White as Snow, Black as Night – Similes in Old Serbian Literary Texts

UDC 811.163.41’322.2
DOI 10.18485/infotheca.2021.21.2.6

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

ABSTRACT: This paper outlines the use of simile rhetorical figure in the Serbian subcollection SrpELTeC, a part of the ELTeC collection. We analyze their use by different criteria: authors, time periods, author’s gender, novel’s size and novel’s popularity. We also analyze adjectives, nouns and markers used in similes found in SrpELTeC, as well as entities to which they apply. We briefly compare this results on an *ad hoc* sample of contemporary Serbian novels.

KEYWORDS: literary corpora, simile, local grammars, ELTeC, Serbian language

PAPER SUBMITTED: 22 November 2021
PAPER ACCEPTED: 28 December 2021

1 About Simile

Veale and Hao (2008, 253) describe similes as “a window to the folk knowledge, since explicit similes make use of highly evocative and inference-rich concepts to ground comparisons and make unfamiliar seem familiar.” They add that “the simile [...] is one common vehicle for folk wisdom, one that uses explicit syntactic means” (454). To illustrate comparison as rhetorical figure in Serbian language, and corroborate these claims, we will give an exquisite example: *U njenim očima on je bio: visok kao bor, mio kao proleće, dobar kao Anđeo hranitelj, mlad kao rujna zora, beo kao labud, lep kao prolećni dan, hrabar kao Obilić!* (In her eyes he was: tall as a pine, dear as spring, good as Angel fosterer, young as ruddy dawn, white as a swan, beautiful as a spring day, brave as Obilić!).[1]

[1] This sentence using 7 similes in sequence is from the novel *Vojnik Stojan: nedovršen ratni roman* (Soldier Stojan: an unfinished war novel) by Dragomir Petrović (1918), which was not used in this research.

In (Krstev, Jaćimović, and Vitas 2020) we presented a preliminary research on the use of simile rhetorical figure in the incomplete version of the SrpELTeC sub-collection,[2] which contained 41 novels. In this paper we will repeat and enhance this research on the 100 novels SrpELTeC sub-collection.

As pointed in (Israel, Harding, and Tobin 2004) both literal comparison and simile have the same recognizable formal structure, a surface form consisting of the following elements: the subject of comparison (TARGET, TOPIC or TENOR), the object of comparison (VEHICLE or SOURCE), a conjunction which signals a comparison (MARKER, in Serbian usually *kao* (as)), and the basis of the comparison implied by the expression (PROPERTY or GROUND), as illustrated by the following example:[3]

reče	*Pavle*	*beo*	*kao*	zid
said	Pavle	white	as	wall
	TARGET	GROUND	MARKER	VEHICLE

As pointed in (Brehmer 2009), the TARGET is usually not a part of a simile (we will corroborate this in Subsection 3.2). Similes usually have a closed structure, containing all three elements: GROUND, MARKER, VEHICLE, as in the previous example, but could also be open, if the attribute is not explicitly stated, but could be derived from the context (MARKER, VEHICLE), as in the following example:

Arhimandrit	*beše*	*(ljut)*	*kao*	*ris*
The archimandrite	was	(angry)	as	a lynx
TARGET		missing GROUND	MARKER	VEHICLE

In this paper we will consider only closed similes.

Besides adjective similes, which represent a class of multi-word expressions (MWE), verbal multi-word expressions (VMWE) are also used for comparison, forming simile figures if conventionalized, as pointed in (Niculae and Yaneva 2013) for English, (Mitrović, Markantonatou, and Krstev 2020) for Greek and (Мршевић-Радовић 1987) for Serbian. In this paper we will not deal with this type of similes, although we will briefly compare the two types at the end of Section 3.

[2] ELTeC is a multilingual collection of novels published in the period from 1840 to 1910. It is developed in the scope of the COST action CA 16204 *Distant Reading for the European Literary History*. SrpELTeC is the sub-collection containing novels in Serbian. More about this sub-collection can be found in (Trtovac, Milnović, and Krstev 2021) in the same issue.

[3] All examples in this paper will be from the SrpELTeC.

This paper is organized as follows. In Section 2 we will present the setting of our research and the methods used. How similes are used in SrpELTeC will be discussed in Subsection 3.1, while characteristics of these similes will be analyzed in Subsection 3.2. A brief comparison of the use of similes in Serbian novels from 19th to early 20th century and novels from the second half of the 20th and the beginning of the 21st century will be given in Section 4. Some directions for future work will be mentioned in Section 5.

