It Takes Two to Tango – Towards a Multilingual MWE Resource

Svetlozara Leseva
Institute for Bulgarian Language
Bulgarian Academy of Sciences
zarka@dcl.bas.bg

Verginica Barbu Mititelu
Research Institute for Artificial Intelligence
Romanian Academy
vergi@racai.ro

Ivelina Stoyanova
Institute for Bulgarian Language
Bulgarian Academy of Sciences
iva@dcl.bas.bg

Abstract

Mature wordnets offer the opportunity of digging out interesting linguistic information otherwise not explicitly marked in the network. The focus in this paper is on the ways the results already obtained at two levels, derivation and multiword expressions, may be further employed. The parallel recent development of the two resources under discussion, the Bulgarian and the Romanian wordnets, has enabled interlingual analyses that reveal similarities and differences between the linguistic knowledge encoded in the two wordnets. In this paper we show how the resources developed and the knowledge gained are put together towards devising a linked MWE resource that is informed by layered dictionary representation and corpus annotation and analysis. This work is a proof of concept for the adopted method of compiling a multilingual MWE resource on the basis of information extracted from the Bulgarian, the Romanian and the Princeton wordnet, as well as additional language resources and automatic procedures.

Keywords: wordnets, Bulgarian, Romanian, derivation, verbal multiword expressions, linked resources

1. Introduction

For almost a decade the development of the Bulgarian and the Romanian wordnets (BulNet and RoWN respectively) has involved shared research interests directed towards the enrichment of the two resources with qualitative information. Another relevant concern has been making linguistic information already existing in the wordnets accessible to computer processing: although the human specialist is able to spot different types of linguistic information in the two resources, it must be encoded in such a way that computer programmes can also easily access and use it. One of the avenues pursued along these lines has been digging derivational relations out of existing synsets and marking them explicitly in the two wordnets. Another strand of research involving joint efforts has been the encoding and exploration of multiword expressions (MWEs), and in particular several types of verbal multiword expressions (VMWEs) in wordnet synsets. The importance of MWEs has been widely acknowledged by linguistics and computational linguistics (Sag et al., 2002), as has been the significance of the ability of language processing systems to access resources in which such information is explicitly marked (Savary et al., 2019).

The results of these past and ongoing efforts have led to the idea of creating a lexical resource presenting a full description of MWEs that unifies the information available in wordnet with detailed morphological, syntactic, semantic, word order, pragmatic and derivational information. The greater goal is, using knowledge about Romanian, Bulgarian and English MWEs, to propose a framework for the description of MWEs that is applicable across languages, while also adaptable to language-specific features. Below we report on the ongoing work for the languages under study – Bulgarian and Romanian – with a recourse to the description of English VMWEs.

We start with a brief presentation of the development of the two wordnets under discussion and their enrichment with further relations (section 2). An interlingual analysis of the results of this enrichment is given in section 3. After that, we present the process of annotating verbal multiword expressions in the two wordnets with several multilingually defined types (section 4), while section 5 contains the results of the comparison between the types and the frequency of these verbal expressions in the two wordnets, as well as their interpretation. The work towards creating a multilingual linked VMWE resource that is currently underway is described in section 6.

2. BulNet and RoWN

The beginning and evolution of the Bulgarian wordnet (BulNet) (Koeva, 2010) and of the Romanian wordnet (RoWN) (Tufiş et al., 2013) were previously presented in Barbu Mititelu et al. (2017). Below we present some of the work carried out and made available through the two wordnets, which has inspired and informed our work on MWEs.

BulNet and RoWN were developed following the expand method (Rodriguez et al., 1998) and in compliance with two main principles: hierarchy preservation and conceptual density. Thus, the two wordnets preserve the structure of the original Princeton wordnet (PWN) and are aligned to it and, consequently, to each other and to any other wordnet aligned to PWN, this being a valuable asset for multilingual research and applications.

The interest in adding derivational relations to the two wordnets sprang up independently in the two teams: Koeva (2008) discusses important theoretical aspects of adding derivational relations to a wordnet, their multilingual relevance in the case of aligned wordnets, and presents the way in which the derivational relations from PWN were transferred and filtered in order to be included in BulNet. Barbu Mititelu (2013a) presents the methodology, heuristics and tools used for adding derivational relations to RoWN, as well as their importance for language applications.

