Journées « Unitex »

Description of Similes, their Retrieval and Annotation with Unitex

Cvetana Krstev

Faculty of Philology, University of Belgrade
Society of Language Resources and Tools – JeRTeh

15-16 Décembre 2021
Paris – Belgrade

Overview

1. About simile figures
2. About the Corpus
3. Simile retrieval and annotation
4. Analysis of the results
5. Conclusion

About simile figures

Similes

Similes are rhetorical figures which play an important role in literary texts.
They are often quite conventionalized, generally known and accepted phrases used by all members of a linguistic community.

- an explicit form of comparison
- essentially figurative, making unexpected connections between literally unlike concepts

An example

What are similes and how are they used is best illustrated with one sentence from one novel (1918):

Example

U njenim očima on je bio: visok kao bor,[1] mio kao proleće,[2] dobar kao Anđeo hranitelj,[3] mlad kao rujna zora,[4] beo kao labud,[5] lep kao prolećni dan,[6] hrabar kao Obilić![7]

'In her eyes he was: tall as a pine,[1] dear as spring,[2] good as Angel fosterer,[3] young as ruddy dawn,[4] white as a swan,[5] beautiful as a spring day,[6] brave as Obilić![7]'

Simile structure

Similes have the multi-word structure, recognizable surface form consisting of the following elements:

Formal three-part structure (*closed simile*)

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

Formal two-part structure (*open simile*)

[She]	was	[as]	[a bird].
tenor		marker	vehicle

Automated processing of similes

Similes have a structure that appears fairly amenable to automated processing.
Simile analysis has become a particularly appealing topic of interest in the field of computational linguistics and corpus studies in recent years.
Automatic simile recognition can be divided into partial (retrieving specific simile patterns, either complete expressions consisting of all simile elements, or only preselected grounds and vehicles) and full (extraction and analysis of all sentences containing a simile marker in unstructured texts) simile identification.

Goals

- 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.
- to provide a method for describing simile figures and to build tools for retrieving them in arbitrary texts.
- to identify the most frequent similes and their components (such as grounds and vehicles), and to analyse their use.

About the Corpus

COST ACTION CA16204 (2017–2022)

compilation of a multilingual European Literary Text Collection (ELTeC)

expected to comprise around 2,500 full-text novels in at least 10 different languages

all texts from this corpus have to fulfill the following criteria:

- they should be originally written in a language of the sub-collection to which they belong,
- their first publication date should fall between 1840-1920 (preferably appearing as a book and not published in installments) and
- they should be at least 10,000 word tokens long.

srpELTeC

- The Serbian sub-collection is almost finalized, and some of the prepared novels might not become part of the final collection due to the balancing criteria that have to be met.[1]
- For this research 116 novels that are candidates for the Serbian sub-collection of the ELTeC corpus were used.
- This particular collection contains novels of exceptional value for the history of Serbian literature.

[1] *https://distantreading.github.io/ELTeC/*

srpELTeC distribution

Table 1: Corpus distribution

Period	Number	Length	Number	Sex	
1840-1859	2	short (<50K)	57	Male	91
1860-1879	17	medium	38	Female	8
1880-1899	41	long (>100K)	5	Unknown	1
1900-1920	42				

Reprint	Number	Authorship	Number
high	38	3 novels per author	11
low	62	1 novel per author	48

Simile retrieval and annotation

Simile description

To retrieve as many similes as possible from our corpus, we have adopted two approaches using the Unitex system:

- we looked for similes already recorded in the Serbian morphological dictionary of MWUs
- we applied a simple regular pattern to spot simile occurrences

DELAC (entry *gladan kao vuk* 'hungry as a wolf')

```
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
...
kao vuk gladnog,gladan kao vuk.A:adms4v
kao vukovi gladni,gladan kao vuk.A:aemp1g
```