2 Research Methods

ground	**marker**	**modification**	**vehicle**	**modification**
hitar	*kao*		*jelen*	
hitar	*kao*	*mlad*	*jelen*	
fast	as	(a young)	deer	
slobodan	*kao*		*ptica*	
slobodan	*kao*		*ptica*	*u gori*
free	as		a bird	(in a wood)
beo	*kao*		*sneg*	
beo	*kao*	*najbelji*	*sneg*	*u planini*
white	as	(the whitest)	snow	(in a mountain)

Table 1. Modifications of a vehicle in similes

The basic structure of simile figures – GROUND, MARKER, VEHICLE – can sometimes be modified, and modification concerns mostly the vehicle (or source). Two most frequent types of simile modifications are represented in Table 1.

Similes can also occur in variants, which can be the result of different pronunciation (Ekavian or Ijkevaian), use of dialects or variant forms, diminutives, etc. Some cases of variants are represented in Table 2.

In addition to that, similes do not always appear in a text in the expected word order (adjective – conjunction – noun), for instance:

...	*kao*	*sneg*	*belih*	*grudi*
...	as	snow	white	bossom
	MARKER	VEHICLE	GROUND	TARGET

Also, the main components of similes are not always contiguous, since insertions are possible:

ground	**marker**	**source**	**type**
beo	*kao*	*sneg*	Ekavian
bijel	*kao*	*snijeg*	Ijkevaian
white	as	snow	
hladan	*kao*	*led*	
ladan	*kao*	*led*	variant (non-literal)
cold	as	ice	
mlad	*kao*	*kap*	
mlad	*kao*	*kaplja*	synonym
mlad	*kao*	*kapljica*	diminutive
young	as	a drop	
beo	*kao*	*zid*	
beo	*kao*	*duvar*	synonym
white	as	wall	

Table 2. Variations of similes

…	*a*	*tanak*	*je*	*kao*	*prut*
…	and	slander	is	as	a twig
		GROUND		MARKER	VEHICLE

Some similes are not complete because occasionally two similes are contracted when they use the same vehicle:

brzo	*i*	*vešto*	*kao*	*mačka*
fast	and	skillful	as	a cat
GROUND		GROUND	MARKER	VEHICLE

Having all these in mind we used two methods to retrieve similes from SrpELTeC:

- We used local grammars in the form of finite-state automata implemented in Unitex,[4] which take care about all modifications, variations and possible changes in the text mentioned before and which were developed earlier on the smaller sample of novels (Krstev, Jaćimović, and Vitas 2020). The set developed within mentioned research comprised of 243 local grammars for the recognition of similes in Serbian texts.
- we used simple patterns to retrieve other possible occurrences of similes. These patterns were implemented in Unitex as well, and they rely on the

[4] Cross-platform corpus processing suite Unitex/Gramlab.

Serbian electronic dictionaries (Krstev 2008). One such simple pattern is:

```
<A>
(<jesam.V>+<E>)
(kao+ko+k('+')o+ka('+ka')+ka+nalik+poput+kano)
```

The pattern is composed of an adjective followed by the conjunction *kao* (in various forms) or other prepositions, with an optional form of the auxiliary *jesam* (to be) in between. The obtained results were manually filtered to reject false recognition.

3 Analysis of Results

3.1 Distribution of Similes in SrpELTeC

Using methods described in the previous section we retrieved 1,051 occurrences of the simile rhetoric figure, an average of 10.5 occurrences per novel. In five novels no simile figures were found: *Jedna ženidba* (SRP18620), *Jurmusa i Fatima* (SRP18790), *Srbin i Hrvatica* (SRP18921), *Pokojnikova žena* (SRP19022), *Stradija* (SRP19025). The highest number of occurrences – 59 – were retrieved from the novel *Hajduk Stanko* (SRP18963), followed by *Nove* (SRP19120) – 39 – and *Novac* (SRP19060) – 37. When sorted by relative frequencies,[5] the novel *Hajduk Stanko* still remains on the top – 5.98 – followed by *Radetića Mara* (SRP18940) – 5.75 – and *Borci* (SRP18891) – 5.48. It is interesting to note that the first and the third novel are written by the same author, Janko Veselinović, and his third novel in this corpus *Seljanka* (SRP18932) is ranked as 16th with the relative frequency 4.32. There are 29 novels with less than one simile per 10,000 words, and the relative frequency of similes in the whole collection is 2.20.

