Analysis of Similes in Serbian Literary Texts (1860-1920) using computational methods

Cvetana Krstev
University of Belgrade
Faculty of Philology
cvetana@matf.bg.ac.rs

Jelena Jaćimović
University of Belgrade
School of Dental Medicine
jelena.jacimovic@stomf.bg.ac.rs

Duško Vitas
University of Belgrade
Faculty of Mathematics
vitas@matf.bg.ac.rs

Abstract

Similes are rhetorical figures which play an important role in literary texts. This paper presents a finite-state methodology developed for the description of adjectival similes, which enables their retrieval and annotation in Serbian novels written in the mid-19th and early 20th centuries. The results of a textometric analysis reveal the most frequent adjectival similes and the specificity of their usage, with respect to the author, title, or publication date, in a subset of the SrpELTeC corpus.

Keywords: rhetorical figures, literary corpus, simile figure, multi-word units

1. Introduction

In the history of rhetoric simile holds a particular place. Although it is one of the oldest recognized figures of speech, from the very beginning, simile has often been taught and studied in conjunction with metaphor. Ever since ancient times, many researchers have been reconsidering the status of simile and its convergence with other familiar figures, treating it either as a literal comparison, a weaker form of metaphoric expression or as a completely distinct figure of speech. Indeed, simile is essentially a rhetorical figure presented, unlike metaphor, as an explicit form of comparison. On the other hand, in contrast to literal comparison, simile is also essentially figurative, making unexpected connections between literally unlike concepts (Israel et al., 2004).

Similes rely on comparisons, semantic figures which bring two different entities together based on a shared feature (Israel et al., 2004), implying a certain likeness between them. Both literal comparison and simile have the same recognizable formal structure, the surface form consisting of the following elements: the subject of comparison (**tenor**, **target**, or **topic**), the object of comparison (**vehicle** or **source**), a conjunction which signals a comparison (**marker**, usually *as ... as*, *as* or *like* in English, *kao* in Serbian, or *comme* in French), and the basis of the comparison implied by the expression (**ground**, **property**, or **tertium comparationis**) (Example 1.1).

Example 1.1.	[She]	was	[free]	[as]	[a bird].
	tenor		**ground**	**marker**	**vehicle**

However, the subject of comparison (**tenor**) most often does not form part of a simile (Brehmer, 2009). Therefore, similes are multi-word expressions (MWE) that, as introduced in (Beardsley, 1981), can either be *closed* (represented with a three-part structure **ground + marker + vehicle**, as given in Example 1.1) or *open* if the shared attribute is not explicitly stated, but could be derived from the context (**marker + vehicle**), as illustrated in Example 1.2 where the shared property of being free is left implicit.

Example 1.2.	[She]	was	[as]	[a bird].
	tenor		**marker**	**vehicle**

In addition to the aforementioned multi-word structure, another formal characteristic of similes is that they are often quite conventionalized, generally known and accepted phrases used by all members of a linguistic community. Even though their lexical composition is highly stable, consisting of at least two or three components, it is not absolute having numerous variants of the essential simile elements. As far as their semantic features are concerned, similes are characterized as being idiomatic and remarkably expressive, which is the result of a powerful connotation, for instance, positive or negative sentiment toward something, and the picturesqueness of their essential parts.

The widespread presence of similes in everyday language stands to reason since they rely on comparing, a fundamental human cognitive activity, producing a particular image in a person's mind (Mpouli, 2016). In view of their evocative power and descriptive capacity, similes are the most attractive comparative structures to investigate in literary texts. As rhetorical instruments, they can easily be combined with other figures of speech (Israel et al., 2004) and used for stylistic effects. As a part of an author's imagery, the similes used can uncover and define the personality and experiences of the author, the tonality of a particular text, or even a literary period. Hence, identifying all simile varieties in a novel appears vital for stylistic examination.

Similes have a structure that appears fairly amenable to automated processing (Niculae and Yaneva, 2013). Still, in computational linguistics, which is particularly interested in figurative language, similes have been overlooked in favor of metaphor even more than in linguistics. Simile analysis has become a particularly appealing topic of interest in the field of computational linguistics and corpus studies in recent years (Niculae, 2013; Yoshimura et al., 2015; Qadir et al., 2015; Qadir et al., 2016; Hu et al., 2017). One of the main tasks is automatic simile recognition, which can be divided into partial and full simile identification. Partial simile identification principally involves retrieving specific simile patterns, either complete expressions consisting of all simile elements, or only preselected grounds and vehicles. Furthermore, this process depends on heuristics or human reasoning to recognize the difference between similes and literal comparisons. On the other hand, full simile recognition involves extraction and analysis of all sentences containing a simile marker in unstructured texts, and subsequent identification of the separate components of each potential simile. For the simile recognition task, various methods have been proposed. Most of them can be classified as feature-based (Niculae and Danescu-Niculescu-Mizil, 2014), pattern-based (Niculae and Yaneva, 2013; Niculae, 2013), or neural network-based (Liu et al., 2018; Zhang et al., 2019).

