Multiword Expressions – Comparative Analysis Based on Aligned Corpora

Cvetana Krstev
Association for Language
Resources and Technologies
Belgrade, Serbia
CvetanaJK@gmail.com

Ranka Stanković
University of Belgrade
F. of Mining and Geology
Belgrade, Serbia
ranka@rgf.bg.ac.rs

Aleksandra Marković
Institute for the
Serbian Language SASA
Belgrade, Serbia
malexa39@gmail.com

Relevant UniDive working groups: WG1, WG2

1 Introduction

The aim of our paper is to research inter- and intra-linguistic similarities/differences in the use of simile rhetorical figures (Niculae and Yaneva, 2013). We aimed at differences in the source concept in different languages, as well as in different means for expressing the comparison (simile, superlatives, compounds etc.). Our research relies on four bilingual aligned corpora containing mostly literary texts and involving English, French (Stanković et al., 2017), German (Andonovski et al., 2019) and Italian (Perišić et al., 2023) as one of the languages and Serbian as the other.[1]

Similes have the recognizable formal structure; their surface form consists of the subject of comparison, the object of comparison, a conjunction which signals a comparison, and the basis of the comparison implied by the expression. In our previous research (Mitrović et al., 2020; Krstev, 2021) we collected a set of 558 similes from Serbian literary texts and presented their structure in a form of finite-state automata (FST), which facilitates our present research. Each FST describes possible lexico-syntactic variants of a simile.

In the present study we focus on similes and their translations. Similes in Serbian texts were retrieved with the high precision (close to 100%) using the set of FSTs. We estimate that the recall is significantly lower, but the goal of this research was not to retrieve all similes in analysed texts. Similes in other languages were retrieved using CQL (corpus query language) incorporated in systems that support aligned corpora. The correspondence between similes in one language and its source or translation in another language were established by using the "close reading" technique.

2 Similarities Across Languages

We established previously that the most frequent simile in Serbian literary texts is *beo kao sneg* 'white as snow' (Krstev, 2021:126). Therefore it is not surprising that this figure occurs in all our corpora.

1984-en: ...with everything forgiven, his soul *white as snow*;

1984-fr: Tout était pardonné et son âme était *blanche comme neige*;

1984-sr: ...gde mu je sve bilo oprošteno, gde mu je duša bila *bela kao sneg*;

Eco-it: Chi aveva parlato era un monaco curvo per il peso degli anni, *bianco come la neve*;

Eco-sr: Te je reči izgovorio monah poguren pod teretom godina, *beo kao sneg*;

Jelinek-de: ...diese lang vergessene Weiblichkeit mit der Haut *so weiß wie Schnee* und dem Haar so schwarz wie Ebenholz.

Jelinek-sr: ...ta dugo zaboravljena ženstvenost sa kožom *belom kao sneg* i kosom crnom kao abonos.

Although the popularity of *beo kao sneg* 'white as snow' in Serbian was already established, in our aligned corpora it occurred only in translations to Serbian. Therefore, in all presented examples the phrase was used in the original; only **1984-fr** is the translation from English.

Today it is still natural to compare the pure white color to snow. In some cases, like *zdrav kao dren* 'healthy as dogwood', the motivation for using this particular tree for a comparison is blurred. In the French original and Italian translation different choices were observed: *chêne* 'oak' in French and *pesce* 'fish' in Italian.

Flaubert-fr: Quant à lui il se portait toujours *comme un chêne*;

Flaubert-sr: Što se njega tiče, on je *zdrav kao dren*;

Andrić-sr: ...i svaki uverava da je *zdrav kao dren* i da nema nikakve veze sa kolerom;

[1] English/Serbian (4.4MW) and French/Serbian (1.7MW) are available at Korpus, German/Serbian (1.6MW) and Italian/Serbian (1MW) at Biblisha, and English/Serbian and Italian/Serbian at Noske as well. These corpora are available to registered users.

Andrić-it: ...ognuno assicurava di essere *sano come un pesce* e di non aver nulla a che vedere col colera.

3 Differences Across Languages

The analysis of translations of retrieved Serbian similes in the German/Serbian corpus revealed that German compounds were in many cases translated to Serbian as simile figures: (a) *schneeweiß* 'snow-white' → *beo kao sneg*; (b) *gertenschlank* 'whip-slim' → *tanak kao trska* 'thin as a reed'.

Dor-de: Zu einem *schneeweißen* Hemd trug er eine schwarze Hose und ein schwarzes Gilet;

Dor-sr: Pored *kao sneg bele* košulje imao je na sebi samo crne pantalone i crno prsluče;

Grass-de: ...*gertenschlank* solo oder neben ihren gleichfalls zierlichen Ballettmeister gestellt;

Grass-sr: ...*tanka kao trska* solo ili pored takođe sitnog baletana...

The similar analysis based on the Italian/Serbian corpus showed that in a number of cases Italian superlatives were translated using simile figures: (a) *pallidissimo* 'very pale' → *bled kao krpa* 'white as a cloth'; (b) *sanissimo* 'very healthy' → *zdrav kao dren* 'healthy as dogwood'.

Pirandello-it: ...mi parve *pallidissimo*.

