ~10 tokens in both the sentence-paragraph and the phrase-sentence pairs. The average phrase length is ~3 tokens, while the average paragraph length is ~29 tokens. Overall, the code comments are approximately half the length of the newswire text items.

Although our initial aim was again to construct a dataset balanced across the range of similarity scores, this proved to be impossible with our selection of source texts, since they pertained to a wide range of programming projects with different purposes and implemented using diverse programming paradigms and languages. This made the construction of pairs with high similarity scores very problematic. We therefore abandoned the goal of obtaining a balanced score distribution, but still instructed the programmers to compile as many highly similar pairs as possible with the given source content. Each programmer was tasked with the construction and scoring of 500 pairs of each length.

The similarity scoring of the text pairs was performed on the basis of guidelines similar to the ones used in the newswire domain, but with a new set of three examples per score and length pairing, drawn from the code comment domain; a subset of phrase-sentence pair examples is shown in Table 2. After the code comment text pairs were constructed, they were forwarded to the same annotators who worked on the *CLSS.news.sr* dataset, in order to obtain multiple parallel annotations. Since this work is still in progress, our linguistic analyses of *CLSS.codecomments.sr* in this paper will be based on the individual similarity scores assigned by the two programmers who constructed the text pairs.

## 4. Linguistic analysis

The NLP algorithms used in automatic treatment of semantic similarity rely on different types of information, including linguistic features. While state-of-the-art models such as *multilingual BERT* and *BERTić* reach performances that correlate highly with human scores, with coefficients $r,\rho > 0.9$ for CLSS on Serbian newswire texts (Batanović and Miličević Petrović, 2022), they lack linguistic transparency and are of limited help in understanding the relative contributions of different levels of language structure and different specific features. Since one of the aims of the AVANTES project is to combine NLP with linguistic knowledge, we conduct two types of linguistic analyses on the datasets. A preliminary qualitative analysis is performed to gain initial insight into the data and help decide on the specifics of detailed annotation of semantic similarity indicators (to be followed by a quantitative analysis of the annotated datasets).

### 4.1. A qualitative overview

A qualitative linguistic analysis was performed on a random sample of ten text pairs per score, for both *CLSS.news.sr* and *CLSS.codecomments.sr*, and for both phrase-sentence and sentence-paragraph pairs. In the case of newswire texts, items that received the same score by all annotators were selected; an approach focused on clear-cut cases was deemed useful as a first step in the analysis given its goals of verifying both the linguistic relevance of the similarity scores and the taxonomy for more detailed linguistic annotation. For comments, the initial scores

assigned by programmers were used for selection. The analysis consisted in a comparison of information content between the pairs' components, as well as a study of vocabulary overlaps (or lack thereof). Its goal was to get an initial grasp of the data and help define a taxonomy to base a more elaborate analysis on.

For both corpora and both types of comparisons, the pairs marked 4 are characterised by the occurrence of the same distinctive vocabulary items: personal names and/or numbers (newswire), or specialised terms (comments). The form is often not identical, but the items involved are clearly relatable on morphological grounds (e.g., they are inflectional forms of the same noun, as in *Kragujevcu*.LOC – *Kragujevca*.GEN 'Kragujevac', *parametre*.ACC – *parametrima*.INS 'parameters', or a noun and a denominal adjective, as in *Vlasotincu*.N – *vlasotinačkom*.ADJ '(of) Vlasotince')[3]. The shared numbers are mostly large and either quite specific or used in a collocation (e.g., *100.620*, or *3.000 dinara* '3000 dinars'). Overlaps in common lexical words are also frequently based on morphologically related rather than identical forms (e.g., *stiglo*.PAST.PART – *stići*.INF 'arrive', *novozaraženih* 'newly infected' – *novih slučajeva zaraze* 'new cases of infection', *filtriranje* 'filtering' – *filtar* 'filter'). A number of synonyms are found (*potvrda – sertifikat* 'certificate', *promenljiva – varijabla* 'variable'), sometimes involving a Serbian and an English word (*mreža – grid* 'grid'), and sometimes within different collocations based on the same term (e.g., *toplotni talas – talas vrućina* 'heat wave', *zoom levela – stepena zoom-a* 'zoom level'). Overall, most lexical words from the smaller unit are present in the larger one, which also contains other elements that describe the situation in more detail, but without adding entirely new topics (*u Londonu* 'in London' – *u centralnom delu Londona* 'in central London'; *funkcija sa parametrima* 'a function with parameters' – *funkcija koja nije f(void), vec prima parametre* 'a function that is not f(void), but accepts parameters').

