Lexicon-Based Exploration of Emotion Dynamics in TEI Novels: Case Study on *SrpELTeC* Corpus

Ranka Stanković[1] [0000-0001-5123-6273], Milena Šošić[2] [0000-0002-9949-9686]

[1]University of Belgrade, Faculty of Mining and Geology, Serbia,
ranka.stankovic@rgf.bg.ac.rs

[2] Language Resources and Technologies Society – JeRTeh, Serbia,
milena.sosic@jerteh.rs

Scientific paper DOI: 10.18485/judig.2026.2.ch8

Abstract

This paper presents a hierarchical analysis of emotional dynamics in the *SrpELTeC* corpus of Serbian novels from the period 1850–1920s. The study combines TEI-encoded texts with a Serbian-tailored emotion lexicon *EmoLex.SR*, based on Plutchik's model of basic emotions. Emotional signals from the corpus were extracted through lexical matching and analyzed across multiple textual hierarchically organized segments – words, sentences, paragraphs, chapters, literary works and the full corpus, which enabled the exploration of both local variation and global patterns of emotional signals. The approach combines emotional words density and intensity measures with hierarchical and chronological grouping, to examine short- and long-term changes in emotional structure across the corpus. First empirical results showed a stable baseline dominated by `trust` and `anticipation` emotions, alongside with the significantly increased emotional variation in the early twentieth century. These findings indicate that emotional organization in Serbian prose evolves in both magnitude and structural complexity over time and that proposed methodology enables consistent comparison and clear computational visibility across literary work segments and time periods.

Keywords: emotions, Plutchik model, lexicon, Serbian literary corpus

1. Introduction

The increasing availability of digitized literary collections has transformed computational literary studies by enabling systematic quantitative investigation of narrative structures, stylistic patterns and latent psychological dimensions embedded within textual corpora. Among these research directions, computational emotion analysis has emerged as a particularly important area because emotional structures represent fundamental mechanisms through which literary narratives construct characters, organize conflicts, communicate cultural values and shape reader interpretation.

While traditional literary scholarship primarily approaches emotional expression through close reading methodologies, computational approaches allow investigation of affective phenomena across substantially larger textual collections and multiple levels of narrative organization. However, literary texts introduce substantial methodological challenges for computational analysis due to complex narrative structures, figurative language, varying stylistic conventions and strong dependencies between local and global textual contexts.

This study presents a hierarchical methodology for computational emotion analysis of the Serbian literary corpus *SrpELTeC* (Krstev, 2021; Stanković et al., 2024). The analysis is grounded in Plutchik's theory of basic emotions (Plutchik, 2001), which defines eight primary emotion categories: `anger`, `anticipation`, `disgust`, `fear`, `joy`, `sadness`, `surprise` and `trust`. These categories are operationalized through the Serbian emotion lexicon *EmoLex.SR*, which provides a structured mapping from lexical items to Plutchik's emotional space (Šošić et al., 2026). In contrast to conventional sentiment analysis approaches that reduce emotional content to aggregated polarity measures, the proposed methodology models emotional structures across multiple textual scales, including corpus, literary work, chapter, paragraph, sentence and lexical lemmas levels. Such hierarchical representation enables investigation of both global affective tendencies and local emotional transitions across the *SrpELTeC* corpus.

In psychology, linguistics and affective computing, the term *affective* is generally used as a broader concept than *emotional*. While *emotional* typically refers to discrete and explicitly identifiable emotions such as `anger`, `joy`, `fear` or `sadness`, *affective* encompasses a wider spectrum of subjective states, including emotions, moods, sen-

timents, attitudes and general emotional tone or atmosphere. Consequently, emotion analysis usually focuses on recognizing specific emotional categories, whereas affective analysis often addresses broader experiential and interpretative dimensions of human feeling and emotional expression. This distinction is particularly important in literary and narrative analysis, where emotional signals may function not only as isolated emotions but also as components of larger affective structures and atmospheres within texts (Scherer, 2005).

The primary contribution of this work lies in introducing a multi-scale emotion detection in TEI-encoded novels, s a case-study for emotion lexicon use. The proposed methodology combines linguistically adapted emotional resource with hierarchical text segmentation (novel, chapter, paragraph, sentence) in order to systematically analyze emotional distributions and the diversity of emotional expression across the examined literary corpus.

The remainder of this paper is organized as follows. Section 3. describes the *SrpELTeC* corpus as main dataset, *EmoLex.SR* as the main lexical resource, followed by lexical emotions mapping. Section 4. presents the micro-level emotional patterns and macro-level emotional tendencies, including results on emotional word density and normalized emotional vector distributions in the *SrpELTeC* corpus. Section 5. discusses that hierarchical lexicon-based emotional analysis captures both local and larger-scale emotional dynamics, revealing increasing emotional diversification and narrative specialization of the novel, while also highlighting limitations related to the required detailed evaluation. Finally, Section 6. summarizes the main contributions of the study and outlines directions for future work, including the evaluation of automatic emotion annotation quality, training and fine-tuning of different models and refinement of the lexicon.

2. Related Work

Emotion analysis has become an important research direction within computational linguistics, natural language processing and digital humanities (Kim and Klinger, 2021). Early approaches primarily focused on polarity-based sentiment analysis, reducing textual affect to binary or ternary positive-negative representations (Stanković et al., 2022; Nešić et al., 2024). The fine-grained emotional modeling (Mohammad and Turney, 2013) has shifted toward more recent

research capable of capturing complex affective structures in text, including multi-label emotion annotations (Šošić et al., 2026) and contextualized affect representations (Rei and Mladenić, 2023). The Mehrazar et al. (2026) showed that although LLMs achieve high emotion classification performance, their generated explanations are often not fully faithful to the actual reasoning process, with prompting strategies influencing explanation quality and revealing the need for more reliable evaluation methods in multilingual and low-resource settings.

Within literary studies, computational approaches have increasingly employed emotion analysis to investigate narrative progression, character development, genre distinctions and stylistic variation (Jockers, 2013; Nešić, 2025). The RAGE (Roman And Greek Emotions) corpus (Riemenschneider et al., 2026) introduces a fine-grained, multi-layered annotation framework that enables quantitative analysis of emotions, their participants, intensity and narrative context across classical literary texts. Large-scale studies of literary corpora have shown that emotional trajectories within literary texts can reveal meaningful narrative structures and support quantitative literary interpretation demonstrating that "emotional arc" of stories exhibit recurring structural patterns that correlate with reader engagement and reception (Reagan et al., 2016).