For Romanian, Barbu Mititelu (2013b) presents in details the steps taken in the process of adding derivational relations among words of all parts of speech in RoWN (nouns, adjectives, verbs, adverbs). An initial phase of automatic identification of possible derivationally related pairs of any part of speech makes use of an exhaustive list of Romanian affixes. However, the resulting pairs require manual investigation for two reasons: on the one hand, some pairs contain false positives given that the beginning of a word can be misinterpreted as a prefix or the ending as a suffix, when this is a mere coincidence: consider the pair *val* 'wave' – *aval* 'downriver'; the latter can be morphologically misanalyzed as being formed from the former with the prefix *a-*, but this is not the case: *aval* is a French borrowing, where it is formed from Latin elements *ad-* 'at' and *vallis* 'valley'. On the other hand, manual validation is necessary because there must be a semantic connection between the words in a derivational relation; as such, there is a derivational relation between the words *drive* and *driver* when they are considered with the senses 'operate or control a vehicle' and 'the operator of a motor vehicle', respectively. There is also a derivational relation between them when they are considered with the senses 'push, propel, or press with force' and 'someone who drives animals that pull a vehicle', respectively. However, there is no such relation between them when considered with the senses interchanged (see Barbu Mititelu (2012) for a more detailed explanation).

As part of an effort along a similar avenue and in line with the theoretical considerations and the analyses proposed in Koeva (2008), Dimitrova et al. (2014) report on the steps, decisions and the theoretical motivation involved in the process of adding verb-noun derivational relations to BulNet. The adopted procedure was to start from the morphosemantic relations encoded in PWN (Fellbaum et al., 2009) and, using morphology-based heuristics, to identify and validate the derivational pairs in the corresponding BulNet synsets. Similar issues have been observed as the ones described for Romanian, particularly false positives and other errors due to overgeneration or failure of the procedures to capture different phonetic variants. An example of a false positive is represented by the pair *pod-slon-ya* 'give shelter' and *slon* 'elephant'. The results of the automatic procedures were therefore validated manually.

Besides marking the formal (i.e. morphological) relation between the words, their semantics was also described in terms of a set of predefined relations. For the noun-verb pairs these relations were bor-

rowed form PWN: Agent, Body-part, By-means-of, Destination, Event, Instrument, Location, Material, Property, Result, State, Undergoer, Uses, Vehicle. They all apply to the Bulgarian and Romanian pairs. With a view to discovering more derivational relations and attaching semantics to them in the already adopted framework, Koeva et al. (2016) proposed a machine learning method for automatic identification and classification of morphosemantic relations between pairs of potentially derivationally related verbs and nouns. The method employs the previously validated verb-noun derivationally related pairs and a number of linguistic features derived from the training data. The method is applicable to classifying MWEs as well, to the extent that the morphosemantic relations between single words would hold for MWEs headed by these single words.

3. Interlingual Comparison between Noun-Verb Relations in BulNet and RoWN

Annotating the morpho-semantically related noun-verb pairs in the two languages offered important insights into the derivational morphology of Bulgarian and Romanian as reflected in the respective wordnets (Tarpomanova et al., 2014): quantitatively, we found a richer system of suffixes in Bulgarian, as well as richer polysemy displayed by them. However, an important number of similarities could also be identified. Firstly, the same relations tend to be best or better represented in both wordnets: Agent and Event are the best represented from two perspectives: number of suffixes involved and frequency in the networks. The latter could also be regarded as a result of the similar objectives followed when deciding on the the wordnets development (see section 2). Almost all the other relations have a similar distribution in the annotated data for both languages[1].

Secondly, polysemous suffixes occurring in both languages are specialized for a certain set of relations, but have a preferred reading: e.g.: Bg. *-tel* forms nouns from verbs that bear the semantic relations Agent, Material, Instrument, By-means-of, Undergoer, and Uses, but Agent is by far the prevalent one; Ro *-(ă)tură* creates nouns that establish one of the following semantic relations – Event, Result, By-means-of, Instrument, Material, Uses – with the root verb, with Event being the best represented. There are suffixes occurring in both languages and showing high similarity in their semantics[2]: e.g. the suffix *-tor* is productive in both languages and in the vast majority of cases serves to derive nouns expressing the relation Agent, but may also be found with other relations such as: Instrument, Material, By-means-of, Uses.

