Hear about Verbal Multiword Expressions in the Bulgarian and the Romanian Wordnets Straight from the Horse's Mouth

Verginica Barbu Mititelu
RACAI
Bucharest, Romania
vergi@racai.ro

Ivelina Stoyanova
DCL, IBL – BAS
Sofia, Bulgaria
iva@dcl.bas.bg

Svetlozara Leseva
DCL, IBL – BAS
Sofia, Bulgaria
zarka@dcl.bas.bg

Maria Mitrofan
RACAI
Bucharest, Romania
maria@racai.ro

Tsvetana Dimitrova
DCL, IBL – BAS
Sofia, Bulgaria
cvetana@dcl.bas.bg

Maria Todorova
DCL, IBL – BAS
Sofia, Bulgaria
maria@dcl.bas.bg

Abstract

In this paper we focus on verbal multiword expressions (VMWEs) in Bulgarian and Romanian as reflected in the wordnets of the two languages. The annotation of VMWEs relies on the classification defined within the PARSEME Cost Action. After outlining the properties of various types of VMWEs, a cross-language comparison is drawn, aimed to highlight the similarities and the differences between Bulgarian and Romanian with respect to the lexicalization and distribution of VMWEs.

The contribution of this work is in outlining essential features of the description and classification of VMWEs and the cross-language comparison at the lexical level, which is essential for the understanding of the need for uniform annotation guidelines and a viable procedure for validation of the annotation.

1 Introduction

The work on the Bulgarian and the Romanian wordnets (BulNet and RoWN, respectively) has started within the BalkaNet project (Tufiș et al., 2004). The approach adopted relies on the Base Concept set approach and the top-down extension (Rodriguez et al., 1998): the initial Base Concept set of the EuroWordNet (1,218 synsets) is extended by transferring all direct or indirect descendant synsets from Princeton WordNet (Miller, 1995; Fellbaum, 1998) (PWN) into the wordnets under development. The literals are then translated and their list is enriched with the help of synonymy and other dictionaries; the synsets are supplied with the appropriate glosses either by translating the English gloss or by constructing a new one; the synsets identification numbers are the same as in PWN. In addition, over 400 concepts considered specific to the Balkan area are included in the wordnets and for them a merge approach is followed: synsets are created for the new concepts, glosses are added, a specific identification number is assigned and a hypernym for each of them is found among the synsets already implemented in the Balkan wordnets to which they are linked as hyponyms (Tufiș et al., 2004).

After BalkaNet's completion the enrichment of BulNet has been directed towards providing lexical coverage of a subset of a reference corpus annotated with word senses from BulNet in the course of a word-sense annotation task (Koeva et al., 2011). Currently, BulNet contains 92,910 manually verified synsets comprising a total of 164,418 literals (representing 76,285 unique ones), out of which 63,930 literals (57,791 unique ones) are multiword expressions, accounting for 28.3% of the total number of literals (i.e., 43.1% unique ones). In recent years the work has expanded towards covering and automatically labelling verb-noun derivational and morphosemantic relations (Koeva, 2008; Dimitrova et al., 2014; Leseva et al., 2015; Koeva et al., 2016), verbal multiword expressions annotation and encoding within the PARSEME project (Ramisch et al., 2018), enhancing BulNet with various semantic and syntactic relations from other resources such as FrameNet and VerbNet (Leseva et al., 2018).

The further quantitative enrichment of RoWN targeted the lexical coverage of various corpora collected over time (Tufiș and Mititelu, 2014). At the moment RoWN contains 59,348 synsets in

Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 2–12
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

which 85,277 literals (representing 50,480 unique ones) occur, out of which 20,031 (i.e., 17,816 unique ones) are multiword literals, accounting for 23.5% of the total number of literals (i.e., 35.3% unique ones). The qualitative enrichment focused on in-line importing of the SUMO/MILO concept labels (Niles and Pease, 2001), connotation vectors for synsets (Tufiș and Ștefănescu, 2012), derivational relations (Barbu Mititelu, 2013) and annotation of verbal synsets with labels specific to various types of multiword expressions, adopting the same framework (the PARSEME annotation guidelines) (Barbu Mititelu and Mitrofan, 2019).

A detailed overview of the work on the two wordnets individually and in parallel is provided in (Barbu Mititelu et al., 2017).

RoWN can be queried at http://relate.racai.ro/, while the BulNet user interface http://dcl.bas.bg/bulnet/ provides access to both BulNet and RoWN, among other languages, as well as parallel visualization of corresponding synsets in two wordnets (Rizov et al., 2015).

In this paper we present the types of VMWEs existing in each language, as they are reflected in the respective corpora created within PARSEME (section 2). We continue with the presentation of and quantitative data about the types of VMWEs in each of the two wordnets (section 3). They constitute the basis for the comparative analysis of VMWEs (in section 4), after which we draw the conclusions and envisage some directions for further work.

2 Bulgarian and Romanian VMWEs in the Multilingual PARSEME Corpus

The multilingual PARSEME Corpus (version 1.1) of verbal multiword expressions contains subcorpora for 20 languages in which verbal MWEs have been manually annotated according to universal guidelines (Ramisch et al., 2018). For most languages, morphological and syntactic annotation was provided, including parts of speech, lemmas, morphological features and/or syntactic dependencies.

