A cross-linguistic study on Greek and Serbian fixed similes and enrichment of lexical resources via crowdsourcing

Jelena Mitrović[1], Stella Markantonatou[2], Cvetana Krstev[3]

[1] Faculty of Computer Science and Mathematics, University of Passau, Germany
[2] Institute for Language and Speech Processing / Athena RIC, Athens, Greece
[3]Faculty of Philology, University of Belgrade

Abstract

This paper is about the Greek and Serbian multiword expressions (MWEs) that belong to the rhetorical figure simile. We use a corpus-driven crowdsourcing method to identify the most commonly used similes in Serbian and Greek. We attempt a first comparison of the two sets of data and discuss issues of simile encoding in lexical resources that are useful in Natural Language Processing.

Keywords: similes; fixed expressions; multiword expressions; crowdsourcing; lexical resources

1. Introduction

Similes have been studied from a variety of perspectives. Hanks (2005) elaborates on the question whether similes are metaphors and claims that they are not. Other researchers consider them part of the folklore culture of a nation with a strong cross-linguistic and cross-cultural dimension (Aasheim, 2012; Χιώτη, 2010). Mladenović & Mitrović (2013) and Mladenović et al. (2016) have studied Serbian similes as rhetorical figures; they have studied simile's effect on language and how rhetorical figures can be exploited by Natural Language Technology (NLP). Μπόλλα-Μαυρίδου (1996) has offered an extensive comparative study of Greek and English similes. In this paper we attempt a first comparative study of fixed similes in Greek (EL) and Serbian (SR) and discuss novel types of exploiting similes in NLP.

We are interested in fixed similes only. Hanks (2005) observes that similes are often quite conventionalised and that they provide MWEs. Following Χιώτη (2010), we assume that some similes can be classified as MWEs because of their endurance in time and the fact that they form a rather closed set in a language. To these properties we would add their semantic particularities that are characteristic of idiomatic language, for instance, (EL) άσπρος σαν το πανί, Lit. white like the cloth, is predicated of humans and body parts exclusively and is used to describe a very pale complexion (see Section 3.3 for a brief discussion of certain aspects of the idiomatic selectional restrictions of similes).

Similes appear in a number of syntactic patterns; here we will talk about the most basic one, namely the pattern "Adjective + comparator + (Det) + Noun". We will adopt the terminology used by Hanks (2005) and will use the term *property* to refer to the Adjective, the term *vehicle* to refer to the Noun and the term *tenor* to refer to the entity of which the simile is predicated.

Section 2 reports on the corpus-driven crowdsourcing method used to collect the fixed similes in the two languages. Section 3 reports on linguistic observations made on the data and on some cross-language comparison results. Section 4 gives an overview of the lexical resources and tools that

benefit from the results of our research. Section 5 describes the usage of similes to detect irony in Twitter. Section 6 suggests some directions of future work on similes. In the Appendix, the Modern Greek similes that were identified with the corpus-driven crowdsourcing method are listed.

2. The corpus-driven crowdsourcing method

Our aim was to collect fixed similes used in contemporary, "live" Greek language; to this end we conducted a corpus-driven crowdsourcing experiment, based on the methodology described in Mladenović et al. (2016).

An extensive collection of similes of the type "Adjective+σαν+(Det)+Noun" is given in Μπόλλα-Μαυρίδου (1996). Μπόλλα-Μαυρίδου retrieved the properties (adjectives, participles) from lexicographic material and developed questionnaires addressed to young native speakers; the speakers drew inspiration from the properties to supply similes. The final list included the similes that were supplied by more than 8 native speakers. However, some of these similes, although they are recognizable, they are not in use: for instance, on a Google search, the simile κίτρινος σαν το φλουρί, Lit. yellow like the lire, returns occurrences retrieved from non-contemporary literary sources only. Furthermore, several similes in the Μπόλλα-Μαυρίδου list return few to none Google search results. Indicatively: πάμπλουτος σαν κροίσος, Lit. rich like Croesus, παλιός σαν αντίκα, Lit. old like antique, χαδιάρα σαν γυναίκα, Lit. cuddling like woman, χαζός σαν χάνος, Lit. silly like comber (=a small fish), have returned no results on a Google search. In everyday language, κροίσος and αντίκα are used on their own in order to assign the respective property to an entity: Ο κροίσος Άγγλος βιομήχανος Sir James Dyson στη Σκιάθο, Lit. the croesus(=extremely rich) English manufacturer Sir James Dyson in Skiathos. Χάνος, on the other hand, is widely used in the verb MWE κοιτώ σαν χάνος, Lit. to stare like a comber, `to stare like a silly person'. Μπόλλα-Μαυρίδου (1996) herself identifies this problem in the responses she received, namely that the speakers were often instigated by the adjective or the participle (the property) and created a simile that does not exist in the language; in such cases only the vehicle, e.g. *croesus, comber*, is used and the property, e.g. "very rich", "silly", is inferred.

We aimed at discovering the similes that native speakers would use in their everyday exchange. We have drawn our material from corpora and have checked it with crowdsourcing methods. For Greek, we adopted the method introduced by Mladenović et al. (2016) who have tried to identify the most frequently used constructs in the Corpus of Contemporary Serbian Language (Utvić, 2014) and to check their status in the contemporary language with a crowdsourcing experiment on Facebook; in this experiment, native speakers would say if they really use a certain simile or not.

We extracted 2000 phrases containing potential Greek "Adjective+σαν+(Det)+Noun" similes from the Hellenic National Corpus (HNC)[1] and from a corpus (110M words) obtained with web crawling (Mastropavlos & Papavasiliou, 2011). Similes are not frequent in these corpora, for instance, for άσπρος σαν το πανί, Lit. white as cloth, that has returned 500+ unique hits in Google searches, HNC returns only four examples while the corpus collected with crawling returns no examples at all.[2]

The number of "Adjective+σαν+(Det)+Noun" structures extracted from the two corpora was refined to

[1] Hellenic National Corpus (HNC) http://hnc.ilsp.gr/

[2] HNC was searched via the provided interface that allows for collocation identification based on word/lemma proximity and part-of-speech (PoS) information. The corpus collected with crawling is annotated for PoS and was searched for the pattern Adjective+σαν+(Det)+Noun with NLP tools.