Score 3 items are distinguished by similar properties in terms of shared lexis and especially personal names and specialised terms, but with entirely new information in the longer item, and/or partly different information in the components of the pair, leading to a less marked overall vocabulary overlap (e.g., *Neuralna mreza* 'neural network' – *vanila neuralna mreza koja se obucava pomocu genetskog algoritma* 'vanilla neural network which is trained via a genetic algorithm'). Near-synonyms appear to be more common in score 3 pairs (*reč* 'word' – *termin* 'term', *nov ugovor* 'new contract' – *produžetak saradnje* 'extension of collaboration'). In both score 4 and score 3 items, the head noun of the phrase tends to appear as the subject or the object of the sentence predicate, or it is a deverbal noun that corresponds to the predicate (*unos*.N – *unosi*.V 'input'). The predicate is typically the same in sentence-paragraph pairs, with additional predicates in the paragraph item.

Among less similar pairs, those marked 2 are somewhat mixed, as they either contain different personal names/ specialised terms and similar common vocabulary, or vice versa (*Tropski pakao u Beogradu* 'tropical hell in Belgrade' – *I sutra će u Novom Sadu biti veoma toplo* 'It will again be very warm in Novi Sad tomorrow'; *prekid rekurzije* 'interruption of recursion' – *ako ima decu onda idemo rekurzivni poziv* 'if it has children then we do a recursive call'). The predicate of the sentence item is typically not related to the head noun of the phrase item. The pairs marked 1 and 0 contain barely any overlapping personal names or specialised terms. Score 1 items do share some common lexical words, but synonyms, near-synonyms, and terms from the same wider semantic field are more present than words that are identical or morphologically closely related (e.g., *tragedija* 'tragedy' – *nesreća* 'accident', *pljuskovi* 'showers' – *kiša* 'rain'). Items marked 0 typically do not share any lexical words.

When it comes to differences between the two corpora, in *CLSS.news.sr* it is often the case that the relatedness of lexical items in the pair is based on real world knowledge (largely about something happening at the time of writing) rather than on linguistic information (e.g., *vakcinacija* 'vaccination' – *virus korona* 'corona virus', *Tokio* 'Tokio' – *Olimpijske igre* 'Olympic games'), especially in items assigned a score below 3. *CLSS.codecomments.sr*, on the other hand, is characterised by various non-standard features, such as inconsistent spelling (*popup* vs. *pop-up*), missing diacritics (*cita* for *čita* 'reads'), inflectional endings on English words inconsistently spelt with/without a dash (*zoom-a*, *workspace-u* vs. *levela*), non-standard abbreviations (*f-ja* for *funkcija* 'function)', or phonetic transcription of English terms (*eksepšn* 'exception').[4]

## 4.2. Linguistic annotation

Using the preliminary analysis outlined above and the existing paraphrase typologies (primarily Vila Rigat, 2013; Vila et al., 2014; also Milićević, 2007; Mel'čuk, 2012), we propose a taxonomy of semantic similarity types and indicators, shown and illustrated in Table 3; most examples are taken directly or adapted from our corpora (examples for two clear indicators are omitted to save space). The initial focus is on the nature of information that similarity is based on, and a core distinction is made between linguistic, quasi-linguistic and extralinguistic similarity types. This is at the same time one of the main points of divergence between our approach and the one by Vila Rigat (2013) and Vila et al. (2014), who acknowledge the existence of non-linguistic paraphrase, but do not include it in their core typology; we rely on Milićević (2007) and Mel'čuk (2012) for these types. Another difference with respect to previous work is that our taxonomy makes reference to similarity *indicators*, while *changes* are invoked in previous work, due to paraphrase being perceived as involving a source and a target item.