2.1 Types of Annotated VMWEs

The types of VMWEs from the PARSEME classification (Savary et al., 2018) applicable to Bulgarian and/or Romanian are:
(1) universal categories, i.e., types of VMWEs existing in all natural languages (participating in the PARSEME corpus annotation action):

- **light verb constructions** (LVCs) are made up of a verb and a predicative noun (directly following the verb or being introduced by a preposition) (Tu and Roth, 2011; Nagy et al., 2013). Depending on the semantics of the verb, two subtypes are identified:
  - LVC.full – these are expressions in which the verb's contribution to the expression's semantics is (almost) null (we call the verb "light"), e.g., EN *pay a visit*, BG *davam podslon* (give shelter), RO *lua o decizie* (make a decision);
  - LVC.cause – in these expressions the verb has a causative meaning, i.e. it identifies the subject as the cause or source of the event or state expressed by the noun in the expression, e.g., EN *grant rights*, BG *hvărlyam văv văztorg* (throw into rapture, "excite"), RO *da bătăi de cap* (give pains of head, "give headaches");

- **verbal idioms** (VIDs) – they have a verb head and at least one dependent component, and their meaning is non-compositional to a certain degree (Sag et al., 2002; Baldwin et al., 2003; Vincze et al., 2012), e.g., EN *kick the bucket*, BG *komandvam parada* (command the parade, "call the shots"), RO *trage pe sfoară* (pull on rope, "cheat");

(2) quasi-universal categories, i.e., existing only in some of the languages (in the PARSEME shared task annotation):

- **inherently reflexive verbs** (IRVs) – these are verbs that are accompanied by a pronoun with a reflexive meaning (usually a clitic), e.g., EN *help oneself*, BG *usmihvam se* ("smile"), RO *se preface* ("pretend");

- **inherently adpositional verbs** (IAVs) – a combination of a verb or a VMWE and a preposition or postposition that is either always required or changes the meaning of the verb significantly and is an idiosyncratic part of the VMWE, e.g., EN *rely on*, BG *zastavam zad* (stand behind, "support, back"). For Romanian, this category was not annotated, although the phenomenon is registered in the language: RO *consta în/din* (consist of/in).

2.2 Corpora

Bulgarian and Romanian corpora were developed for both edition 1.0 (Savary et al., 2017) and edition 1.1 (Ramisch et al., 2018) of the PARSEME shared task on automatic identification of VMWEs, but the discussion here focuses on the latter edition, for which the guidelines were enhanced (Savary et al., 2018) and larger corpora were used as compared with the first edition.

The Bulgarian subcorpus consists of news articles and comprises 480,413 tokens in 21,599 sentences, covering 6,704 annotated VMWEs. The Romanian corpus is also compiled of journalistic texts, containing 56,703 sentences with 1,015,623 tokens and with 5,891 VMWEs annotated. We can notice the higher density of VMWEs in the Bulgarian corpus in comparison with the Romanian one (see discussion below, subsection 3.2).

Both annotated corpora are available for download and use under the Creative Commons BY 4.0 license[1].

The distribution of the types of VMWEs in the two corpora is presented in Table 1. Although the corpora are not parallel and we cannot discuss directly correspondences in the distribution of VMWEs, both corpora consist of news texts and some comparisons between the two languages can be drawn. We notice the high frequency of reflexive verbs (IRV) in both of them. LVCs are much better represented in the Bulgarian corpus. This is easy to explain considering the greater number of “light” verbs identified in Bulgarian. The reverse is observed for VIDs, which can be due to: (i) the different coverage of the phenomenon in the two languages, (ii) the types of texts: even if both corpora are journalistic, the targeted audience, the types of articles, etc. influence the authors’ lexical choices, and hence, the linguistic characteristics of the corpora, (iii) the different treatment of borderline cases. The percentage of LVC.cause is similar in both corpora.

We will try to answer the questions related to the differences observed at the text level (the PARSEME corpora) and the lexical level (the wordnets) in the sections to follow.

3 VMWEs in BulNet and RoWN

The annotation of the two corpora was a stepping stone towards the analysis of the behavior of VMWEs in the two languages. The envisaged comparative approach could only be imagined in connection to the two wordnets, as they are aligned lexical resources (see section 1). In what follows, we discuss the distribution of the VMWE types at the lexicon level in the two languages, always having in mind the fact that the two wordnets are not complete, they do not offer a comprehensive image of the lexical richness and diversity of the two languages.

Type of VMWEs	BG		RO	
	#	%	#	%
VID	1,260	18.8	1,611	27.3
LVC.full	1,909	28.5	313	5.3
LVC.cause	222	3.3	183	3.1
IRV	3,223	48.1	3,784	64.2
IAV	90	1.3	-	-
TOTAL	6,704	100	5,891	100

Table 1: Distribution of VMWEs types in the BG and RO corpora.

The teams involved in the annotation of MWEs in BulNet and RoWN are to a large degree the same as the language teams involved in the PARSEME project, so the current work is a continuation of our joint efforts focused on establishing a suitable representation of VMWEs at the lexicon level. Achieving a uniform and consistent annotation strategy of VMWEs in Bulgarian (as a Slavic language) and Romanian (as a Romance language) will be a step towards a largely language independent description which can support ongoing efforts in the field of MWEs. What is more, these teams are also the ones involved in the development of the two wordnets, thus they are very familiar with the characteristics and intricacies of the two lexical resources.