The relative frequency of simile in SrpELTeC novels per four corpus composition criteria[6] is represented in Figure 1. One can observe that the authors of T2 novels used less simile figures than the authors in other periods – data for T1 period is not really comparable since there are only two novels in that

[5] The relative frequency is calculated as the number of similes per 10,000 words.

[6] For the composition criteria of ELTeC collection and distribution of novels from SrpELTeC according to them see (Trtovac, Milnović, and Krstev 2021) in this issue.

period. Female authors tend to use fewer simile figures than male authors – data for authors of unknown gender is not significant since there is only one such author. It seems that longer novels use less simile figures, but it is hard to find an explanation for the dependency of the number of similes on a novel's size. One has to bear in mind that there are only 5 long novels in the whole corpus. It seems that the binary reprint parameter, representing the popularity and presumably the quality of novels, is not correlated with the use simile figures.

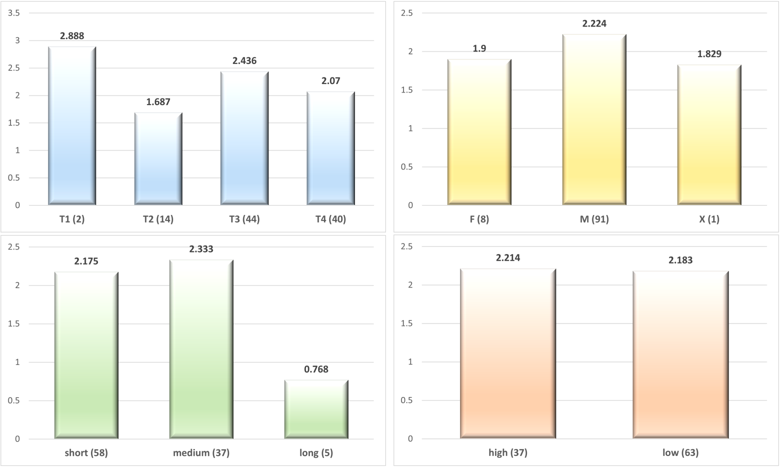

Figure 1. The relative frequency of simile in SrpELTeC novels per four corpus balance criteria: time slot, gender, size, reprint

When the novels are ordered by the year of publication and grouped into groups of 10, the average numbers of similes in groups appears to be uncorrelated with the year of publication (Figure 2, left). When the same is done for the size of novels, measured by number of words,[7] the result is the same – no correlation (Figure 2, right).

We ranked the authors who are represented in SrpELTeC by at least 3 novels according to the relative frequency of simile occurrences in all their novels, and the results can be seen in Figure 3.

[7] All novels are sorted by their size, and then grouped in groups of 10 according to their rank in the sorted list.

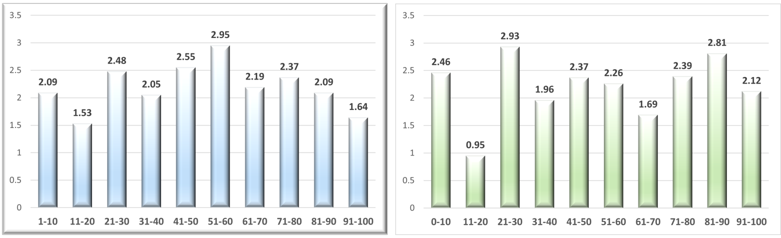

Figure 2. The relative frequency of simile in groups of 10 per: (a) year of the first publication; (b) number of words.

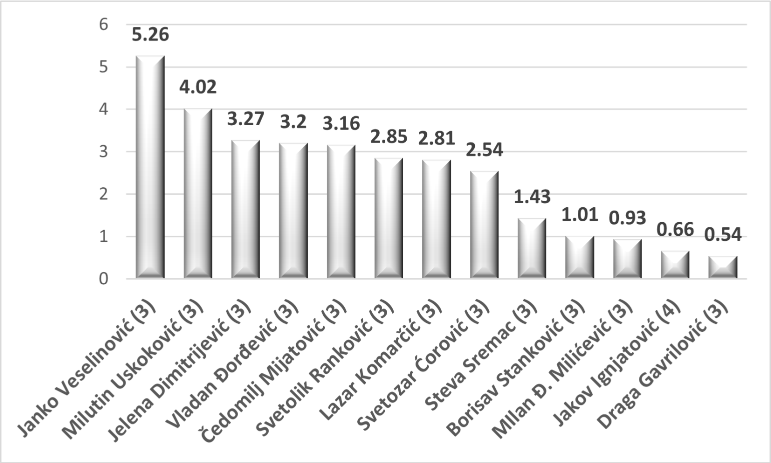

Figure 3. The relative frequency of simile for authors that have more than 3 novels in SrpELTeC.