154 candidate similes, because that is the number that was used in the Serbian crowdsourcing experiment – this decision was made in order to ensure compatibility of the statistical analysis results. To this end, several structures were not included drawing on Google search results.

Next, we circulated Google Forms via Facebook and asked native speakers of Greek to let us know whether they would use certain constructs in their everyday linguistic interaction. Speakers were presented with a form that allowed them to click on a construct or not. Table 1 gives an overview of the number of constructs and participants per each Google form.

Google form	**Number of constructs per form**	**Participants per form**
1	30	67
2	42	85
3	41	79
4	41	59
Total	154	290

Table 1: Distribution of constructs and participants in the crowdsourcing experiment for Greek similes.

Inter-annotator agreement was checked with Krippendorff Alpha Coefficient (Krippendorff, 2012) and the results are shown in Table 2. An inter-annotator agreement between 0,65 and 0,8 would be considered "good". In order to facilitate statistical analysis, all Google forms, except for one, were split into two parts so that 30 or less constructs were included in each form. We located the five most reliable participants, based on the performance of a pairwise t-test (the common statistical evaluative test used to determine which agents present the most similar results). Next, the Krippendorff Alpha Coefficient was used. As seen in Table 2, inter-annotator agreement was unexpectedly good. 83 out of the 154 constructs have been annotated with "Yes"; this is a reliable indication that the "Yes"-similes are used in the everyday language. The fixed similes collected with the corpus-driven crowdsourcing experiment are listed in the Appendix.

Our method has yielded similes that can be traced with Google searches; the number of unique hits for similes selected with the crowdsourcing experiment ranges from 5 (βρώμικος σαν το γουρούνι, Lit. dirty like the pig) to 1500+ (άσπρος σαν το χιόνι, Lit. white like the snow, γλυκός σαν μέλι, Lit. sweet like honey). On the other hand, certain similes that were not selected with the crowdsourcing experiment also returned fair numbers of Google search unique hits (Markantonatou & Mitrović, 2017). For instance, κίτρινος σαν το φλουρί, Lit. yellow like the lire, returned 150 unique hits and γερός σαν ταύρος, Lit. strong/healthy like bull, 111 unique hits. The first of these two rejected similes, as mentioned in Section 2, returned mainly occurrences in the literature of the past, so it seems that speakers rejected it not because they did not recognize it but because they thought it to be "old-fashioned". On the other hand, the second rejected simile is rather colloquial. These facts indicate that the crowdsourcing method we used may be over-selective for reasons that require further analysis.

Form set	No of participants	No of questions	Alpha value	No of questions annotated with "Yes"
1	5	30	α = 1*	20
2a	5	21	α = 0.736*	11
2b	5	21	α = 0.69*	13
3a	5	21	α = 0.735*	10
3b	5	20	α = 0.696*	19
4a	5	21	α = 0.697*	12
4b	5	19	α = 0.698*	9
Total		154		94

Table 2: Results of the crowdsourcing experiment for Greek similes.

3. Cross-language comparisons and linguistic observations

We attempted a first comparison of Serbian and Greek similes drawing on the data collected with the corpus-driven crowdsourcing experiment in the two languages. Such comparisons presuppose (or reveal) a classification of the similes. A classification of English similes is offered in Hanks (2005) and it is based on the semantics of the vehicle. A classification of Greek similes is offered in Μπόλλα-Μαυρίδου (1996) who also offers a classification of English similes along with a comprehensive discussion of other classifications proposed in the literature. Μπόλλα-Μαυρίδου classifies similes on two axes, that of the property and that of the vehicle. The classification according to the property contains the following classes: property is understood through vision, listening, touching, taste, smell, many senses together, property concerns some biological, behavioural, spiritual state and property is related to space and to time. The classification according to the vehicle contains the following classes: living entities, natural products and materials, manufactured products and materials, social life, supernatural/historical/political entities, natural phenomena. In these classifications animals and plants have been found to play an important role (Χιώτη, 2010).

In our data, we sought translational and conceptual equivalences and derivational similarities. We also offer some first observations concerning the idiomatic semantics of similes. Below we list and briefly discuss the equivalences we identified in our data.

3.1 Translational equivalents

We sought translational and/or conceptual equivalents in our data. The results can be seen in the Appendix. 38 Serbian equivalents of Greek similes were identified, that is a 46% of the identified contemporary Greek fixed similes have Serbian equivalents; this considerable overlap might be expected given the many centuries that the two nations have been in close contact.

The class of similes with adjectives denoting colours (1)-(6), i.e. classification of similes by property, and the class of similes with vehicles denoting animals, i.e. classification of similes by vehicle, provide nearly half (18 out of 38) of the word-by-word translational equivalents (indicative examples are given in (7)-(9), more examples can be found in the Appendix):

Similes with properties in the semantic field of colour (visual property):

(1) EL άσπρος σαν το γάλα - SR beo kao mleko
EL aspros san to gala
'white like milk'– positive sentiment in both languages

(2) EL άσπρος σαν το χιόνι – SR beo kao sneg
EL aspros san to chioni
'white like snow' – positive sentiment in both languages

(3) EL άσπρος σαν το πανί – SR beo (bled) kao krpa –
EL aspros san to pani
'very pale' – negative sentiment in both languages

(4) EL κίτρινος σαν το λεμόνι– SR žut kao limun
EL kitrinos san to lemoni
'yellow like the lemon' – negative sentiment in both languages

(5) EL κόκκινος σαν {αίμα, παπαρούνα} – SR crven kao {krv, bulka}
EL kokinos san {ema, paparoyna}
'red like {blood, poppy}' – varying sentiment in both languages

(6) EL μαύρος σαν {το κάρβουνο, τον χάρο} – SR crn kao {ugalj, smrt}
EL mavros san {to karvoyno, ton charo}
'very black' – negative sentiment in both languages

Similes related to characteristics attributed to animals according to folk wisdom (the vehicle denotes an animal):

(7) EL πονηρός σαν αλεπού – SR lukav kao lisica
EL poniros san alepoy
'sly like a fox'– negative sentiment in both languages

(8) EL πιστός σαν σκύλος/σκυλί – SR veran kao pas
EL pistos san skylos /skyli
'faithful like a dog' – varying sentiment in both languages

(9) EL βρώμικος σαν γουρούνι – SR prljav kao svinja
EL vromikos san goyroyni
'dirty like a pig'– negative sentiment in both languages

Other translational equivalents may have vehicles denoting manufactured objects (10) or natural objects (11):

(10) EL χοντρός σαν βαρέλι – SR debeo kao bure
EL chontros san vareli
'fat like a barrel'– negative sentiment in both languages

(11) EL γλυκός σαν μέλι – SR sladak kao med
EL glykos san meli
'sweet like honey'– positive sentiment in both languages

There are concepts that are expressed with similes in both languages, but the similes are different. The concept "something is easy" is expressed in both languages with a simile but in Greek it is expressed with a verb simile and in Serbian with an "Adjective as Noun" simile.