The research related to rhetorical figures and their automatic processing for the Serbian language started with building the Ontology of Rhetorical Figures (Mladenović and Mitrović, 2013). A method of automatic recognition and classification of rhetorical figures, including similes, that uses ontological inference rules in an ontology based on Serbian WordNet (SWN), was also developed (Mladenović, 2016). In (Mitrović et al., 2019), the authors applied a corpus-driven crowdsourcing method for enrichment of lexical resources with Serbian and Greek similes. A corpus of similes used in modern Serbian language was produced based on a methodology for semi-automated collection of similes from the World Wide Web using text mining techniques (Milošević and Nenadić, 2016; Milošević and Nenadić, 2018).

The main goal of this paper is to provide an analysis of adjectival similes in Serbian novels written in the mid-19th and early 20th centuries, retrieved through automatic recognition and annotation. Moreover, it aims to identify the most frequent similes and their components (such as grounds and vehicles), using the textometric method for analysis and visual presentation of results.

2. About the Corpus

One of the main objectives of the *Distant Reading for European Literary History* (COST Action CA16204) project[1] is compilation of a multilingual European Literary Text Collection (ELTeC). This work is still in progress, but before it ends, the project is expected to comprise around 2,500 full-text novels in at least 10 different languages. All texts from this corpus have to fulfill the same criteria: they should be originally written in a language of the subcollection to which they belong, their first publication date should fall between 1840-1920 (preferably appearing as a book and not published in installments) and

[1]Distant reading `https://www.distant-reading.net/`

they should be at least 10,000 word tokens long. Each language subcollection will eventually comprise up to 100 novels that fulfill certain balancing criteria:[2]

- each of the four twenty-year periods should be represented by approximately the same number of novels;
- at least 10%-50% of the works featured should be written by female authors;
- 9 to 11 authors should be represented by exactly three novels (other authors should be represented by one novel);
- at least 20% should be short novels (10-50k word tokens), at least 20% should be long novels (>100k word tokens);
- at least 30% should be highly popular novels and at least 30% should be novels that are not known to the general public.

For this research 41 novels that are candidates for the Serbian subcollection of the ELTeC corpus were used.[3] The characteristics of the sample corpus having a size of 1,471,141 word tokens are represented in Table 1. It becomes apparent that the Serbian subcollection will not be able to meet all the balancing criteria: presently, there are neither novels from the 1840–1859 time period, nor those that exceed 100,000 word tokens.

Period	**Number**	**Length**	**Number**	**Sex**	**Number**
1840-1859	0	short	30	Male	34
1860-1879	3	medium	11	Female	7
1880-1899	16	long	0		
1900-1920	22				

Table 1: Corpus distribution

This particular collection contains novels of exceptional value for the history of Serbian literature. Besides well-known novels, which introduce a modern narrative structure, this corpus contains novels by forgotten authors, like Dragomir Šišković and Stevan Mamuzić, as well. Moreover, the first Serbian science-fiction novel *Jedna ugašena zvezda* (*An Extinguished Star*) by Lazar Komarčić is part of the srpELTeC corpus too, and so is the novel *Babadevojka* (*Old Maid*) by Draga Gavrilović, the first female author who wrote a novel in the Serbian patriarchal society of the time. A complete list of the novels used in this research can be found in Appendix A. The dimensions of the srpELTeC corpus used on this occasion and partitioned based on authorship are presented in Figure 1.

3. Simile retrieval and annotation

The first step of our research consisted of an attempt to retrieve as many similes as possible from our corpus. Two approaches have been adopted for this purpose: first, we looked for simile figures in the electronic morphological dictionary of Serbian (SMD), and then we applied a simple regular pattern to spot simile occurrences. In both cases, we used the Unitex system and the incorporated SMD (Krstev, 2008).[4]

At present, SMD contains 68 multi-word expressions that represent similes. In our corpus, we retrieved 98 occurrences of these already identified simile figures, or 33 different forms among which *bled kao smrt* 'pale as death' (n=14) was the most frequent one. Based on a regular expression `<A> (<jesam.V>+<E>) (kao+ko+k('+')o)` – an adjective followed by the conjunction *kao*, or some

[2]Encoding Guidelines for the ELTeC: level 1 `https://distantreading.github.io/Schema/eltec-1.html`

[3]The Serbian subcollection is still under construction, and some of the prepared novels might not become part of the final collection due to the balancing criteria that have to be met `https://distantreading.github.io/ELTeC/`.