Linguistic similarity is based on language-internal information at the word/lexical unit level (i.e., the morpho-lexicon), the level of structural organisation, and the level of meaning (i.e., semantics). The first two types have two subtypes each: morphology- and lexicon-based and syntax- and discourse-based indicators respectively; the indicator types and subtypes thus follow the classical organisation in formal levels of linguistic analysis. Finally, the indicator names in the last column of Table 3 denote specific mechanisms through which semantic similarity is established. Following Vila et al. (2014), our assumption is that the indicators reveal what triggers semantic similarity at the micro level. In other words, unlike the similarity

[3] Abbreviations used: LOC – locative; GEN – genitive; ACC – accusative; INS – instrumental; ADJ – adjective; N – noun, PAST.PART – past participle; INF – infinitive; V – verb.

[4] Many of the features found in code comments are shared with computer-mediated communication in Serbian (see Miličević Petrović et al., 2017).

scores assigned to pairs of items as wholes (i.e., to entire phrases, sentences, or paragraphs), the linguistic taxonomy targets individual phenomena that cumulatively contribute to the overall score, where such individual elements are not mutually exclusive and several can be co-present.

Looking more closely at the indicator subtypes, morphology-based indicators concern the morphological form of words, capturing complete equivalence, as well as inflectional and derivational relations, i.e. different forms of the same word or changes of category via derivational

| **Similarity type** | **Indicator type** | **Indicator subtype** | **Indicator (*example*)** |
|---|---|---|---|
| Linguistic | Morpholexicon-based | Morphology-based | - Identical (*požar – požar* 'fire')<br>- Inflectional (*parametre*.ACC – *parametrima*.INS 'parameters')<br>- Derivational (*Vlasotincu*.N – *vlasotinačkom*.ADJ '(of) Vlasotince') |
| | | Lexicon-based | - Spelling and format (*pop-up – popup*)<br>- Synthetic/analytic (*novozaraženih* 'newly infected' – *novih slučajeva zaraze* 'new cases of infection')<br>- Same polarity<br>-- Synonymy (*potvrda – sertifikat* 'certificate)<br>-- Near-synonymy (*reč* 'word' – *termin* 'term')<br>-- Hyponymy (*škoda* 'Škoda' – *automobil* 'car')<br>-- Meronymy (*Vašington* 'Washington' – *SAD* 'USA')<br>- Opposite polarity (*izgubio* 'lost' – *nije uspeo da pobedi* 'failed to win')<br>- Converse (*pogibija dva pešaka* 'death of two pedestrians' – *usmrtio pešake* 'killed the pedestrians') |
| | Structure-based | Syntax-based | - Diathesis alternations (*opljačkali su stan* 'robbed the flat' – *stan je opljačkan* 'the flat was robbed')<br>- Coordination changes<br>- Subordination and nesting changes |
| | | Discourse-based | - Punctuation (*Potpis dana - Aleksandar Kolarov!* 'Signature of the day - Aleksandar Kolarov!' – *Aleksandar Kolarov potpisao novi ugovor* 'Aleksandar Kolarov signed a new contract')<br>- Direct/indirect style (*Bilčik ocenjuje da vežbe ne pomažu* 'Bilčík states that the military exercises do not help' – *Bilčik ukazuje da vesti o vežbi "nisu od pomoći"* 'Bilčík points out that the news of a military exercise "is not helpful"')<br>- Sentence modality (*maske više nisu obavezne?* 'masks no longer compulsory?' – *neće biti obavezne zaštitne maske* 'protective masks will not be compulsory') |
| | Semantics-based | | (*Tropski pakao* 'tropical hell' – *biti veoma toplo* 'be very warm') |
| | Miscellaneous | | - Change of order (*klasa singleton – Singleton patern* 'singleton class/pattern')<br>- Addition/deletion (*funkcija za sortiranje* 'sorting function' – *metoda koja sortira uzetu matricu* 'the method that sorts the given matrix') |
| Quasi-linguistic | Pragmatic | | (*Scattered showers are very likely – Bring your umbrella*; Mel'čuk, 2012: 60) |
| Extralinguistic | Situational | | (*Besplatno kroz Severnu Makedoniju od danas* 'Free travel through North Macedonia from today' – *Novina od 15. juna* 'New rules from 15 June') |
| | Encyclopaedic | | (*Italija* 'Italy (the team)' – *ekipa sa Apenina* 'the team from the Apennine Mountains') |
| | Logical | | (*Još pola dinara za veknu hleba* 'Half a dinar more for a loaf of bread' – *Cena hleba visa za 20%* 'The price of bread higher by 20%'; Milićević, 2007: 145) |

