Chapter 3

A uniform multilingual approach to the description of multiword expressions

Svetlozara Leseva[a], Verginica Barbu Mititelu[b], Ivelina Stoyanova[a] & Mihaela Cristescu[c]

[a]Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences [b]Research Institute for Artificial Intelligence, Romanian Academy [c]Faculty of Letters, University of Bucharest

In this chapter we describe a linked bilingual (Bulgarian and Romanian) computational lexicon of multiword expressions, a new resource which encompasses lexical, morphological, semantic and stylistic information, in an independent, though unified way. The lexicon is a bilingual lexicographic resource, originating in the wordnets for the two languages, and is made up of self-contained monolingual lexicons of multiword expressions, which may be expanded to cover other levels and features of linguistic description, as well as other languages.

1 Introduction and main objectives

Along with the efforts in the domain of traditional lexicography, various developments towards the compilation of lexicons of multiword expressions (MWEs) for the needs of computational lexicography and computational linguistics have also been undertaken. As emphasised in a position paper (Savary et al. 2019) that emerged from the PARSEME[1] initiative (Savary et al. 2015), devising syntactic

[1]PARSEME was a COST Action (2013–2017) focusing on parsing and MWEs. Some of its major results were the creation of annotation guidelines for verbal MWEs for more than 20 languages from various language families, a multilingual journalistic corpus annotated according to these guidelines made publicly available and a series of shared tasks on the identification of MWEs in texts, in which the previously mentioned corpus was used for training and testing the participating systems. See https://typo.uni-konstanz.de/parseme/.

Svetlozara Leseva, Verginica Barbu Mititelu, Ivelina Stoyanova & Mihaela Cristescu. 2024. A uniform multilingual approach to the description of multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 73–116. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998635

MWE lexicons was recognised as a prerequisite for advancing research in MWE identification and other MWE-related tasks.

We propose an electronic bilingual MWE lexicon that comprises morphological (inflectional and derivational alike), syntactic (including word order) and semantic description in an independent, though unified, way. We build upon the one proposed by Leseva et al. (2020), itself inspired by the MWE description in Koeva et al. (2016). Our goal is to create a linked bilingual lexicographic resource consisting of self-contained monolingual lexicons of MWEs that may be expanded to other levels of linguistic description and to other languages.

Our work has the following main contributions: (i) an overview of several approaches for the description of MWEs with interest in language-independent, cross-lingual, bilingual, and/or multilingual representation, and especially in the features used in the MWE description – see §2. §3 briefly describes the wordnets[2] for the two languages in focus and their characteristics that allow for the creation of the linked lexicon presented here, along with the compilation of the datasets of verbal multiword expressions (henceforth VMWEs) involved in the linguistic analysis. The features previously mentioned serve as a starting point in designing the structure of the MWE lexicons for Bulgarian and Romanian, linked into one resource, described in this work; (ii) the presentation of a uniform framework for the construction of a linked resource consisting of two MWE lexicons (for Bulgarian and Romanian) that takes into consideration the advantages and challenges posed by the existing approaches and practices – see §4. This is a first step in the creation of a multilingual resource for the lexicographic description of MWEs, both in structural and semantic perspectives; (iii) the exploration of the lexicographic representation of MWEs in the context of aligned general lexical, semantic and morpho-syntactic resources not exclusively compiled for MWEs, this step being an important prerequisite for various Natural Language Processing (NLP) applications – see §5. We show that a uniform description of MWEs is possible for two languages from different families, highlighting language similarities, but also ensuring the mechanisms that allow for the description of language specificities.

[2]We write *wordnet* when refering to a "lexical knowledge base for a given language, modeled after the principles of Princeton WordNet" (see http://www.dblab.upatras.gr/balkanet/journal/20_BalkaNetGlossary.pdf). We write *Wordnet* when refering to a particular such resource, here the Bulgarian Wordnet and the Romanian Wordnet; the form *WordNet* is used only with reference to the trademarked Princeton WordNet (see https://wordnet.princeton.edu/).

2 Advances in computational lexicography with a recourse to MWEs

Most of the times, MWEs are recorded in general language dictionaries, where they are usually only semantically described, i.e., their meaning is explained. Large computational lexical resources also make provisions to incorporate MWEs (Chiarcos et al. 2024 [this volume]). Even valence dictionaries focused on the general language can contain descriptions of MWEs: see Walenty (Przepiórkowski et al. 2014b), which was extended to accommodate properties of MWEs (Przepiórkowski et al. 2014a).

However, dedicated lexicons do exist for MWEs in some languages and various grammatical formalisms were adopted in their description: the Lexicon-Grammar framework (Gross 1975, 1982), which spurred substantial advances in the formal linguistic description, including the treatment of MWEs, was more recently used in the description of Italian MWEs (Vietri 2014b, Monti 2014); Lexical-Functional Grammar (LFG, Bresnan 1978, Dalrymple 2023) was applied in the development of a Norwegian MWE resource (Dyvik et al. 2019); Head-driven Phrase Structure Grammar (HPSG, Pollard & Sag 1987, 1994, Müller et al. 2021) was adopted in the LinGO project[3] for the creation of a lexicon including both simplex entries and MWEs (Villavicencio et al. 2004b); Frame Semantics was used to provide shallow semantic representation of multiword predicates (Giouli et al. 2024 [this volume]); Meaning-Text Theory (Mel'čuk 1981) was employed in Mel'čuk's (2006) Explanatory Combinatorial Dictionary, while the work by Schafroth (2015) offers a learner-centered description of Italian idioms based on the theoretical principles of Construction Grammar (Fried & Östman 2004).

Most MWE lexicons are monolingual resources (Fellbaum & Geyken 2005, Grégoire 2007, Odijk 2013, Shudo et al. 2011, Villavicencio et al. 2004b, Vietri 2014b, Schafroth 2015, Mel'čuk 2006, Markantonatou et al. 2019, Skoumalová et al. (2024 [this volume])). Others boast multilinguality as an important feature. However, multilingual support is ensured in different ways in different projects. Villavicencio et al. (2004a) report on MWEs in a source language that are manually given their equivalents in a target language, thus ensuring semantic equivalence between MWEs in the two languages, while the lexical and syntactic equivalences have to be decided upon by the user. Konbitzul[4] (Iñurrieta et al. 2018) is a bilingual Spanish-Basque verb-noun lexicon of MWEs. Besides containing MWE equivalents in the two languages, it also offers morphosyntactic information about the MWEs in both languages, which is introduced either manually

[3] https://www-csli.stanford.edu/groups/lingo-project

[4] http://ixa2.si.ehu.eus/konbitzul/

or semi-automatically. The Genoese-Italian phraseological dictionary[5] describes Genoese MWEs, including their Italian equivalent(s) (Autelli 2020).