[4]Unitex/Gramlab, the multilingual corpus processing suite `https://unitexgramlab.org/`.

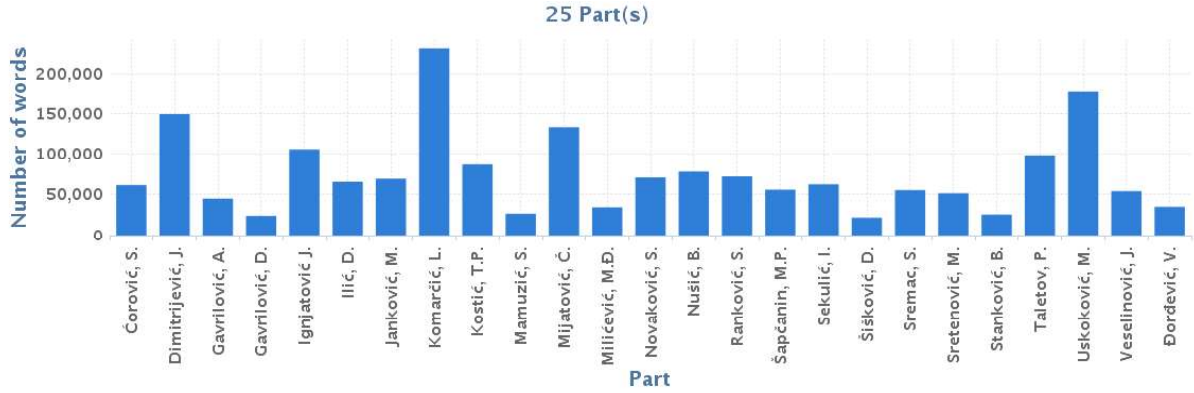


Figure 1: Dimensions of the srpELTeC corpus parts created based on authorship

of its irregular variants, with a possible auxiliary *jesam* 'to be' in between, we obtained a list of possibilities from which we extracted 267 simile occurrences, or 225 distinct forms where *žut kao vosak* 'yellow as wax' (n=5) was the most frequent case. In 4 novels out of 41 in the corpus, no simile was retrieved (all of them were "short" novels).

As mentioned above, some similes have already been recorded in the Serbian morphological dictionary of MWUs (Krstev et al., 2013). This format is consistent with the morphological dictionary of simple words and it allows a description of the various properties of an MWU, besides its morphological features. In the case of simile figures, a dictionary description can specify:

- morphological behavior of an adjective – it does not inflect in degree, it is always used in the positive form;
- morphological behavior of a noun (vehicle) – whether it changes in number to agree with a noun (tenor) or not;
- the order of constituents which can be *A kao N* or *kao N A*.

The example 3.1 illustrates this by way of the entry *gladan kao vuk* 'hungry as a wolf'. DELAC entry is used to produce all inflected and variant forms of an MWU (Savary, 2009).

Example 3.1.

```
DELAC:  gladan(gladan.A18:akms1g) kao vuk(vuk.N128:ms1v)
DELACF: gladnog kao vuk,gladan kao vuk.A:adms4v
        gladnome kao vuk,gladan kao vuk.A:adms7g
        gladni kao vuk,gladan kao vuk.A:aemp1g
        gladni kao vuci,gladan kao vuk.A:aemp1g
        gladni kao vukovi,gladan kao vuk.A:aemp1g
        kao vuk gladnog,gladan kao vuk.A:adms4v
        kao vukovi gladni,gladan kao vuk.A:aemp1g
        ...
```

However, with this representation, a number of deviations occurring in the use of similes cannot be described, such as:

1. variations that may occur in all constituents of similes:
   (a) variation in the ground: for instance, three different forms (near synonyms) *mek*/*mehak*/ *mekan* in the figure *mek kao pero* 'soft as a feather';
   (b) variation in the vehicle: for instance, three different forms (near synonyms) *pero*/*perce*/ *paperje* in the figure *mek kao pero* 'soft as a feather';

(c) variation in the marker: the conjunction *kao* can also be written as *k'o*, *ko*, *ka'*, etc. Although only the first form is sanctioned by the Serbian orthography, other forms occur frequently in literary texts.

2. the vehicle can be modified:
   (a) with an adjective, for instance, *crven kao rak* and *crven kao pečen rak* 'red as a (fried) crayfish'. One should note that the choice of an adjective is not free, in this case it can be just *pečen/kuvan* 'fried/cooked';
   (b) with an adjunct, for instance, *slobodan kao ptica*, *slobodan kao ptica na grani*, *slobodan kao ptica u gori* 'free as a bird (on a bough/in a wood)'. The possible adjuncts are also limited;
   (c) with a determiner (adjectival pronoun), for instance, *mudar kao kakav pop* 'wise as some priest'.