As stated before, we retrieved 1,051 similes from SrpELtEC, of which 556 were different. We treated as equal similes that have the same property and the vehicle but can differ in the marker or due to modifications (see Table 1). We also treated as equal similes with different property and/or vehicle if that difference comes from different pronunciation or variant form (see first two rows in Table 2). There were 426 similes that occurred only once. Top ten similes in SrpELTeC are presented in Table 3. If we take the number of novels in which a simile appears as a sign of its popularity, then the obtained results show that the top 10 most frequent similes are also the most popular, although their rank is changed slightly, as seen in Table 3.

simile	**translation**	**absolute frequency**	**popularity in novels**	**rank by popularity**
beo kao sneg	white as snow	48	30	1
bled kao krpa	pale as a cloth	37	22	2
bled kao smrt	pale as death	37	14	4
hladan kao led	cold as ice	23	15	3
crven kao krv	red as blood	14	12	5
jasan kao dan	clear as day	13	9	8
mlad kao kaplja	young as a drop	13	11	6
plav kao nebo	blue as sky	12	9	9
crven kao rak	red as a crab	11	8	10
ljut kao ris	angry as a lynx	11	11	7

Table 3. The most frequent and the most popular similes

3.2 Characteristics of Similes

Ground – Adjectives A total of 202 different adjectives appear as properties among all extracted similes, with 102 of them in only one simile. The most frequent adjectives and nouns with which they combine and occur more than once are presented in Table 4. It can be seen that all most frequent adjectives combine with a number of nouns (column **No.** in Table 4 displays the number of nouns with which an adjective combines), and that the most frequently used noun barely exceeds 50% of all occurrences (the number in column **%** in Table 4 represents the percentage of appearances of the most frequent noun among all occurrences). Moreover, similes in which an adjective combines with only one noun never occur more than twice.

adj.	**freq.**	**No.**	**%**	**nouns**
bled	103	16	35.9	37: krpa, smrt, 7: vosak, 5: mrtvac, 3: kip, 2: ljiljan, samrtnik, senka
pale				37: cloth, death, 7: wax, 5: dead person, 3: statue, 2: lily, dying person, shadow
beo	83	21	57.8	48: sneg, 9: mleko, 3: alabaster, ovca, zid, 2: krin
white				48: snow, 9: milk, 3: alabaster, sheep, wall, 2: lily
crn	52	22	13.5	7: gar, 6: noć, 5: ugljen, zift, 4: gavran, zemlja, 3: gak, trnjina, 2: ugalj
black				7: soot, 6: night, 5: coal, tar, 4: raven, earth, 3: grey heron, blackthorn, 2: coal
hladan	43	15	55.8	24: led, 5: stena, 2: grob
cold				24: ice, 5: rock, 2: grave
crven	37	10	37.8	14: krv, 11: rak, 4: paprika, 2:vatra
red				14: blood, 11: crab, 4: paprika, 2: fire

Table 4. The most frequent adjectives and nouns with which they combine

Similes are often used to describe colors. There are 9 adjectives representing colors in our corpus: *beo* (with Ijekavian variant *bijel*) (white), *crn* (black), *crven* (red), *plav* (blue), *zelen* (green), *žut* (yellow), *siv* (gray), *modar* (livid), *rumen* (ruddy), with 3 additional that are derivatives or compounds: *žućkast* (yellowish), *bledomrk* (pale dark), *bledožut* (pale yellow).

Vehicle - Nouns In extracted similes 356 different nouns appear as vehicles, 209 of them in only one simile. The most frequent nouns and adjectives with which they combine and occur more than once are presented in Table 5. It can be seen that even most frequent nouns combine with just a few adjectives, and that the most frequently used adjectives always exceed $^2/_3$ of all occurrences. Moreover, the noun *krpa* (cloth), in the third row on the list, is used with only one adjective, *bled* (pale).[8] A similar situation is with nouns *led* (ice) and *krv* (blood), which are used with two variants or two synonymous adjectives only, of which one is strongly preferred.