Simile in DELAC

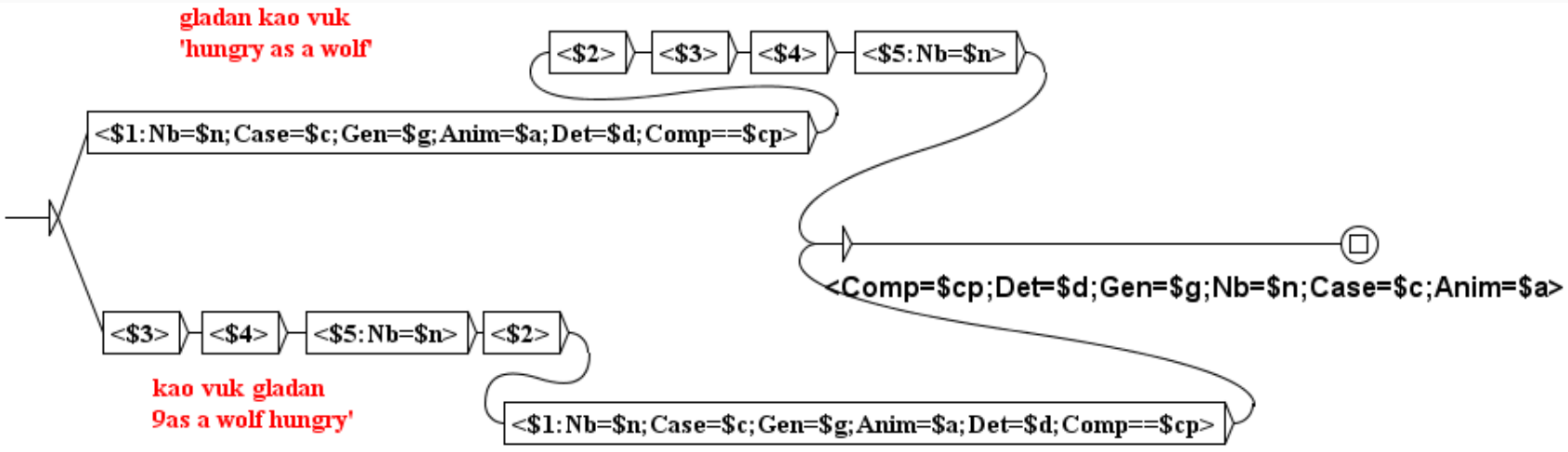


Figure 1: The inflectional graph for MWU of the type “A as N” – ground and vehicle can change places

DELAC description is not enough

- Descriptions of simile with inflectional graphs assume that there are no insertions between the parts of simile.
- They cannot capture variations in the parts of simile.

Example

mudar kao kakav pop ‘wise as some priest’
hladna je kao led ledeni ‘cold is (she) as ice icy’

Simile fishing

To retrieve as many similes as possible from our corpus we applied a simple regular pattern to spot simile occurrences

Regular expression

```
<A~Comp> (<јесам.V>+<E>)
(као+ко+к('+’)о+ка(’+ка')+ка+налик+попут+кано)
```

Positive examples

bled kao smrt ‘pale as death’

žut kao vosak ‘yellow as wax’

False examples

veseo kao nikad dotle ‘joyfull as never before’

(Ja sam) ogorčena kao žena ‘(I am) indignant as a woman’

Simile variations

A number of variations occurring in the use of similes cannot be described with e-dictionaries:

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	

Simile variations

A number of modifications occurring in the use of similes cannot be described with e-dictionaries:

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)

Similes in texts

Similes in text do not always occur int the form A kao N:

- variations due to the free word order kao led hladna ‘as ice cold’;
- insertion of an auxiliary, for instance, crvena **si** kao zreli nar ‘you **are** red as a ripe pomegranate’;
- insertion of a subject (tenor), for instance, vreo **dah** kao plamen ‘ hot **breath** as a flame’;
- insertion of a pronoun (clitic), for instance, privržen **mu** kao pašče ‘attached **to him** as a dog’;
- variations resulting from rephrasing, for instance, žut kao **što je** slama ‘yellow as straw **is**’.

Generally, parts of simile are not very distant in texts.

General simile description

Our solution is based on the fact that there are modification that can apply to all simile figures and there are some that are specific for particular simile figures.

- A generic finite-state automata (FSA) was developed that describes modifications that can apply to simile in general;
- All detected similes were described in a table – lexical variations and additions;
- A table and a generic graph were used to automatically produce a specific graph for each simile figure.

Local grammars in the form of finite-state automata are implemented in Unitex/Gramlab.

A generic simile graph

Generic graphs in some nodes refer to values in particular cells of a simile table.