3. variations due to the free word order:
   (a) insertion of an auxiliary, for instance, *crvena si kao zreli nar* 'you are red as a ripe pomegranate';
   (b) insertion of a subject (tenor), for instance, *vreo dah kao plamen* 'breath hot as a flame';
   (c) insertion of a pronoun (clitic), for instance, *privržen mu kao pašče* 'attached to him as a dog';

4. variations resulting from rephrasing, for instance, *žut kao što je slama* 'yellow as straw is'.

The presented variations can be described by local grammars in the form of finite-state automata. One such automaton that recognizes the figure *beo/bjel kao sneg/snijeg [u planini]* 'white as snow [in the mountain]' is presented in Table 2a).[5] The production of such graphs for each individual figure would be impractical. For that reason, generic automata were constructed (Table 2b) in which the information in certain nodes is filled with specific information stored in the table describing all similes (Table 2c). For instance, the upper left node containing `%A,B,C%` in the generic automaton, is replaced by the content of the cells A, B, and C from the simile table to obtain, for the first table line, the corresponding node in the specific automaton: `<beo.A:a>+<bjel.A:a>`, meaning adjective *beo* or *bjel* in the positive form.[6]

The upper path in the generic graph recognizes similes with a regular word order, while the lower part recognizes figures with a reverse order. Variations in the ground and the vehicle (see items 1(a) and 1(b) from the list above) are figure specific and the information filling the appropriate graph node is obtained from the table, while marker variations (item 1(c) are common to all figures and they are coded in the graph. Most optional additions are also figure specific (items 2(a) and 2(b)) and for them, the information is transferred from the table; some are common (item 2(c)), including word order and other insertions (items 3 and 4) and they are coded in the generic graph. In this way, 243 graphs are produced for simple similes (one adjective) and 44 for complex figures (two adjectives).

Even though most researchers tend to mark only phrases representing similes (Mpouli, 2017), we have decided to annotate all recognized similes both at phrase and word levels. Each identified simile has been enclosed within the tag `<simile>`, specifying the type of the rhetorical figure in question and its range. Furthermore, each simile basic element, namely ground, marker, and vehicle, has also been marked with the corresponding tag. Example 3.2 illustrates the annotation of one simile. The annotation process at the moment includes neither annotation of entire sentences containing similes nor additional information regarding simile semantic features.

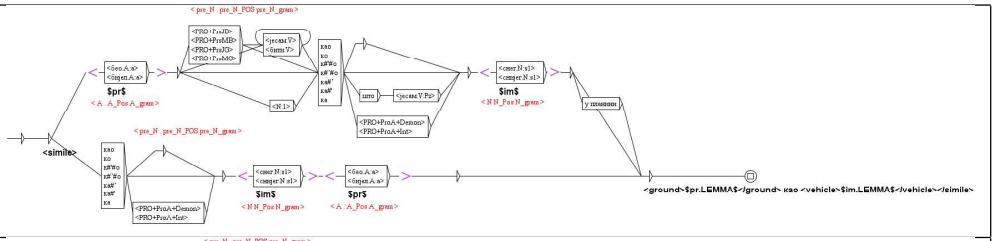
Example 3.2. *zdrav i rumen kao jabuka* 'healthy and ruddy as an apple'

```
<simile><ground>zdrav</ground> i <ground>rumen</ground>
<marker>kao</marker> <vehicle>jabuka</vehicle></simile>
```

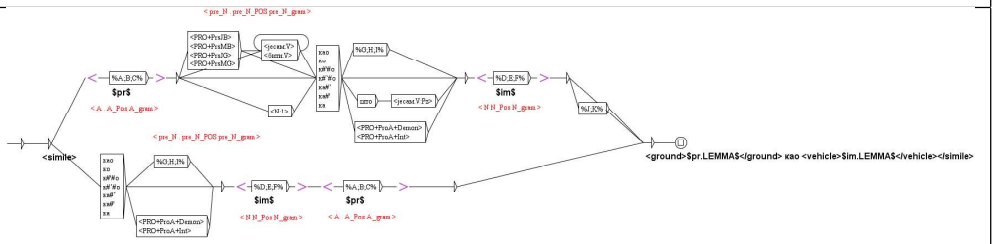
[5] These graphs are implemented in Unitex/Gramlab and they use SMD information implemented in the same environment.

[6] All examples in this paper are given in the Latin script. Most of the novels were published in the Cyrillic script, with just a few in the Latin script. Specific automata were constructed automatically for both scripts.

a



b



c

A	B	C	D	E	F	G	H	I	J	K
beo,bjel	A	:a	sneg,snijeg	N	:s1				u planini	X
blažen	A	:a	dete,djete	N	:s1	mali,malen				
crn	A	:a	gavran	N	:s	zlokoban				
lep,lijep	A	:a	upisan	A	:as1					

Table 2: a) specific finite-state automata; b) generic finite-state automata; c) data describing specific similes.