Moreover, in this specific discourse, emotional signals appear as distributed structures spanning sentences, paragraphs, chapters and entire narrative arcs. Consequently, emotional analysis requires representations capable of preserving hierarchical relationships between textual components. It was also reported that computational literary studies have demonstrated that emotional structure in novels is inherently multi-level, ranging from local sentence-level affect to global narrative arcs (Kim and Klinger, 2021; Öhman et al., 2024).

The later research has significantly expanded this direction by introducing large-scale datasets of literary texts annotated with emotional trajectories. For instance, *EmotionArcs* provides emotional arc representations for over 9,000 novels and enables large-scale computational analysis of emotional structure and its relationship to literary quality and reception (Öhman et al., 2024). Similarly, recent work on emotion dynamics in novels has moved beyond single-level modeling by distinguishing between narration and character dialog, showing that emotional trajectories differ across narrative components and

are more accurately captured at the character level (Vishnubhotla et al., 2024).

Despite growing interest in computational literary analysis, research focused on emotion analysis in South Slavic languages remains relatively limited. Morphological complexity, orthographic variation and limited availability of language resources for emotions create substantial challenges for large-scale emotional analysis of Serbian literary corpora. The present work addresses these limitations through adapted lexical resources for Natural Language Processing (NLP) applications utilizing the novel methodology on hierarchical analysis of literary works on emotion dynamic exploration.

3. Methodology

3.1. SrpELTeC Corpus

The *SrpELTeC*[1] corpus (Krstev et al., 2023) employed in this research contains Serbian literary novels originally published between 1840 and 1920, covering a wide range of authors and narrative styles (Krstev, 2021). The corpus was created within the framework of the COST Action *CA16204: Distant Reading for European Literary History*[2] (2018–2022) and encoded following the TEI (Text Encoding Initiative) standards[3]. *SrpELTeC* represents the Serbian component of the *European Literary Text Collection* (ELTeC), a multilingual collection of 100-novel-corpora designed to ensure comparability in structure, scope and quality across various European languages to enable innovative, sophisticated, data-driven, computational methods of literary text analysis.

The so-called Level-1 TEI annotation of *ELTeC* is described in the document: "Encoding Guidelines for the *ELTeC*: level 1"[4]. It offers a standardized structural representation of literary texts, covering elements such as chapters, divisions, paragraphs, headings and core bibliographic metadata, thus enabling interoperability and supporting corpus-oriented research. Consequently, besides complete novels, the corpus texts are organized hierarchically into several textual layers, including chapters and paragraphs.

[1] https://github.com/COST-ELTeC/ELTeC-srp
[2] https://www.distant-reading.net/
[3] https://tei-c.org/
[4] https://distantreading.github.io/Schema/eltec-1.html

The Level-2 annotation of *ELTeC* is specified in document: "Encoding Guidelines for the *ELTeC*: level 2"[5]. It further enriches the TEI encoding with automatically produced sentence boundaries, linguistic annotations, such as tokenization, lemmatization, morphosyntactic tagging and named entity recognition, thereby enabling advanced textometric, corpus-linguistic and semantic analyses (Stanković et al., 2024).

Tokenization and sentence segmentation were conducted using the *Unitex*[6]) corpus processing platform together with Serbian language resources (Krstev, 2008). Morphosyntactic tagging and lemmatization were performed using a *TreeTagger* (Schmid, 1994) tool and model for Serbian Krstev et al. (2021), as explained in (Stanković et al., 2022). Named entity recognition was then applied with *SrpNER*, a Serbian NER system implemented within the *Unitex* framework[7]. In order to maintain compatibility with existing literary corpora and annotation conventions, the resulting *SrpNER* annotations were additionally normalized and aligned with the *ELTeC* NER tagset (Nešić et al., 2022) in accordance with the *ELTeC* Annotation Guidelines.

The *SrpELTeC* corpus (Krstev et al., 2023) comprises 100 Serbian novels enriched with hierarchical structural annotation, enabling analysis across multiple textual levels. The corpus contains 2,047 chapters, 20,831 pages and 149,402 paragraphs, further segmented into 368,156 sentences and more than 4.9 million word tokens. In addition, the corpus includes 1,145,701 punctuation marks, 2,873 verse segments, 787 foreign-language segments and 420 references to mentioned literary works.

These statistics demonstrate both the structural richness and the substantial scale of the corpus, making it one of the largest structurally annotated collections of Serbian literary prose currently available. On average, each novel contains approximately 49,000 words, 3,700 sentences and 1,500 paragraphs, providing sufficient textual diversity and volume for large-scale quantitative, corpus-linguistic and textometric analyses.

The TEI-based annotation capturing multiple structural and specialized textual layers, supports advanced literary, linguistic, se-

[5] https://distantreading.github.io/Schema/eltec-2.html
[6] https://unitexgramlab.org/
[7] http://ner.jerteh.rs/

mantic, NLP analyses. More specifically, for the use case presented in this paper, this type of complex text preparation is valuable for the research on emotional phenomena in Serbian literary texts.

3.2. EmoLex.SR Lexicon

Reliable emotion analysis of Serbian literary texts requires lexical resources adapted to both linguistic characteristics of Serbian and the stylistic properties of literary discourse. Existing emotion lexicons developed for English cannot be directly transferred to Serbian due to substantial grammatical, semantic and cultural differences. To address these limitations, this research employs the *EmoLex.SR*, a Serbian emotion lexicon developed through a multi-stage adaptation process designed to preserve emotional semantics while accounting for language-specific and cultural characteristics (Šošić et al., 2026).

The construction of *EmoLex.SR* consisted of three consecutive phases. The first phase involved automatic translation of lexical entries from the existing emotion resources developed for English organized according to Plutchik's model of basic emotions (Mohammad and Turney, 2013). This stage produced an initial Serbian emotional vocabulary aligned with eight primary emotional categories: `anger`, `anticipation`, `disgust`, `fear`, `joy`, `sadness`, `surprise` and `trust`. Since direct lexical transfer introduces substantial semantic inconsistencies, the second phase introduced linguistic adaptation, including Part-of-Speech (PoS) tagging, lemmatization, correction of translation artifacts, removal of ambiguous entries and adjustment to Serbian morphological characteristics (Krstev, 2008).