[8] It is interesting to note that an analogous simile exists in Greek (Mitrović, Markantonatou, and Krstev 2020), translated to English as “white as cloth”. Hanks (2004) does not mention *cloth* or anything similar among artefacts used in similes in English.

noun	freq.	No.	%	adjectives
sneg	53	3	90.6	48: beo, 4: čist, 1: nedotaknut
snow				48: white, 4: clean, 1: untouched
smrt	45	8	82.2	37: bled, 2: lagan, 1: beo, hladan, jak, nem, nepomičan, ukočen
death				37: pale, 2: light, 1: white, cold, strong, mute, immovable, stiff
krpa	37	1	100.0	37: bled
cloth				37: pale
led	24	2	83.3	20: hladan, 4: ladan
ice				20: cold, 4: cold
krv	18	2	77.8	14: crven, 4: rumen
blood				14: red, 4: ruddy
nebo	18	5	66.7	12: plav, 2: vedar, 1: navodnjen, širok, taman
sky				12: blue, 2: clear, 1: watery, wide, dark

Table 5. The most frequent nouns and adjectives with which they combine

Nouns that refer to animals and plants are often used in similes,[9] although there are no such nouns among the most frequently used nouns in similes presented in Table 5. There are as many as 70 nouns referring to animals, and 34 referring to plants in similes in SrpELTeC, and the most frequent nouns of this kind and adjectives with which they combine are presented in tables 6 and 7.

One can see that an animal or a plant are sometimes used to denote a specific quality (examples are *ris* (lynx) and *dren* (dogwood)), while sometimes they are associated with various qualities (examples are *mačka* (cat) and *jabuka* (apple)). Animals and plants used in similes are mostly those with which users of Serbian are familiar, although *lav* (lion) and *tigar* (tiger) are also mentioned: *strašan kao lav* (dreadful as a lion) and *brz kao tigar* (quick as a tiger).

As far as proper names are concerned, they appeared in only 4 cases: *dobar kao Hristos* (good as Christ), *crn kao Arapin* (black as an Arab), *brz kao Grk* (quick as a Greek), and *strog kao Turčin* (strict as a Turk).

Marker – conjunctions The conjunction *kao* (as) is by far the mostly used as a marker in similes retrieved from SrpELTeC. Sometimes it is used in a

[9] Hanks (2004) notes that animals often occur as “secondary subjects” in similes and presents a list of 33 animals appearing in conventionalized similes in English.

noun.	**freq.**	**No.**	**%**	**nouns**
jagnje	13	6	46.2	8: miran, 1: blag, dobar, poslušan, smiren, umiljat
lamb				8: calm, 1: mild, good, docile, calm, amiable
ris	12	2	91.7	11: ljut, 1: ljutit
lynx				11: angry, 2: angry
rak	11	1	100	11: crven
crab				11: red
mačka	10	5	40.0	4: oprezan, 3: brz, 1: lagan, pakostan, vešt
cat				4: careful, 3: quick, 1: light, spiteful, dexterous
ovca	10	2	70.0	7: sed, 3: beo
sheep				14: gray-haired, 3: white

Table 6. The most frequent nouns referring to animals and adjectives with which they combine.

noun.	**freq.**	**No.**	**%**	**nouns**
jabuka	14	5	35.7	5: rumen, 4: pun, 3: zdrav, 1: jedar, okrugao
apple				8: ruddy, 4: plump, 3: healthy, 1: sturdy, round
bor	11	4	36.4	5: prav, 4: visok, 1: zdrav, dičan
pine	11			5: upright, 4: tall, 1: healthy, worthy
dren	7	1	100.0	7: zdrav
dogwood				7: healthy
jela	6	4	33.3	2: prav, vit, 1: vitak, izrastao
fir				2: upright, slender, 1: slender, grown
paprika	5	2	80.0	4: crven, 1: ljut
paprika				4: red, 1: angry

Table 7. The most frequent nouns referring to plants and adjectives with which they combine.

modified form to convey the spoken language in informal speech: *ka'*, *k'o*, *ko*, *ka*, *kano*. Prepositions *nalik na* and *poput* can also be used sometimes, but that was rare in SrpELTeC corpus: we retrieved only two cases using *poput*: *blijed poput krpe* (pale as a cloth) and *raširen poput lepeze* (spread as a fan).

Target The target of the large part of similes extracted from SrpELTeC is a person (see Figure 4, left): a man (335), a woman (184), a child (8), or a group of people (46). A person's appearance (body part) is often referred to, as shown in Figure 4 (right) for all body parts occurring more than once. A person's cloths are mentioned as well: cloths in general (*odelo*, *odeća*, *ruho*) 10 times, *košulja* (shirt) 6 times, *haljina* (dress) 3 times.