(12) EL κάτι είναι παιχνιδάκι – SR prosto kao pasulj
EL kati ine pechnidaki
'something is a piece of cake'

We have not encountered any similes that are word-by-word equivalents but denote different concepts.

Below, some linguistic observations are listed concerning the similes of the type "Adjective as Noun". Here, we only record the phenomena and leave a detailed study for future research.

3.2 Similes related with morphological derivation

The issue of morphological derivation in the domain of MWEs has received little attention so far (for an extensive discussion on the issue, see Mititelu & Leseva, 2018). The data we present here are interesting because they indicate that derivation of MWEs from MWEs may present cross-linguistic regularities. Further research on the issue is required.

Our data concern the similes with properties in the semantic field of colour (see Section 3.1). In Serbian there are pairs such as *crven kao bulka* 'red as a poppy' – *pocrveneo kao bulka* 'blush as a poppy' (5),(13). Similarly, Greek de-adjectival verbs such as *ασπρίζω* 'whiten'*, κοκκινίζω* 'redden' give structures of the type "Verb as Noun" (5), (14); the "as Noun" part appears in a simile of the type "Adjective as Noun" and the Adjective is related with the verb morphologically.

(13) SR Plavi mladić u trećem redu **pocrveni kao bulka**.
'The young, blonde boy in the third row blushed like a poppy.'

(14) EL Εκείνη **κοκκίνισε σαν παπαρούνα** κι έσκυψε το κεφάλι ντροπαλά.
EL ekini kokkinise san paparoyna ki eskypse to kefali ntropala
'She blushed like a poppy and bowed her head timidly.'

Serbian has no participles. Greek has participles and they often assume functions of the adjective. Several Greek fixed similes with participles as properties (15) are morphologically related with 'Verb as Noun' MWEs (16).

(15) EL Γυρίζει από την αλάνα **πεινασμένος σαν λύκος.**
EL gyrizi apo tin alana pinasmenos san lykos
'He is starving when he comes back from the sandlot.'

(16) EL Μετά το θέατρο που **πεινάω σαν λύκος**, προτιμώ το καλό φαγητό.
EL meta to theatro pou pinao san lykos, protimo to kalo fagito

'After the theater when I am starving, I prefer decent food.'

3.3 Gender preferences

Certain Greek similes of the type "Adjective+σαν+(Det)+Noun" seem to impose restrictions on the entity they select as a tenor and, even more, on the gender of the tenor. Table 3 shows the distribution of the supersense PERSON for some of the Greek similes listed in the Appendix that were formed with the adjectives απαλός (*apalos*) 'soft', άσπρος (*aspros*) 'white', γερός (*geros*) 'strong', κόκκινος (*kokinos*) 'red', κρύος (*kryos*) 'cold', στολισμένος (*stolismenos*) 'adorned'. The supersense PERSON is identical with that of WordNet (Schneider et al., 2013). For instance, στολισμένος σαν φρεγάτα, Lit. adorned like a frigate, strongly selects for the female gender to the point that if the simile is predicated of a human male entity, the overall structure sounds like a comment on the entity's masculinity. However, not all similes show such preferences and the distribution of PERSON and gender values are not exceptional as it can be seen in Table 3.

It should be noted that such preferences do not follow from the structure of the simile and the semantics of its parts. We hypothesize that the PERSON/gender selectivity may reflect the "prototype" nature of the vehicle (Hanks, 2005) and that it indicates the idiomaticity of the simile, which is reflected on the semantics of the selected tenor, and has to be recorded in an MWE lexicon.

	Simile	Person	Masculine	Feminine	Neuter
1	*apalos san xadi*	5.1	**75.0**	25.0	0.0
2	*aspros san to pani*	**82**	57	40.4	2.6
3	*aspros san to gala*	**47.0**	32.5	**64.6**	2.9
4	*aspros san to chioni*	19.6	48.8	42.7	8.5
5	*geros san tavros*	**93.1**	**76.8**	18.9	4.2
6	*grigoros san astrapi*	**54.7**	**82.8**	13.9	3.3
7	*kokinos san astakos*	**86.5**	**65.6**	32.2	2.2
8	*kokinos san paparoyna*	**60.7**	34.3	**61.6**	7.6
9	*kryos san pagos*	36.9	50.5	45.5	4.0
10	*stolismenos san fregata*	**93.8**	6.7	**93.3**	0.0

Table 3: Distribution of the supersense PERSON and distribution of GENDER within the simile population selected with this supersense.

4. Lexical resources

In this section, we take up the issue of incorporating MWEs that are rhetorical figures in lexical resources dedicated to MWEs and in lexical resources of the general language.

4.1 Multiword resources

Multiword Expressions have been represented and used in the scope of Serbian Morphological Dictionaries (Krstev et al., 2010) where they represent 32,5% of the total entries. Current attempts to include MWEs in lexical resources in a non-conventional way add new semantic relations in the Serbian WordNet and build an Ontology of Rhetorical Figures; both these activities will be discussed in some detail below.

On the Modern Greek front, there has been extensive work on MWE collection and study that, in this

volume, is detailed in the introductory chapter and in the contribution by Anastasiadis et al. As regards the collected similes, they will be included in IDION, a web-based lexicographic tool tailor-made to the needs of the multi-dimensional documentation of Modern Greek MWEs (Markantonatou et al., in this volume).

4.2 WordNet, Serbian WordNet and Greek WordNet

WordNet (Fellbaum, 1998) is a set of approximately 117.000 concepts interconnected with semantic relations that form a semantic network. Concepts in WordNet are represented by a set of English synonyms that form the so-called "synsets". The syntactic categories represented in WordNet are noun, verb, adjective and adverb.