The third phase focused on cultural adaptation and lexical extension. During this stage, lexicon entries were expanded through synonym expansion, semantic refinement and manual verification to accurately capture emotional expressions characteristic of Serbian language use. Multiple synonym generation strategies were combined, including automatic synonyms expansion using Large Language Models (LLMs), Serbian WordNet lexical resource Krstev et al. (2004); Stanković et al. (2018) and manual expert curation. This iterative process resulted in two lexicon versions: *EmoLex.SR-v1* (4,179 lemma-PoS entries), which preserved intermediate lexical expansion metadata and synonym provenance and *EmoLex.SR-v2* (13,584 lemma-PoS entries), representing the unfolded version used for corpus-scale

emotional analysis through this study (Šošić, 2025). Figure 1 presents an entry as an example of a multi-stage adaptation process.

```
{
    "lemma": "dragocen",
    "pos": "ADJ",
    "synonym_lists": {
        "synonyms_gpt_3_5": "vredan,skup,vredan",
        "synonyms_gpt_4_1": "cenjen,neprocenjiv,skup,skupocen,vredan,značajan",
        "synonyms_swn": "",
        "synonyms_incorrect": "izveštačen,afektiran",
        "synonyms_manual": "skupocen"
    },
    "weights": {
        "anger": 0.0,
        "anticipation": 0.25,
        "disgust": 0.0,
        "fear": 0.0,
        "joy": 0.25,
        "surprise": 0.25,
        "sadness": 0.0,
        "trust": 0.25,
        "neutral": 0.0
    },
    "label": "anticipation|joy|surprise|trust"
},
```

```
{
    "lemma": "dragocen",
    "pos": "ADJ",
    "weights": {
        "anger": 0.0,
        "anticipation": 0.25,
        "disgust": 0.0,
        "fear": 0.0,
        "joy": 0.25,
        "surprise": 0.25,
        "sadness": 0.0,
        "trust": 0.25,
        "neutral": 0.0
    },
    "label": "anticipation|joy|surprise|trust"
},
```


Figure 1: An example lexical entry for the lemma *"dragocen"* (eng. *"precious"*, ADJ) from the *EmoLex.SR-v1* lexicon (left), alongside with the consolidated entry in the *EmoLex.SR-v2* lexicon (right).

The resulting lexicon follows a probabilistic representation where each lexical entry is associated with a normalized distribution of emotional weights across eight emotional categories and an additional `neutral` category (Figure 1). Emotional weights represent relative association strengths between a lexical item and emotional categories, while the performed normalization ensures comparability across lexical entries. Unlike categorical emotion dictionaries that assign words exclusively to a single and discrete emotion category, this representation allows modeling of emotional ambiguity and overlapping affective associations. Lexical entries in the lexicons (v1/v2) contain a lemma, PoS annotation, emotional weight distribution and categorical emotional labels derived from the most dominant emotional weights. For example, the lemma *"dragocen"* is associated simultaneously with `anticipation`, `joy`, `surprise` and `trust`, demonstrating that emotional meaning frequently emerges from multiple overlapping emotional dimensions rather than isolated emotional categories. Such representations enable richer modeling of emotional signals in literary texts, where emotional expression is often implicit, context-dependent and semantically layered.

3.3. Hierarchical Emotion Annotation and Aggregation

Emotion annotation is performed through direct string-matching between the lemma of the corpus word and the headword lemma of the entry in the *EmoLex.SR* lexicon. Since both resources employ compatible lemmatization procedures and Universal Dependencies (UD) PoS annotation schemes[8], an additional transliteration step from Cyrillic to Latin script was performed to enable alignment between the entries in both resources. Each matched token is associated with the corresponding emotional weight vector from the lexicon, which is used as an indicator of emotional expression and as a weighting factor for measuring emotional density and intensity in the subsequent emotion analysis.

Let each lexical token t_i be associated with an emotional weight vector:

$$\mathbf{e}(t_i) = [e_c]_{c \in C} \tag{1}$$

where $|C| = 9$, e.g. C={ `anger`, `anticipation`, …, `neutral`}. Emotional representations of the hierarchically organized segments in the corpus are constructed through recursive aggregation:

$$E(S) = \frac{1}{|S|} \sum_{s_i \in S} E(s_i) \tag{2}$$

where S denotes a textual segment and $E(S)$ its aggregated emotional representation. The cardinality $|S|$ denotes the number of words contained in the segment, excluding all other textual characters. Depending on the emotional representation, $E(s_i)$ may encode either weighted emotional scores or occurrence counts. In the former case, the aggregation captures emotional intensity, whereas in the latter it represents emotional density. The aggregation hierarchy follows the textual organization of literary works following the TEI structure:

$$\texttt{Token} \rightarrow \texttt{Sentence} \rightarrow \texttt{Paragraph} \rightarrow \texttt{Chapter} \rightarrow \texttt{Novel} \rightarrow \texttt{Corpus} \tag{3}$$

The structured and hierarchical design of the *SrpELTeC* corpus enables emotional analysis across multiple levels of textual granularity, preserving local emotional variation while simultaneously allowing global narrative analysis. After each aggregation step, emotional vectors are normalized to maintain numerical consistency and comparability across textual units of different lengths and structural

[8] https://universaldependencies.org/

complexity. For interpretability, aggregated emotional representations may be transformed into dominant emotional labels by selecting categories associated with the maximal weight values. When multiple emotional categories share the maximum value, multiple labels are retained to preserve emotional ambiguity.

The analytical pipeline, therefore, produces two complementary representations. At the macro level, literary works are transformed into normalized multidimensional emotional profiles suitable for comparative analysis across authors, genres and historical periods. At the micro level, hierarchical emotional annotations preserve local emotional transitions across textual segments, which enable fine-grained investigation of emotional dynamics throughout narrative progression. Such multi-scale representation allows emotional signals to be analyzed as evolving narrative structures rather than the single-level properties.

4. Experimental Results

4.1. Micro-Level Emotional Patterns

A micro-level analysis examines how emotional structures emerge locally within narrative flow through sentences and paragraphs used as primary analytical units.

Sentence-level analysis captures fine-grained emotional variation and short-range affective fluctuations throughout literary texts. Sequential sentence representations reveal localized emotional peaks, transitional regions and periods of emotional stabilization, which allows the reconstruction of emotional trajectories across narrative progression (Figure 2).

Rather than being uniformly distributed, emotional signals appear clustered around specific narrative contexts and frequently accumulate near conflict escalation, emotionally salient interactions or transitional narrative segments. To further investigate localized affective concentration, the analysis additionally extracts sentences with the highest emotional density values across the corpus (Equation 2), which enables qualitative inspection of emotionally salient narrative fragments and their narrative function (Table 1). The obtained results indicate that local emotional structures frequently remain invisible when analysis is performed exclusively at the level of complete novel.