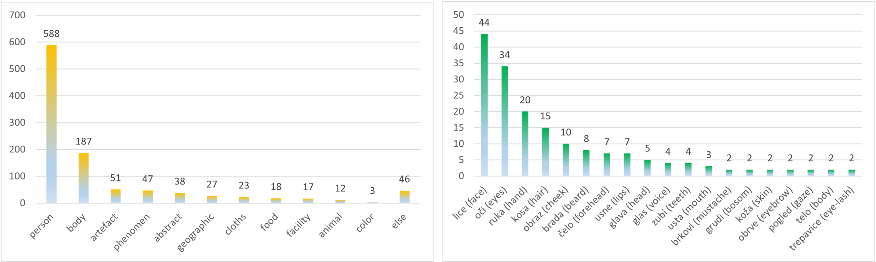

Figure 4. Frequency of targets classified in large groups (left); frequency of body parts used as targets (right).

Extracted data show that some similes are used to describe a large variety of entities. For instance, the most frequent simile *beo kao sneg* (white as snow) connects to many different targets: men, women (individually and collectively), body parts, cloths, animals, household items etc. On the other hand the second on the list, *bled kao krpa* (pale as a cloth) is used only for persons – a man (21), a woman (12), a group of people (2) and a person's face (2), similarly as the third one *bled kao smrt* (pale as death) – a man (20), a woman (14), a group of people (1) and a person's face (2). There are some similes that are used to describe only persons, e.g. *sed kao ovca* (gay-haired as a sheep), while others are used to describe natural or weather phenomena, e.g. *gust kao testo* (thick as dough) is used in SrpELTeC to desribe: *magla* (fog), *mrak*, *pomrčina*, *tama* (darkness), *nebo* (sky), *noć* (night).

Conventionalized similes vs. metaphors One can presume that among similes that occur only once in SrpELTeC there are many which are not conventionalized, or fixed similes, but rather introduced by authors. We tried to analyze the use of this non-conventionalized or infrequent similes by authors who are represented with at least 3 novels in SrpELTEC, and counted the number of these similes among all similes used by a particular author. The results are presented in Figure 5. One can observe that some authors (Svetozar Ćorović and Milutin Uskoković) avoid the use of conventionalized similes, while the others prefer them (Lazar Komarčić and Čedomilj Mijatović). The other authors tend to use both conventionalized and unique similes to a different extent; however, Milan Đ. Milićević, Borisav Stanković and Draga Gavrilović use few similes so it is not possible to establish their preference. On the average, these thirteen authors used almost as many unique similes, as similes shared with other authors – 50.8% of unique similes among all similes used.

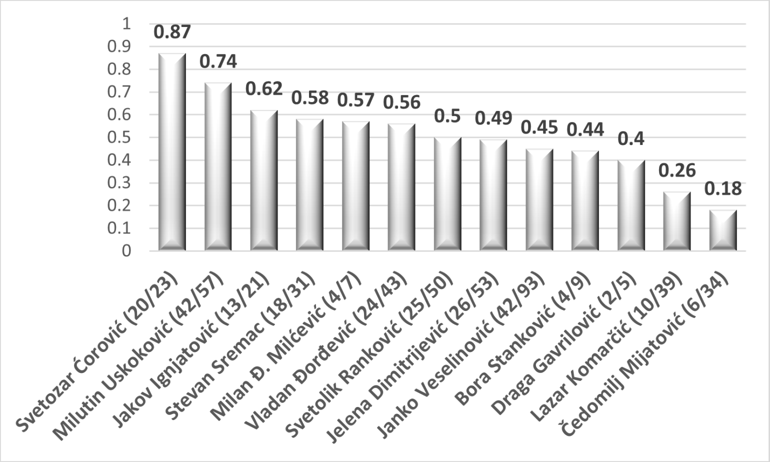

Figure 5. The use of unique similes by authors represented by at least 3 novels in SrpELTeC.