3.1 Annotation Procedures and Conventions

For annotating VMWEs in BulNet and RoWN each team extracted the verbal synsets in the two wordnets containing at least one multitoken literal. Each such literal was manually assigned a label from the set defined in PARSEME (VID, LVC.full, LVC.cause, IRV, for both languages, and IAV for Bulgarian). One would say that the IRV label could have been automatically assigned. However, both in Bulgarian and in Romanian the reflexive pronoun *se* (with all its inflected forms) is ambiguous – besides the reflexive value, it can also:

[1] https://gitlab.com/parseme/sharedtask-data/tree/master/1.1

(a) have an impersonal meaning in Romanian, e.g. RO *se înţelege* (SE understand "everyone understands") – these cases are encoded as type NONE (see below), (b) express passive in both languages, e.g., BG *primerite se broyat răchno*, RO *exemplele se numără manual* (examples are counted manually) – these cases are not included in either wordnet, or (c) be part of a larger VID expression, e.g. RO *se sparge în figuri* (SE break in figures, "boast") or BG *broya se na prăsti* (to be counted on fingers, "be in very small numbers") – and encoded as VID. Thus, manual annotation was necessary.

Wordnet principles of knowledge representation as well as the expand method for the development of BulNet and RoWN necessitated two additional labels: NONE and NO_LEX. The first label (NONE) was introduced for those cases where the multitoken verbal literals are free phrases with a literal, compositional meaning not exhibiting the (semantic and morphosyntactic) characteristics of the VMWE classes, such as EN *find fault*, RO *culege nuci* (pick nuts), equivalent to the PWN synset {*nut:1*} (gloss: gather nuts), or BG *tantsuvam dzhayv* (dance jive) corresponding to the PWN synset {*jive:1*} (gloss: dance to jive music). The implementation of synsets containing free multitoken phrases was adopted in the cases where these constitute good or conventional translation equivalents to the respective lexicalized English concepts; in many cases these phrases qualify as collocations, or, at least, are likely to appear in running text in the given form.

The second label (NO_LEX) is reserved for cases where a certain concept existing in PWN is not familiar in the languages under discussion and therefore could not be supplied with an exact correspondence (such as a VMWE or a conventional free phrase or collocation). These synsets have been annotated differently in Bulgarian and Romanian. In BulNet, a descriptive, gloss-like literal has been constructed which presents the concept but is unlikely to appear in running text, e.g., BG {*bera drebni bezkostilkovi plodove:1*}, EN {*berry:1*} (gloss: pick or gather berries). The Romanian team has decided on a different approach, leaving these literals empty, but adding a descriptive gloss to them. While the phenomenon of lexicalization is beyond the scope of the current study, we included these cases in the data in order to examine lexical gaps.

The Bulgarian team has annotated two additional categories: (i) cases with a mandatory pronominal accusative or dative clitic (*ACCT/DATT*), e.g., BG {*sărbi me:1*} (itch.3SG me.ME.1SG.ACC) – EN {*itch:2*} (gloss: have or perceive an itch); and (ii) borderline cases (*OTH*), e.g., BG *razpăvam na krăst* (spread on the cross "nail to the cross"), which is used both literally and figuratively. In the literal sense, each of the elements bears its own semantic load and the meaning is easily construable as compositional, thus not a VID, but nevertheless understood as a whole. The same may be observed with MWE terms which are more likely to be marked as VID: BG {*povdigam na kvadrat:1*} (raise to a square) and RO {*ridica la pătrat*} (raise to square), both corresponding to EN {*square:2*} (gloss: raise to the second power).

The *ACCT/DATT* and *OTH* categories fall outside the scope of this study due to the fact that they have not been part of the PARSEME annotation process, have not been consistently described as VMWEs and are not annotated in RoWN.

3.2 Distribution of VMWEs in BulNet and RoWN

The types of VMWEs in BulNet and RoWN and their distribution across categories are presented in Table 2. Unlike Romanian, Bulgarian verbs have the category of aspect, which means that for a given synset there may be two (or more) Bulgarian VMWEs with roughly the same meaning, e.g., *izpera pari* (perfective) – *izpiram pari* (imperfective), to which there is only one RO *spăla bani* and one EN {*launder:2*} (gloss: convert illegally obtained funds into legal ones) counterpart. Prefixation may also result in the formation of aspectual pairs/triples, as almost all verbal prefixes may have a semantically bleached sense with predominantly aspectual meaning. In fact, in the above example, there is such a triple: BG *pera pari* (imperfective), *izpera pari* (perfective, formed by prefixation), *izpiram pari* (secondary imperfective, formed by suffixation from the perfective). In the context of VMWEs, this question has been discussed by Barbu Mititelu and Leseva (2018). This is one of the main reasons for the greater number of VMWEs in Bulgarian as compared to Romanian (see columns BulNet (all) and RoWN in Table 2). This is why for Bulgarian we also present the number of VMWEs where aspec-

Type of VMWEs	BulNet (all)		BulNet (asp. gr.)		RoWN	
	#	%	#	%	#	%
VID	1,177	24.0	775	23.9	614	35.1
LVC.full	675	13.8	465	14.4	102	5.8
LVC.cause	112	2.3	63	1.9	42	2.4
IRV	2,779	56.7	1,822	56.3	989	56.4
IAV	54	1.1	39	1.2	-	-
OTH	51	1.0	31	1.0	-	-
ACCT/DATT	53	1.1	42	1.3	-	-
Ambiguous	-	-	-	-	5	0.3
TOTAL	4,901	100	3,237	100	1,752	100

Table 2: Distribution of VMWEs types (excluding 'NONE' and 'NO_LEX') in BulNet and RoWN. For BulNet, we present number of total VMWEs (all) as well as data where aspectual verb pairs are grouped and counted as a single VMWE (asp. gr.).

tual pairs (suffix-based only) are counted as single VMWEs (columns BulNet (asp. gr.) of Table 2) in order to facilitate comparison between the two languages.