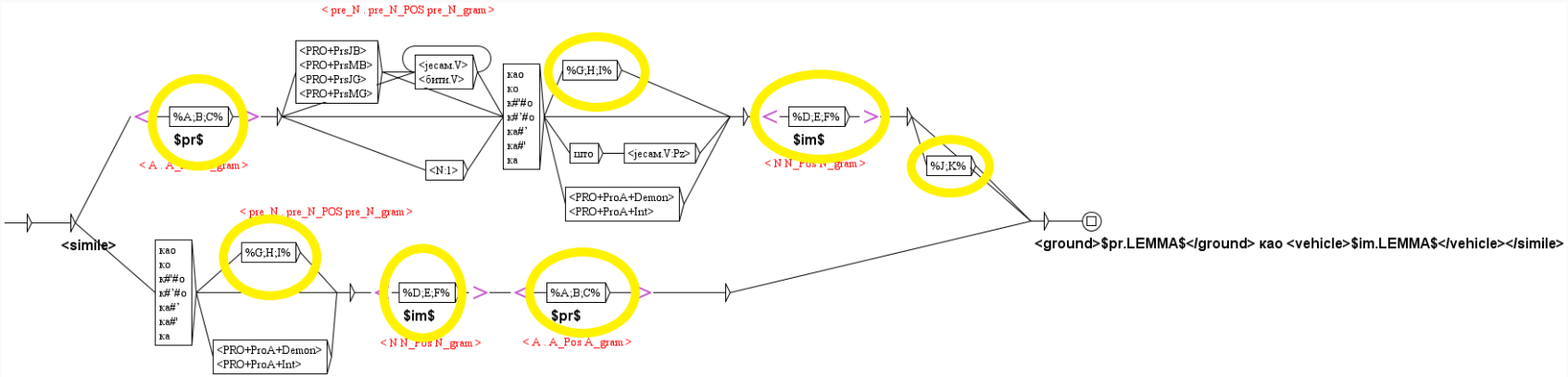


Figure 2: Generic graph that describes possible occurrences of simile in texts.

Tabular description of similes

Each line in the table describes one particular simile.

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	A	:as1		
crn	A	:a	gavran	N	:1	zlokoban	A	:as		
lep, lijep	A	:a	upisan	A	:as1					

A specific graph

In specific graphs all nodes are either literals or references to e-dictionaries of Serbian. It is produced automatically (Python script) from a generic graph and one table row.

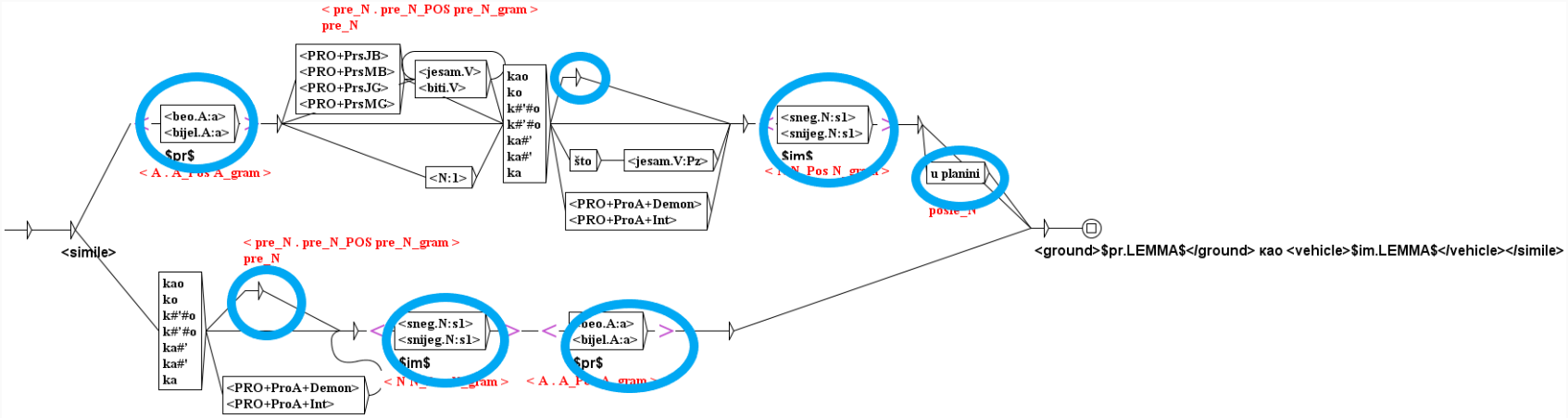


Figure 3: One automaton that recognizes the figure beo/bjel kao sneg/snijeg [u planini] 'white as snow [in the mountain]'.