4. Identifying and Classifying VMWEs in BulNet and RoWN

Wordnets contain both simple words and word combinations. The manual inspection of the latter has led to distinguishing (Barbu Mititelu et al., 2019), on the one hand, between free combinations with a compositional meaning (annotated with the label NONE) and expressions and, on the other hand, among several types of verbal multiword expressions, which were defined in the PARSEME shared task 1.0 (Savary et al., 2017) and then refined for shared task 1.1 (Ramisch et al., 2018). The VMWE labels used for annotating the VMWEs in BulNet and RoWN are: VID (verbal idioms), LVC.full (light verb constructions in which the verb is semantically bleached), LVC.cause (light verb constructions in which the verb has a causative meaning), IRV (inherently reflexive verbs), for both languages, and IAV (inherently adpositional verbs) only for Bulgarian (although such verbs also exist in Romanian, but the preposition remains underspecified in synsets). Table 1 illustrates all the types of labels with examples from the two wordnets.

[1]Further analysis of the data (Barbu Mititelu et al., 2015) involved comparison of the annotated noun-verb pairs in Bulgarian and Romanian with the corresponding English ones and that revealed the same tendency in the productivity of relations in all three languages, with Event, Agent and Result being the best represented. At the same time, the data confirmed the tendency towards conversion displayed by English.

[2]Larger data could further confirm these results as well as refine them.

VMWE type	Example from BulNet	# in BulNet	Example from RoWN	# in RoWN
VID	*cheta mezhdu redovete* 'read between the lines'	775	*citi printre rânduri* 'read between the lines'	614
LVC.full	*vzema uchastie* 'take part'	465	*lua parte* 'take part'	102
LVC.cause	*hvărlyam văv văztorg* 'cause to go into ecstasies'	63	*lăsa loc* 'allow for'	42
IRV	*gnevya se* 'become angry'	1,822	*[se] înfuria* 'become angry'	989
IAV	*razbiram ot* 'be good at'	39	not annotated	-

Table 1: Types of VMWEs in BulNet and RoWN and their distribution.

These types of VMWEs were defined within a multilingual context involving almost thirty languages from different families and displaying various characteristics. However, the annotation of the corpora participating in the shared tasks did not involve parallel corpora and neither was any interlingual analysis of VMWEs at the sense level made. Should there be such annotation available in the two wordnets, it would be possible to study interlingual equivalents.

5. An Interlingual Account of VMWEs

Previous analyses on VMWEs as represented in BulNet and ROWN, cf. Barbu Mititelu et al. (2019), have shown a number of parallels and differences between Bulgarian and Romanian VMWEs. In the paper under discussion, we report on 3,656 multitoken literal-to-literal pairs in corresponding synsets. These include VMWEs proper and multitoken free phrases (marked as NONE); their distribution is presented in Table 2 (for the purpose 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 2: Distribution of VMWE literal-to-literal correspondences between BulNet and RoWN.

5.1. Interlingual Analysis of the Data

With a big overlap of 72.7% reported between the VMWE types in the data under discussion, there are also plenty of examples of the same meaning being lexicalized by different types of VMWEs across the two languages and even in the same language (different-type MWE literals in one synset), a trend that is even more relevant when comparing or describing multiple languages.

As discussed therein, the interlingual correspondence is most consistent in the category of reflexive verbs (IRVs), which is to be expected given the similar semantics attached to reflexive verbs in the two languages (Slavcheva, 2006). In light of a dictionary-based approach to accounting for MWEs, IRVs do not pose considerable difficulties, as they constitute a recognized part of the vocabulary in both languages and their description follows the general guidelines for single words, as the reflexive component does not vary.

Other consistent correspondences are found in the class of verbal idioms (VIDs), which, as the authors admit, might be due to the fact that the choice of VIDs to encode was more or less influenced by internationally established idioms (respectively, calques in the two languages) already implemented

in PWN, such as {*read between the lines*:1}, {*send a message*:1}, among others. Nonetheless, these are expressions that are established in the languages under discussion and observe their morphological and syntactic peculiarities.