For both languages the distribution of the types of VMWEs in the wordnets correlates with that in the corpus: data distribution in the lexicon mainly confirms language use. In both BulNet and RoWN the IRVs are the most numerous, ∼56%, followed by VIDs. It can be seen from the data (Table 2) that the percentage of VIDs in Romanian is higher than in Bulgarian, in both the corpora (27.3 in Romanian to 18.8 in Bulgarian) and the wordnets (35.1 in Romanian to 24.0 in Bulgarian), while the opposite tendency is observed for LVCs.

The relatively small number of LVCs in the two wordnets, and especially in Romanian, is largely explainable by the fact that this type of VMWEs is not well established in the wordnet structure, and the teams who have worked on the two wordnets throughout the years have followed different conventions.

Moreover, the adopted principle within the WordNet framework has been to define separate synsets to account for the "light" verb senses, such as: {*give:5, pay:7*} (gloss: convey, as of a compliment, regards, attention, etc.; bestow), as in *Don't pay him any mind*, *pay attention*; *give the orders*, *Give him my best regards*; and {*make:1, do:1*} (gloss: engage in), as in *make love, not war*, *make an effort*, *make revolution*; *do research*, *do nothing*. As a result, the inclusion of LVCs in PWN was more of an exception rather than a rule. In view of the approach adopted in the initial stages of creation of RoWN and BulNet, LVCs were introduced primarily where no lexicalized verbs were found as a counterpart for the respective English synset, e.g., BG {*postavyam v shah:1, davam shah:1,...*}, RO {*da şah:1*}, EN {*check:19*} (gloss: place into check). Well-established LVCs (frequently used in the language) were also added, especially if they have counterparts in PWN (see subsection 4.5).

The existing lexicographic tradition has also played a part in the decisions made by the teams. For instance, Bulgarian dictionaries tend to encode primarily VIDs and IRVs and have largely neglected LVCs and IAVs (the existence of the latter is subject to debate). Only a few researchers outside the computational linguistics community have acknowledged the need for systematic lexicographic description and treatment of LVCs (cf. for instance (Korytkowska, 2008)). The situation is similar in the Romanian lexicography: IRVs and VIDs are systematically recorded, and the latter also benefit dedicated dictionaries. Among them, (Mărănduc, 2010) is the most permissive and many phrases, LVCs among them, found their place in it.

Most VMWEs belong to only one type, irrespective of the number of their occurrences (i.e., synsets to which they belong) in one wordnet. However, in RoWN there are some literals which are annotated differently when belonging to different synsets, i.e., when having different meanings: e.g. *scoate fum* (give out smoke) is annotated as NONE when being in the synset corresponding to the English {*fume:4; smoke:4*} (gloss: emit a cloud of fine particles) and it is annotated as VID when belonging to the synset corresponding to the

English {*steam:3*} (gloss: get very angry).

4 Comparative Analysis of VMWEs in BulNet and RoWN

In this section we look comparatively at the VMWEs in the two wordnets: our interest is in the concepts which the two languages tend to lexicalize as VMWEs and, going even a bit further, to what degree the concepts in the two languages are lexicalized by the same type of VMWEs. As far as we are aware, this is the first time such a linguistic comparison is made, at least at the lexicon level. Although bilingual dictionaries (Kaldieva-Zaharieva, 1997) show such correspondences, there have not been any studies dedicated to this aspect.

4.1 Overview

Table 3 offers an overview of the number of synsets containing VMWEs in BulNet and RoWN. Out of the total number of verbal synsets, we show how many contain at least one VMWE, then the number of synsets in the set intersection of the sets of synsets containing VMWEs in the two wordnets. In the last row we calculated the number of synsets which in one language contain at least one VMWE, while in the other they contain none.

Number of:	**BulNet**	**RoWN**
# verbal synsets in WNs	7,172	10,397
# synsets with VMWEs in each WN	2,362	2,087
# synsets with VMWEs in both WNs	944	944
# synsets with VMWEs in only one WN	1,418	1,143

Table 3: Synsets with VMWEs in BulNet and RoWN

The set intersection of the synsets containing VMWEs in BulNet and in RoWN comprises 944 synsets, which represents 40% of the verbal synsets containing MWEs in BulNet and 45% of those in RoWN, thus showing a substantial overlap between the two wordnets, already indicative of some common tendencies in the two languages with respect to the way in which verbal concepts are lexicalized. In comparison, the intersection of synsets containing VMWEs between BulNet and PWN is 664, and between RoWN and PWN is 656, counting the multiword literals in PWN synsets, as VMWEs are not annotated in PWN.