Similes as Verbal MWEs Similes as verbal MWEs consist of, according to (Qadir, Riloff, and Walker 2015), four components: besides the subject of the comparison, the object of the comparison, and the comparator (marker), there is a fourth element, an EVENT, ACT or STATE. Here is an example:

Binko	*ciknu*	*kao*	*guja*	kad ugleda popa.
Binko	squealed	like	a snake	when he saw the priest
TARGET	EVENT	MARKER	VEHICLE	

A GROUND can be optionally included as well; however, it is rare since the intended meaning is usually conveyed without it:

Binko	*ciknu*	(*strašno*)	*kao*	*guja*	kad ugleda popa.
Binko	squealed	(terribly)	like	a snake	when he saw the priest
TARGET	EVENT	(GROUND)	MARKER	VEHICLE	

We have extracted 1,067 verbal similes from SrpELTeC using a regular expression analogous to the one for property similes (method 2 mentioned in Section 2). In this paper we will not analyze this type of similes in depth; we will limit our presentation to only two aspects. The first is the extent to which verbs derived from names of colors are used in verbal similes. The results are presented in Table 8. One can observe that for each color that is not a compound nor a derivative and is used in adjective similes in SrpELTeC, except *siv* (gray), a derived verb participates in verbals similes as well, using in many cases the same vehicles.

Finally, we analyzed to what extent adjective and verbal similes used the same nouns. We found that they have 141 nouns in common. From six most frequently used nouns in adjective similes (Table 5), *led* (ice) and *nebo* (sky) were not used by verbal similes. The other four, *sneg* (snow), *smrt* (death), *krpa* (cloth), *krv* (blood), were used in verbal similes as well, conveying similar meanings, although *sneg* and *smrt* also appear in a different context: *raskraviti se kao sneg* (to loosen up like snow), *kositi kao smrt* (to mow like death), *zvoniti kao smrt* (to ring like death).

All most frequently used animals in adjective similes appear also in verbal similes, conveying the same or similar meaning. For instance, *rak* (crab) is used both in adjective and verbal similes only for its red color. The example of *jagnje* (lamb) is interesting. In adjective similes it is used to characterize someone who is calm, mild, good, etc. While there is no lexical connection between the adjectives and the verbs used, the verbal simile *spavati kao jagnje* (to sleep like a lamb) conveys a similar meaning: only somebody calm, mild, good, etc. can sleep sound.

Plants *jela* (fir) and *dren* (dogwood) do not appear in verbal similes. *Paprika* is used for its red color, *jabuka* (apple) for its red or ruddy color, while *bor* (pine) is used in *porasti kao bor* (to grow up as a pine), meaning to become tall, and consequently upright and slender.

adj.	**verbs**	**nouns**
beo	BELETI SE, *pobeleti*	*sneg**, *kreč**, *visibaba*
white	to be, to become white	snow, lime, snowdrop
crn	CRNETI SE, *pocrneti*	*zift**, *zemlja**, *strnjište*, *Ciganin*
black	to be, to become black	tar, earth, sttuble-field, Gypsy
crven	CRVENETI SE, *pocrvneti*, *zacrveneti se*	*rak**, *paprika**, *jabuka*, *ruža*, *trešnja*,...
red	to be, to become red	crab, paprika, apple, rose, cherry,...
plav	PLAVITI SE	*čivit**
blue	to be blue	indigo
zelen	*pozeleneti*	*trava**, *žuć*, *gušter*
green	to become green	grass, bile, lizard
žut	ŽUTETI SE, *požuteti*	*limun**, *vosak**, *smilje**, *dukat*, *ćilibar*
yellow	to be, to become yellow	lemon, wax, immortelle, gold coin, amber
modar	MODRETI SE, *pomodreti*	*čivit**, *more*, *smokva*
livid	to be, to become livid	indigo, sea, fig
rumen	RUMENETI SE, *porumeneti*, *zarumeneti se*	*jabuka**, *ruža**, *jagoda**, *gvožđe*, *žeravica*,...
ruddy	to be, to become pink	apple, rose, strawberry, iron, ember,...

Table 8. Verbs derived from colors used in verbal similes in SrpELTeC; verbs in small caps are imperfective; nouns with an * are also used in adjective similes with the corresponding color.

4 Simile in Contemporary Serbian Novels

In order to compare the use of similes in SrpELTeC, which comprises Serbian novels from 1840-1920, with contemporary novels, we compiled an *ad hoc* collection of 22 novels published from 1954 to 2010, which we will call Novels22.[10] The collection contains almost 1.6M words, an average of 72,281 words per novel.[11]

In order to retrieve similes from this collection we used local grammars in the form of finite-state graphs, as explained in Section 2 and in (Krstev, Jaćimović, and Vitas 2020). We enhanced the initial set of local grammars, developed for similes of the incomplete version of SrpELTeC, on the basis of retrieved similes from the complete SrpELTeC, so that now it contains graphs for 557 different similes, while each of these graphs takes care about modifications and variations listed in tables 1 and 2.