Some of the discussed MWE initiatives supply translation equivalents to the described units in other languages (either MWEs, if available, or free phrases) (Markantonatou et al. 2019, 2020, 2024). This feature is especially useful for dictionaries of less-spoken languages where the use of English as a metalanguage increases the usability and understandability of the resource.

Some of the projects developing MWE resources focused on harvesting them from corpora, providing consistent representation of the MWE system within a language, as well their extensive description at various linguistic levels.

Harvesting of MWEs from corpora was done (i) automatically, either from corpora annotated with MWEs (Grégoire 2007) or from corpora lacking such annotation (Fellbaum & Geyken 2005, Odijk 2013); or (ii) manually (Dyvik et al. 2019, Shudo et al. 2011, Odijk et al. 2024).

Given the characteristics of MWEs (e.g., discontinuity, inflection of components, word order variation, etc.), the automatic analysis of corpora is prone to errors, hence it is usually followed by a manual inspection and selection of MWEs. Automatic identification of MWEs in corpora benefits from the morphosyntactic annotation and lemmatisation of the texts (Odijk 2013). Some authors combine the extraction of MWEs from corpora with selecting MWEs from available idiom or general-purpose dictionaries or lists. In such cases, examples from corpora and/or the web serve to supplement the dictionaries with new entries, to confirm and exemplify the uses and various phenomena concerning MWEs (Hnátková et al. 2019, Markantonatou et al. 2019, 2020, Skoumalová et al. 2024).

Describing the system of MWEs within a language concerns the paradigmatic aspect of MWEs, a topic that is more rarely touched upon in the dedicated literature. Grégoire (2007) discusses the organisation of Dutch MWEs in classes (called "equivalence classes") according to syntactic characteristics, the inner structure of MWEs and the possibility for them to have modifiers; Villavicencio et al. (2004b) use "meta-types" to organise the MWEs in classes and to map "the semantic relations between the elements of the MWE into the appropriate grammar dependent features" (Villavicencio et al. 2004b).

With respect to the way in which MWEs are described in lexicographic resources, two trends were dominant in the literature. In one of them, all MWEs are entries in a lexicon: their description is made either by specifying a class to which they belong (Grégoire 2007) or by enumerating their characteristics, with

[5]https://romanistik-gephras.uibk.ac.at

special focus on idiosyncrasies (Gross 1996, Shudo et al. 2011, Al-Haj et al. 2013, Markantonatou et al. 2019, 2020).

In a different approach, Villavicencio et al. (2004b) propose a description of MWEs adjusted to their decomposable or non-decomposable types. Thus, fixed (i.e., non-decomposable) MWEs should be treated as simplex entries: their orthography, syntactic and semantic type as well as morphological inflection of components are specified. Flexible or decomposable expressions are also lexical entries encoded in three stages: (i) their components are registered as idiomatic entries associated with the non-idiomatic entries from which they inherit their grammatical characteristics; (ii) over-generation is avoided by defining the context of use for these idiomatic entries: for each MWE the components are listed, along with their obligatory or optional status; (iii) MWEs are assigned to a metatype.

Similarly, Al-Haj et al. (2013) include MWEs as entries in their lexicon, alongside entries of simple words. Each component of a MWE contains a pointer to the corresponding simple entry in the lexicon. In a way similar to Villavicencio et al. (2004b), they propose adding fossil words[6] as entries, which are not assigned a part of speech, but are marked as "fossil", which is an indication of their occurrence only as components of MWEs.

Alternatively, in the *Explanatory combinatorial dictionary* (Mel'čuk 2006) different types of MWEs are treated differently: idioms and quasi-idioms are allotted separate entries (also cross-referenced with their components' entries) with their own fully-fledged description, whereas the so-called semi-phrasemes are described in the entry of their base, which, in the case of light verb constructions (LVCs) (a type of semi-phrasemes), is most often a noun serving as the semantic head of the expression. The combinatorial properties of semi-phrasemes are represented lexicographically by means of a special lexical function. Equivalent meanings formed on different support verbs are listed together.

Given that no standard was defined for it (yet), an important aspect of the linguistic description of MWEs is that it should not be framework-specific and should allow for its reuse by any system (Odijk 2013). There is agreement among researchers that MWEs must be explicitly marked as such in lexicons (Fellbaum & Geyken 2005, Mel'čuk 2006, Al-Haj et al. 2013, Dyvik et al. 2019, Hnátková et al. 2019, Markantonatou et al. 2019, 2020).

Taking as a point of departure the above mentioned lexicographic resources that focus on or include MWEs, below we summarise the levels of description we consider relevant for our work: lexical, derivational, morphological, syntactic,

[6] *Fossil* words are those that only occur in MWEs; they are also known as *cranberry* words.

semantic, contextual, stylistic.[7] A detailed description of the complex multilevel representation of a broad range of MWEs and MWE types in Czech (another morphologically rich language), which shares many commonalities with the approach adopted herein is presented in Skoumalová et al. (2024 [this volume]). A different, though not contradicting approach to a rich multilayered description for Bulgarian MWEs is adopted in Osenova & Simov (2024 [this volume]). We defer the discussion as to which levels of description are implemented (and how) in the proposed Bulgarian-Romanian VMWE lexicon to §4, where we also provide an explanation for favouring a particular decision or approach over another.

2.1 Lexical level

The lexical level contains information about:

- the list of lexemes that can substitute components in the multiword expressions (Villavicencio et al. 2004b, Grégoire 2007, Przepiórkowski et al. 2014a, Hnátková et al. 2019, Markantonatou et al. 2019, 2020, Skoumalová et al. 2024). The variations may be handled uniformly regardless of the status of the component affected (i.e., as alternative realisation within the same citation form) or differently, according to certain criteria, e.g., whether the verbal head or an invariable component is concerned, cf. the treatment by Markantonatou et al. (2019, 2020);
- cross-references from the dictionary entries of each of the components of the MWEs (except for function words) to the entry/ies of the MWEs in which they occur (Villavicencio et al. 2004b, Mel'čuk 2006).[8]

2.2 Derivational information

Expressions that are derivationally related to the MWEs, e.g., nominal expressions derived from VMWEs (Mel'čuk 2006, Hnátková et al. 2019, Monti 2014), are recorded in the dictionary, thus providing links to other parts of the language's lexicon, including MWEs and one-word compounds.

[7] For a discussion of lexical encoding formats for MWEs that can be used in NLP systems, see Lichte et al. (2019).

[8] In Mel'čuk (2006) it is not clear if all lexical entries of a MWE component contain references to the respective MWE or only that which reflects the meaning it has in the MWE, although the author admits the semantic non-compositionality of some idioms.