Correspondences are less marked in the domain of light-verb constructions (LVCs) for a couple of reasons: first of all, this class of VMWEs is not consistently represented in BulNet and ROWN – LVCs have usually been implemented to make up for lexical gaps; secondly, the teams working on the two wordnets have adopted different strategies, including a considerable difference in the number of light verbs recognized – 118 verbs for Bulgarian and 21 verbs for Romanian. Nonetheless, as the annotated data from the PARSEME corpora show, LVCs are pervasive in the two languages, so one of the objectives in proposing a dictionary-based resource is to properly account for this category of MWEs that is underrepresented in the two wordnets, as well as in many dictionaries.

As the type of VMWE that lexicalizes a particular sense is largely idiosyncratic, we would like the VMWE type to be an integral part of the entry of each individual VMWE: it should be assigned to or validated (if already available) individually for each VMWE literal (if there are more than one in a synset) and should be accessible for processing as the VMWE type enables the prediction of certain morphological, syntactic, word-order and other properties of the respective unit. Therefore, we have encoded the VMWE type as one of the features for description at the semantic level.

Given these observations, our efforts are directed primarily to capturing the linguistic features of LVCs and VIDs.

6. Towards a Multilingual Linked VMWEs Resource

The description of the various linguistic levels below is based on a proposal for the semi-automatic compilation of a MWE dictionary made in Stoyanova et al. (2016), which was further expanded to accommodate: (i) a stand-off format with links to wordnet synsets and literals; (ii) other levels of description, such as the VMWE types adopted in PARSEME, information about the connotation and the derivational potential of VMWEs; (iii) a multilingual testing setting for the description of VMWEs (Stoyanova et al., 2019).

The linked VMWE resource proposed harnesses several previously developed resources: (i) the three wordnets discussed above: RoWN, BulNet and PWN, which inform the general framework and provide a substantial part of the VMWE inventory as well as rich semantic and pragmatic information for the VMWEs included in them; (ii) VMWE annotated corpora for the two languages developed under the PARSEME initiative(Ramisch et al., 2018); (iii) single-word derivational patterns and instances for Romanian (Barbu Mititelu, 2013b) and Bulgarian (Dimitrova et al., 2014; Koeva et al., 2016) and MWE-to-MWE patterns for the two languages (Barbu Mititelu and Leseva, 2018).

6.1. Levels of Description

Below is a summary of the levels of description proposed.

1. **Technical information**. The technical information supports the linking between the dictionary entries and the respective wordnets, particularly through the unique synset ID and an additionally employed VMWE ID which serves both to identify a VMWE as part of a particular synset and to distinguish it from other VMWEs in the same synsets or from identical VMWE literals in other synsets. This allows us to: (a) access all the synset-level linguistic information provided; (b) make references to a particular VMWE uniquely, e.g., in the description of derivatives.

2. **Morphological description** which includes several types of information:

   - The lemma of the VMWE (non-abstract lemma) and a lemmatized form of each component of the VMWE (abstract lemma) (Savary, 2008). The parallel use of both types of lemma is motivated as follows: the non-abstract lemma is the human readable lemma, while the abstract one helps identifying VMWEs in a lemmatized corpus and assigning each such corpus occurrence of a VMWE a linguistically proper lemma that will link it to the wealth of information associated with the respective dictionary entry.

- A regular morphosyntactic representation which consists of the unrestricted set of forms of the expression's head and the unrestricted set of forms of the non-head components. This type of description is relevant for VMWEs with a full paradigm of the verbal head and its dependents and is typical of IRVs and IAVs, as well as of many LVCs. This set can be obtained from the in-house morphologic lexicons each team has.
- Restrictions on the paradigmatic realization of the verbal head with respect to one or more morphosyntactic features, such as person, number, tense, mood, polarity, etc., e.g. RO *nu privi cu ochi buni* (not watch with eyes good, 'regard with disfavour') is always used with the negative marker *nu* 'not'; the same goes for BG *ne iskam akăl nazaem* (not want brains to borrow, 'to not need unsolicited advice').
- Restrictions on the inflected forms of the dependent components of a VMWE. Such a field is defined for each dependent and is used to explicitly encode any restrictions on a dependent's possible forms as part of the VMWE. Considering the above example, the noun *ochi* ('eyes') is restricted to the plural indefinite form, while the BG *akăl* ('brains') is restricted to the singular indefinite form.