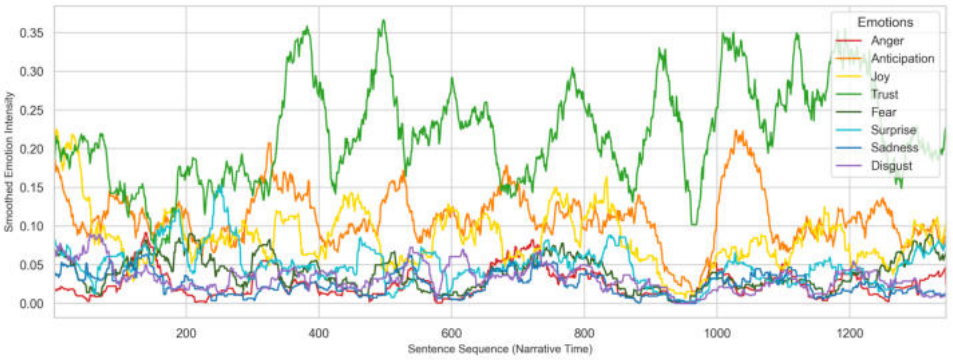

Figure 2: Comparative overview of emotions dynamics in the novel *"Stradija"* (Radoje Domanović, 1902), smoothed using a 40-sentences sliding window moving average through the narrative flow.

Although sentence-level representations provide the highest analytical granularity, hierarchical segmentation further enables aggregation at larger textual scales. Paragraph-level grouping serves as an intermediate representation suitable for identifying emotionally dense narrative regions through a composite Emotional Climax Score (ECS = density × total words), which filters out brief exclamations while preserving local contextual continuity. It applies specific structural constraints of a minimum length of 25 words and at least 4 emotional tokens per paragraph. Such a multi-tiered approach successfully extracts sustained emotional peaks and isolates their core semantic drivers at the sentence level (Table 2).

On the other hand, chapter-level aggregation enables observation of broader emotional organization throughout complete narrative structures, making it possible to analyze larger affective transitions and long-range emotional development within literary works.

Figure 3 with radar plots for 12 novels reveal relatively stable emotional profiles across the analyzed novels, with `trust`, `joy` and `anticipation` consistently representing the dominant emotional categories. At the same time, noticeable variation can be observed in the intensity and balance of secondary emotions such as `sadness`, `fear` and `surprise`, indicating stylistic and narrative differences between individual novels. Certain works exhibit more concentrated emotional distributions, while others display broader and more heterogeneous affective structures, suggesting differences in narrative tone, thematic focus and emotional complexity. Overall, the visu-

Table 1: Top 10 sentences with the highest emotions in the *SrpELTeC* corpus.

Novel	SntID	Sentence text	Dens.(%)
"Grofica Agnesa"	4051	— *Molim* , *molim* , *gospodine* , za *dokaze* ; *dokaze* *molim* .	85.71
"Dragocena ogrlica"	2451	« *Majko* , *majko* , *slatka* *majko* .... *probudi* se! ...»	83.33
"Kaluđer i hajduk"	1143	*Zlo* , veliko *zlo* nas *čeka* , *dragi* *prijatelju* !	71.43
"Kočina krajna"	708	Možda *pevahu* *anđeli* na *nebu* *slaveći* *Boga* ? ...	71.43
"Moj kočijas"	792	Mi *tražimo* *stvarnu* *slobodu* ; mi *tražimo* *jednakost*	71.43
"Hajduk Stanko"	441	*Radost* , *žalost* , *zlo* , *dobro* — **delili** su zajednički.	71.43
"Gila"	1960	*Dete* bogatog *oca* , uči *školu* — *čudna* *čuda* ! ...	71.43
"Legija smrti"	8213	Ovako *mlad* , *žalim* *umreti* ovako *očiglednom* *smrću* ...	71.43
"Grofica Agnesa"	2007	*Prijateljstvo* *daje* *sreći* veći *sjaj* : *utehom* i *saučešćem* *neprilike* olakšava.	70.00
"Legija smrti"	2043	Od *oficira* su *glavni* *mlađi* *zaverenici* .	66.67

alizations demonstrate that radar-based emotional representations provide an intuitive overview of both shared emotional tendencies and distinctive emotional patterns across literary texts.

4.2. Macro-Level Emotional Tendencies

Macro-level analysis investigates corpus-wide emotional structures and long-term historical changes by treating literary works as primary analytical units. Table 3 summarizes the aggregated emotional distribution across the complete corpus of one hundred

Table 2: Top paragraphs exhibiting the highest Emotional Climax Scores (ECS) across the *SrpELTeC* corpus

Rank 1: *"Car Dubrovnika"* (Vid Vuković Vrčević, 1901)		
Work ID: SRP19011, Chapter ID: C12		
ECS:	5042.60	
Paragraph:	37	Density: 13.27% (380/2863)
Sentence:	651	Density: 40.0% (8/20)
Sentence Text:	*"Ljubav je od Boga, a mržnja od hudobe, pa će se carstvo zemaljsko podignuti ljubavi dobra vladara, a srušiti mržnjom..."*	
Emo. Lemmas:	*"ljubav", "Bog", "mržnja", "ljubav", "dobar", "vladar", "srušiti", "mržnja"*	
Rank 2: *"Nove"* (Jelena Dimitrijević, 1912)		
Work ID: SRP19120, Chapter ID: C13		
ECS:	2913.14	
Paragraph:	1069	Density: 12.89% (226/1753)
Sentence:	4045	Density: 37.5% (3/8)
Sentence Text:	*"Ja na majku ličim telom, na oca dušom..."*	
Emo. Lemmas:	*"majka", "otac", "duša"*	
Rank 3: *"Bez Oca i Majke"* (Pera Rajčević, 1875)		
Work ID: SRP18752, Chapter ID: C1.1		
ECS:	1915.20	
Paragraph:	22	Density: 15.20% (126/829)
Sentence:	68	Density: 33.33% (5/15)
Sentence Text:	*"Bože!" — povikah u očajnom strahu svome, dršćući kao prut na vatri — "Moja majka sada gori!"*	
Emo. Lemmas:	*"očajan", "strah", "drhtati", "vatra", "majka"*	

Note: Paragraph-level filtering required a minimum of 25 words and at least four emotional lexical units, while sentence-level filtering required a minimum sentence length of three words.