Pirandello-sr: ...izgledao mi je *bled kao krpa*.

Ferrante-it: ...a parte la necessità di prendere un po' di tranquillanti, risultò *sanissimo*.

Ferrante-sr: ...osim što mu je propisana upotreba blagih sedativa, ispostavilo se da je *zdrav kao dren*.

French comparisons using nouns instead of adjectives to indicate properties were also translated as simile figures in Serbian: (a) *un noir d'ébène* 'a black of ebony' → *crn kao abonos* 'black as ebony'; (b) *la solidité du roc* 'the solidity of rock' → *čvrst kao stena* 'solid as a rock'.

Verne80days-fr: ...de simples civils, chevelure lisse et *d'un noir d'ébène*, tête grosse, ...

Verne80days-sr: ...obični građani sa zalizanom kosom *crnom kao abonos*, velikom glavom, ...

VerneBalloon-fr: ...plus d'une île a disparu ainsi, qui paraissait avoir *la solidité du roc*.

VerneBalloon-sr: Mnogo je ostrva nestalo na taj način iako su izgledala *čvrsta kao stena*.

The analysis of English/Serbian corpus revealed that English noun(source)-adjective(property) constructions were translated in Serbian as similes: *blood-red* → *crven kao krv* 'red as blood'; (b) *sky-blue* → *plav kao nebo* 'blue as a sky'.

Clark-en: The light ... whose *blood-red* hues paled swiftly...;

Clark-sr: Svetlost ... čije su *kao krv crvene* boje stale naglo da blede...;

Stendhal-en: ...a red waistcoat, her little *sky-blue* jacket with its silver braid...

Stendhal-sr: ...u crvenom kaputu, odelu *plavom kao nebo*, ukrašenom srebrnim gajtanima...

The means of conveying the same meaning as similes presented here are not exclusive to one particular language, and can also be used in Serbian, as shown by the following examples.

Kiš-sr: ...s kraljevskim prstenjem na rukama svojim *prebelim*;

Kiš-de: ...mit königlichen Ringen auf ihren *schneeweißen* Händen;

Tišma-sr: ...koji su prosto leteli s punim *snežnobelim* jedrima;

Tišma-de: ...die *schneeweißen* Wolken der Segel;

Ferrante-sr$_1$: ...skoro neprepoznatljiv zbog neuredne *zift crne* brade;

Ferrante-it$_1$: ... quasi irriconoscibile per la barba incolta, *nerissima*.

In the **Kiš** example, the used form corresponds to the Italian *bianchissimi* although here translated with the German compound *schneeweißen*. The same German compound is used in **Tišma** to correspond to the Serbian compound *snežnobeo*. The noun *zift* 'tar' is used as an attribute of the black color in **Ferrante**$_1$ example.

4 Other considerations

Besides adjective similes the verbal similes are used as well. The following example illustrates the use of a verbal simile *crneti se kao zift* 'to look black as tar' in the Serbian text which was translated with the adjective simile *neri come il carbone* 'black as coal' in Italian.

Bora-sr: Oko joj još toplo, kosa joj *se* još *kao zift crni* i svetli.

Bora-it: I suoi occhi erano ancora ardenti e i capelli *neri* e splendenti *come il carbone*.

In cases when equivalent similes are used in the original and in the translation the question is whether the translation is literal or an MWU is used:

Ferrante-it$_2$: L'unità centrale della macchina è *grande come un armadio a tre porte*;

Ferrante-sr$_2$: Sama osovina mašine *velika* je *poput trokrilnog ormana*;

Albahari-sr: ...svako je *go kao pištolj* i svako drhti za sebe;

Albahari-de: ...jeder ist nackt wie eine Pistole, jeder ... zittert um sich selbst.

Ferrante$_2$ example uses a strange MWU 'big as a three-door cabinet', so one might think that the translation is literal; however; the same MWU is actually used in Serbian.[2] The MWU 'naked as a gun' is translated literally *nackt wie eine Pistole* in **Albahari**. Does this MWU exist in German? And if it does has it the same meaning as in Serbian: 'to be extremely poor, having nothing to wear'?

Finally, a simile in one language can be translated into another language with different type of MWE as demonstrated by the next example. A simile figure *as cheap as dirt* was translated from English as *ne ceniti ni koliko prebijenu paru* 'lit. do not value as much as a beaten coin'.

Austen-en: ...but all the honour of the family he held *as cheap as dirt*.

Austen-sr: ...ali čast porodice *nije cenio ni koliko prebijenu paru*.

5 Future Work

The obvious next step is to experiment with different techniques for establishing equivalences across languages. Presently, our corpora are POS-annotated and lemmatized. We plan to add annotations in concordance with Actions recommendations that would enable more refined search over all involved languages. This, along with the enhancement of existing corpora and the addition of new languages will enable not only to deepen the presented research, but also to expand it to finding another ways to express comparison, e.g. by using constructions with genitive *brzinom munje* → *brz kao munja* 'fast as lightning', other types of MWUs and rhetorical figures.