The (English) WordNet has been used as a basis upon which other wordnets, for other world languages, have been developed and continue to develop. The most commonly used model for creating other wordnets is called the *expand* model, in which synsets from WordNet are translated into a target language. This model was used for creating both the Serbian and Greek[3] wordnets (Mladenović, 2014; Vossen, 1997). An advantage of the expand-model is that the developed wordnets are all connected through the Inter-Lingual-Index (ILI) that links similar concepts between languages, which is highly advantageous for various multilingual applications. ILI was first introduced in the scope of EuroWordNet (Vossen, 1997). The Serbian WordNet (SWN) is connected with WordNet and other important lexical and semantic resources such as the SUMO ontology, Serbian Morphological Dictionaries (SrpMD) (Krstev, 2008) and SentiWordNet (Esuli & Sebastiani, 2006).

The BalkaNet[4] project (Tufiş et al., 2004) initiated the development of both the Serbian and the Greek wordnets. Currently, the SWN contains approximately 22,000 synsets including synsets in particular conceptual domains such as biological biomedicine, religion, law, linguistics, literature, librarianship, computer science, the culinary domain, sentiment analysis domain, etc. (Krstev, 2008) and tools for the SWN maintenance and further development were built (Mladenović et al., 2014).

In order to enable automatic detection of rhetorical figures, an addition was made to SWN enabling the encoding of the figure Simile (Mladenović et al., 2016). A pair of new mutually inverse semantic relations named *specificOf/specifiedBy* were added: the relation *specificOf* has adjectives as its domain and nouns as its range while the relation s*pecifiedBy* has nouns as its domain and adjectives as its range, e.g. Poppy is *specifiedBy* Red; Red is *specificOf* Poppy.

4.3 Ontology of rhetorical figures

The formal domain ontology of rhetorical figures for Serbian, the *RetFig ontology* (Mladenović & Mitrović, 2013) was a first step in developing tools for automated analysis of rhetorical figures in Serbian. The two top concepts of the RetFig ontology subsume a rhetorical figure because a rhetorical figure is both a linguistic and a rhetorical phenomenon. As a linguistic concept, rhetorical figure is represented by *Linguistic Scope* and by *Linguistic Object*. The former defines the textual structure (paragraph, strophe, verse, sentence, phrase, word) in which the rhetorical figure appears. Linguistic Object (verse, sentence, phrase, word), on the other hand, is situated inside the Linguistic Scope. When a transformation of Linguistic Object is caused by some linguistic operation, a structure that can

[3] Greek WordNet contains 18.461 synsets; it has been published under the principals of Linked Data and is included in the Open Multilingual WordNet (Bond & Paik, 2012) and is available for scientific usage under a CC-BY-SA license.

[4] BalkaNet was funded by the EU (September 2001 – August 2004)

be recognized as a rhetorical figure emerges. Figure 1 depicts the relations among Linguistic Scope and Linguistic Object in the case of the detection of the rhetorical figure Aphaeresis.

Rhetorical figures are grouped in *Rhetorical Groups*. All the figures in a Rhetorical group share some features – the part of text they affect, the meaning they convey, the linguistic elements and operations that are used in their formation etc. *Tropes* is such a group that contains figures such as Simile, Metaphor, Hyperbole, Metonymy, Oxymoron. Automatic detection of these figures in text leads to better results in discourse analysis, sentiment analysis and opinion mining.

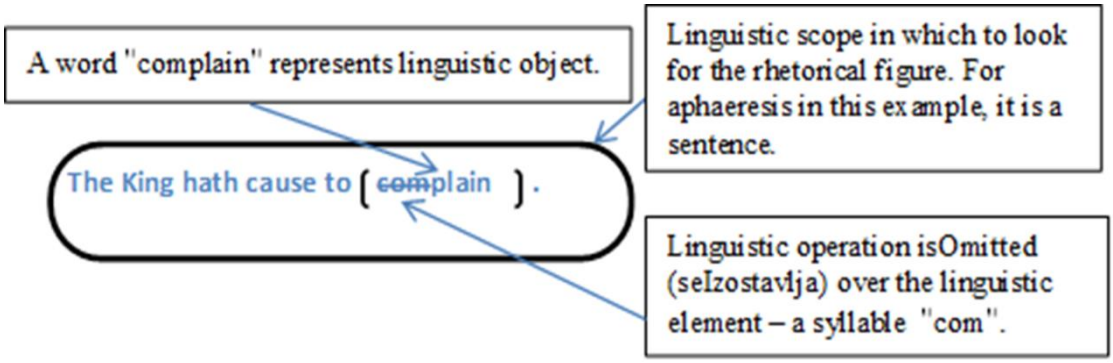

Figure 1: Aphaeresis detection via the RetFig ontology.[5]

5. Using lexical resources to detect rhetorical figures

A method of detecting rhetorical figures in the group Tropes of RetFig in text in the Serbian language was proposed in Mladenović (2016). This method uses SWRL[6] rules defined in the ontology based on the SWN (enriched with the*specificOf/specifiedBy* pair of mutually inverse relations) to identify whether a linguistic structure extracted from text satisfies some of the rules that would identify it as a rhetorical figure. The proposed method used (i) the RetFig ontology for building the taxonomy of Tropes classes and (ii) SWRL rules defined in the SWN ontology for identifying candidates for rhetorical figures (belonging to the classes Simile, Irony, Periphrasis and Oxymoron).

The digital Corpus of the contemporary Serbian language was used for semi-automatic and automatic SWN ontology learning related with different forms of rhetorical figures from authentic examples. Simile recognition experiments were performed on a collection of texts in Serbian comprising 10 digitized writings of various genres (children's songs, fairy tales, comedies, novels and essays). This ontology-based method for the recognition of the rhetorical figure simile achieved accuracy 51.8%.

An example of the power of the machinery that has been built on figurative language recognition is the detection of irony in Twitter. The overall effect of the irony figure is that the meaning of the structure used is the opposite of the structure's compositional meaning; as a result, machines that rely on the

[5] Figure taken from Mladenović & Mitrović (2013).