Serbian novels. The obtained results reveal a highly non-uniform emotional profile characterized by a clear dominance of positive oriented emotional categories (Figure 4). Category `trust` represents

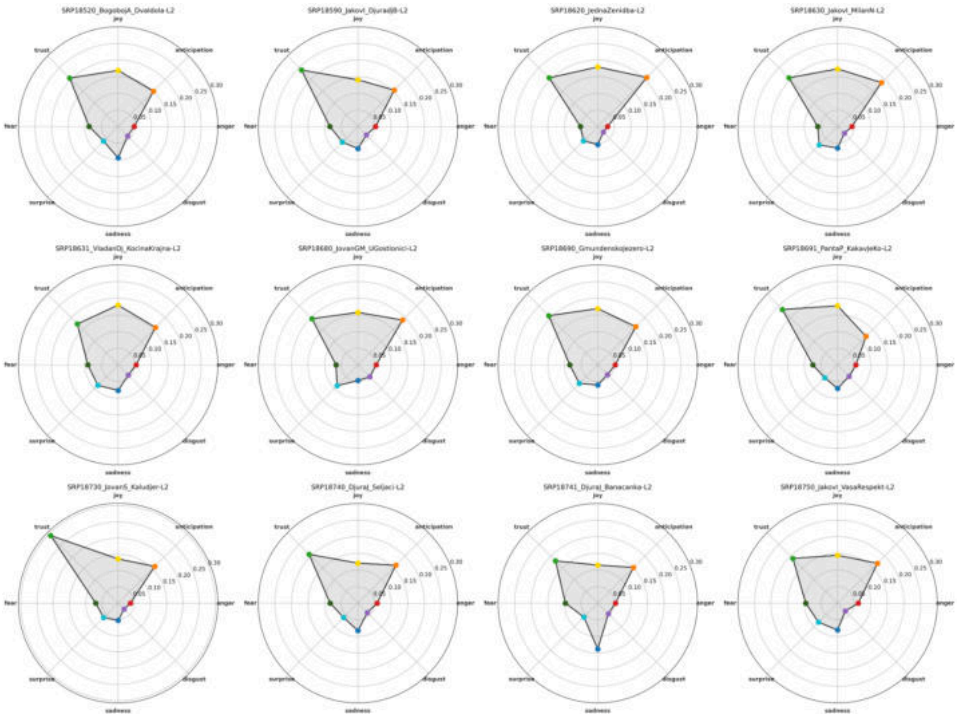

Figure 3: Comparative overview of the emotions dynamics in the sample of 12 novels.

the most prominent emotional dimension ($\mu = 0.223$), followed by `anticipation` ($\mu = 0.174$) and joy ($\mu = 0.146$). The predominance of these categories suggests, on the basis of these initial results, that Serbian prose from the nineteenth and early twentieth century relies strongly on emotionally constructive lexical structures associated with interpersonal relations, expectation, social interaction and narrative progression, although further, more detailed research is needed to substantiate this interpretation. A more detailed analysis of the most frequent lexicon-matched words from the `trust` category, revealed two distinct groups of lexical items. The first consists of general narrative words, such as *"reći"* (eng. *"to say"*), *"doći"* (eng. *"to come"*), *"čuti"* (eng. *"to hear"*), *"reč"* (eng. *"word"*) and *"početi"* (eng. *"to begin"*), whose high frequency primarily reflects the narrative style of literary texts. Although these words are associated with moderate weights, their frequent occurrence leads to a substantial cumulative contribution to the `trust` category. The second group comprises emotionally and culturally salient words, including *"Bog"* (eng.

"God"), *"otac"* (eng. *"father"*), *"crkva"* (eng. *"church"*), *"duša"* (eng. *"soul"*), *"ljubav"* (eng. *"love"*) and *"istina"* (eng. *"truth"*), which more directly express interpersonal, religious, familial and ethical concepts closely related to the trust emotion. Therefore, the predominance of the `trust` category reflects both the cumulative effect of highly frequent narrative vocabulary and the presence of semantically salient lexical items characteristic of the literary and cultural context of the corpus.

Table 3: Macro emotional statistics including average emotional weights and words density of the *SrpELTeC* corpus.

Emotion	$\mu(e_c)$	$\sigma(e_c)$	$min(e_c)$	$max(e_c)$
Anger	0.050	0.011	0.026	0.084
Anticipation	0.174	0.019	0.121	0.209
Disgust	0.039	0.010	0.024	0.071
Fear	0.075	0.015	0.045	0.118
Joy	0.146	0.024	0.079	0.195
Sadness	0.071	0.015	0.037	0.138
Surprise	0.074	0.009	0.053	0.097
Trust	0.223	0.029	0.136	0.296
Emotional Density (%)	10.75	1.42	5.41	14.42

Negative emotional categories appear substantially less represented. Categories `fear` ($\mu = 0.075$) and `sadness` ($\mu = 0.071$) occupy intermediate positions, while `anger` ($\mu = 0.050$) and `disgust` ($\mu = 0.039$) remain comparatively sparse across the corpus. This distribution indicates that emotionally negative vocabulary primarily functions as a localized narrative mechanism associated with conflict episodes rather than constituting the dominant affective baseline.

Standard deviation values reveal that variability is not evenly distributed across emotional categories. Category `trust` shows the highest dispersion across literary works, indicating that social and relational emotional structures vary substantially between authors and literary periods. In contrast, emotions associated with direct conflict, such as `anger` and `disgust`, remain relatively constrained due to their lower overall frequency within the corpus.

The average emotional density across the corpus equals 10.75%, indicating that approximately one out of ten lexical units can be associated with explicit emotional content through lexicon matching. Nevertheless, relatively large variance values demonstrate

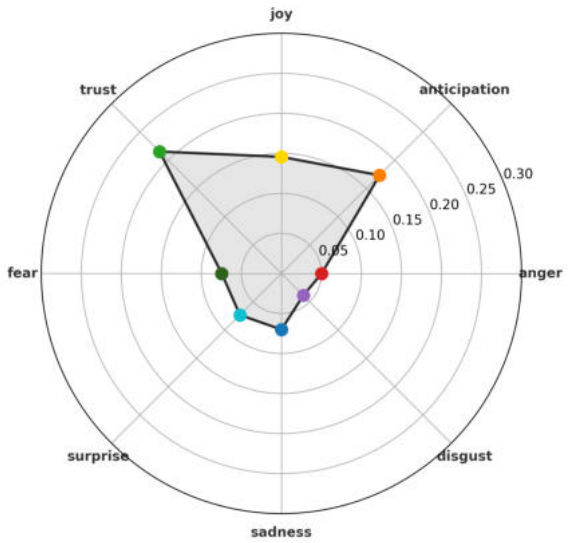

Figure 4: Radar plot of normalized emotional vector distributions in the *SrpELTeC* corpus across Plutchik's emotion categories.

substantial stylistic heterogeneity among literary works. Emotional density ranges from only 5.41% in *"Đul-Marika"* (Jelena Dimitrijević, 1903) to 14.42% in *"Stradija"*, suggesting considerable differences in emotional narrative across authors and literary movements.