The literals from corresponding BulNet and RoWN synsets are considered translation equivalents. There are 3,656 such literal-to-literal relations where the literals are VMWEs or multitoken free phrases (marked as NONE), and their distribution is presented in Table 4 (in the current section, for the purposes of comparison, suffix-based aspectual pairs in Bulgarian are counted as a single VMWE).

		BulNet			
		VID	LVC	IRV	NONE
RoWN	VID	**192**	16	99	140
	LVC	41	**44**	75	138
	IRV	151	64	**2,023**	148
	NONE	49	5	96	**263**

Table 4: Distribution of VMWE literal-to-literal correspondences between BulNet and RoWN

Table 5 reflects the number of synsets where there is a direct correspondence between VMWE types (cf. Table 4 which shows the number of all literal-to-literal relations, including multiple cases within the same synsets). Such cases represent 72.7% of the synsets in the intersection. That is indicative of the two languages' strong tendency of lexicalizing the same concepts by means of the same type of VMWEs.

	# BG-RO literal pairs of the same type				
Type	**1**	**2**	**3**	**4+**	**Total**
IRV	289	123	30	16	458
VID	54	15	1	-	70
LVC.full	13	1	-	-	14
LVC.cause	1	-	-	-	1
NONE	131	11	1	-	143

Table 5: Number of synsets with literal-to-literal correspondence of VMWE types in BulNet and RoWN

In what follows, we analyze the cases where there is asymmetry between the Romanian and the Bulgarian data – cases where there is a VMWEs only in one of the languages but not in the other (section 4.2), and the specifics of the non-VMWE multitoken phrases and their place in the WordNet structure (section 4.3).

We further illustrate equivalent synsets representing the three most frequent categories – IRV (section 4.4), LVC (section 4.5) and VID (section

4.6), and discuss the similarities as well as the differences in expressing the relevant concepts in the two languages. The corresponding PWN synsets are also presented to facilitate the understanding of the Bulgarian and the Romanian synsets.

4.2 VMWEs from either Wordnet with no VMWE correspondence in the other one

There are 1,418 synsets in BulNet which contain a VMWE with no corresponding VMWE in RoWN, and 1,143 synsets with a VMWE only in RoWN. Several cases were identified:

(i) the synset in one of the languages contains at least one VMWE, while its counterpart in the other language consists of simple-word literals: **805** cases in the BulNet data and **955** cases from RoWN: BG {*spya leten săn:1, estiviram:1*}, RO {*estiva:1*}, EN {*estivate:1, aestivate:1*} (gloss: sleep during summer);

(ii) the synset in one of the languages contains at least one VMWE of the types adopted in the PARSEME project or is expressed by a free phrase (marked as NONE or NO_LEX), while in the other language it is not lexicalized: **129** synsets in the BulNet data – out of which 39 are VMWEs (LVC, VID, IRV or IAV), 45 are conveyed by a free phrase marked as NONE and the remaining 45 – as descriptive phrases marked as NO_LEX: BG {*implantiram se:1*}, RO {*no correspondence*}, EN {*implant:2*} (gloss: become attached to and embedded in the uterus); no such cases are found in the RoWN data since the non-lexicalized synsets in BulNet are supplied with a descriptive literal (marked as NO_LEX) and thus are present in the BulNet dataset;

(iii) the synset in one of the languages contains a VMWE, but its counterpart in the other wordnet has not yet been implemented so there is no information regarding its lexicalization – **342** cases in the BulNet data and **188** cases in the RoWN: BG {*izmivam si rătsete:1, izmiya si rătsete:1*}, RO {*not implemented*}, EN {*wash one's hands:1*} (gloss: to absolve oneself of responsibility or future blame);

(iv) the synset is language specific (denotes a concept which is typical of one of the languages and is not present or at least not implemented in the other wordnet or in PWN) and contains a VMWE – **142** cases in the data from BulNet and none in RoWN: BG {*edva se dărzha na kraka:1*} (can barely stand on one's feet (with fatigue)).

4.3 Non-VMWE Multitoken Phrases and Lexicalization

As discussed earlier (section 3), non-VMWE multitoken phrases have been encoded in the two wordnets (1,209 in BulNet and 1,217 in RoWN). We may even argue that a number of literals in PWN also fall in this category, e.g., EN *make pure:1*, *use of goods and services:1*, *make unnecessary:1*, although the precise number of these cases is unknown.

In the two languages under discussion non-VMWE multitoken literals have been implemented largely by way of compensating for lexical gaps where a free phrase constitutes a widely used translation equivalent. Although the differences in the lexicalization patterns across languages may be quite idiosyncratic, certain trends have emerged from the analysis of the data. Here we illustrate one such trend: in a number of cases where the Bulgarian and the Romanian wordnet teams have resorted to encoding non-VMWE multitoken phrases, we find in PWN a lexicalization pattern typical for English where an argument is incorporated in the conceptual structure of a verb and the name of this argument gives the name of the respective predicate (Jackendoff, 1990). Such verbs are found across classes of verbs more or less systematically. The example below illustrates incorporated Theme-argument verbs – the item undergoing some influence or change (bearing the semantic role of Theme) gives the name of the predicate relation: BG: {*săbiram perli:1, săbera perli:1*} (lit. gather pearls), RO {*pescui perle:1*} (lit. fish pearls), EN: {*pearl:1*} (gloss: gather pearls, from oysters in the ocean). Apart from this synset, there are a number of other synsets in the same local WordNet tree (synsets with a common hypernym {*gather:1, garner:3, collect:3, pull together:1*}) whose common definition may be posited as "gather X...", where X is nuts/clams/oysters,..: {*nut:1*}, {*clam:1*}, {*oyster:1*}, respectively. The Bulgarian and the Romanian counterparts of these verbs are combinations of the type V + object NP, where the NP corresponds to the English incorporated Theme-argument. The productivity of this pattern is reflected in the productivity of zero derivation.