2.3 Morphological description

The following information pertain to this level:

- lemma (i.e., canonical) form of all the components (Dyvik et al. 2019, Grégoire 2007, Odijk 2013, Odijk et al. 2024, Osenova & Simov 2024, Skoumalová et al. 2024);
- restrictions on the inflection of components that can help automatically generate all the possible forms of the MWE (Grégoire 2007, Al-Haj et al. 2013, Markantonatou et al. 2019, 2020, 2024, Osenova & Simov 2024, Skoumalová et al. 2024).

2.4 Syntactic level

This level contains the following information:

- syntactic category of the expression (e.g., nominal, verbal, adjectival, etc.) (Shudo et al. 2011, Al-Haj et al. 2013, Dyvik et al. 2019, Markantonatou et al. 2019), sometimes referred to indirectly, by means of reference to the class to which the MWE belongs (Grégoire 2007, Odijk 2013);
- internal syntactic structure of the expression (Dyvik et al. 2019, Grégoire 2007, Hnátková et al. 2019, Markantonatou et al. 2019, 2020, Shudo et al. 2011, Przepiórkowski et al. 2014a, Villavicencio et al. 2004b, Mel'čuk 2006) represented in terms of one of various theoretical frameworks: dependency structures (Hnátková et al. 2019, Odijk 2013, Villavicencio et al. 2004b, Markantonatou et al. 2024, Osenova & Simov 2024, Skoumalová et al. 2024), Lexicon-Grammar (Gross 1982), HPSG (Villavicencio et al. 2004b), LFG (Dyvik et al. 2019), constituent structures (Skoumalová et al. 2024 [this volume]), among others;
- possible modifiers of components (Fellbaum & Geyken 2005, Markantonatou et al. 2019, 2020, Grégoire 2007, Shudo et al. 2011, Al-Haj et al. 2013, Markantonatou et al. 2024, Osenova & Simov 2024, Skoumalová et al. 2024);
- clear indication of the optional and obligatory components (Fellbaum & Geyken 2005, Markantonatou et al. 2019, 2020, Villavicencio et al. 2004b, Markantonatou et al. 2024, Skoumalová et al. 2024);
- word order of the components with respect to each other (Al-Haj et al. 2013, Markantonatou et al. 2019, 2020, 2024) or marking of specific or anomalous

word order Markantonatou et al. (2024 [this volume]), Skoumalová et al. (2024 [this volume]), see also the approach adopted below;

- valency information about the MWE which determines its realisation in text (Giouli et al. 2024 [this volume]), (Osenova & Simov 2024 [this volume]), (Skoumalová et al. 2024 [this volume]);
- combinatorial possibilities of the expression extracted from corpora, such as possible subjects, complements, pre- or post-modifiers, etc. (Odijk 2013, Mel'čuk 2006), sometimes with their frequency (Odijk 2013, Odijk et al. 2024);
- other syntactic variations such as passivisation, causative-inchoative alternations, long-distance dependencies, alternative forms of the MWE (Dyvik et al. 2019, Fellbaum & Geyken 2005, Markantonatou et al. 2019, 2020, Vietri 2014a, Markantonatou et al. 2024, Skoumalová et al. 2024), but only when they violate the rules of the grammar (Mel'čuk 2006).

2.5 Semantic description

The information contained at this level consists of:

- a paraphrase, a definition or an explanation of the meaning of the MWEs (Villavicencio et al. 2004b, Markantonatou et al. 2019, 2020, Mel'čuk 2006, Osenova & Simov 2024, Markantonatou et al. 2024, Skoumalová et al. 2024);
- relations to other idioms, such as synonymy (Autelli 2020, Osenova & Simov 2024, Markantonatou et al. 2024), antonymy (Fellbaum & Geyken 2005, Markantonatou et al. 2019, 2020, 2024), hypernymy and hyponymy (Fellbaum & Geyken 2005), as well as other relations that serve to define a network of VMWEs expressing a concept (Markantonatou et al. 2019, 2020): causative-inchoative or stative relations, verb alternations, lexical variants, etc., making it possible to group MWEs in synonym sets (Markantonatou et al. 2019, 2020, 2024);
- semantic domain (Fellbaum & Geyken 2005, Monti 2014), by means of cross-references to other entries in the dictionary having the same or related meaning (Mel'čuk 2006).

2.6 Contextual information

This level contains information such as:

- examples of sentences (extracted from corpora) containing the respective MWE (Grégoire 2007, Markantonatou et al. 2019, 2020, Odijk 2013, Autelli 2020, Osenova & Simov 2024, Markantonatou et al. 2024, Skoumalová et al. 2024). When the MWEs in the lexicon originate from corpora, the information extracted from the corpus (such as context of occurrence, frequency, etc.) is kept track of by a reference from the lexicon entry to the file storing the respective information (Grégoire 2007);
- contextual restrictions on the occurrences of MWEs, such as co-occurrence with specific syntactic phrases (Shudo et al. 2011) or with semantically specific adverbs or other external modifiers (Fellbaum & Geyken 2005);
- frequency of occurrence of MWEs in corpora (Odijk 2013).

2.7 Stylistic information

The label "stylistic" encompasses all kinds of information about the style or language register in which a MWE is typically used, such as "ironic", "disparaging", "humorous" (Fellbaum & Geyken 2005); "formal", "colloquial", "offensive" (Markantonatou et al. 2019, 2020); "vulgar", "negative connotation", "disused" (Autelli 2020), or other similar descriptions (Skoumalová et al. 2024 [this volume]).

2.8 Other information

Besides the linguistic types of information already mentioned, some lexicons also include the following:

- diachronic information: changes in the form and meaning of the VMWEs over time (Fellbaum & Geyken 2005);
- translation into other languages such as English (Al-Haj et al. 2013, Markantonatou et al. 2019, 2020) and French (Markantonatou et al. 2019, 2020, 2024);
- the emphatic function of MWEs (Fotopoulou et al. 2014, Markantonatou et al. 2019, 2020).

The overview of the types of linguistic information encoded about MWEs shows that the lexicons referenced above contain relevant descriptions and partially overlapping types of information, distributed over several linguistic levels. One of our aims when developing the linked Bulgarian-Romanian bilingual lexicon of MWEs was to provide a consistent and uniform framework for the representation of MWEs that would take into account the various levels of linguistic description and the approaches to tackle them in line with the findings of the theoretical analysis as well as the specific requirements of the bilingual (and by extension – multilingual) representation of data.

3 Compilation of a Bulgarian-Romanian MWE lexicon

We first describe the lexical resources that the lexicon is derived from, i.e., the Bulgarian and Romanian wordnets. We then present the different levels of linguistic description in comparison with other frameworks and initiatives.