To examine temporal evolution, literary works were grouped according to publication decades and analyzed using normalized emotional density measures. Figure 5 illustrates emotional words density distributions across decades for texts from the 1840–1920 period in the *SrpELTeC* corpus.

Earlier literary periods, particularly the 1850s, exhibit relatively narrow distributions characterized by smaller inter-quartile ranges and lower variance in emotional words distribution, because there are only 2 novels in this period. Beginning with the late nineteenth century, emotional words distributions become progressively wider with a median value gradually increasing. This trend becomes particularly pronounced after 1900, where median emotional density increase substantially. These results suggest that emotional evolution can be described as movement toward emotional expansion and increasing affective diversity, but this should be further analyzed since the number of novels in decades is not the same.

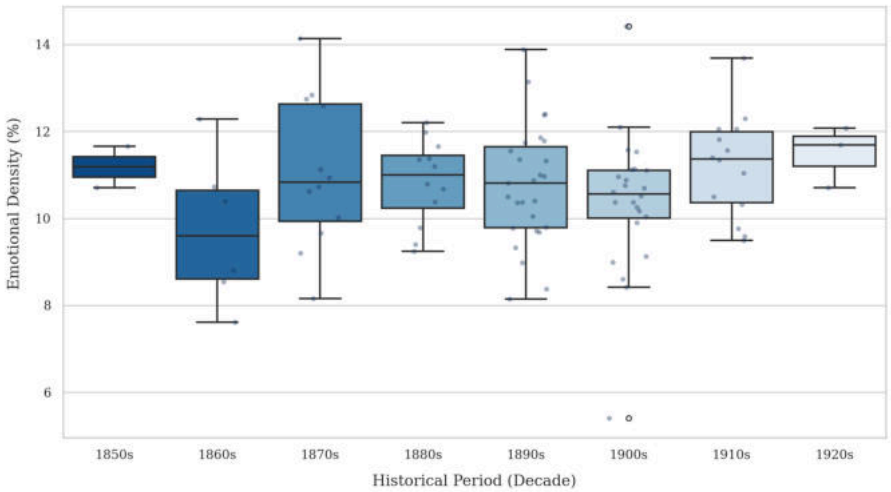

Figure 5: Shift in emotion density distribution across *SrpELTeC* literary work first publication decades.

Figure 6 illustrates pairwise correlations between novel emotional representations. Strong positive associations appear between `trust` and `joy`, indicating that socially cohesive emotional categories frequently co-occur within similar narrative contexts. Conversely, negative emotional clusters involving `anger`, `disgust`, `fear` and `sadness` demonstrate stronger internal associations while exhibiting weaker relationships with positive emotional categories.

The third cluster comprises the emotions `surprise` and `anticipation`, which exhibit a strong positive correlation. Interestingly, `surprise` also shows a weaker but noticeable association with emotions from the negative cluster, suggesting that its emotional role in literary discourse may be context-dependent and not strictly aligned with positive affect. These correlation structures suggest that emotional organization in literary prose is not driven by isolated emotional dimensions but by recurring combinations of emotional states that collectively shape the narrative atmosphere.

To investigate higher-order emotional organization, emotional representations were projected into two-dimensional space using Principal Component Analysis (PCA) algorithm (Figure 7). The two principal components explain 73.9% of total variance, with PC_1 accounting for 49.0% and PC_2 for 24.9%.

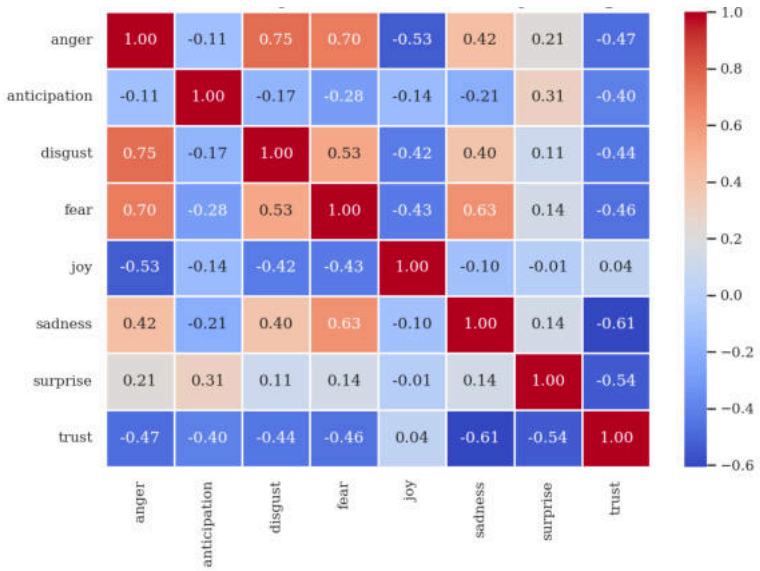

Figure 6: Correlation matrix between Plutchik's emotion categories on the whole *SrpELTeC* corpus.

Table 4: Contribution of emotion dimensions in the two principal components.

Emotion	PC1	PC2
anger	-0.185	-0.216
anticipation	-0.183	0.154
disgust	-0.125	-0.130
fear	-0.254	-0.270
joy	0.177	0.865
sadness	-0.281	-0.031
surprise	-0.121	0.065
trust	0.853	-0.294

The first principal component primarily captures a continuum between negatively oriented emotional configurations and `trust`-oriented emotional profiles, while the second component is primarily driven by `joy` emotion dimension (Table 4). Literary works located at negative values ($PC_1 < -0.05$), including *"Bespuće"* (Veljko Milićević, 1912) and *"Jarani"* (Svetozar Ćorović, 1911), exhibit emotional structures characterized by relatively stronger contributions from `sadness`, `fear` and `anger`. In contrast, texts occupying pos-

itive regions ($PC_1 > 0.07$), such as *"Car Dubrovnika"* (Vid Vuletić Vukasović, 1906), demonstrate stronger contributions from `trust` and related positive emotional dimensions.