A super-graph

A super-graph invokes all specific graphs and it is also produced (Python script) - it does not look very nice.

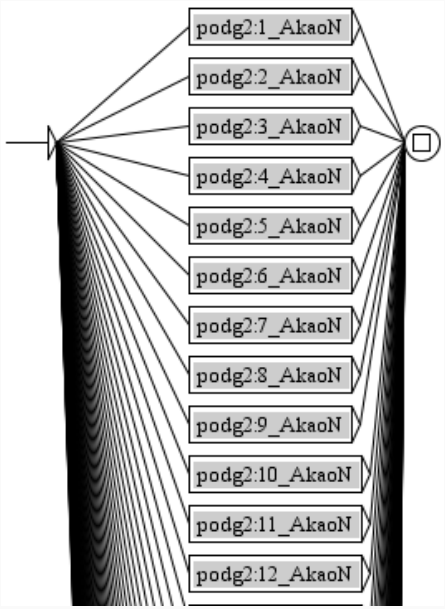


Figure 4: A super graph that invokes all specific graphs.

Simile annotation

Produced specific FST enable both retrieval of simile figures and their annotation at phrase and word levels.

The source text

...ridj i podbuo od pića a sad, radi posta, ljut kao ris...
red and puffy from the drink and now, because of the fast, angry as a lynx...

The annotated text

<simile>
<ground>ljut</ground>
kao
<vehicle>ris</vehicle>
</simile>

A generic graph for similes with two grounds

Sometimes two grounds are connected with one vehicle e.g. *zdrav i rumen kao jabuka* 'healthy and ruddy as an apple'