Another visible trend is for English synsets to contain a one-word compound or a metaphor which in the languages under discussion is conveyed by a free phrase: BG {*parkiram uspo-*

redno:1, parkiram paralelno:1}, RO {*parca lateral:1*}, EN {*parallel-park:1*} (gloss: park directly behind another vehicle).

4.4 Analysis of IRVs – Correspondences and Differences

As straightforwardly visible from the data, IRVs are by far the most represented category in the RoWN – BulNet intersection, which is to be expected, taking into account the semantics of the reflexive verbs in the two languages (Slavcheva, 2006).

Analyzing the semantic primes (Koeva et al., 2016) of these IRVs, we notice that more than a quarter of them are verb.change (125). The next well represented semantic prime is verb.motion (81). Others are verb.stative (48), verb.social (45), verb.communication (31), etc. Here is an example of VMWEs of another semantic prime, verb.emotion (with 22 expressions altogether), in two rich synsets: RO {*[se] înfuria*:2 IRV, *[se] enerva*:1 IRV, *[se] irita*:1 IRV, *[se] mânia*:1 IRV, *[se] supăra*:1 IRV}, BG {*yadosvam se*:3 IRV, imperf., *yadosam se*:3 IRV, perf., *razsărdvam se*:1 IRV, imperf., *razsărdya se*:1 IRV, perf., *gnevya se*:1 IRV, imperf., *razgnevyavam se*:1 IRV, imperf., *razgnevya se*:1 IRV, imperf.}, EN {*anger*:2, *see red*:1 VID} (gloss: become angry).

4.5 Analysis of LVCs – Correspondences and Differences

Besides what has already been discussed in section 3.2, the reasons for the difference in the number of LVCs in the two wordnets and the respective PARSEME corpora are due to the number and frequency of "light" verbs involved in the LVCs in the two languages. In the PARSEME corpus, we find 9 different verbs heading Romanian LVCs, most of them with a considerable number of occurrences, while in the Bulgarian corpus, the verbs that head LVCs are more than 100, and approximately half of them have more than 5 occurrences. Similarly, in RoWN we see 21 light verbs, only 5 of them having more than 5 occurrences, and 118 in BulNet, of which 32 have relatively high frequency.

It has become apparent that different teams construe the scope of the light verbs differently. Although the PARSEME project outlines some guidelines for identifying LVCs, the judgment of a verb as semantically bleached, which is a key point in the LVC identification process, remains subjective. It is the approach for many languages to identify a limited set of highly frequent verbs and consider them as most likely light verb candidates in combination with a predicative noun. The Bulgarian team have considered a broader range of high frequency verbs and their synonyms (in BulNet) as possible heads of LVCs and have applied manual verification to LVC candidates (Stoyanova et al., 2016). The attempt has been to uncover the true extent of the phenomenon in the language without limiting it beforehand.

It is a well-known fact that LVCs often have a single verb counterpart which is derivationally related to the eventive noun in the respective VMWE, e.g. BG *resha/V – reshenie/N – vzemam reshenie/VMWE*, EN *decide/V – a decision/N – make a decision/VMWE*, RO *decide/V – o decizie/N – lua o decizie/VMWE*. Bearing in mind the structure of wordnets and other factors pointed out in section 3.2, in many cases wordnet developers have given preference to the single verb and have left out possible LVCs conveying the same meaning: for example, the LVC RO *face o vizită* (pay a visit), although synonymous with the verb *vizita*, is not included in any synset in which *vizita* occurs, in spite of their identical meaning(s).

Due to the above reasons, we find considerable discrepancy in the numbers of LVCs in RoWN and BulNet – only 44 cases of LVC–LVC correspondences (Table 4).

An example of LVC–LVC correspondence is provided by the following synsets, in which the choice of VMWE literals is supported by the PWN data: RO {*lua parte*:2 LVC.full, *participa*:5}, BG: {*uchastvam*:2, *vzemam uchastie*:1 LVC.full,*vzema uchastie*:1 LVC.full}, EN {*participate*:1, *take part*:1 LVC.full} (gloss: share in something).

With the prerequisites made so far, in the majority of the cases found in the data, an LVC in one of the languages under discussion corresponds to a free phrase collocation in the other.

4.6 Analysis of VIDs – Correspondences and Differences

Due to their characteristics VIDs are both easily recognizable and well-represented in lexical resources, including in PWN, which has most likely influenced the choice of VIDs to encode in BulNet and ROWN: BG {*cheta mezhdu redovete*:1 VID, *prochitam mezhdu redovete*:1 VID, *procheta mezhdu redovete*:1 VID}, RO {*citi printre rânduri*:1 VID}, EN {*read between the lines*:1

VID} (gloss: read what is implied but not expressed on the surface).