3.1 BulNet and RoWN: Sources of MWEs for the lexicon

A wordnet is a semantic network: its nodes are represented by synonym sets (synsets), which contain one or more linguistic items (called "literals") that lexicalise a concept; literals may be single words or multiword combinations alike.[9] The edges connecting the nodes are semantic relations that hold between a pair of synsets. Only words belonging to content parts of speech are usually represented in such language resources: nouns and verbs have a hierarchical organisation, descriptive adjectives are organised in clusters created around a pair of antonymic adjectives, relational adjectives and adverbs have no organisation. The first such network, Princeton WordNet (WordNet, Miller 1995), was developed for English; wordnets for other languages have been subsequently developed,[10] most of which are aligned with WordNet, i.e. the synsets in different wordnets with equivalent meanings are mapped to each other.

The development of the Bulgarian Wordnet (BulNet, Koeva 2010) and the Romanian Wordnet (RoWN, Tufiș & Barbu Mititelu 2014) started in the BalkaNet project (Tufiș et al. 2004), which had as one of its objectives the implementation of a set of synsets common to all languages in the project. The construction of the two wordnets adopted the "expand" approach, which involves translation of the

[9] For a discussion on the representation of figurative language, proverbs and idioms in WordNet, see Fellbaum (1998a).

[10] For a list of existing wordnets in the world, see http://globalwordnet.org/resources/wordnets-in-the-world/.

literals in the English synsets and automatic transfer (and possibly revision) of the semantic relations from WordNet (Fellbaum 1998b) to BulNet and RoWN. The content of the synsets and associated information (literals, gloss, usage examples, stylistic notes, etc.) were devised by native language experts, who consulted relevant monolingual and bilingual dictionaries. These decisions and work methods led to the creation of wordnets aligned to WordNet and thereby to each other (via WordNet),[11] on the other. Figure 1 shows the interlinking among the wordnets, in which the English, Romanian and Bulgarian synsets contain verbal idioms: (bg) *давам най-доброто от себе си davam nay-dobroto ot sebe si* (lit. 'give the best of oneself'), *давам всичко от себе си davam vsichko ot sebe si* (lit. 'give all of oneself') – (ro) *da totul* (lit. 'give all'), *da ce e mai bun* (lit. 'give the best'), *da tot ce e mai bun* (lit. 'give all the best').

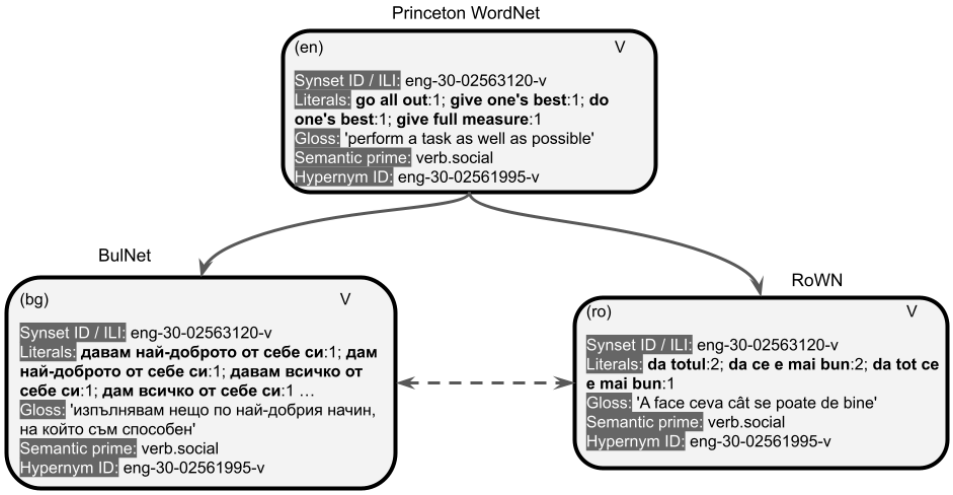

Figure 1: Interlinking wordnets.

After the end of BalkaNet, each team continued the development of the respective wordnet independently, with different interests in the conceptual coverage of their resources. The development of the wordnets for Bulgarian and Romanian (as well as for any other language constructing a wordnet using the expand approach) is naturally biased towards English, as WordNet provided the original inventory of senses. While this fact was acknowledged, it was not considered a serious concern, as no resource could be absolutely unbiased, on the one hand, and because of the fact that concepts are shared by different languages, which made the alignment among wordnets possible, on the other. MWEs were not

[11]They are also aligned to any wordnet that is aligned to WordNet.

a particular focus of the development of BulNet and RoWN; however, as they are treated on a par with single words, MWEs were included whenever relevant for a synset. The current versions of the two wordnets do not cover the lexical inventory of the languages thoroughly.[12]

3.2 Dataset construction

The features of the bilingual resource outlined in the following sections were described on the basis of linguistic analysis aiming at delineating the common linguistic characteristics and the differences between the two languages that need to be taken into account in such a lexicon. This analysis is based on 3,656 multitoken literal-to-literal pairs in corresponding synsets in BulNet and RoWN. These include VMWEs proper, as well as multitoken free phrases with purely compositional meaning. We filtered out the latter and were left with 2,705 VMWE-to-VMWE pairs. As the VMWEs under discussion are part of pairs of corresponding aligned synsets, they are treated as possible translation equivalents to each other, cf. the synset counterparts in (1), and are included in the constructed bilingual resource. As part of the VMWE bilingual lexicon, each VMWE is analysed and described on the morphological, syntactic, semantic, stylistic, connotational and derivational level individually. The linguistic information which is common to all the members of a synset, e.g. the gloss, is also assigned to each VMWE in the relevant synset, as each VMWE is a separate unit in the VWME lexicon. In addition, all the VMWEs belonging to the same synset share the same synset ID and are thus identifiable as part of the synset. We did not implement any further linking beyond the alignment at the synset level, which was performed while the individual wordnets were being constructed.

The verbal multiword literals in BulNet and RoWN were manually annotated with the VMWE types from the PARSEME 1.2 guidelines:[13] verbal idioms (VID), light verb constructions whose verb is semantically totally bleached (LVC.full), light verb constructions in which the verb adds a causative meaning to the noun (LVC.cause), inherently reflexive verbs (IRV), for both languages, while the category inherently adpositional verbs (IAV) was annotated only for Bulgarian (Barbu Mititelu et al. 2019b).

The compilation of the lexicon started with those synsets that are lexicalised by VMWEs of the same type in both wordnets: 192 VID examples, 44 LVC ones and 2,023 IRVs. IRVs are also of interest for comparative studies, but will be part

[12]We used Princeton WordNet – 3.0 aligned with Bulgarian and Romanian wordnets. BulNet consists of 85,954 synsets created by expert linguists (Koeva 2021), while RoWN contains 59,348 synsets (Tufiș & Barbu Mititelu 2014).

[13]https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/

of future work. Thus, the set of VMWEs currently included into the lexicon and subject to description is made up of 2,259 pairs of corresponding VMWEs.