Finally, the annotated corpus was imported into the TXM program environment (Heiden et al., 2010; Heiden, 2010) for a quantitative and qualitative analysis of the recognized similes. Based on the total number of simile occurrences in the whole corpus (F), the total number of simile occurrences in the texts in a particular part of the corpus (f) and the *specificity score* (S), significantly common or significantly rare occurrences of adjectival similes in distinct parts compared to the whole corpus were identified, as well as the specific use cases of simile adjectival and nominal elements.

4. Analysis of results

The results of the study show that in the corpus consisting of 41 Serbian novels written between 1860 and 1920, 404 occurrences of 251 distinct adjectival similes are found. Among them there are 392 similes with one adjectival ground and 12 represent examples with two adjectives used as simile ground (as shown in Example 3.2). This figure of speech appears most frequently in the texts penned by Uskoković M. (f=58), Komarčić L. (f=56), Dimitrijević J. (f=54), and Taletov P. (f=36). Nevertheless, the results reveal that adjectival similes are extremely specific to the part of the corpus written by Sretenović M. (S=4.2), despite lower absolute frequency (f=26) compared to the previously mentioned authors. The specificity distribution of the recognized adjectival similes in the corpus partitioned based on authorship is presented in Figure 2. On the other hand, if the total number of simile occurrences in the whole corpus is taken into account, for the part dedicated to the works by Komarčić L, where a high simile frequency is recorded, the observed *specificity score* is 0.7, which indicates common rather than specific uses of adjectival similes. Adjectival similes are less represented lexical units in the part of the corpus written by Ignjatović J. (f=7, S=-4). These figures are also significantly rare in corpus parts authored by Gavrilović A. (f=1), Sremac S. (f=3) and Milićević M.Đ. (f=1), with the *specificity score* of -3.2, -2.6 and -2.3, respectively. The novels characterized by a significantly high frequency of use of adjectival similes are *Radetića Mara* (*Mara of the Radetic's*) (S=4.2), *Jedna ugašena zvezda* (*An Extinguished Star*) (S=4.1), *Došljaci* (*Newcomers*) (S=3.2), *Novac* (*Money*) (S=2.9) and *Nove* (*New Women*) (S=2.6). Moreover, if we look at the adjectival simile use in the corpus partitioned by decades, significantly common use of similes occurs in the novels published in the first decade of the 20th century (S=3.8).

The syntactic pattern Adjective + Conjunction + Noun is the most frequent (86.9%) of all adjectival simile occurrences in the corpus (F=404). We have also recorded other syntactic variants and examples

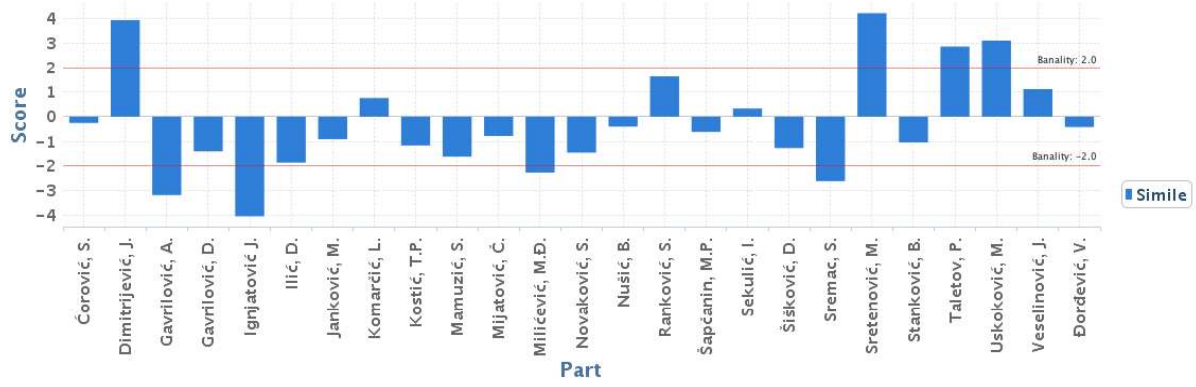


Figure 2: The specificity of adjectival simile use in the srpELTeC corpus by authors

where nominal (10.6%) and adjectival phrases (2.5%) are used instead of a noun as a vehicle of the recognized simile. Besides closed similes, we have also found cases of open simile syntactic patterns such as Conjunction + Noun. In these situations, the connection between the adjective and the noun is very strong making it possible to omit the adjective and still retain the meaning, for instance, *Arhimandrit beše (ljut) kao ris* 'Archimandrite was (angry) like a lynx' where the adjective in the parentheses is omitted. An open simile occurred in two more cases: *(mali) kao makovo zrno* '(small) as a poppy seed' and *(vredan) kao pčela* '(hard working) as a bee'.