More interesting cases are represented by mismatches in the two languages. Here is an example illustrating the situation when a Romanian synset contains a VID and the Bulgarian has an expression annotated as NONE. The Romanian expression has the structure transitive verb + direct object realized as a definite noun, *vârsta*, and it answers positively the test for lexical inflexibility from the annotation guidelines. The Bulgarian counterpart includes both V + direct object NP and V + AP with the literal meaning of "reach majority" or "become a major": BG {*(do)stigam pălnoletie*:1; *(do)stigna pălnoletie*:1; *navărshvam pălnoletie*:1; *navărsha pălnoletie*:1; *stavam pălnoleten*:1; *stana pălnoleten*:1}, RO {*avea vârsta*:1 VID} (have the age), EN {*come of age*:1} (gloss: reach a certain age that marks a transition to maturity).

The next example illustrates the case of a Bulgarian VID whose equivalent in Romanian is a free word combination made up of the verb *fi* (be) and the adjectival locution *de ajutor* (of help "helpful"). The Bulgarian counterpart consists of the verb *davam* (give, lend) or *udryam* (hit) and the noun *ramo* (shoulder), with a possible insertion of *edno* (one) ("give a/one shoulder"): BG {*davam ramo*:2 VID; *dam ramo*:1 VID; *davam edno ramo*:1 VID; *dam edno ramo*:1 VID; *udryam edno ramo*:1 VID; *udarya edno ramo*:1 VID}; RO {*fi de ajutor*:1}, EN {help out:1} (gloss: be of help, as in a particular situation of need).

5 Conclusions and Future Work

The comparative overview of the representation of VMWEs in BulNet and RoWN can be a starting point for drawing conclusions about the scope and the distribution of VMWEs in Bulgarian and Romanian, as well as for establishing good practices for the description of VMWEs in wordnets in general.

The work presented here has helped in determining essential features of the description and classification of VMWEs with a view to facilitating the future applications of the resources: morphosyntactic and inflectional description, which enables the recognition of VMWEs in running text, description of VMWE variants (e.g., aspectual verb pairs, prefixed verbs, possible modification of components, etc.), derivational information to identify VMWE derivatives, etc.

Analyzing VMWEs comparatively at the lexical level as reflected in the two wordnets under discussion gives a new, outsider's perspective at the annotation of VMWEs and allows for studying not only the similarities and dissimilarities between languages, but also the understanding and application of annotation guidelines cross-linguistically and emerges as a viable procedure in the validation of the annotation performed for a given language.

As a multilingual lexical-semantic resource, wordnets have numerous applications in machine and machine-aided translation. Addressing the issues of VMWEs in a unified way across wordnets, will widen the possibilities of their use.

Beyond translation, it will provide language material for the study of lexicalization, cross-linguistic semantic analysis of VMWEs, metaphors, etc.

6 Acknowledgments

Part of this work has been undertaken under the bilateral project *Enhancing Multilingual Language Resources with Derivationally Linked Multiword Expressions* (2018–2020) between the Institute for Bulgarian Language at the Bulgarian Academy of Sciences and the Research Institute for Artificial Intelligence at the Romanian Academy. Another part has been undertaken under the project *Multilingual Resources for CEF.AT in the legal domain (2017-EU-IA-0136) - MARCELL*. Both teams would like to thank the four anonymous reviewers for the valuable suggestions and hard work. The Bulgarian team would also like to thank Valentina Stefanova for participating in the annotation phase.

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96. ACL.

Verginica Barbu Mititelu. 2013. Increasing the effectiveness of the romanian wordnet in nlp applications. *CSJM*, 21(3):320–331.

Verginica Barbu Mititelu and Svetlozara Leseva. 2018. Derivation in the domain of multiword expressions. In Manfred Sailer and Stella Markantonatou, editors, *Multiword expressions: Insights from a multi-*

lingual perspective, Phraseology and Multiword Expressions, pages 215–246. Language Science Press.

Verginica Barbu Mititelu, Svetlozara Leseva, and Dan Tufis. 2017. The Bilateral Collaboration for the Post-BalkaNet Extension of the Bulgarian and the Romanian Wordnets. In *Proceedings of the International Jubilee Conference of the Institute for Bulgarian Language Prof. Lyubomir Andreychin (Sofia 15 – 16 May 2017)*, pages 192–200. Institute for Bulgarian Language Prof. Lyubomir Andreychin.

Verginica Barbu Mititelu and Maria Mitrofan. 2019. Leaving no stone unturned when identifying and classifying verbal multiword expressions in the romanian wordnet. In *Proceedings of the 10th Global WordNet Conference*, page in press, Wroclaw, Poland.

Tsvetana Dimitrova, Ekaterina Tarpomanova, and Borislav Rizov. 2014. Coping with Derivation in the Bulgarian WordNet. In *Proceedings of the Seventh Global Wordnet Conference (GWC 2014)*, pages 109–117.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Ray S. Jackendoff. 1990. *Semantic Structures*. MIT Press, Cambridge.

Stefana Kaldieva-Zaharieva. 1997. *Romanian-Bulgarian Phrasal Dictionary*. BAS Marin Drinov.

Svetla Koeva. 2008. Derivational and morphosemantic relations in Bulgarian Wordnet. *Intelligent Information Systems*, pages 359–368.