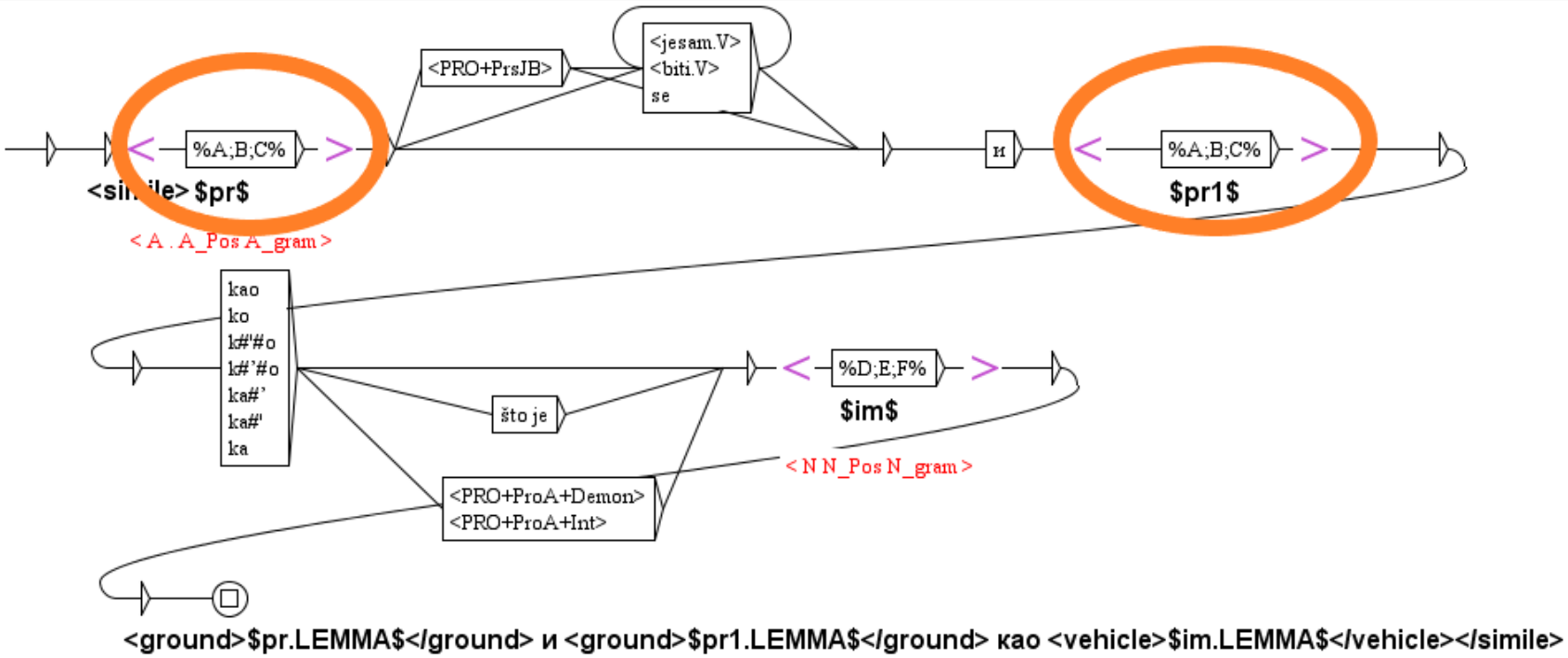


Figure 5: Generic graph that describes possible occurrences of similes with two grounds in texts.

Tabular description of similes with two grounds

It is the same table, but now sorted by the ground, which enables to have together all descriptions using the same noun.

A	B	C	D	E	F
jedar,	A	:a	jabuka,	N	:s1
pun	A	:a	jabuka	N	:s1
rumen	A	:a	jabuka,	N	:s1
zdrav	A	:a	jabuka	N	:s1

similes with jabuka 'apple' as a vehicle

{jedar, pun, rumen. zdrav} kao jabuka
'{firm, plump, ruddy, healthy} as an apple

A specific graph for similes with two grounds

In specific graphs all nodes are either literals or references to e-dictionaries of Serbian. It is produced automatically (Python script) from a generic graph and all table rows having the same noun in corresponding column.

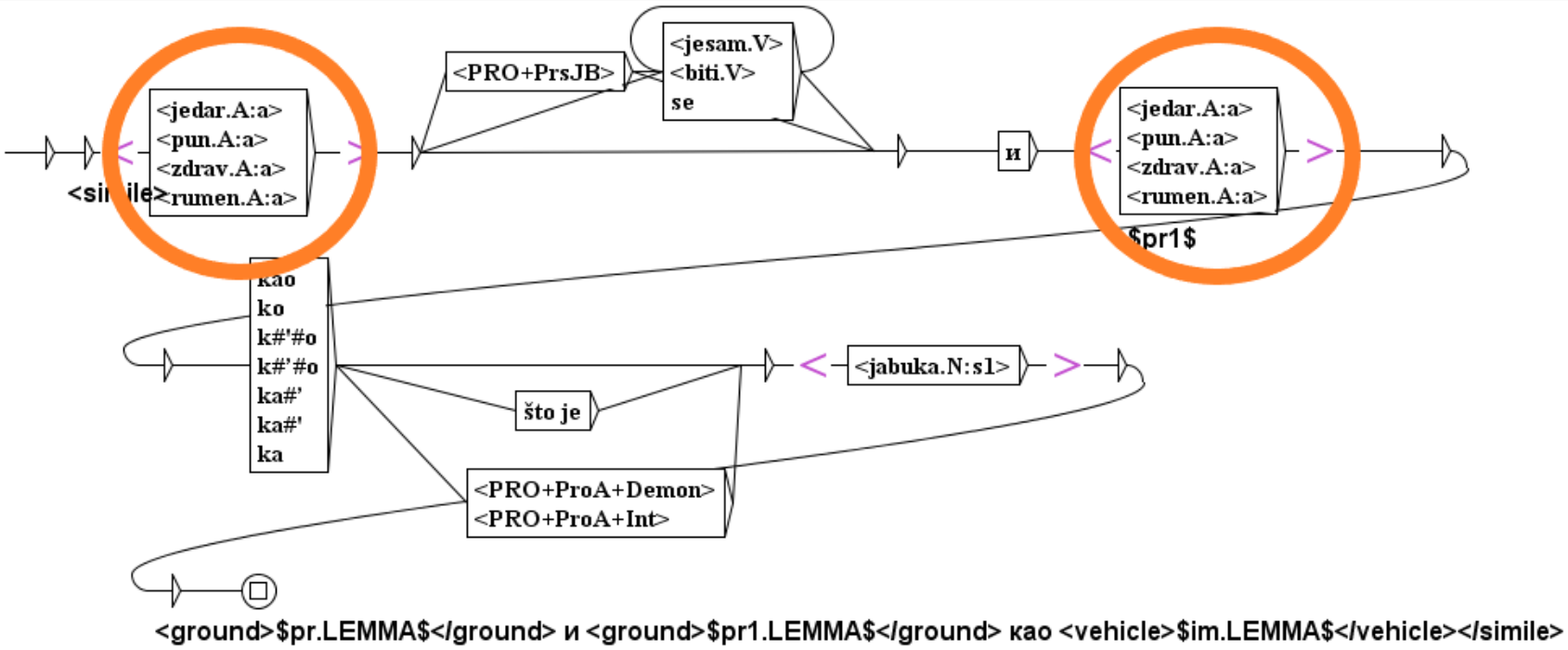


Figure 6: Generic graph that describes possible occurrences of similes with two grounds in texts.