[6] SemanticWeb Rule Language (SWRL) is the language used in SemanticWeb for presenting formal logical expression, combining the features of OWL DL language (Web Ontology Language for Description Logic) and RML language (Rule Markup Language).

recognition of meaning, such as emotion recognition with NLP methods, are misled. Examples of verbal irony are based on several types of semantic relations (Bryant et al. 2002); commonly used forms of irony in Serbian (and Greek) are based on the usage of an adjective instead of its opposite (antonym) adjective, as in the examples:

– "He's just brilliant!" (the hidden meaning of the claim being that someone is really stupid).

– "See how skinny he is! (the hidden meaning of the claim being that someone is fat).

Another form of irony consists of a noun and an adjective whose hidden meaning is opposite to the meaning of an adjective that is prototypically associated with the particular noun: turtles are the prototypes of slow motion and rabbits are the prototypes of weakness or cowardness.

– "Fast as a turtle."

– "Brave as a rabbit!"

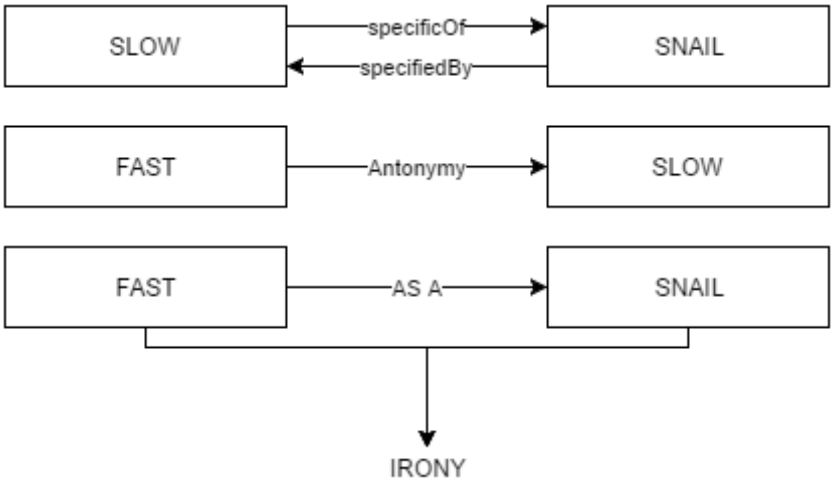

Figure 2: Irony detection using Serbian WordNet knowledge.

A corpus consisting of 2,127 tweets was used for detection of such ironic statements in the Serbian language (Mladenović et al., 2017). One part of the machine learning classification system that was applied to this corpus is presented in Fig. 2. Semantic relations *specificOf/specifiedBy* and *near_antonym* in a WordNet allow for usage of an inference system that can detect irony. Irony is represented as the antithesis between an inferred simile of the type 'Adjective as Noun' and the usage of the antonym of the 'Adjective' of the inferred simile. Therefore, the implicit relation representing irony is a relation between an instance of a noun synset class and an instance of adjective synset class. In the SWN ontology such rules can be expressed as OWL rules.

6. Future work

The observations made in Section 3 show that there is a lot to discuss, model and encode about this particular type of adjectival MWEs that show varying degrees of fixedness. Furthermore, resources containing fixed similes can be of use in NLP. In Section 4 we discussed how they can be useful in

detecting irony in Twitter. Qadir et al. (2015) have used similes characterized about their affective polarity to perform sentiment analysis in Twitter. In the light of the above, a corpus-driven detailed description and encoding of similes of the "Adjective as Noun" type (and of other types) in IDION, the Greek MWE web-based tool, is due. Furthermore, crowdsourcing experiments may be enriched with degree of acceptance measurements (as was kindly suggested by an anonymous reviewer).

The cross-language aspect of this research can be pursued by enriching the Ontology of Rhetorical Figures for Serbian (Mladenović & Mitrović, 2013) with Greek similes. The RetFig ontology has already been connected with the Serbian WordNet proving the feasibility of the enterprise and its potential for NLP purposes. Populating the ontology with Greek examples will enable easier cross-lingual detection of new rhetorical figures as well as irony. Furthermore, the Greek part of the ontology could be easily connected with the Greek WordNet or other Greek lexical ontologies, such as Ekfrasis (Fotopoulou et al., 2014).

7. **Περίληψη**

Οι παρομοιώσεις έχουν μελετηθεί ως προς τη σχέση τους με το φαινόμενο της μεταφοράς (Hanks, 2005), την λαογραφική παράδοση (Χιώτη, 2010), τη ρητορική και την Επεξεργασία της Φυσικής Γλώσσας (ΕΦΓ, Mladenović & Mitrović, 2013; Mladenović και λοιποί, 2016). Η Μπόλλα-Μαυρίδου (1996) κάνει μια λεπτομερή σύγκριση των παγιωμένων παρομοιώσεων της Ελληνικής και της Αγγλικής. Σε αυτό το άρθρο γίνεται προσπάθεια για μια πρώτη σύγκριση των παγιωμένων παρομοιώσεων της Σερβικής και της Ελληνικής. Επιπλέον, γίνεται μια σύντομη παρουσίαση γλωσσικών πόρων καθώς και εφαρμογών ΕΦΓ που περιλαμβάνουν παγιωμένες παρομοιώσεις.

Το ενδιαφέρον μας επικεντρώνεται στις παγιωμένες παρομοιώσεις που ανήκουν στο σύνολο των πολυλεκτικών εκφράσεων της Ελληνικής (Χιώτη, 2010). Μελετάμε μόνον έναν τύπο παρομοίωσης, συγκεκριμένα τον τύπο «Επίθετο+σαν+(Άρθρο)+Όνομα». Υιοθετούμε την ορολογία που χρησιμοποιεί ο Hanks (2005) και χρησιμοποιούμε τον όρο *ιδιότητα* για να αναφερθούμε στο Επίθετο, τον όρο *όχημα* για να αναφερθούμε στο Όνομα και τον όρο *φέρων* για να αναφερθούμε στην οντότητα την οποία προσδιορίζει η παρομοίωση.