The most frequently used adjectival simile in the sample corpus of Serbian novels is *beo kao sneg* 'white as snow' (F=28), followed by the simile *bled kao smrt* 'pale as death' (F=15), which also turns out to be the most frequent simile in the British and French corpora consisting of novels written between the mid-19th and early 20th century (Mpouli and Ganascia, 2015). The adjectives *beo* 'white' and its synonym *bled* 'pale' often appear in the most frequent similes in Serbian novels, and they occur in British and French literary texts from a similar period as well.

There are 416 occurrences of 111 distinct adjectives used as simile ground in this corpus. Adjectives occurring in the retrieved simile figures are frequently connected to different nouns and vice versa. The adjectives and nouns that show the widest variety in connections are represented in Table 3. In some cases, two adjectives are explained by the same noun as in *dobar i miran kao jagnje* 'good and quiet as lamb'. There were 12 such cases.

***beo* 'white' (45)**	***bled* 'pale' (33)**	***stena* 'rock' (10)**	***jagnje* 'lamb' (9)**
sneg (28) 'snow'	*smrt* (15) 'death'	*hladan* (6) 'cold'	*miran* (3) 'quiet'
mleko (7) 'milk'	*krpa* (9) 'cloth'	*nem* (1) 'mute'	*dobar* (2) 'good'
krin/ljiljan (3) 'lily'	*vosak* (4) 'vax'	*neosetljiv* (1)	*blag* (1) 'mild'
alabaster (1)	*senka* (2) 'shadow'	'impassive'	*poslušan* (1) 'docile'
list hartije (1) 'sheet (of paper)'	*mrtvački* (1) 'deathly'	*nepomičan* (1)	*smiren* (1) 'serene'
sir (1) 'cheese'	*sveća* (1) 'candle'	'motionless'	*nevin* (1) 'innocent'
ovca (1) 'sheep'	*zemlja* (1) 'ground'	*silan* (1) 'strong'	
ruža (1) 'rose'			
šećer (1) 'sugar'			
srebro (1) 'silver'			

Table 3: The most frequent adjectives and nouns and their connections

Besides, the same adjective can be used in two different similes, either to describe a physical characteristic (*čist kao sneg* 'clean as snow') or a person's character trait (*čist kao suza* 'pure as a tear'). Moreover, one and the same simile can be used in both senses: *mek kao pamuk* 'soft as cotton'.

A simile can undergo numerous variations. An adjective or a noun can be used in either Ekavian or

Iekavian form,[7] such as, for instance, Ekavian variant *beo kao sir* vs. Iekavian variant *bijel kao sir* 'white as cheese'. Some other variations can be observed as well: the use of diminutive forms of nouns (*lak kao pero* vs. *lak kao perce* 'light as feather') or collective nouns for plural forms (*nevin kao jagnje* vs. *nevini kao jagnjad* 'innocent as a lamb/innocent as lambs). In some cases, a non-literary form of either adjectives or nouns is used: *hladan kao led* vs. *ladan kao led* 'cold as ice' and *slobodan kao ptica* vs. *slobodan kao tica* 'free as a bird'. Finally, (near) synonyms are used as well: *oštar kao zmija* vs. *oštar kao guja* 'sharp as a snake' and *velik kao jaje* vs. *golem kao jaje* 'big as an egg'.

The similes retrieved from the srpELTeC corpus can be classified into the following groups based on the ground:

- figures referring to physical characteristics of objects or people (F=184): *zdrav kao jabuka* 'healthy as an apple', *čist kao sneg* 'clean as snow', *okrugao kao pun mesec* 'round as the full moon';
- figures referring to colors (F=133): *crn kao gavran* 'black as a raven', *crven kao krv* 'red as blood', *plav kao more* 'blue as the sea';
- figures used for describing a person's character or abilities (F=69): *ljut kao ris* 'angry as a lynx', *pljašljiv kao srna* 'timid as a roe deer' *čist kao suza* 'pure as a tear';
- figures representing tastes (F=2): sladak kao šećer 'sweet as sugar';
- other figures (F=28): *skup kao šafran* 'expensive as saffron', slobodan kao ptica 'free as a bird'.

Among the most commonly used adjectival grounds in similes are lexemes denoting color concepts, which can designate not only the color of an object, but also someone's emotional, mental, or physical state. Such frequency of use is expected since colors evoke the vividness of visual images, having a wide range of connotative meanings culturally associated with them (Mpouli, 2016; Filipović Kovačević, 2019). With respect to the whole period covered by the corpus based on the *specificity score* values, novels published in the 20$^{\text{th}}$ century are distinguished by a significantly positive use of colors as adjectival grounds, especially yellow (S=1.4), white (S=1.3), red (S=1), grey (S=0.8), black (S=0.6) and blue (S=0.4).