All most frequent similes in SrpELTeC (listed in Table 3) except one (*mlad kao kaplja* (young as a drop)) appear also in Novels22, some of them several times: *beo kao sneg* (white as snow), *bled kao krpa* (pale as a cloth), *crven kao krv* (red as blood). However, besides most frequent similes we retrieved in Novels22 some similes that appeared in SrpELTeC only once, e.g. *crn kao Arapin* (black as an Arab) and *lak kao dim* (light as smoke).

The graphs retrieved 69 similes from Novels22, that is, 3.14 per novel (compared to 10.5 in SrpELTeC), or 0.434 per 10,000 words in a novel (compared to 2.20 in SrpELTeC). In interpreting these results one should keep in mind that these graphs can retrieve only similes that were confirmed in SrpELTeC. However, if we consider that, on the average, authors of SrpELTeC used as many unique similes as those used by other authors (Subsection 3.2), an estimate of the average number of similes used by Novels22 authors would be approximately double (6.28), which is still considerably bellow the average use in SrpELTeC. This can suggest that either authors of modern novels do not use similes as much as their predecessors or that the repertoire of similes has changed over time. In order to confirm or reject either of these

[10] This collection contains 14 novels from the German-Serbian parallel corpus (Andonovski, Šandrih, and Kitanović 2019). The corpus comprises 7 novels originally written in Serbian and 7 novels written in German and translated to Serbian. The remaining 8 novels were taken from the Anthology of Serbian Literature.

[11] According to ELTeC classification the collection comprises 8 short, 12 medium sized, and 2 long novels.

hypotheses a systematic research of simile use in contemporary novels is needed.

5 Conclusion

The results presented here will serve as the basis for a future database of similes in Serbian, from which local grammars that recognize and tag them in a text can be automatically produced. Our future work will go in two directions. We will examine the use of simile figures in the Serbian language in literary and other texts by using both general corpora and literary corpora covering different time periods. Also, we will expand our research to similes differing in structure, e.g. *čvrst kao od čelika* (firm as from steel), and to verbal similes like *sevati kao munja* (to blaze as lightening) and *liti kao iz kabla* (to pour as from a bucket).

References

Andonovski, Jelena, Branislava Šandrih, and Olivera Kitanović. 2019. "Bilingual lexical extraction based on word alignment for improving corpus search." *The Electronic Library.*

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

Hanks, Patrick. 2004. "Similes and Sets: the English Preposition like." In *Jazyky a jazykoveda (Languages and Linguistics : Festschrift for Professor Fr. Čermák,* edited by R. Blatná and V. Petkevič. Prague: Philosophy Faculty of the charles University.

Israel, Michael, Jennifer Riddle Harding, and Vera Tobin. 2004. "On simile." *Language, culture, and mind* 100.

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

Krstev, Cvetana, Jelena Jaćimović, and Duško Vitas. 2020. "Analysis of Similes in Serbian Literary Texts (1840-1920) Using Computational Methods." In *Proc. of the 4^{th} Int. Conference Computational Linguistics in Bulgaria (CLIB 2020),* edited by Svetla Koeva, 31–41. Sofia, Bulgaria: Institute for Bulgarian Language "Prof. Lyubomir Andreychin", Bulgarian Academy of Sciences.

Mitrović, Jelena, Stella Markantonatou, and Cvetana Krstev. 2020. “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,* edited by Stella Markantonatou and Anastasia Christofidou, 241–262. Research Centre for Scientific Terms / Neologism.

Niculae, Vlad, and Victoria Yaneva. 2013. “Computational considerations of comparisons and similes.” In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop,* 89–95.

Qadir, Ashequl, Ellen Riloff, and Marilyn Walker. 2015. “Learning to recognize affective polarity in similes.” In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* 190–200.

Trtovac, Aleksandra, Vasilije Milnović, and Cvetana Krstev. 2021. “The Serbian Part of the ELTeC Collection – from the Empty List to the 100 Novels Collection.” *Infotheca - Journal for Digital Humanities* 21 (2): 7–25. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.1.

Veale, Tony, and Yanfen Hao. 2008. “Enriching WordNet with folk knowledge and stereotypes.” In *Proceedings of GWC,* 453–461.

Мршевић-Радовић, Драгана. 1987. *Фразеолошке глаголско-именичке синтагме у савременом српскохрватском језику.* Београд: Филолошки факултет Универзитета у Београду.