Table 3: Overview of the taxonomy of semantic similarity (the examples are drawn from *CLSS.news.sr/CLSS.codecomments.sr*, or from the literature).

affixes. The *identical* indicator is not present under the morphology heading in Vila Rigat (2013) and Vila et al. (2014), who categorise it as a "paraphrase extreme", which is a special type in their taxonomy, capturing longer chunks of text; we add it based on the preliminary analysis presented in Section 4.1, which revealed that identical individual words are common in highly similar items in CLSS. Additional information that could prove useful concerns parts of speech, the distinction between personal and common nouns, as well as information on general vs. specialised vocabulary. Given that the identification of specialised terminology would require work that goes beyond the scope of the current project, we are still evaluating the possibility of including it in the analysis.

Lexicon-based indicators are somewhat more varied, ranging from different spellings of the same words, to syntactic and analytic expressions of the same meaning, and to lexical semantic relations in the narrow sense. Same polarity items constitute the most complex group of lexical relations, comprising synonymy as a similarity relation *par excellence*, near-synonymy, hyponymy (the relationship between superordinate/more general and subordinate/more specific lexical items), and meronymy (a part-whole relation). Opposite polarity relations are based on antonym pairs with opposite comparative words, or with one of the components negated. Finally, a converse relation captures complementary actions whose arguments are inversed.

Syntax-based indicators capture those relations that imply a syntactic reorganisation in the sentence; they can be found within single sentences, or in the way multiple sentences are connected. Specific cases include instances of diathesis alternations (such as the active/passive alternation), coordination (where coordinated units are present in one member of the pair, but not in the other), and subordination or nesting (where subordinate/nested elements are present in only one item). The second subtype of structural changes, discourse-based indicators, do not affect the sentential arguments, but are instead related to elements such as punctuation and formatting (beyond single lexical units), affirmative vs. interrogative sentence modality, and direct vs. indirect speech.

The semantics-based subtype is also distinguished by going beyond the level of individual lexical items, as it concerns phrase/sentence-level meaning. No subtypes of specific indicators are singled out, as this level of analysis refers generally to the distribution of semantic content across lexical units, and it can involve multiple and varied formal changes that lead to different lexicalisations of the same meaning units. The boundaries between semantics-based similarity and lexicon-based similarity indicators are not always clear-cut, but it is generally the case that lexicon-based indicators concern individual words or multiword units, while semantics-based similarity relies on multiple lexical items.