The nouns most commonly used for the description of physical characteristics of objects or people are *smrt* 'death' (*bled kao smrt* 'pale as death'), *led* 'ice' (*hladan kao led* 'cold as ice'), and *krpa* 'cloth' (*bled kao krpa* 'pale as cloth'), as well as an adjective *upisan* 'inscribed' (*lep kao upisan* 'beautiful as inscribed' ('pretty as a picture')), while a person's character or abilities are most frequently compared to the nouns *jagnje* 'lamb' (*nevin kao jagnje* 'innocent as a lamb'), *stena* 'rock' (*hladan kao stena* 'cold as a rock'), *anđeo* 'angel' (*čist kao anđeo* 'pure as an angel') or *devojka* 'girl' (*stidan kao devojka* 'bashful as a girl'). The nouns that name animals, used as vehicles in adjectival similes, are especially interesting because of their expressiveness, connotations and picturesqueness. The animals that are the most frequently featured in similes are *jagnje* 'lamb', *ovca* 'sheep', *srna* 'roe deer', or *detlić* 'woodpecker'.

5. Conclusion

This paper presents the use of the current version of the SrpELTeC corpus, consisting of Serbian prose works published between 1860 and 1920, in order to retrieve and annotate the instances of rhetoric figures, namely, similes and analyze their usage. As a result, we developed a method for the description of these figures, based on finite-state transducers that makes their retrieval and annotation in Serbian texts possible. The annotated texts were used to study their specific use with respect to the author, title, or publication date. In the future, we will collect other types of simile figures, for instance, those that use prepositional phrases instead of nouns, e.g. *težak kao od olova* 'heavy as if it were made out of lead', as well as verbal similes, e.g. *rikati kao vo* 'roar like a bull'. Besides, we plan to enrich the current annotation scheme with the attributes indicating semantic characteristics of the recognized similes. Our ultimate goal is to publish a database of simile figures used in Serbian novels written between 1860 and 1920.

[7] Two different pronunciations in Serbian.

Acknowledgements

This research was made possible through the support of COST Action CA 16204 *Distant Reading for European Literary History*. We would like to thank numerous volunteers from the Society for Language Resources and Technologies *Jerteh*[8] who helped the production of the SrpELTeC corpus by correcting and annotating the novels.

6. Appendix A. List of the novels from the srpELTeC corpus used in the research

Author	Title	Publication Year
Ćorović, Svetozar	Ženidba Pere Karantana	1905
	Brđani	1919
Dimitrijević, Jelena	Fati-Sultan	1907
	Nove	1912
Đorđević, Vladan	U front	1913
Gavrilović, Andra	Prve žrtve	1893
Gavrilović, Draga	Babadevojka	1887
Ignjatović, Jakov	Jedna ženidba	1862
	Vasa Rešpekt	1875
	Pojeta i advokat	1882
Ilić, Dragutin	Hadži Đera	1904
Janković, Milica	Pre sreće	1918
	Kaluđer iz Rusije	1919
	Neznani junaci	1919
Komarčić, Lazar	Dragocena ogrlica	1880
	Moj kočijaš	1887
	Jedna ugašena zvezda	1902
	Prosioci	1905
Kostić, Tadija	Gospoda seljaci	1896
	Prvo veselje	1903
Mamuzić, Stevan	Nejednaka braća	1896
Mijatović, Čedomilj	Ikonija, vezirova majka	1891
	Rajko od Rasine	1892
	Knez Gradoje od Orlova grada	1899
Milićević, Milan	Jurumusa i Fatima	1879
	Deset para	1881
Novaković, Stojan	Kaluđer i hajduk	1913
Nušić, Branislav	Opštinsko dete	1902
Popović Šapčanin, Milorad	Sanjalo	1888
Ranković, Svetolik	Porušeni ideali	1900
Sekulić, Isidora	Đakon Bogorodičine crkve	1919
Šišković, Dragomir	Jedan od mnogih - roman iz prestoničkog života	1920
Sremac, Stevan	Ivkova slava	1895
Sretenović, Mihailo	Radetića Mara – pripovetka iz seoskog života	1894
Stanković, Borisav	Uvela ruža	1899
	Pokojnikova žena	1902
Taletov, Pera	Novac - roman iz beogradskog života	1906
Uskoković, Milutin	Došljaci	1910
	Potrošene reči	1911
	Čedomir Ilić	1914
Veselinović, Janko	Seljanka	1893

References

Beardsley, M. C. (1981). *Aesthetics: Problems in the Philosophy of Criticism*. Indianapolis: Hackett Publishing.