Annotation of similes with two grounds

The source text

...Станко прић̈е дрвету, пљуну у длане и поче се пузати брзо и вешто као мачка.

Stanko approached the tree, spat in his palms and began to crawl quickly and skillfully like a cat.

The annotated text

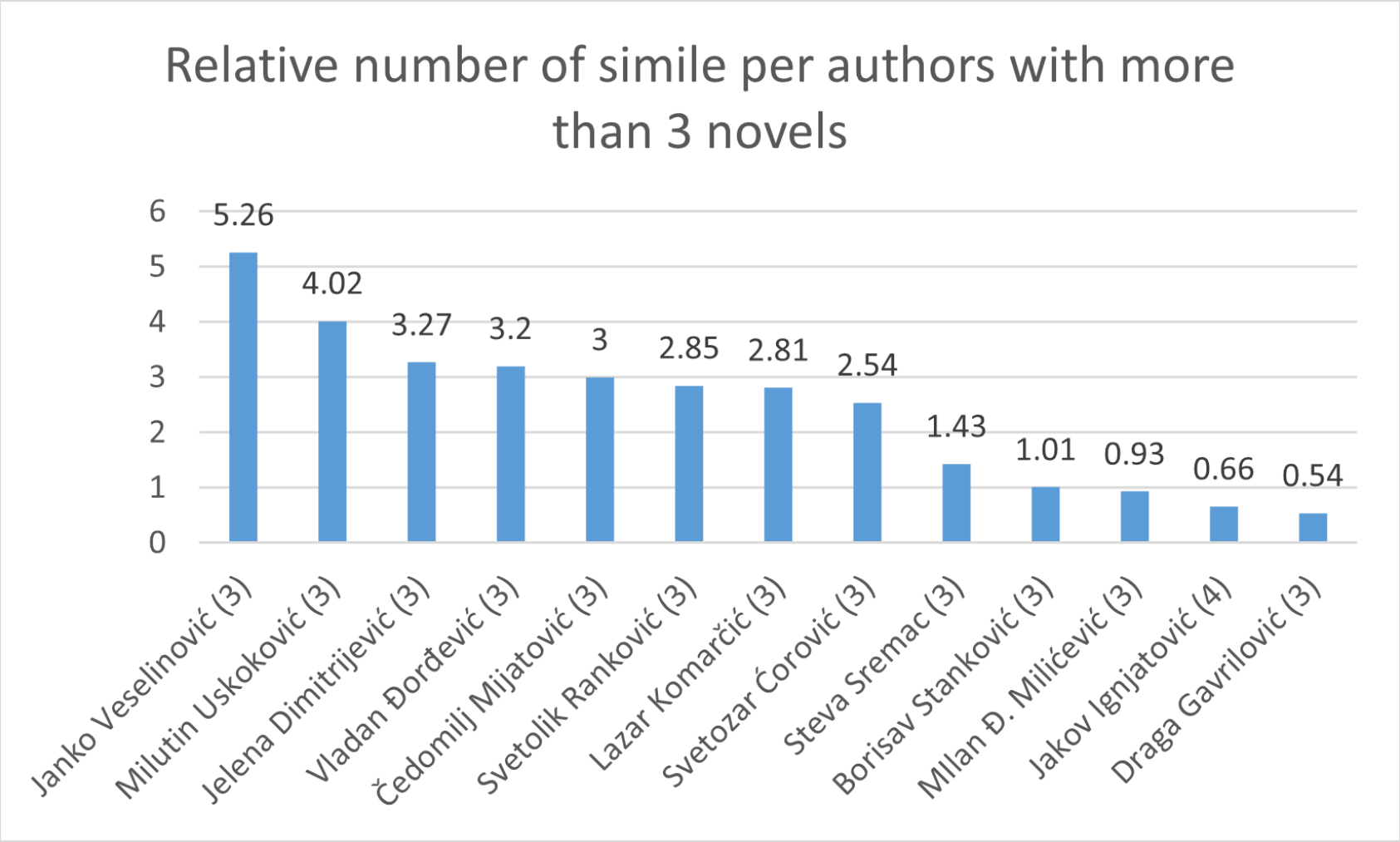
<simile>
*<ground>*brz*</ground>*
i
*<ground>*vešt*</ground>*
као
*<vehicle>*mačka*</vehicle>*
</simile>