3. **Syntactic description**. The syntactic description is based on the UD framework as it aims at achieving universality, while offering the possibility to define language characteristics in the same framework (`https://universaldependencies.org/u/dep/index.html`).

   - Internal syntactic structure of the VMWE, which describes the number of components, the syntactic category of each them and the syntactic relations between the components. The most common structures found across Romanian and Bulgarian VMWEs as reflected in the analysed data are illustrated in Table 3.

Relation	**Description of relation**	**Example RO**	**Example BG**
V + obj	linking V to the entity acted upon or undergoing change	*da declarație* (give declaration, 'declare')	*vzemam reshenie* ('make a decision')
V + obl	linking V to a nominal as a non-core (oblique) argument or adjunct	*inceta din viață* (cease from life, 'die')	*poemam v svoi rătse* ('take into one's own hands')
V + advmod	linking V to a (non-clausal) adverb or adverbial phrase that modifies the predicate	*da afară* (give outside, 'remove from job, fire')	*vzemam predvid* ('take into account')
V + nsubj	the VMWEs is made up of a verb and a subject	*fura somnul* (steal sleep-the, 'fall asleep')	*zvezdata mi izgryava* ('one's star is rising')
V + nsubj + obl	linking V to a subject and a non-core (oblique) argument or adjunct	*îngheța sângele în vine* (freeze blood-the in veins, 'get cold feet')	*krăvta zamrăzva v zhilite mi* (blood-the freezes in my veins, 'get cold feet')
V + obj + obl	linking V, an object and a non-core (oblique) argument or adjunct	*găsi drumul în viață* (find road-the in life, 'find one's way in life')	*tsepya stotinkata na dve* (split the penny in two, 'be stingy')

Table 3: Types of syntactic structures across Romanian and Bulgarian VMWEs.

   - Possible dependents of the MWE elements. Some MWE components may have dependents, others may not. These dependents are either arguments, that is, obligatory dependents of the VMWE components, or adjuncts, i.e. dependents that are not required for the sentence to be grammatical.
     Argument dependents are usually predetermined by the head verb's argument structure. Consider, for instance RO *citi printre rânduri*, BG *cheta mezhdu redovete* and EN *read between*

the lines, which have an identical syntactic structure: the verbal head takes dependents of the type subject (nsubj or csubj in UD terminology) and direct object (obj in UD) in order to form a grammatical sentence and these positions need to be posited as slots in the VMWE description that need to be filled by a suitable phrase in order for the VMWE to form grammatical utterances. The dependent MWE component, in this case the prepositionally introduced noun, cannot be or is rarely modified by another word.

In contrast, the dependents of some VMWEs, light-verb constructions in particular, may readily take adjuncts of their own, e.g. BG *vzemam reshenie* ('make a decision') > *vzemam vazhno reshenie* ('make an important decision'), where the dependent noun, which is an object, may be modified (nmod in UD).

Both types of possible dependents must be encoded in the syntactic description, especially as many of them may intervene between the components of the VMWE; keeping track of them may provide useful information about the distance between the individual elements of a MWE in running text – a peculiarity that affects the automatic recognition of MWEs.

- Any restrictions on the word order of the VMWE components and of the possible dependents are encoded in this field. For example, in the RO VMWE *da ortul popii* (give coin-the to-priest-the, 'die') the obj *ortul* always precedes the indirect object *popii* (iobj in UD); in the BG VMWE *na star krastavichar krastavitsi prodavam* (to an old cucumber-seller cucumbers sell, 'to try to cheat someone with experience') the obl *na star krastavichar* precedes the obj *krastavitsi*, while the verb can be either at the front or at the end. This information is useful in MWE recognition.

4. **Semantic description**. The semantic description unites the idiosyncratic information about the basic properties of MWEs that predetermine their morphosyntactic behaviour, with lexical, usage and pragmatic information available from wordnet and possibly from other resources.