Η μελέτη μας στηρίζεται σε σύνολα παγιωμένων παρομοιώσεων της Ελληνικής και της Σερβικής, η συλλογή των οποίων έγινε με την μεθοδολογία που προτάθηκε από τους Mladenović και λοιποί (2016). Σε αντίθεση με την μεθοδολογία που χρησιμοποίησε η Μπόλλα-Μαυρίδου (1996), η οποία στηρίχτηκε σε αποδελτίωση λεξικών και σε ερωτηματολόγια, εμείς συλλέξαμε παρομοιώσεις από σώματα κειμένων και επιλέξαμε τις παγιωμένες παρομοιώσεις με πληθοπορισμό αξιοποιώντας τα μέσα κοινωνικής δικτύωσης. Στόχος μας ήταν να συγκεντρώσουμε τις παγιωμένες παρομοιώσεις που χρησιμοποιούν οι ομιλητές της Ελληνικής στον καθημερινό τους λόγο (και όχι παρομοιώσεις που γνωρίζουν από την λογοτεχνία, για παράδειγμα).

Η μεθοδολογία που εφαρμόσαμε για τη συλλογή των παγιωμένων παρομοιώσεων είναι η ακόλουθη: από τον ΕΘΕΓ και από ένα σώμα κειμένων 110 εκατομμυρίων λέξεων που έχει δημιουργηθεί με crawling (Mastropavlos & Papavasiliou, 2011) παραλάβαμε 2000 φράσεις που περιείχαν την δομή «Επίθετο+σαν+(Άρθρο)+Όνομα» (της οποίας η συχνότητα εμφάνισης είναι μάλλον μικρή). Επιλέξαμε 154 παρομοιώσεις (όσες επέστρεφαν παραπάνω από ένα παράδειγμα σε κατάλληλη αναζήτηση με το Google). Στη συνέχεα αναπτύξαμε φόρμες Google που διαδώσαμε μέσω του Facebook και ζητήσαμε από τους ομιλητές να επιλέξουν τις παρομοιώσεις που χρησιμοποιούν στην καθημερινή τους ομιλία.

Στο πείραμα συμμετείχαν 290 ομιλητές (Πίνακας 1). Η συμφωνία μεταξύ των ομιλητών ελέγχθηκε με τον Krippendorff Alpha Coefficient (Krippendorff, 2012). Τα αποτελέσματα αναγράφονται στον Πίνακα 2: είναι εμφανές ότι υπήρχε σημαντική συμφωνία μεταξύ των ομιλητών. 94 παρομοιώσεις ξεχώρισαν ως καθημερινής χρήσης (βλέπε Παράρτημα). Η μέθοδός μας ίσως είναι υπερβολικά επιλεκτική καθώς δεν ανέδειξε ως καθημερινής χρήσης παγιωμένες παρομοιώσεις όπως *κίτρινος σαν το φλουρί* και *γερός σαν ταύρος*. Η πρώτη παρομοίωση μπορεί να θεωρηθεί πεπαλαιωμένη και άρα όχι καθημερινής χρήσης, πράγμα που όμως δεν ισχύει για την δεύτερη.

Συγκρίναμε τα σύνολα παγιωμένων παρομοιώσεων της Σερβικής και της Ελληνικής για να βρούμε μεταφραστικά ισοδύναμα και ισοδύναμες έννοιες. Μεταφραστικά ισοδύναμα εντοπίστηκαν σε παρομοιώσεις όπου η ιδιότητα ανήκει στο σημασιολογικό πεδίο του χρώματος (παραδείγματα (1) – (4)), επίσης σε παρομοιώσεις όπου το όχημα δηλώνει ζώο (παραδείγματα (5)-(7)). Εντοπίστηκε μόνο ένα μεταφραστικό ισοδύναμο με όχημα μη φυσικό αντικείμενο (παράδειγμα (8)). Η έννοια «κάτι είναι εύκολο» στην Ελληνική εκφράζεται με την ρηματική παρομοίωση *κάτι είναι παιχνιδάκι* ενώ στη Σερβική με την επιθετική παρομοίωση prosto kao pasulj (παράδειγμα (9)). Δεν εντοπίσαμε παρομοιώσεις που είναι λέξη προς λέξη μεταφράσεις η μία της άλλης αλλά δεν εκφράζουν την ίδια έννοια.

Και στις δύο γλώσσες παρατηρούνται σχέσεις μορφολογικής παραγωγής μεταξύ ρηματικών και επιθετικών παρομοιώσεων, ιδίως όσον αφορά επίθετα και ρήματα από το σημασιολογικό πεδίο του χρώματος (παραδείγματα (9)-(10)). Στην Ελληνική το φαινόμενο είναι εντονότερο γιατί υπάρχει η μετοχή η οποία χρησιμοποιείται ως επίθετο και σε παγιωμένες παρομοιώσεις (παραδείγματα (11)-(12)).

Όσον αφορά τις σημασιολογικές ιδιαιτερότητες των παγιωμένων παρομοιώσεων, οι πρώτες παρατηρήσεις σε υλικό που συλλέχθηκε από το Διαδίκτυο δείχνουν ότι ορισμένες παρομοιώσεις επιλέγουν το γένος των οντοτήτων που προσδιορίζουν, για παράδειγμα η παρομοίωση *κόκκινος σαν αστακός* προτιμά το αρσενικό γένος ενώ η παρομοίωση *στολισμένος σαν φρεγάτα* προτιμά το θηλυκό. Υποθέτουμε ότι αυτές οι προτιμήσεις είναι εκδηλώσεις της ιδιωματικότητας των παγιωμένων παρομοιώσεων και θεωρούμε ότι πρέπει να καταγράφονται στα λεξικά ως ιδιότητες των συγκεκριμένων πολυλεκτικών εκφράσεων.

Το άρθρο ολοκληρώνεται με την παρουσίαση λεξικών πόρων που περιέχουν πολυλεκτικές εκφράσεις και παγιωμένες παρομοιώσεις για την Σερβική (Serbian WordNet-SWN) και μίας μεθόδου αξιοποίησης των πόρων αυτών στην επεξεργασία της φυσικής γλώσσας και συγκεκριμένα στην ανίχνευση ειρωνείας στο Twitter (Mladenović και λοιποί, 2016). Η ενσωμάτωση των παγιωμένων παρομοιώσεων ως πολυλεκτικών εκφράσεων και ως ρητορικού σχήματος στο SWN γίνεται μέσω της ειδικής οντολογίας RetFig ontology (Mladenović & Mitrović, 2013) που αναπτύχθηκε ακριβώς για αυτόν τον σκοπό.