The description of VMWE literals was performed independently for each of the two languages according to a common set of features and their possible values. IRVs have regular structure, word order and syntactic properties, so our work is focused only on VID and LVC cases, which pose a number of challenges for their description and the analysis of their properties.

As a result, we obtain a new resource, a self-contained bilingual MWE lexicon where each VMWE in each of the languages is described individually, but each VMWE is described by filling in the relevant fields in the predefined template of a language-independent lexicon entry. In the following section we delve into the types of information included in each dictionary entry and how these are handled in practical terms.

4 The content of a lexicon entry

Following one of the dominant trends in MWE lexicon crafting, we adopt the approach of encoding VMWEs explicitly as distinct entries instead of describing the rules of combining their components. This makes it possible to reflect and access in a straightforward way the morphosyntactic, syntactic, semantic and derivational information associated with a particular entity that may not be readily obtainable from the combination of its components. In (1), we illustrate three aligned synsets in WordNet, BulNet and RoWN.[14] We notice that in the same synset there may be MWEs based on a different support verb (as in (2a) for Bulgarian) or a different semantic head (as in (2b) for Romanian).[15]

(1) a. form:8; take form:1; take shape:1; spring:6 (en)
Synset ID: eng-30-02623906-v
Definition: 'develop into a distinctive entity'

b. образувам се:1, оформям се:2, оформя се:2, формирам се:1, obrazuvam se:1, oformyam se:2, oformya se:2, formiram se:1, приемам форма:1, приема форма:1, добивам форма:1, добия форма:1, кристализирам:1 (bg)
priemam forma:1, priema forma:1, dobivam forma:1, dobiya forma:1, kristaliziram:1

c. se forma:1; se contura:1; prinde contur:1; prinde formă:1 (ro)

[14]The synset ID and definition are rendered only for WordNet.

[15]For brevity, we do not give literal translations where they are similar or identical to the idiomatic translation.

(2) a. приемам форма, добивам форма (bg)
priemam forma, dobivam forma
adopt shape, obtain shape

b. prinde contur, prinde formă (ro)
catch outline, catch shape

While it is obvious that some literals are closer correspondences to each other in terms of structure and/or semantics – e.g. (bg) *formiram se* – (ro) se forma ('form') and (bg) *priemam forma* – (ro) *prinde formă* – (en) *take form* – we do not attempt to connect to each other such stricter correspondences found within the same pairs of synsets; instead, we take all literals on one side to be relevant translation equivalents of the literals on the other side, as translation choices may be guided by factors other than structural or semantic similarity.

In the following subsections we present the levels of description of VMWEs adopted in the resource presented. Given one of the organisation principles of WordNet, i.e., each synset stands for a concept and each word/expression can occur a number of times equal to its number of senses, it is clear that all information provided for a MWE pertains to one of its senses, in case it is polysemous.

4.1 Technical information

This level of description serves two main purposes: the unique identification of the VMWE lexicon entry within the dataset for one language, as well as pairing the VMWE entries across languages. For this, we employ wordnet indexing with additional identification elements which serve both to identify a VMWE as part of a particular synset (via synset ID) and to distinguish it from other VMWEs in the same synsets, or from identical VMWE literals in other synsets (via literal IDs, see (3)). For Bulgarian we also include a verb aspect identifier, which allows us to refer jointly or separately to aspectual pairs lexicalising the VMWEs – this is useful when comparing languages that differ with respect to the verb aspect or where the aspectual systems are organised differently.[16] The identification system allows us to: (i) access all the synset-level linguistic information provided; (ii) make references to a particular VMWE uniquely, e.g., in the description of derivatives (e.g., (3b) as derived from (3a) and not from its aspectual counterpart *snema otpechatatsi*, literal ID: bg_2330, nor its synonym *vzemam otpechatatsi*, literal ID: bg_2327); (iii) extract translation equivalents of VMWEs from wordnets

[16]This feature is only relevant for Bulgarian. Romanian lacks a lexico-grammatical verb aspect (i.e. marked on separate lexemes) and aspectual distinctions are expressed by other means.

for different languages; (iv) use the rich relational structure of WordNet for the purposes of the semantic description of VMWEs.

(3) a. снемам отпечатъци (bg)
snemam otpechatatsi
take fingerprints
Synset ID: eng-30-01748748-v, Literal ID: bg_2329, Aspect: IPFV

b. снемане на отпечатъци (bg)
snemane na otpechatatsi
taking of figerprints (the act of fingerprinting)
Synset ID: eng-30-00152338-n

4.2 Morphological description

4.2.1 Lemma of the VMWE

Savary (2008) considers two main approaches to lemma representation that have become dominant: (i) an abstract lemma, where a citation form that generates all the possible forms of the relevant single word is assigned to each component; (ii) a non-abstract lemma in which each of the components is represented by the form that is part of the relevant MWE, and the MWE lemma is associated with a formalised description of the grammatically possible combinations of forms of the MWE components, thus avoiding overgeneration. Even though the latter approach is linguistically more justified and was adopted by other authors (see §2.3 above), the former allows recognition and retrieval of MWEs from corpora where MWEs are not annotated, thus possibly being capable of recognising MWEs not included in lexicons. Still others (Fellbaum & Geyken 2005) determine MWE lemmas on the basis of the frequency of occurrence, maintaining multiple citation forms where two or more dominant forms are relatively equally distributed. Such an approach accounts for the fact that the non-abstract MWE lemmas are often not morphologically unmarked and that they may occur preferentially in particular forms but not in others.

We adopt a two-way approach by assigning each MWE both a non-abstract lemma and an abstract one. The function of the former is to represent the most neutral form in which the components occur in the language. It is this lemma that we consider in determining the inflection of the MWE components that reflects the actual morphological restrictions imposed on the forms that need to be described in the fields dedicated to morphosyntactic restrictions. Consider the following examples of non-abstract lemmas:

(4) затварям си очите (bg)
zatvaryam si ochite
close self.CL eye.PL.DEF
lit. 'close one's eyes'
'turn a blind eye'

(5) închide ochii (ro)
close eye.PL.DEF
lit. 'close the eyes'
'turn a blind eye'

In examples (4) and (5), the verbal head's inflection is unrestricted, whereas the nominal complement is only found in its plural definite form. In addition, in Bulgarian the reflexive possessive pronoun is in its short (clitic) form (which is invariable). The description of the relevant restrictions in the dictionary prevents the overgeneration of non-existing forms.