Analysis of the results

Adjectival similes in ELTeC

- In the corpus consisting of 100 Serbian novels from SrpELTeC, 1,051 occurrences were retrived.
- Approximately 10.5 similes per novel, from 0 similes (in five novels) to 59 in one of the novels.
- On average, the relative frequency (number of similes per 10,000 words) is 2,20, 29 novels had a relative frequency < 1, and the novels with the highest absolute frequency had also the highest relative frequency 5.98.
- The number of different similes is 556 - this number corresponds to the number of rows in the table and consequently the number of specific graphs, that is beo kao sneg and bijel kao snijeg are counted as one.

The use of adjectival similes per authors



The most frequent similes

Popularity - in how many novels a simile was used

simile	translation	abs. freq.	popul. in novels	rank by popul.
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

The most frequently used adjectives

202 different adjectives

adj.	freq.	No.	nouns
bled pale	103	16	***37: krpa***, smrt, 7: vosak, 5: mrtvac, 3: kip, 2: ljiljan, samrtnik, senka... 37: cloth, death, 7: wax, 5: dead person, 3: statue, 2: lily, dying person, shadow...
beo white	83	21	***48: sneg***, 9: mleko, 3: alabaster, ovca, zid, 2: krin... 48: snow, 9: milk, 3: alabaster, sheep, wall, 2: lily...
crn black	52	22	***7: gar***, 6: noć, 5: ugljen, zift, 4: gavran, zemlja, 3: gak, trnjina, 2: ugalj... 7: soot, 6: night, 5: coal, tar, 4: raven, earth, 3: grey heron, blackthorn, 2: coal...

The most frequent used nouns

356 different nouns

noun	freq.	No.	adjectives
sneg	53	3	***48: beo***, 4: čist, 1: nedotaknut
snow			48: white, 4: clean, 1: untouched
smrt	45	8	***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	***37: bled***
cloth			37: pale
led	24	2	***24: hladan***
ice			24: cold
krv	18	2	***14: crven***, 4: rumen
blood			14: red, 4: ruddy