Svetla Koeva, Svetlozara Leseva, Borislav Rizov, Ekaterina Tarpomanova, Tsvetana Dimitrova, Hristina Kukova, and Maria Todorova. 2011. Design and development of the Bulgarian Sense-Annotated Corpus. In *Proceedings of the Third International Corpus Linguistics Conference (CILC), 7-9 April 2011, Valencia, Spain*, pages 143–150. Universitat Politecnica de Valencia.

Svetla Koeva, Svetlozara Leseva, Ivelina Stoyanova, Tsvetana Dimitrova, and Maria Todorova. 2016. Automatic prediction of morphosemantic relations. In *Proceedings of the Eighth Global Wordnet Conference*, pages 168–176. University Al. I. Cuza Publishing House.

Małgorzata Korytkowska. 2008. Sachetaniyata glagol plyus sashtestvitelno kato leksikolozhki i leksikografski problem. / verb–noun combinations as a lexicological and lexicographic problem. *Izsledvaniya po frazeologiya, leksikologiya i leksikografiya (v pamet na prof. Keti Ankova-Nicheva) / Studies in Phraseology, Lexicology and Lexicography (in memory of Prof. Keti Ankova-Nicheva)*, pages 227–232.

Svetlozara Leseva, Ivelina Stoyanova, and Maria Todorova. 2018. Classifying Verbs in WordNet by Harnessing Semantic Resources. In *Proceedings of CLIB 2018*, pages 115–125, Sofia, Bulgaria.

Svetlozara Leseva, Maria Todorova, Tsvetana Dimitrova, Borislav Rizov, Ivelina Stoyanova, and Svetla Koeva. 2015. Automatic classification of wordnet morphosemantic relations. In *Proceedings of BS-NLP 2015, Hissar, Bulgaria*, pages 59–64.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Cătălina Mărănduc. 2010. *Dicţionar de expresii, locuţiuni şi sintagme ale limbii române*. Corint, Bucharest.

Istvan Nagy, Veronika Vincze, and Richard Farkas. 2013. Full-coverage Identification of English Light Verb Constructions. In *Proceedings of the International Joint Conference on Natural Language Processing, Nagoya, Japan, 14-18 October 2013*, pages 329–337. University of Hamburg.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*, pages 2–9.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240.

Borislav Rizov, Tsvetana Dimitrova, and Verginica Barbu Mititelu. 2015. Hydra for web: A multilingual wordnet viewer. In *Proceedings of the 11th International Conference Linguistic Resources and Tools for Processing the Romanian Language*, pages 19–30, Iaşi, Romania.

Horacio Rodriguez, Salvador Climent, Piek Vossen, Laura Bloksma, Wim Peters, Antonietta Alonge, Francesca Bertagna, and Adriana Roventini. 1998. The top-down strategy for building eurowordnet: Vocabulary coverage, base concepts and top ontology. *Computers and the Humanities*, 32(2-3):117–152.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Third International Conference on*

Computational Linguistics and Intelligent Text Processing, CICLing '02, pages 1–15. Springer-Verlag.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejek, Fabienne Cap, Slavomir pl, Silvio Ricardo Cordeiro, Glen Eryiit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskait, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartn, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. 2017. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, Phraseology and Multiword Expressions, pages 87–147. Language Science Press.

Agata Savary, Carlos Ramisch, Archna Bhatia, Claire Bonial, Marie Candito, Fabienne Cap, Silvio Cordeiro, Vassiliki Foufi, Polona Gantar, Voula Giouli, Carlos Herrero, Uxoa Iñurrieta, Mihaela Ionescu, Alfredo Maldonado, Verginica Mititelu, Johanna Monti, Joakim Nivre, Mihaela Onofrei, Viola Ow, Carla Parra Escartín, Manfred Sailer, Renata Ramisch, Monica-Mihaela Rizea, Nathan Schneider, Ivelina Stonayova, Sara Stymne, Ashwini Vaidya, Veronika Vincze, and Abigail Walsh. 2018. Annotation guidelines of the PARSEME shared task on automatic identification of verbal MWEs – edition 1.1. http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/.

Mileva Slavcheva. 2006. Semantic descriptors: The case of reflexive verbs. In *Proceedings of the 5th Language Resources and Evaluation Conference*, pages 1009–1014.

Ivelina Stoyanova, Svetlozara Leseva, and Maria Todorova. 2016. Towards the Automatic Identification of Light Verb Constructions in Bulgarian. In *Proceedings of CLIB 2016*, pages 28–37, Sofia, Bulgaria.

Y. Tu and D. Roth. 2011. Learning English Light Verb Constructions: Contextual or Statistical. In *Proceedings of MWE 2011, Portland, Oregon, USA*, pages 31–39. ACL.

Dan Tufiș and Dan Ștefănescu. 2012. Experiments with a differential semantics annotation for wordnet 3.0. *Decision Support Systems*, 53(4):695–703.

Dan Tufiș, Dan Cristea, and Sofia Stamou. 2004. BalkaNet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information Science and Technology Special Issue*, 7(1-2):9–43.

Dan Tufiș and Verginica Barbu Mititelu. 2014. *The Lexical Ontology for Romanian*, pages 491–504. Springer.

Veronika Vincze, Attila Almsi, and Janos Csirik. 2012. Multiword verbs In WordNets. In *Proceedings of the 6th International Global Wordnet Conference*, pages 337–381.