For the automatic recognition of MWEs, we also encode an abstract lemma for each MWE (see examples (6) and (7) corresponding to the VMWEs in (4) and (5), respectively); this means representing the nominal complement in its citation form, i.e., singular indefinite for both languages, respectively, and, in Bulgarian, representing the reflexive possessive in its base form, i.e. masculine singular indefinite, which changes in terms of the number, gender and definiteness of the possessed entity.[17]

(6) затварям свой око (bg)
zatvaryam svoy oko
close self.REFL.POSS.M.SG eye.SG.INDEF
lit. 'close one's eye'
'turn a blind eye'

(7) închide ochi (ro)
close eye.SG.INDEF
lit. 'close eye'
'turn a blind eye'

The abstract lemma is invoked when a sequence of words corresponding to a MWE in the lexicon is recognised as such in a lemmatised corpus (i.e. where,

[17]The long form of the reflexive possessive pronoun does not denote person, and the categories it inflects for (gender, number, definiteness) are features not of the possessor but of the entity possessed.

most often, lemmas are assigned to single words); it is itself a sequence of forms that will not be found in the language in an idiomatic meaning or is completely impossible, as the abstract lemma in example (6) above: *zatvaryam svoy oko*.

The abstract lemma thus matches the lemmas assigned in the corpus and allows for each occurrence of the relevant MWE in the corpus to be associated with the dictionary entry and the information it contains.

The components of the MWE are numbered and identified with respect to their position in the lemma and the abstract lemma. In this way the morphological features, the restrictions on a component's paradigm, as well as the blocking of modifiers and external elements between particular components can be precisely defined.

4.3 Syntactic description

The syntactic variability of VMWEs is much greater than expected despite the traditional understanding about the relatively fixed nature of the structure and linearity of VMWEs. In particular, many (V)MWEs exhibit the regular syntactic behavior of free phrases, including the possibility of intervening external elements that modify a particular element of the VMWE or the entire expression/sentence, various semantic-syntactic transformations, alternative complement expression, long-distance dependencies, etc. That is why we chose to describe only the deviations from the regular syntactic behavior of the MWEs.

The syntactic description of the VMWEs in the lexicon is based on the Universal Dependencies[18] (UD) framework (de Marneffe et al. 2021). The choice for this framework was natural, in order to ensure a consistent treatment of the VMWEs in the wordnets and in the Bulgarian (Savary et al. 2018) and Romanian (Barbu Mititelu et al. 2019a) corpora created (alongside those for other languages) within PARSEME, and annotated with the same types of VMWEs. These corpora were automatically syntactically annotated using UDPipe (Straka 2018), with the syntactic relations defined in UD (Savary et al. 2023).

There are several types of syntactic information recorded in our resource: the internal structure of VMWEs, their valence frames, word order restrictions on their components and the possibility of other words to occur within the expressions. They are all discussed in what follows.

[18]https://universaldependencies.org/

4.3.1 Internal syntactic structure

The syntactic annotation of the VMWEs in the two wordnets with UD relations was done manually, with the aim of describing the number of components within each VMWE and the syntactic relations between them. The representation of the VMWE structure follows this convention: the head of the expression (i.e., the verb) followed by the UD relations that the other components of the VMWEs establish with the head or with other components. In the description of the internal structure of VMWEs, the order of these relations reflects the linear order of the components in the expression. For example, the internal structure of the VMWE (en) *kick the bucket* is `V + [det + obj]`. The square brackets indicate that the determiner (`det`) and the direct object (`obj`) are not both attached to the verb, but only the `obj`, whereas the other depends on it.

Table 1 shows only some of the most frequent syntactic structures that have correspondences in the analysed VMWEs in Bulgarian and Romanian, but variants of these structures are omitted. For example, patterns such as `V + obj` and `V + case + obl` can have as variants `V + obj + amod` and `V + [case + obl + amod]` in Romanian and `V + [amod + obj]` and `V + [case + amod + obl]` in Bulgarian, `V + [nummod + obj]` in both languages, where the word order variations arise from the structural differences in the two languages, i.e., in Romanian modifiers usually follow the nominal head, whereas in Bulgarian they precede it.[19]

We did include several parallel patterns. They are given a somewhat different analysis – i.e., we construe the possessive clitic in their structure as `expl:poss` in Romanian and as `det` in Bulgarian. But, in fact, they correspond to each other and translate in the same way. Such an example is illustrated by the pattern `V + expl:poss/det + obj`. Leaving the linguistic discussion aside, we treat them as equivalent, thus aiming at pointing out the essential commonalities instead of the less important differences.

When correlating the PARSEME VMWEs types with their valence frames we notice the following. According to PARSEME guidelines, a characteristic of LVCs is the fact that they are made up of a verb and a noun, the latter determining the semantics of the expression. In Romanian and Bulgarian most expressions of the type LVC.full have the internal structure `V + obj`, consider the chess term (ro) *da șah* and its counterpart (bg) *davam shah*, both literally meaning 'give check' and translated as 'place into check', or `V + [case + obl]` – (ro) *lua în serios* (lit. 'take in serious') 'treat seriously' and (bg) *stigam do sporazumenie* (lit. 'reach to an agreement') 'come to an agreement'. The instances of Romanian LVC.cause

[19]Where such syntactic patterns are presented, we stick to a uniform way of encoding them, e.g., `V + [obj + amod]`, disregarding the differences between the two languages.

Table 1: Frequent syntactic structures within VMWEs in Bulgarian and Romanian.

Syntactic pattern	Romanian example	Bulgarian example
`V+obj`	*avea grijă* lit. 'have care' 'take care'	*imam grizha* lit. 'have care' 'take care'
`V+ [case+obl]`	*citi printre rânduri* lit. 'read among lines' 'read between the lines'	*cheta mezhdu redovete* lit. 'read between the lines' 'read between the lines'
`V+ expl:poss/det +obj`	*își ține gura* lit. 'keep one's mouth' 'shut one's mouth'	*zatvaryam si ustata* lit. 'close one's mouth' 'shut one's mouth'
`V+obj+ [case+obl]`	*arunca praf în ochi* lit. 'throw dust in eyes' 'throw dust in the eyes'	*hvarlyam prah v ochite* lit. 'throw dust in eyes' 'throw dust in the eyes'
`V+ expl:poss/det + [case+obl]`	*își ieși din fire* lit. 'escape from one's temper' 'flip one's lid'	*plyuya si na petite* lit. 'spit on one's heels' 'head for the hills'
`V+ expl:poss/det +obj+advmod`	*își lua cuvintele înapoi* lit. 'take back one's words' 'take back one's words'	*vzemam si dumite nazad* lit. 'take back one's words' 'take back one's words'
`V+nsubj+ [case+advmod]`	*lua gura pe dinainte* lit. 'the mouth takes on ahead' 'let the cat out of the bag'	–
`V+advmod+ det+nsubj`	–	*mnogo mi znae ustata* lit. 'my mouth knows a lot' 'have a big mouth'

display two types of internal structures, i.e., `V + xcomp` and `V + [case + obl]`. The same structures are also found in Bulgarian, compare (ro) *face public* 'make public' and (bg) *pravya raven* 'make equal', as well as (ro) *pune în circulație* and (bg) *puskam v obrashtenie* 'put into circulation'. In Bulgarian we also attested LVC.cause with the structure `V + obj`, e.g., (bg) *pravya upoyka* (lit. 'administer anesthesia') 'put under, anesthesise'.