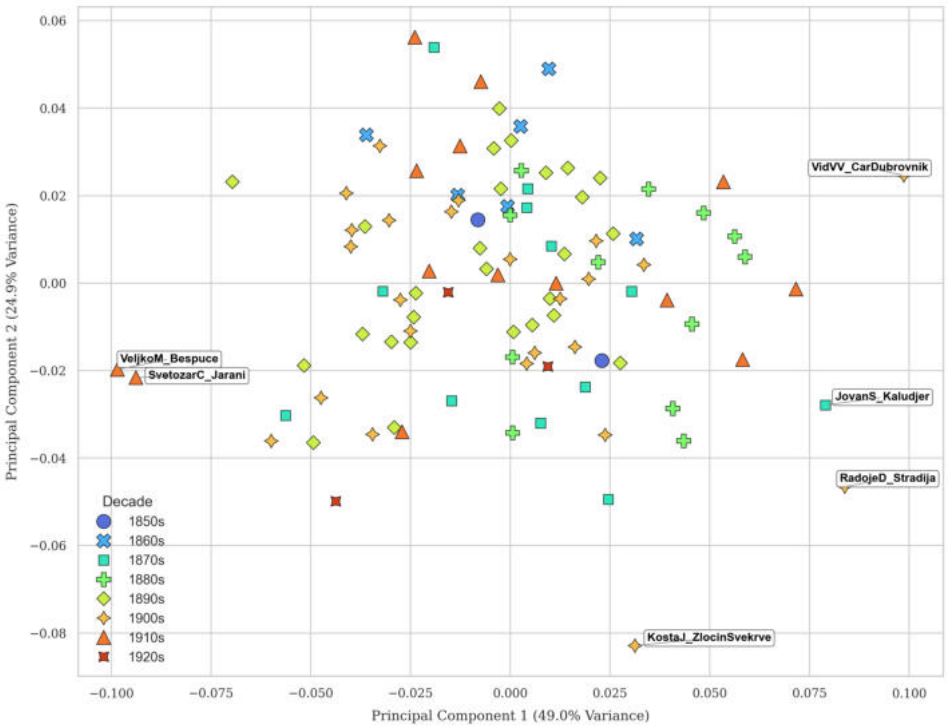

Figure 7: Two-dimensional PCA projection of literary works from the *SrpELTeC* corpus based on eight-dimensional emotional representations.

The second component is strongly dominated by `joy` and separates joy-oriented emotional structures from emotionally negative configurations associated with `fear`, `anger` and lower positive affect. Several statistically isolated observations emerge within latent space. The novel *"Zločin svekrve"* (Kosta Jovanović, 1906) occupies the lower boundary of the second component ($PC_2 = -0.083$), reflecting emotional profiles with comparatively low joy and stronger negative emotional contributions. Similarly, *"Stradija"* appears in the lower-right quadrant ($PC_1 = 0.084$, $PC_2 = -0.046$), combining trust-oriented emotional structure with comparatively lower positive affect. The existence of such outliers indicates that emotional representations capture stylistic and genre-specific properties rather than simple lexical frequency effects.

5. Discussion

The micro-level analysis demonstrates that lexicon-based emotional representations can capture emotional fluctuations across multiple hierarchical textual structures. The results indicate that emotional signals can be identified not only at document level but also within smaller textual units such as sentences, paragraphs and chapters, enabling the detection of emotionally charged segments and localized emotional variation throughout literary texts. These findings suggest that hierarchical analysis provides additional interpretative resolution compared to approaches relying exclusively on single-level representations.

The macro-level results indicate that emotional evolution in Serbian literary prose is characterized primarily by increasing emotional diversification rather than simple shifts in emotional intensity. Using the *EmoLex.SR* emotional lexicon, dominant emotional categories within the corpus, such as `trust`, `anticipation` and `joy` were identified together with recurring correlation patterns between emotions across literary periods. Earlier literary periods exhibit relatively compact emotional organization and shared emotional baselines, consistent with more standardized literary conventions, whereas literary production approaching the twentieth century displays increasingly dispersed emotional representations, reflecting the emergence of multiple stylistic directions.

The PCA results further support that earlier literary works cluster near the center of latent space, suggesting relatively homogeneous emotional structures, whereas later works occupy increasingly dispersed regions. This dispersion reflects growing stylistic individuality rather than uniform emotional change. The isolation of texts such as *"Stradija"* and *"Zločin svekrve"* demonstrates that emotional representations capture meaningful literary properties associated with genre, satire, psychological intensity and narrative style. Similarly, emotionally specialized narrative forms, including matrimonial narratives with elevated anticipation scores, further indicate that emotional organization reflects narrative function rather than lexical frequency alone. These findings suggest that emotional signals operate as structured narrative mechanisms whose interpretation depends strongly on analytical scale. Hierarchical analysis therefore provides a substantially richer representation of literary emotional organization than standard single-level sentiment approaches.

The proposed approach remains constrained by the size and lexical coverage of the emotional lexicon itself, limiting recognition performance for low-frequency words and multi-word expressions (MWEs). Additional limitations arise from difficulties in recognizing archaic lexical forms that diverge substantially from contemporary standardized Serbian vocabulary, particularly within old literary texts. Furthermore, the normalization pipeline introduces additional uncertainty since lemmatization and PoS tagging techniques may propagate errors that affect emotional category assignment. Consequently, emotional representations presented here should be interpreted as approximations of underlying emotional organization rather than complete reconstructions of literary emotional expression.

6. Conclusion

This paper presents a hierarchical computational approach for emotion analysis in nineteenth-century Serbian literary corpora based on a Serbian emotion lexicon grounded in Plutchik's emotional theory. By combining linguistically adapted lexical resources with multilevel structural analysis, the proposed approach enables exploration of emotional dynamics across multiple narrative scales.

The presented methodology contributes both computationally and theoretically to digital humanities research by providing scalable tools for investigating emotional structures embedded within literary narratives. Future work will focus on contextual emotion modeling and the integration of semantic knowledge representations into emotion analysis pipelines for literary works.

Acknowledgements

This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA https://tesla.rgf.bg.ac.rs/.

Language Resources

Krstev, Cvetana and Stanković, Ranka and Trtovac, Aleksandra and Milnović, Vasilije Vitas, Duško and Obradović, Ivan. 2023. *The Serbian novel collection for the ELTeC: The European Literary Text Collection*. European Language Grid (ELG). Available through

GitHub and the European Language Grid: https://live.european-language-grid.eu/catalogue/corpus/23621.