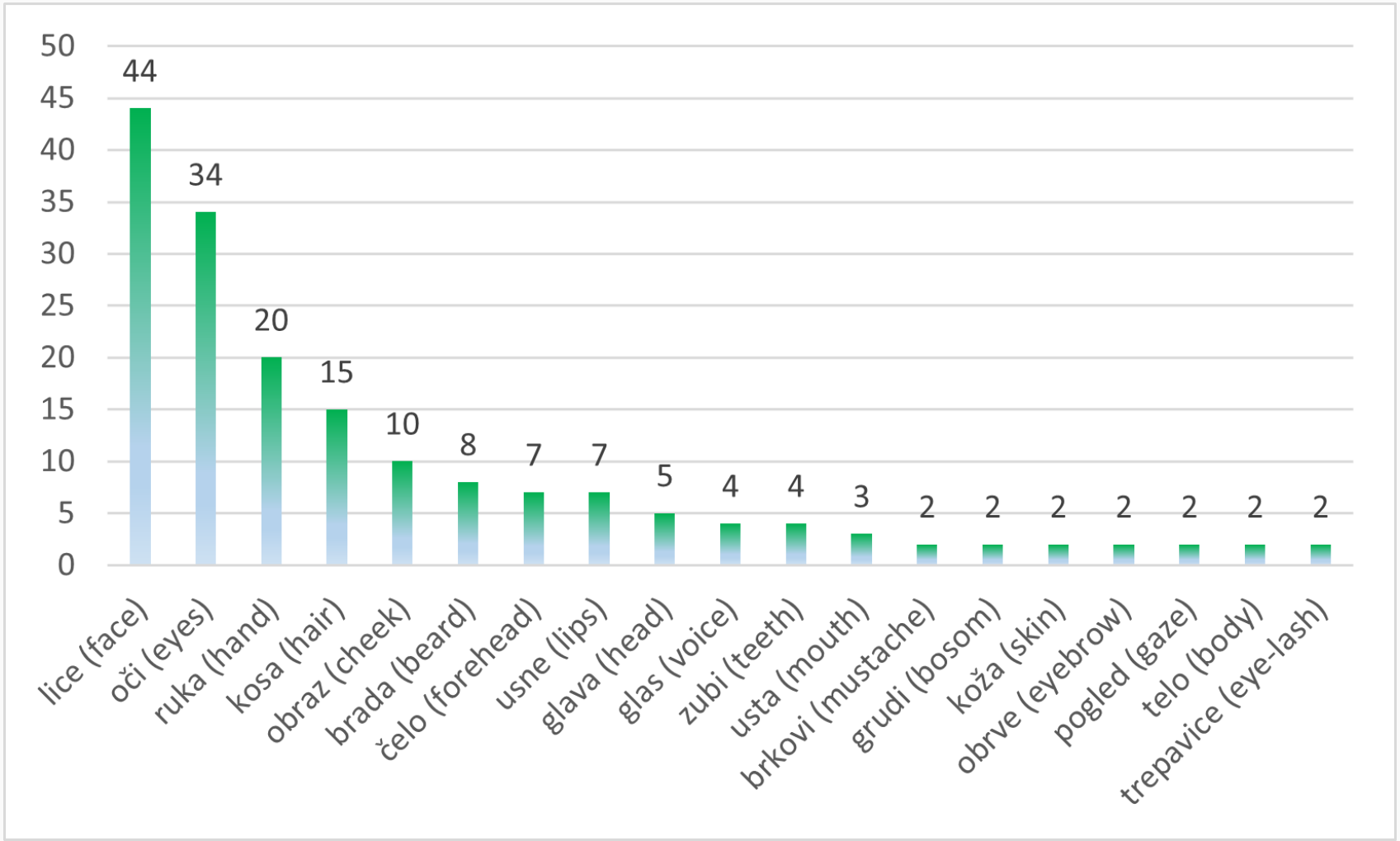
The most frequent animals and plants

70 nouns referring to animals, 34 to plants

noun.	freq.	No.	nouns
jagnje	13	6	***8: miran***, 1: blag, dobar, poslušan, smiren, umiljat
lamb			8: calm, 1: mild, good, docile, calm, amiable
ris	12	2	***11: ljut***, 1: ljutit
lynx			11: angry, 2: angry
rak	11	1	***11: crven***
crab			11: red
jabuka	14	5	***5: rumen***, 4: pun, 3: zdrav, 1: jedar, okrugao
apple			8: ruddy, 4: plump, 3: healthy, 1: sturdy, round
bor	11	4	***5: prav***, 4: visok, 1: zdrav, dičan
pine	11		5: upright, 4: tall, 1: healthy, worthy

The most frequent tenors

The target similes are persons (588), person's appearance (body part) (187), artefacts (51), phenomena (47), ...



The most frequent

description of physical characteristics of objects or people

smrt ‘death’ (*bled kao smrt* ‘pale as death’)
led ‘ice’ (*hladan kao led* ‘cold as ice’)
krpa ‘cloth’ (*bled kao krpa* ‘pale as cloth’)
upisan ‘inscribed’ (*lep kao upisan* ‘beautiful as inscribed’ (‘pretty as a picture’))

person’s character or abilities

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’)
devojka ‘girl’ (*stidan kao devojka* ‘bashful as a girl’)

Conclusion

Summary

- We developed a method for the description of these figures that makes their retrieval and annotation in Serbian texts possible.
- Now, we can apply our super-graph to arbitrary texts to retrieve simile figures (557 recorded in the table).
- For instance, we used it on a collection of 22 contemporary Serbian novels and find out that all top 10 similes from old novels (except one) are still in use.
- Also, it will help us enhance our table with new similes.

Future work

- We are collecting 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’ and drhtati kao u groznici ‘to tremble like in fever’;
- Now we have to describe these other types of similes and prepare generic graphs and tables for them;
- We will publish a database of simile figures used in Serbian novels written between 1860 and 1920.

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*[2] who helped the production of the SrpELTeC corpus by correcting and annotating the novels.

The full paper describing this research will be published in Infotheca 2021(2) that is dedicated to SrpELTeC collection – development and use.

[2] *http://jerteh.rs/*

Thank you

Maybe you have some questions?