Acknowledgements

We thank the anonymous reviewers and the volume editors who contributed to making this text clearer. All remaining mistakes and inadequacies are ours.

Abbreviations

Abbreviated Form	Full Form
ΕΦΓ	Επεξεργασία Φυσικής Γλώσσας
Det	Determiner
EL	Greek
ILI	Inter-Lingual-Index
MWEs	Multiword Expressions
NLP	Natural Language Processing
PoS	Part of Speech
SR	Serbian
SWRL	Semantic Web Rule Language
SWN	Serbian WordNet

E-mail:
Jelena Mitrović: jelena.mitrovic@uni-passau.de
Stella Markantonatou: marks@ilsp.athena-innovation.gr
Cvetana Krstev: cvetana@poincare.matf.bg.ac.rs

REFERENCES IN GREEK

Μπόλλα-Μαυρίδου, Β. (1996). *Αντιπαραθετική εξέταση των στερεοτύπων παρομοιώσεων της Ελληνικής και Αγγλικής γλώσσας* (Διδακτορική διατριβή) Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Θεσσαλονίκη.

Χιώτη, Α. (2010). *Οι παγιωμένες εκφράσεις της Νέας Ελληνικής: ιστορική διάσταση, ταξινόμηση και στερεοτυπικότητα* (Διδακτορική διατριβή) Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Θεσσαλονίκη.

REFERENCES IN OTHER LANGUAGES

Aasheim, I. (2012). *A contrastive study of similes in English and Norwegian* (Master Thesis). University of Oslo, Oslo.

Bond, F. & Paik, K. (2012). A survey of wordnets and their licenses. *Proceedings of the 6th Global WordNet Conference (GWC 2012)* (pp. 64-71). Matsue, Japan.

Bryant, G. A. & Fox Tree, J. E. (2002). *Recognizing verbal irony in speech. Metaphor and Symbol, 17,* 99-117.

Esuli, A. & Sebastiani, F. (2006). SENTIWORDNET: A publicly available lexical resource or opinion mining. *Proceedings of 5th Conference on Language Resources and Evaluation (LREC)* (pp. 417-422). Genoa, Italy.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

Fotopoulou, A, Markantonatou, S. & Giouli, V. (2014). Encoding MWEs in a conceptual lexicon. *Proceedings of the 10th Workshop on Multiword Expressions (MWE), European Association of Computational Linguistics (EACL).*Gothenburg, Sweden.

Hanks, P. (2005). Similes and Sets: the English Preposition "like". In Blatna, R. & Petkevic, V. (Eds.). *Languages and Linguistics: Festschrift for Professor Fr. Cermak*. Prague: Philosophy Faculty of the Charles University.

Krippendorff, K. (2012). *Content Analysis: An Introduction to Its Methodology* (3rded.). Thousand Oaks, CA: Sage.

Krstev, C. (2008). *Processing of Serbian – Automata, Texts and Electronic Dictionaries*. Faculty of Philology, University of Belgrade.
Krstev, C., Stanković, R., Obradović, I., Vitas, D. & Utvić, M. (2010). Automatic Construction of a Morphological Dictionary of Multi-Word Units. In Loftsson, H., Rögnvaldsson, E. & Helgadóttir, S. (Eds.). *Proceedings of the 7th International Conference on NLP*, (pp. 226-237). IceTAL, Reykjavik, Iceland, August 16-18, Lecture Notes in Computer Science 623. Berlin,Heilderberg: Springer.
Markantonatou, S. & Mitrović, J. (2017). *Corpus Informed Enrichment of Lexical Resources (with Fixed Similes) via Facebook-enabled Crowdsourcing.* Between Corpora and Dictionaries / Crowdsourcing and Gamification. COST ENeL WG3 meeting, 24-25 February 2017. Budapest, Hungary.
Markantonatou, S., Zakis, G., Minos, P., Kolleti, E., Margariti, E., Stripelli, A. & Samaridi, N. (2017). In Markantonatou, S. & Christofidou, A. (Eds.). *Special issue on MWEs in Greek and other languages: from theory to implementation, Bulletin of Scientific Terminology and Neologisms of the Academy of Athens*. Athens: Academy of Athens.
Mastropavlos, N. & Papavassiliou, V. (2011). Automatic Acquisition of Bilingual Language Resources. *Proceedings of the 10th International Conference of Greek Linguistics*. Komotini, Greece.
Mititelu, V. & Leseva, S. (2018). Derivation in the domain of multiword expressions. In Sailer, M. & Markantonatou, S. (Eds.). *Multiword Expressions: Insights from a multi-lingual perspective*. Phraseology and Multiword Expressions Series (pp. 215-246). Berlin: Language Science Press.
Mitrović, J., Mladenović, M. & Krstev, C. (2015). *Adding MWEs to Serbian Lexical Resources Using Crowdsourcing.* Poster presentation at the 5th General PARSEME meeting. Iasi, Romania.
Mladenović, M. & Mitrović, J. (2013). *Ontology of Rhetorical Figures for Serbian.* Lecture Notes in Computer Science and Artificial Intelligence. 8082 (pp. 386-393). Berlin Heidelberg: Springer-Verlag.
Mladenović, M., Mitrović, J. & Krstev, C. (2014). Developing and Maintaining a WordNet: Procedures and Tools. *Proceedings of the 7th Global WordNet Conference*. Tartu, Estonia.
Mladenović, M. (2016). Ontology-based rhetorical figures recognition. *Infotheca - Journal for Digital Humanities, 16*(1-2).
Mladenović, M., Mitrović, J. & Krstev, C. (2016). A Language-independent Model for Adding New Semantic Relations to a WordNet. *Proceedings of the 8th Global WordNet Conference*. Bucharest, Romania.
Mladenović, M., Krstev, C., Mitrović, J. & Stanković, R. (2017). Using Lexical Resources for Irony and Sarcasm Classification. *8th Balkan Conference in Informatics*. Skopje, FYROM.
Qadir, A., Riloff, E. & Walker, M. A. (2015). Learning to recognize affective polarity in similes. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 190-200). Lisbon, Portugal.
Schneider, N., Mohit, B., Oflazer, K. & Smith, N. A. (2013). Coarse lexical semantic annotation with supersenses: an Arabic case study. *Proceedings of NAACL-HLT 2013* (pp. 661–667). Atlanta, Georgia.
Tufiş, D., Cristea, D. & Stamou, S. (2004). BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. *Romanian Journal of Information Science and Technology, 1*(1-2), 9-43.
Utvić, M. (2014). *Liste učestanosti korpusa savremenog srpskog jezika [Corpus of Contemporary Serbian Frequency Lists]*. Naučni sastanak slavista u Vukove dane.
Vossen, Piek. (2007). *EuroWordNet: a multilingual database for information retrieval*. DELOS workshop on Cross-language Information Retrieval. Zurich, Switzerland.