The last type of linguistic indicators is classified as miscellaneous, given that it captures phenomena that do concern the linguistic structure of items, but do not clearly belong to a single level of linguistic analysis. Change of order and addition/deletion are found here as specific indicator types, the former involving units with the same content expressed using different word orders, and the latter based on added or omitted information. Both indicators concern at least syntax and discourse; given the cross-level setup, the latter is particularly important for our datasets.

Beyond the linguistic structure, the quasi-linguistic domain captures inference-based similarity that relies on pragmatic information. The core linguistic meanings and the extralinguistic referents are different in this case, but the meaning of one element in the pair can still be inferred from the meaning of the other. Given the nature of our texts, this type of similarity is expected to be infrequent, and we have so far not identified any examples; however, we leave this category in our taxonomy to possibly be applied in the annotation phase. The extralinguistic domain also entails inequality of linguistic meaning, but it involves information equivalence between two texts, i.e. reference to the same real-world situation. It requires knowledge external to language for similarity to be recognised; this knowledge can be situational (containing elements such as *today* or *here*), encyclopaedic (involving general knowledge), or logical (requiring calculations or other similar operations). Based on the initial analyses of our datasets, this is a common type of similarity, especially in newswire texts.

Keeping the above definitions in mind, the outlined taxonomy will be applied to the *CLSS.news.sr* and *CLSS.codecomments.sr* corpora. Detailed guidelines are currently being developed, and the texts (initially from *CLSS.news.sr*) are being prepared for word/segment-level annotation with semantic similarity indicators, within the identified pairs. The annotation will be performed by the project researchers, first as a double procedure on a smaller sample, and then individually once a satisfactory level of agreement is reached. The initial phase will at the same time enable us to verify the appropriateness of the taxonomy, and adapt it should the need arise. The annotated datasets will be used for empirically validating the taxonomy, for gaining a better understanding of the linguistic factors that carry the most weight in cross-level semantic similarity in different text genres, and for learning how this kind of information can be taken into account in NLP models. Based on previous work on paraphrase and a preliminary exploration of our data at text level (with entire pairs marked for indicator presence/absence), morphological indicators, addition/deletion and same polarity items are expected to be particularly prominent.

## 5. Concluding remarks

In this paper, we have described the first non-English CLSS corpora, *CLSS.news.sr* and *CLSS.codecomments.sr*. The focus was on the methodology used to construct and annotate the data, as well as on their initial linguistic analysis. We believe these two datasets to be an important resource for Cross-Level Semantic Similarity research, not only in virtue of representing a new language, but also due to introducing an underexplored text genre (source code comments), and due to dedicating substantial attention to the linguistic properties of the datasets.

Our planned next steps are to complete the CLSS annotation of code comments, implement the proposed linguistic taxonomy of semantic similarity in the annotation of both datasets, conduct a more extensive linguistic analysis based on the annotated data, and examine the impact of linguistic traits on the performances of automatic CLSS models. Another goal is to compare the results to those obtained on similar datasets for English, using the *SemEval* dataset for newswire, and our own dataset (which is currently being created) for source code comments.

## 6. Acknowledgements

The AVANTES project (*Advancing Novel Textual Similarity-based Solutions in Software Development*) is supported by the Science Fund of the Republic of Serbia, grant no. 6526093, within the "Program for Development of Projects in the Field of Artificial Intelligence". The authors would like to thank Jelica Cincović and Dušan Stojković for constructing the code comment text pairs, as well as Bojan Jakovljević, Lazar Milić, Marija Lazarević, Ognjen Krešić, and Vanja Miljković for annotating the corpora with semantic similarity scores.

## 7. References

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

Vuk Batanović. 2020. *A Methodology for Solving Semantic Tasks in the Processing of Short Texts Written in Natural Languages with Limited Resources*. Ph.D. thesis, University of Belgrade.

Vuk Batanović. 2021. Semantic similarity and sentiment analysis of short texts in Serbian. In: *Proceedings of the 29th Telecommunications forum (TELFOR 2021)*, Belgrade, Serbia, IEEE.