Krstev, Cvetana and Vitas, Duško and Stanković, Ranka and Škorić, Mihailo. 2021. *SrpMD4Tagging - Serbian Morphological Dictionaries for Tagging*. European Language Grid (ELG), https://live.european-language-grid.eu/catalogue/lcr/9294, 1.0.

Šošić, Milena. 2025. *EmoLex.SR: Serbian Lexicon of Emotional Words*. European Language Grid (ELG), https://live.european-language-grid.eu/catalogue/lcr/23616, 1.0.

References

Matthew L Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Evgeny Kim and Roman Klinger. 2021. A Survey on Sentiment and Emotion Analysis for Computational Literary Studies. *Zeitschrift für digitale Geisteswissenschaften*. First published: 2019-12-16.

Cvetana Krstev. 2008. *Processing of Serbian – Automata, Texts and Electronic Dictionaries*. University of Belgrade, Faculty of Philology, Belgrade.

Cvetana Krstev. 2021. The Serbian Part of the Eltec Collection through the Magnifying Glass of Metadata. *Infotheca-Journal for Digital Humanities*, 21(2):26–42.

Cvetana Krstev, Gordana Pavlović-Lažetić, Duško Vitas, and Ivan Obradović. 2004. Using Textual and Lexical Resources in Developing Serbian Wordnet. *Romanian Journal of Information Science and Technology*, 7(1–2):147–161.

Mobina Mehrazar, Mohammad Amin Yousefi, Parisa Beygi, and Behnam Bahrak. 2026. Can LLMs Faithfully Explain Themselves in Low-Resource Languages? A Case Study on Emotion Detection in Persian. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 560–579, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a Word–Emotion Association Lexicon. *Computational intelligence*, 29(3):436–465.

Milica Ikonić Nešić. 2025. Analysis of Family Relationships of the Characters in the SrpELTeC Corpus Based on Semantic Web Techniques and Textometry. *Infotheca: Journal for Digital Humanities*, 25(1/2):7.

Milica Ikonić Nešić, Saša Petalinkar, Mihailo Škorić, Ranka Stanković, and Biljana Rujević. 2024. Advancing Sentiment Analysis in Serbian Literature: A Zero and Few–Shot Learning Approach Using the Mistral Model. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 58–70.

Milica Ikonić Nešić, Ranka Stanković, Christof Schöch, and Mihailo Škorić. 2022. From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back). In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 7–16.

Emily Öhman, Yuri Bizzoni, Pascale Feldkamp, and Kristoffer Nielbo. 2024. EmotionArcs: Emotion Arcs for 9,000 Literary Texts. In *Proceedings of the 8th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature (LaTeCH-CLfL 2024)*, pages 51–66.

Robert Plutchik. 2001. The Nature of Emotions: Human Emotions have Deep Evolutionary Roots, a Fact that may Explain their Complexity and Provide Tools for Clinical Practice. *American scientist*, 89(4):344–350.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The Emotional Arcs of Stories Are Dominated by Six Basic Shapes. *EPJ data science*, 5(1):31.

Luis Rei and Dunja Mladenić. 2023. Detecting Fine-Grained Emotions in Literature. *Applied Sciences*, 13(13):7502.

Frederick Riemenschneider, Jonathan D. Geiger, Thomas Kuhn-Treichel, and Anette Frank. 2026. RAGE: Roman and Greek Emotions. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 947–957, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Klaus R Scherer. 2005. What are Emotions? And how can they be Measured? *Social science information*, 44(4):695–729.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 1–9, Manchester, UK.

Milena Šošić, Jelena Graovac, and Ranka Stanković. 2026. Building an Emotion Lexicon for Serbian Using Curated Language Resources. *Language Resources and Evaluation*, 60(1):9.

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

Ranka Stanković, Cvetana Krstev, and Duško Vitas. 2024. SrpELTeC: A Serbian Literary Corpus for Distant Reading. *Primerjalna književnost*, 47(2):45–63.

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

Ranka Stanković, Miloš Košprdić, Milica Ikonić-Nešić, and Tijana Radović. 2022. Sentiment Analysis of Sentences from Serbian ELTeC Corpus. In *Proceedings of the SALLD-2 Workshop at Language Resources and Evaluation Conference (LREC), Marseille, France*, pages 31–38.

Krishnapriya Vishnubhotla, Adam Hammond, Graeme Hirst, and Saif Mohammad. 2024. The Emotion Dynamics of Literary Novels. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2557–2574.

Leksičko istraživanje emocionalne dinamike u TEI romanima: studija slučaja na korpusu *SrpELTeC*

Ranka Stanković, Milena Šošić

Sažetak

Ovaj rad predstavlja hijerarhijsku analizu emocionalne dinamike u korpusu srpskih romana *SrpELTeC*. Studija kombinuje TEI-kodirane tekstove u korpusu sa *EmoLex.SR* leksikonom emocionalnih reči za srpski jezik, zasnovanim na Plutčikovom modelu osnovnih emocija. Emocionalni signali iz korpusa izdvojeni su putem leksičkog podudaranja po lemi i vrsti reči, a zatim su analizirani na više hijerarhijski organizovanih nivoa teksta, kao što su reči, rečenice, pasusi, poglavlja, književna dela i celokupni korpus. Ovako strukturisana analiza omogućila je istraživanje i prepoznavanje lokalnih varijacija i globalnih obrazaca emocionalnih signala u književnim tekstovima. Pristup kombinuje mere gustine emocionalnih reči i njihovog intenziteta sa hijerarhijskim i hronološkim grupisanjem, radi ispitivanja kratkoročnih i dugoročnih promena u emocionalnoj strukturi *SrpELTeC* korpusa. Prvi empirijski rezultati pokazali su stabilan osnovni obrazac u kome dominiraju emocije `poverenje` (eng. `trust`) i `iščekivanje` (eng. `anticipation`), uz značajno veći emocionalnu raznolikost u delima sa početka dvadesetog veka. Ovi nalazi ukazuju na to da se emocionalna organizacija u srpskoj prozi razvijala u pogledu intenziteta i strukturne složenosti tokom vremena, kao i da je predložena metodologija u ovom radu omogućila računarsku interpretabilnost emocionalnog aspekta jezika u različitim segmenatima pojedinačnih književnih dela, kao njihovo međusobno poređenje na osnovu vremenskih perioda u kojima su nastala.

Ključne reči: emocije, Plutčikov model, leksikon, srpski književni korpus

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

https://creativecommons.org/licenses/by/4.0/