   - The MWE type – defined according to the guidelines adopted in the PARSEME shared task edition 1.1 (Ramisch et al., 2018).
   - The wealth of semantic information – the explanatory definition (gloss), single-word and MWE synonyms, other semantic and derivational relations, usage examples – that is accessible through the linking to wordnet and pertains to the entire synset.
   - Usage and register information. This field provides relevant restrictions on the usage of VMWEs, which may be automatically retrieved from the respective wordnet, if available, or added by a lexicographer. For example, many idioms are specific to the informal use, e.g. RO *bate la ochi* (beat at eyes, 'catch someone's eye'); BG *udryam kyoravoto*:1 (hit the blind, 'hit the jackpot'), and need to be accordingly marked.
   - The positive, negative or neutral connotation of a given VMWE whose value may be obtained either from available resources, such as SentiWordNet (Baccianella et al., 2010), or supplied manually. In the former case the connotation values are assigned from the respective synsets which have been assigned values from SentiWordNet (transferred automatically to BulNet and RoWN). For instance, the corresponding synsets BG: {*puka mi:1, dreme mi:2, davam pet pari:1, dam pet pari:1, davam puknata para:1, dam puknata para:1*}, RO: {*da doi bani:1, da două parale:1*}, EN: {*care a hang:1, give a hoot:1, give a hang:1, give a damn:1*} are assigned a positive value of +0.125 and a negative value of –0.375. Even so, manual validation is needed as the connotation value of individual literals may be language specific.

5. **Derivational information**. The derivational potential of MWEs has been tackled to a certain extent in the PARSEME initiative. The Romanian and the Bulgarian perspective on VMWE-to-MWE derivation, including a description of the semantic, syntactic and other changes that take place in the process of derivation, has been discussed in Barbu Mititelu and Leseva (2018), which we follow to a great degree. We adopt a verb-centric approach, regardless of the actual direction of the derivation, and currently focus on verb-to-noun derivation such as the one exemplified by the pairs: RO *spăla*

creierul – spălarea creierului, BG *promivam mozăka – promivane na mozăka*, EN *brainwash – brainwashing*. The derivational information is presented as a list of possible derivatives for each VMWE lexicon entry. Derivatives are encoded in the form that occurs in the respective wordnet and are accompanied by the ID of the synset to which they belong. If the derivatives are not implemented in the wordnet yet, then the ID remains unspecified.

6.2. Procedures for Semi-automatic Description

Below we present the baseline VMWE resource, which incorporates various levels of linguistic description for each of the languages. It was compiled using a series of automatic procedures and heuristics. The original VMWE inventory consists of all the synsets in BulNet and RoWN that contain at least one VMWE. Consequently, their correspondences in Princeton wordnet were also included regardless of whether they contain VMWEs. The baseline resource consists of 944 synsets which have a VMWE in both Bulgarian and Romanian with 2,744 literals on the Bulgarian side and 1,533 literals on the Romanian side. For 340 out of the 944 synsets there is a VMWE correspondence in English with a total of 662 VMWE literals.

The automatically retrievable information for each field of the description was assigned. Where possible, default values were determined, which need to be checked manually. The default values depend on a number of factors: (i) the form in which the VMWEs components are found in the lemma of the VMWE: if a component participates in a VMWE in its citation form, its full paradigm is its default value (not considering other factors); if the component in the VMWE's lemma is in a different form, it is most likely restricted with respect to the relevant grammatical category: consider, for instance, the VMWE *make advances* – the nominal component *advances* is in the plural in the lemma of the MWE and is unlikely to be found in the singular; (ii) the type of the MWE – for example, LVCs are more permissible than VIDs with respect to the modification of the dependents. Below we present the types of information that are automatically retrievable from the description of the MWEs in the wordnets under discussion.

1. **Automatic tagging and further morphological analysis**. The MWEs in the three languages are automatically POS-tagged using available programming tools. The BG data were annotated using the Bulgarian Language Processing Chain[3] and the RO and the EN MWEs were processed using the UDPipe with a Romanian and an English language model[4] respectively. The tagging was used in the grammatical description of the MWEs, in particular, for identifying (i) the POS tags of the MWEs components; (ii) the MWE's abstract lemma; (iii) the lexico-grammatical and grammatical features, such as verb aspect (in BG), number and definiteness for nominal components, etc.. As illustrated by the example above (*make advances*), the form in which a component is fixed in the non-abstract lemma, such as the one retrievable from wordnet, helps in predicting the possible variations of this component's grammatical properties (or a part of them).