Brehmer, B. (2009). Äquivalenzbeziehungen zwischen komparativen Phraseologismen im Serbischen und Deutschen. *Südslavistik online*, 1:141–164.

Filipović Kovačević, S. (2019). Metonymy-based Colour Metaphors Expressing Mind and Body States: Evidence from English and Serbian. *Godišnjak Filozofskog fakulteta u Novom Sadu*, 44(1):75–92.

[8] http://jerteh.rs/

Heiden, S., Magué, J.-P., and Pincemin, B. (2010). Txm: Une plateforme logicielle open-source pour la textométrie-conception et développement. In *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010*, volume 2, pages 1021–1032. Edizioni Universitarie di Lettere Economia Diritto.

Heiden, S. (2010). The txm platform: Building open-source textual analysis software compatible with the tei encoding scheme. In *24th Pacific Asia conference on language, information and computation*, pages 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University.

Hu, X., Song, W., Liu, L., Zhao, X., and Du, C. (2017). Automatic recognition of simile based on sequential model. In *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, volume 2, pages 410–413. IEEE.

Israel, M., Harding, J. R., and Tobin, V. (2004). On simile. In Achard, M. and Kemmer, S., Eds., *Language, Culture, and Mind*, pages 123–135. Stanford: Center for the Study of Language and Information.

Krstev, C., Obradović, I., Stanković, R., and Vitas, D. (2013). An approach to efficient processing of multi-word units. In *Computational Linguistics*, pages 109–129. Springer.

Krstev, C. (2008). *Processing of Serbian – Automata, Texts and Electronic dictionaries*. Faculty of Philology, University of Belgrade.

Liu, L., Hu, X., Song, W., Fu, R., Liu, T., and Hu, G. (2018). Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553, Brussels, Belgium, October-November. Association for Computational Linguistics.

Milošević, N. and Nenadić, G. (2016). As Cool as a Cucumber: Towards a Corpus of Contemporary Similes in Serbian. *arXiv preprint arXiv:1605.06319*.

Milošević, N. and Nenadić, G. (2018). Creating a contemporary corpus of similes in Serbian by using natural language processing. *arXiv preprint arXiv:1811.10422*.

Mitrović, J., Markantonatou, S., and Krstev, C. (2019). A cross-linguistic study on Greek and Serbian fixed similes and enrichment of lexical resources via crowdsourcing. In *Multiword expressions: drawing on data from Modern Greek and other languages. Bulletin of Scientific Terminology and Neologisms*, pages 1–17. Academy of Athens.

Mladenović, M. and Mitrović, J. (2013). Ontology of Rhetorical Figures for Serbian. In Habernal, I. and Matoušek, V., Eds., *Text, Speech, and Dialogue*, pages 386–393, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mladenović, M. (2016). Ontology-based rhetorical figures recognition. *Infotheca - Journal for Digital Humanities*, 16(1-2):24–47.

Mpouli, S. and Ganascia, J.-G. (2015). "Pale as death" or " pâle comme la mort": Frozen similes used as literary clichés. In *EUROPHRAS2015:Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, Malaga, Spain, June.

Mpouli, S. (2016). *Automatic annotation of similes in literary texts*. Ph.D. thesis, Université Pierre et Marie Curie.

Niculae, V. and Danescu-Niculescu-Mizil, C. (2014). Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2008–2018, Doha, Qatar, October. Association for Computational Linguistics.

Niculae, V. and Yaneva, V. (2013). Computational considerations of comparisons and similes. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 89–95, Sofia, Bulgaria, August. Association for Computational Linguistics.

Niculae, V. (2013). Comparison pattern matching and creative simile recognition. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, pages 110–114, Trento, Italy, November.

Qadir, A., Riloff, E., and Walker, M. A. (2015). Learning to recognize affective polarity in similes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 190–200, Lisbon, Portugal. Association for Computational Linguistics.

Qadir, A., Riloff, E., and Walker, M. A. (2016). Automatically Inferring Implicit Properties in Similes. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1223–1232, San Diego, California, June. Association for Computational Linguistics.

Savary, A. (2009). Multiflex: a multilingual finite-state tool for multi-word units. In *International Conference on Implementation and Application of Automata*, pages 237–240. Springer.

Yoshimura, E., Imono, M., Tsuchiya, S., and Watabe, H. (2015). A simile recognition system using a common-sense sensory association method. *Procedia Computer Science*, 60:55–62.

Zhang, P., Cai, Y., Chen, J., Chen, W., and Song, H. (2019). Combining part-of-speech tags and self-attention mechanism for simile recognition. *IEEE Access*, 7:163864–163876.