APPENDIX

Greek Fixed Simile	Literal translation in English	Translation Equivalent Serbian Fixed Simile
αδύνατος σαν οδοντογλυφίδα	thin as toothpick	
αδύνατος σαν σκελετός	thin as skeleton	mršav kao kostur
αδύνατος σαν στέκα	thin as beanpole	
αθώος σαν άγγελος	innocent as angel	
αθώος σαν παιδί	innocent as child	
ακλόνητος σαν βράχος	firm as rock	čvrst kao kamen
αλαφρύς σαν πούπουλο	light as plume	
ανάλαφρος σαν αεράκι	light as breeze	
απαλός σαν μετάξι	soft as silk	mekan kao svila
απαλός σαν χάδι	soft as stroke	
αργός σαν χελώνα	slow as turtle	spor kao kornjača
άσπρος σαν το γάλα	white as milk	beo kao mleko
άσπρος σαν το πανί	white as the cloth	beo kao krpa
άσπρος σαν το χιόνι	white as the snow	beo kao sneg
αστραφτερό σαν το διαμάντι	sparkling as the diamond	
βαρύς σαν μολύβι	heavy as lead	težak kao olovo
βρεγμένος σαν παπί	wet as duck	
βρώμικος σαν γουρούνι	dirty as pig	prljav kao svinja
γλυκός σαν μέλι	sweet as honey	sladak kao med
γρήγορος σαν λαγός	fast as hare	brz kao zec
γρήγορος σαν αστραπή	fast as lighting	brz kao munja
γρήγορος σαν τον άνεμο	fast as the wind	brz kao vetar
δειλός σαν κότα	cowardly as hen	plašljiv kao kokoška
δυνατός σαν ταύρος	strong as a bull	jak kao bik
ελαφρύς σαν φτερό	light as a feather	lagan kao pero
ενωμένοι σαν γροθιά	united as fist	
εξαγριωμένος σαν ταύρος	angry as a bull	
έξυπνος σαν αλεπού	clever as fox	
εύθραυστος σαν γυαλί	fragile as glass	
ήσυχος σαν αρνί	quiet as lamb	
ίσιος σαν λαμπάδα	straight as candle	
καθαρό σαν νερό	clear as water	
καθαρός σαν κρύσταλλο	clear as crystal	jasan kao kristal
καθισμένος σαν βασιλιάς	seated as king	
καθιστός σαν τον Βούδα	seated as the Buda	
κατάμαυρος σαν το σκοτάδι	jet black as the darkness	
κίτρινος σαν το λεμόνι	yellow as lemon	žut kao limun
κόκκινος σαν το αίμα	red as blood	crven kao krv
κόκκινος σαν αστακός	red as lobster	
κόκκινος σαν παπαρούνα	red as poppy	crven kao mak
κόκκινος σαν το παντζάρι		
κολλημένοι σαν σαρδέλες	jammed as sardines	

κολλημένος σαν στρείδι	attached as oyster	
κοφτερός σαν λεπίδι	sharp as knife	
κοφτερός σαν ξυράφι	sharp as razor	
κρύος σαν πάγος	cold as ice	hladan kao led
μαλακό σαν πούπουλο	soft as plume	
μαλακός σαν βούτυρο	soft as butter	mekan kao puter
μαύρος σαν το κάρβουνο	black as the coal	crn kao ugalj
μαύρος σαν τον χάρο	black as the death	crn kao smrt
μπλεγμένος σαν κουβάρι	knotted as ball of yarn	
ντυμένος σαν γαμπρός	dressed as bridegroom	
ντυμένος σαν καρνάβαλος	dressed as carnival	
ντυμένος σαν κρεμμύδι	dressed as onion	
όμορφη σα ζωγραφιά	beautiful as painting	lep kao slika
όμορφος σαν άγγελος	beautiful as angel	lep kao anđeo
οπλισμένος σαν αστακός	armed as lobster	
παστωμένοι σαν σαρδέλλες	jammed as sardines	nabijeni kao sardine
πεινασμένος σαν τον λύκο	hungry as a wolf	gladan kao vuk
πεισματάρης σαν μουλάρι	stubborn as a mule	tvrdoglav kao mazga
πεταμένος σαν σκουπίδι	thrown away as rubbish	
πιστός σαν σκυλί	faithful as a dog	veran kao pas
πονηρός σαν αλεπού	sly as a fox	lukav kao lisica
πόνος σαν σουβλιά	painful as pang	
πράος σαν αρνί	meek as lamb	
σκληρός σαν ατσάλι	hard as steel	čvrst kao čelik
σκληρός σαν βράχος	hard as rock	čvrst kao kamen
σκληρός σαν σίδερο	hard as iron	čvrst kao gvožđe
στητός σαν κυπαρίσσι	erect as cypress	uspravan kao čempres
στοιβαγμένοι σαν ζώα	piled as animals	
στολισμένος σαν λατέρνα	adorned as a musical instrument	
στολισμένη σαν φρεγάτα	adorned as frigate	
στολισμένος σαν γαμπρός	adorned as bridegroom	
στριμωγμένοι σαν σαρδέλλες	jammed as sardines	
στριμωγμένοι σαν τα ποντίκια	jammed as the mice	
τρυφερό σαν χάδι	tender as stroke	
φορτωμένος σαν γαϊδούρι	loaded as donkey	natovaren kao magarac
φουσκωμένος σαν διάνος	bloated as turkey	
φουσκωμένος σαν μπαλόνι	bloated as balloon	
φωτεινός σαν ήλιος	bright as sun	
χαρούμενος σαν παιδί	happy as child	
χλωμός σαν φάντασμα	pale as ghost	bled kao duh
χοντρός σαν βαρέλι	fat as barrel	debeo kao bure