2. **Syntactic analysis**. On the basis of the morphosyntactic tagging we derive the linear order of the components and we identify the basic internal syntactic structure of the MWEs, in particular: (i) the head and the dependents; (ii) the possible modifiers of the components (e.g. an NP dependent may take an adjective modifier); (iii) their basic word order and word order variations (e.g., the position of the reflexive particle in IRVs in BG and RO); (iv) the default values for the possible modifiers of the dependents based on the PARSEME type: 'yes' for LVCs, 'no' for VIDs and IRVs.

3. **Semantic description**. We extracted the available semantic information such as the synset ID, the definition, synonyms, semantic relations, register restrictions, etc. from the relevant synsets in the wordnets.

4. **Derivational information**. The derivational information is retrieved from wordnet as well by collecting all the synsets labelled as derivationally related to the one to which the MWE under discussion belongs regardless of the language for which the derivation applies. Further, we select

[3] http://dcl.bas.bg/dclservices/index.php
[4] http://ufal.mff.cuni.cz/udpipe

the multiword derivatives and analyse the matching components between the original MWE verb (literal in the verb synset) and the potential derivatives.

Table 4 shows the linking of corresponding MWE entries in BG *zatvaryam si ochite* 'close one's eyes', RO *închide ochii* 'close the eyes' and EN *turn a blind eye* with the components of their description. As the respective wordnet synsets do not have derivatives encoded, regularly produced derivatives – such as eventive nouns derived from verbs, e.g. BG *zatvarym si ochite* 'close one's eyes' – *zatvaryane na ochite* 'closing of the eyes' – need to be additionally extracted from corpus data or from available (lexicographic) resources.

Feature	**BG**	**RO**	**EN**
PWN ID	eng-30-00801977-v	eng-30-00801977-v	eng-30-00801977-v
MWE ID	bg_427	ro_265	en_3
Lemma ID	*zatvaryam si ochite*	*închide ochii*	*turn a blind eye*
Abstract lemma ID	*zatvaryam svoy oko*	*închide ochi*	EN *turn a blind eye*
Components	1_zatvaryam_V 2_svoy_PronP 3_oko_N	1_închide_V 2_ochi_N	1_turn_V 2_a_DET 3_blind_A 4_eye_N
Syntactic structure	V + obj	V + obj	V + obj
Verbal head	*zatvaryam*	*închide*	*turn*
Gram. features	1_VLITsr1_IMPERF	1_Vmip3s	1_VB
Dependents	2_svoy_PFPZ 3_oko_NCNpd	2_ochi_Ncmpd	2_a_DET 3_blind_A 4_eye_Ns
Restrictions	3_Npd	2_Npd	4_Ns
Modifiers	No	No	No
Word order	V_PronP order changes	–	fixed
PARSEME type	VID	VID	VID
Synonyms	bg_428: *zatvorya si ochite*	–	–
Register	Informal	Informal	Informal
Sentiment	–0.5 / +0.0	–0.0 / +0.0	–0.5 / +0.0

Table 4: An example of linked corresponding MWE entries in BG, RO and EN. (The POS notation is unified across the languages. POS: V – verb, N – noun, A – adjective, Adv – adverb, P – preposition, Pron – pronoun, DET – determiner, etc. The morphological features are partially unified so as to facilitate the use of the uniform notation of restrictions: PERF/IMPERF – verb aspect, s/p – singular/plural, 0/d – indefinite/definite, etc.).

7. Conclusions

The construction of the linked VMWE resource is work in progress and we are currently focused on the manual validation of the entries and the addition of missing linguistic information. Apart from providing description of Romanian and Bulgarian VMWEs in the adopted format, we are also interested in testing the applicability of the description cross-linguistically for capturing language-specific features towards obtaining a more fine-grained typology of syntactic and semantic types of VMWEs.

While the proposal makes use of widely recognized frameworks, such as aligned wordnets, the UD formalism, PARSEME VMWEs types, derivational morphology and semantics, our effort is aimed at accommodating them in a unified, data-driven framework and at providing a linked data formalism.

Acknowledgements

This work was carried out under the project *Enhancing Multilingual Language Resources with Derivationally Linked Multiword Expressions* between the Institute for Bulgarian Language at the Bulgarian Academy of Sciences and the Research Institute for Artificial Intelligence at the Romanian Academy.

References

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010), May 17-23, Valletta, Malta*, pages 2200–2204.