The internal structure of VIDs is more diverse, though, given that they can even be/contain clauses: e.g., (ro) *bate fierul cât e cald* 'strike the iron while it is hot'. The syntactic structures attested in the data are based primarily on a verb-complement or verb-modifier pattern, while the subject and another complement or modifier are part of the VMWE's valence frame. This fact is reflected in Table 1, which shows that only a few examples including a subject are found in the data, cf. the last two rows – (ro) *lua gura pe dinainte* and (bg) *mnogo mi znae ustata*, where the nouns *gura* and *ustata*, respectively, are the subject of the verb.[20]

Besides the patterns in Table 1, the data also contains a number of structures that are less represented in the bilingual lexicon due to its size. In fact, many of them are variations of the ones described in the table, e.g., `V + [case + advmod]` (bg) *izlizam na otkrito* (lit. 'come out in open') 'come to light' is a variant of `V + advmod`; the patterns involving an expletive reflexive (expl:pv), such as `V + expl:pv + [case + obl]` (bg) *makna se po petite* (lit. 'drag oneself on someone's heels') 'tag along', are variations of the respective models based on the pattern `V + obj`, as the expletive blocks the direct object (reflexive verbs are intransitive).

4.3.2 Morphosyntactic description

The morphosyntactic description deals with the morphological properties of the head and the dependent components of the VMWEs and the ways in which each of the components varies morphologically as part of the expression. The way morphological variation is treated depends on the extent of variation, the way the MWE lemma is defined, etc. (see §2). Regarding its variability, each component may be unrestricted (i.e., the MWE component displays the full simple word paradigm), restricted (the MWE component's forms vary grammatically, but it is restricted with respect to one or more grammatical categories) or fixed (the MWE component does not vary morphologically).

We adopt the practice that lack of any morphosyntactic restrictions is the default value for each component and hence not marked, whereas restrictions or invariability are explicitly defined in the respective field of the MWE entry. For instance, in the following equivalent examples – (ro) *pune pe fugă* (lit. 'put on

[20]The empty cells show that the pattern is not attested in the data, though may well be possible in the language.

run'), (bg) *obrashtam v byagstvo* (lit. 'turn into flight') 'rout out, oust, cause to flee' – the MWEs consist of a verbal head and an oblique expressed by a noun introduced by a preposition (`V + [case + obl]`). The verb may be found in any form and is thus unrestricted, prepositions in both languages are invariable, while the noun is only found in its singular indefinite form.

In the analysed data, most often the verbal head's paradigm is unrestricted, with just a few exceptions, e.g., (ro) *lua gura pe dinainte* (lit. 'the mouth takes on ahead') 'let the cat out of the bag'. Examples of such exceptions are the cases where: (i) the nominal subject is part of the MWE and therefore the verbal head agrees with it; or (ii) the subject's referent cannot be a participant in the communication; or (iii) the verb is otherwise restricted as in weather expressions, where it can only be in the third person singular, e.g., (ro) *ploua cu găleata* (lit. 'rains with bucket') and (bg) *vali kato iz vedro* (lit. 'rains as if out of a bucket') 'rain buckets'.

We note that the most frequent restriction found in both languages is the singular indefinite form of the nominal dependent, followed by the singular definite form, etc. (Table 2). These restrictions are found across the most well-represented syntactic patterns – `V + obj` and `V + [case + obl]` as well as in more complex variations of these structures, e.g., `V + [case + amod + obl]`. – (bg) *dokarvam do proseshka toyaga* (lit. 'bring to a beggar's stick') 'beggar, pauperise'. Another frequent variant in patterns with definite nominal dependents features an expletive possessive `V + expl:poss + obj` – (ro) *își rupe spatele* 'break one's back' or reflexive possessive clitic `V + det + obj` – (bg) *iztarvavam si nervite* (lit. 'drop one's nerves') 'lose one's temper'. In both languages the possessive clitic occurs only with definite nouns or noun phrases.

Another relatively frequent pattern, as shown in Table 2, is the one containing an object that is restricted to the singular (definite and indefinite) forms: see the examples (ro) *avea încredere* 'have trust' – (bg) *imam vyara* 'have faith'. A plural object (e.g., (ro) *închide ochii*) is more rarely found in the Romanian data as compared with the singular, although in Bulgarian the patterns with plural definite complements are quite well represented: see examples (bg) *darpam kontsite* and (bg) *hodya po nervite*.

In Bulgarian, unrestricted objects/modifiers are also represented to a certain extent. Examples such as (bg) *iznasyam lektsia* and (bg) *vzemam prisartse* show patterns with a nominal complement unrestricted for number and definiteness, or an adverbial modifier, that is unrestricted for the category of degree (comparative, superlative), which is possible for some MWEs. In Romanian, such examples could not be found in the dataset.

Table 2: The most frequent morphosyntactic restrictions on dependents found with VMWEs in Bulgarian and/or Romanian (literal translation is provided only when it differs from the English equivalent).

Restrictions	Romanian example	Bulgarian example
Number = sg Def = indf `V + obj`	*lua parte* 'take part'	*vzemam uchastie* 'take part'
Number = sg Def = indf `V + [case + obl]`	*pune pe fugă* lit. 'put on running' 'oust, cause to flee'	*obrashtam v byagstvo* lit. 'turn into flight' 'oust, cause to flee'
Number = sg Def = def `V + obj`	*atrage atenția* lit. 'attract attention' 'call attention'	*nasochvam vnimanieto* lit. 'direct attention' 'call attention'
Number = sg Def = def `V + [case + obl]`	*sta la baza* lit. 'stand in the base' 'underlie'	*lezha v osnovata* lit. 'lie in the base' 'underlie'
Number = sg `V + obj`	*avea încredere* lit. 'have trust' 'trust'	*imam vyara* lit. 'have faith' 'trust'
Number = pl Def = def `V + obj`	*închide ochii* lit. 'close eyes' 'turn a blind eye'	*darpam kontsite* 'pull strings'
Number = pl Def = def `V + [case + obl]`	*fi cu ochii* lit. 'be with eyes' 'keep an eye on'	*hodya po nervite* lit. 'walk on the nerves' 'madden'
Unrestricted `V + obj`	–	*iznasyam lektsia* lit. 'present a lecture' 'lecture'
Unrestricted `V + advmod`	–	*vzemam prisartse* 'take to heart'

Def = def
`V+`
`expl:poss/det+obj`
	își rupe spatele
	'break one's back'
		prosya si belyata
		lit. 'beg for my own trouble'
		'ask for trouble'

4.3.3 Valence frames

Another important aspect of the syntactic description of VMWEs is represented by their valence frames, which we encode by the use of the following conventions. First, they are formulated as UD relations: for each MWE, we define the types of relations it establishes within a sentence to ensure its grammatical correctness. For example, the MWE *kick the bucket* has a valence frame containing only the subject, i.e., `nsubj`.