Vuk Batanović and Maja Miličević Petrović. 2022. Cross-Level Semantic Similarity for Serbian Newswire Texts. In: *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France. European Language Resources Association.

Vuk Batanović, Miloš Cvetanović, and Boško Nikolić. 2018. Fine-grained Semantic Textual Similarity for Serbian. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1370–78, Miyazaki, Japan, European Language Resources Association.

Vuk Batanović, Bojan Furlan, and Boško Nikolić. 2011. A software system for determining the semantic similarity of short texts in Serbian. In: *Proceedings of the 19th Telecommunications forum (TELFOR 2011)*, pages 1249–52, Belgrade, Serbia, IEEE.

Costanza Conforti, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Towards automatic fake news detection: Cross-level stance detection in news articles. In: *Proceedings of the First Workshop on Fact Extraction and VERification*, pages 40–49, Brussels, Belgium, Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of NAACL-HLT 2019*, pages 4171–86, Minneapolis, Minnesota, USA, Association for Computational Linguistics.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora. In: *Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–56, Geneva, Switzerland, Association for Computational Linguistics.

Bojan Furlan, Vuk Batanović, and Boško Nikolić. 2013. Semantic similarity of short texts in languages with a deficient natural language processing support. *Decision Support Systems*, 55(3):710–19.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. SemEval-2014 Task 3: Cross-Level Semantic Similarity. In: *Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval 2014)*, pages 17–26, Dublin, Ireland. Association for Computational Linguistics.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Cross Level Semantic Similarity: An Evaluation Framework for Universal Measures of Similarity. *Language Resources and Evaluation*, 50(1):5–33.

Marija Kostić, Aleksa Srbljanović, Vuk Batanović, and Boško Nikolić. 2022. Code Comment Classification Taxonomies. In: *Proceedings of the Ninth IcETRAN Conference*, Novi Pazar, Serbia.

Nikola Ljubešić and Davor Lauc. 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)*, pages 37–42, Kiev, Ukraine, Association for Computational Linguistics.

Igor A. Mel'čuk. 2012. *Semantics. From Meaning to Text*. John Benjamins, Amsterdam.

Maja Miličević Petrović, Nikola Ljubešić, and Darja Fišer. 2017. Nestandardno zapisivanje srpskog jezika na Tviteru: mnogo buke oko malo odstupanja? *Anali Filološkog fakulteta* 29(2):111–36.

Jasmina Milićević. 2007. *La paraphrase*. Peter Lang, Bern.

Maâli Mnasri, Gaël de Chalendar, and Olivier Ferret. 2017. Taking into account Inter-sentence Similarity for Update Summarization. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 204–209, Taipei, Taiwan. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Roberto Navigli. 2015. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228:95–128.

Navid Rekabsaz, Ralf Bierig, Mihai Lupu, and Allan Hanbury. 2017. Toward optimized multimodal concept indexing. In: N. Nguyen, R. Kowalczyk, A. Pinto, and J. Cardoso, eds., *Transactions on Computational Collective Intelligence XXVI*, pages 144–61, Cham, Springer International Publishing.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In: *Proceedings of the Third Workshop on Machine Reading for Question Answering*, pages 149–57, Punta Cana, Dominican Republic, Association for Computational Linguistics.

Marta Vila Rigat. 2013. *Paraphrase Scope and Typology. A Data-Driven Approach from Computational Linguistics*. Ph.D. thesis, University of Barcelona.

Marta Vila, M. Antonia Martí, and Horacio Rodríguez 2014. Is this a paraphrase? What kind? Paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, 4:205–18.

Marie Zemankova and Caroline M. Eastman. 1980. Comparative lexical analysis of FORTRAN code, code comments and English text. In: *Proceedings of the 18th annual Southeast regional conference*, pages 193–97, Tallahassee, Florida, USA, Association for Computing Machinery.