Our findings show that the translation of similes and other rhetorical figures can pose a challenge for translators, because different languages and cultures may have different ways of expressing the same idea and some similes may be based on cultural and/or historical references (or frames, in Fillmore's terms) that are not easy to translate. Even when the same simile exists in different languages, the translator may decide not to use a simile but another way of expressing comparison. A multilingual lexicon with aligned simile MWEs supported by examples can help both human and machine translation.

References

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

Cvetana Krstev. 2021. White as Snow, Black as Night – Similes in Old Serbian Literary Texts. *Infotheca – Journal for Digital Humanities*, 21(2):119–135.

Jelena Mitrović, Stella Markantonatou, and Cvetana Krstev. 2020. A cross-linguistic study on Greek and Serbian fixed similes and enrichment of lexical resources via crowdsourcing. In Stella Markantonatou and Anastasia Christofidou, editors, *Multiword Expressions: Drawing on Data from Modern Greek and Other Languages*, pages 241–262. Research Centre for Scientific Terms and Neologism.

Vlad Niculae 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*, pages 89–95.

Olja Perišić, Ranka Stanković, Milica Ikonić Nešić, and Mihailo Škorić. 2023. It-Sr-NER: Web services for recognizing and linking named entities in text and displaying them on a web map. *Infotheca – Journal for Digital Humanities*, 23(1). To be published.

Ranka Stanković, Cvetana Krstev, Duško Vitas, Nikola Vulović, and Olivera Kitanović. 2017. Keyword-based search on bilingual digital libraries. In Andrea Calì, Dorian Gorgan, and Martín Ugarte, editors, *Semantic Keyword-Based Search on Structured Data Sources: COST Action IC1302 Second International KEYSTONE Conference*, pages 112–123. Springer International Publishing, Cham.

A Translation of non-English examples to English

Translations are literal; they are listed in the order in which they appear in the text.

Eco The speaker was a monk bent with the weight of years, *white as snow*;

Jelinek ...that long-forgotten femininity with skin *as white as snow* and hair as black as ebony;

Flaubert As for him, he always carried himself *like an oak tree*;

[2] Among 22 examples of *trokrilni orman* 'three-door cabinet' in SrpKor21, 4 describe a very big person.

Andrić ...and every one assures that he is *healthy as a dogwood* and has nothing to do with cholera;

Dor Besides his *as white as snow* shirt, he was wearing only black pants and a black vest;

Grass *...willow-slim*, solo or placed next to her equally dainty ballet master;

Pirandello ...he seemed *very pale*;

Ferrante ...apart from the need to take some tranquilizers, he turned out to be *very healthy*;

Verne80days ...simple civilians, hair smooth and of *a black of ebony*, big head, ...

VerneBalloon ...more than one island has disappeared like this, which seemed to have *the solidity of rock*.

Kiš ...with royal rings on his hands *too white...*;

Tišma ...who simply flew with full *snow white* sails;

Ferrante$_1$ ...almost unrecognizable due to his unkempt beard, *very black...*

Bora Her eye is still warm, her hair still *looks black as tar* and bright;

Ferrante$_2$ The central unit of the machine is *as big as a three-door wardrobe*;

Albahari ...each is *naked as a gun* and each trembles for himself...

B The List of Sources of the Examples

The list of literary works and their translations; the language of the original is always given first.

1984 George Orwell en: *1984*;

Albahari David Albahari sr: *Mamac* – de: *Mutterland*;

Andrić Ivo Andrić sr: *Na Drini ćuprija* – it: *Il ponte sulla Drina*;

Austen Jane Austen en: *Persualtion* – sr: *Pod tuđim uticajem*;

Bora Borisav Stanković sr: *Nečista krv* – it: *Sangue impuro*;

Clarke Arthur Clarke en: *2001: A Space Odyssey* – sr: *Odiseja u svemiru 2001*;

Dor Milo Dor de: *Wien, Juli 1999* – sr: *Beč, juli 1999*;

Eco Umberto Eco it: *Il nome della rosa* – sr: *Ime ruže*;

Ferrante Elena Ferrante it: *Storia di chi fugge e di chi resta* – sr: *Priča o onima koji odlaze i onima koji ostaju*;

Flaubert Gustave Flaubert fr: *Bouvard et Pécuchet* – sr: *Buvar i Pekiše*;

Grass Günter Grass: *Im Krebsgang* – sr: *Hodom raka*;

Jelinek Elfriede Jelinek de: *Die Klavierspielerin* – sr: *Pijanistkinja*;

Kiš Danilo Kiš sr: *Peščanik* – de: *Sanduhr*;

Pirandello Luigi Pirandello it: *Uno, nessuno e centomila* – sr: *Jedan, nijedan i sto hiljada*;

Stendhal Stendhal fr: *Vanina Vanini*;

Tišma Aleksandar Tišma sr: *Upotreba čoveka* – de: *Der Gebrauch des Menschen*;

Verne80days Jules Verne fr: *Le tour du monde en quatre-vingts jours* – en: *Around the World in Eighty Days* – sr: *Put oko sveta za 80 dana*;

VerneBalloon Jules Verne fr: *Cinq Semaines en ballon* – en: *Five Weeks in a Balloon* – sr: *Pet nedelja u balonu.*