Barbu Mititelu, V. and Leseva, S. (2018). *Multiword expressions: Insights from a multi-lingual perspective*, chapter Derivation in the domain of multi–word expressions, pages 215–246. Berlin: Language Science Press.

Barbu Mititelu, V., Rizov, B., Tarpomanova, E., Leseva, S., and Dimitrova, T. (2015). Noun-Verb Derivation in the Bulgarian, Romanian and English Wordnets – A Comparative Approach. In *Proceedings of the 11th International Conference "Linguistic Resources and Tools for Processing the Romanian Language"*, pages 53–64.

Barbu Mititelu, V., Leseva, S., and Tufiș, D. (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*, volume 2, pages 192–200.

Barbu Mititelu, V., Stoyanova, I., Leseva, S., Maria Mitrofan, T. D., and Todorova, M. (2019). Hear about Verbal Multiword Expressions in the Bulgarian and the Romanian Wordnets Straight from the Horse's Mouth. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 2–12.

Barbu Mititelu, V. (2012). Adding Morpho-semantic Relations to the Romanian Wordnet. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2596–2601. European Language Resources Association (ELRA).

Barbu Mititelu, V. (2013a). Increasing the Effectiveness of the Romanian Wordnet in NLP Applications. *Computer Science Journal of Moldova*, 21(3(63)).

Barbu Mititelu, V. (2013b). *Reţea semantico-derivaţională pentru limba română*. Editura Muzeul Literaturii Române, Bucharest.

Dimitrova, T., Tarpomanova, E., and Rizov, B. (2014). Coping with derivation in the Bulgarian wordnet. In *Proceedings of the 7th Global Wordnet Conference*, pages 109–117.

Fellbaum, C., Osherson, A., and Clark, P. E. (2009). *Responding to In- formation Society Challenges: New Advances in Human Language Technologies*, volume 5603, chapter Putting Semantics into WordNet's "Morphosemantic" Links, pages 350–358. Springer Lecture Notes in Informatics.

Koeva, S., Leseva, S., Stoyanova, I., Dimitrova, T., and Todorova, M. (2016). Automatic prediction of morphosemantic relations. In *Proceedings of the Eighth Global Wordnet Conference*, pages 168–176. University Al. I. Cuza Publishing House.

Koeva, S. (2008). Derivational and Morphosemantic Relations in Bulgarian Wordnet. *Intelligent Information Systems*, XVI.

Koeva, S. (2010). Bulgarian Wordnet – current state, applications and prospects. In *Bulgarian-American Dialogues*.

Ramisch, C., Cordeiro, S., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., and Walsh, A. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *The Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240.

Rodriguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertagna, F., and Roventini, A. (1998). The top-down strategy for building eurowordnet: Vocabulary coverage, base concepts and top ontology. *Computers and the Humanities*, 32(2-3):117–152.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. pages 1–15.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., and Doucet, A. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*.

Savary, A., Cordeiro, S., and Ramisch, C. (2019). Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91. Association for Computational Linguistics.

Savary, A. (2008). Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, 1(2)).

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

Stoyanova, I., Koeva, S., Todorova, M., and Leseva, S. (2016). Semi-automatic Compilation of a Very Large Multiword Expression Dictionary for Bulgarian. In *Proceedings of GLOBALEX 2016: Lexicographic Resources for Human Language Technology, Workshop at LREC-2016, Portorož, Slovenia, May 24, 2016*, pages 86–95.

Stoyanova, I., Leseva, S., Barbu Mititelu, V., Todorova, M., and Cristescu, M. (2019). Wrapping our Heads Around VMWEs and their Derivatives. In *Proceedings of the 14th International Conference "Linguistic Resources and Tools for Natural Language Processing"*, pages 153–166.

Tarpomanova, E., Leseva, S., Todorova, M., Dimitrova, T., Rizov, B., Barbu Mititelu, V., and Irimia, E. (2014). Noun-Verb Derivation in the Bulgarian and the Romanian WordNet – A Comparative Approach. In *Proceedings of the First International Conference Computational Linguistics in Bulgaria*, pages 23–31.

Tufiş, D., Barbu Mititelu, V., Ştefănescu, D., and Ion, R. (2013). The Romanian Wordnet in a Nutshell. *Language Resources and Evaluation*, 47(4):1305–1314.