The valence frames can contain obligatory, as well as optional relations. The difference between them is that the latter can be absent from the sentence without affecting its grammatical correctness: consider the sentence in (8):

(8) Regizorul i -a **dus** **de nas** pe spectatori cu un scurtmetraj. (ro)
Director them has taken of nose on audience with a short-film
lit. 'The director lead the audience by the nose with a short film.'
'The director pulled the wool over the audience's eyes with a short film.'

The subject *Regizorul* and the object *spectatori* are obligatory relations, but the prepositional object *cu un scurtmetraj* is optional. The optional nature of a relation is marked by means of round brackets around it; thus the valence frame for the VMWE in (8) is: `nsubj, obj, (case{cu}, obl)`.

Third, lexical restrictions on the form of prepositions or markers are rendered between curly brackets immediately after the relevant relation, `case` and `mark`, respectively: e.g., `case{cu}` in the frame presented for example (8).

Fourth, alternative valences are separated by a slash. For example, if two different prepositions occur after a VMWE, they are listed as values of the respective relation in the manner described: `case{împotriva/asupra}`.

Fifth, whenever an alternative consists of at least two elements (e.g., relations, forms, etc.), they are grouped together within square brackets: for example, (ro) *da drumul* (lit. 'give way') 'let go' can take either a prepositional object with the preposition *la* or an indirect object; this is represented as follows: `[case{la}, obl]/iobj`.

Table 3 shows the most frequent valence frames characterising VMWEs in the Bulgarian and Romanian datasets. Most of the encoded valences describe personal verb constructions, thus they require a subject in the frame, unless it is part of the expression, which happens rarely, as mentioned above.

When correlating the PARSEME VMWE types with their valence frames, we notice the following. Besides the subject, the valence frames of all expressions of the type LVC.cause have an obligatory object. This is in line with the definition of this type in PARSEME, according to which the noun in the LVC.cause "has

Table 3: The most frequent valence frames in the two languages.

Valence frame	Romanian example	Bulgarian example
`nsubj`	*o lua la goană* lit. 'her take at rush' 'break away'	*hvashtam gorata* lit. 'take the wood' 'take to the woods'
`nsubj, obj`	*aduce în sapă de lemn* lit. 'bring in hoe of wood' 'pauperise'	*dokarvam do prosiya* lit. 'bring to beggary' 'pauperise'
`nsubj, iobj`	*da frâu liber* lit. 'give rein free' 'unleash'	*davam volya* lit. 'give freedom' 'unleash'
`nsubj, case, obl`	*da piept* lit. 'give breast' 'confront'	*varvya v krak* lit. 'walk in step' 'keep pace'

`nsubj,`
`[case, obl] /`
`[mark, ccomp]`
	da seamă
	lit. 'give count'
	'be responsible for'
		namiram sili
		lit. 'find strength'
		'take heart'

semantic arguments expressed as non-subject elements in the sentence".[21] E.g., (ro) *pune în circulație* (lit. 'put in circulation') 'issue' has the internal structure `V + [case + obl]` and the valence frame `nsubj, obj`, where the `obl` has the `obj` as a semantic argument – see the example: *Banca* ***pune*** *banii* ***în circulație*** (lit. 'Bank puts money in circulation') 'The bank issues money', in which *money* is the semantic argument of *circulație*.

The valence frames of VMWEs of the types LVC.full and VID may contain only the subject or the subject and a nominal (`obj`, `iobj` or `obl`) or a clause: here are some examples: (a) VID (ro) *prinde inimă* (lit. 'catch heart') 'cheer up' takes only a subject: *Copilul a* ***prins inimă*** 'The child cheered up'; (b) VID (ro) *purta sâmbetele* (lit. 'bear Saturdays') 'bear ill will' takes a subject and an indirect object: *Bărbatul îi* ***purta sâmbetele*** *soacrei sale* 'The man was bearing his mother-in-law ill will'; (c) VID (ro) *cădea de acord* (lit. 'fall of agreement') 'reach agreement' takes a subject, an oblique indicating the person with whom agreement is achieved, and a subordinate clause or a prepositional phrase indicating the matter which was the subject of discussion: *Avocatul a* ***căzut de acord*** *cu*

[21] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=050_Cross-lingual_tests/020_Light-verb_constructions__LB_LVC_RB_

clientul [asupra onorariului]/[cât să îl plătească] 'The lawyer has reached agreement with his client [on the fee]/[how much to pay him'; (d) LVC.full (ro) *avea încredere* (lit. 'have trust') 'trust' takes a subject, an oblique denoting the person who the subject trusts, and a subordinate clause indicating with respect to what the subject trusts the other person: *Bărbatul* ***are încredere*** *în avocat că va câștiga procesul* 'The man trusts the lawyer that he will win the trial'.

In addition to these, both languages display valence patterns where one of the elements is an obligatory `nmod` or complement that usually enters the relation `obj` or `obl` with the verb, e.g., (ro) *sta la baza* (lit. 'stand at the base') and (bg) *lezha v osnovata* (lit. 'lie in the base') where the obliques (ro) *baza* and (bg) *osnovata* need a nominal modifier to form a grammatical sentence. These may also be possessive phrases, e.g., (bg) *hodya po nervite* + `nmod`: *na nyakogo* (lit. 'walk on the nerves + `nmod`: of someone') 'madden'.

Empty valence frames are also possible where the VMWEs are headed by impersonal verbs and they do not have obligatory complements or modifiers. In Romanian, this is the case of weather expressions, such as (ro) *plouă cu găleata* (lit. 'rains with bucket') 'it is raining cats and dogs'. The corresponding Bulgarian expression (bg) *vali kato iz vedro*, with the same meaning, may be headed by an impersonal or by a personal verb and thus takes alternatively either an empty or an `nsubj` frame.

4.3.4 Word order variation

Both languages are characterised by a relatively free word order. The manual analysis of the VMWEs and the validation of this linguistic introspection using large corpora show that most VMWEs are no exception to this general rule. Here is an example of a LVC.full in Romanian (9) and in Bulgarian (10) showing this free word order:

(9) a. **Luăm parte** la concert. (ro)
Take part at concert
'We take part in the concert.'

b. **Parte luăm** la concert. (ro)
Part take at concert
'We take part in the concert.'

(10) a. В концерта **взеха участие** известни изпълнители. (bg)
V kontserta **vzeha uchastie** izvestni izpalniteli.
In concert-DEF took part famous performers.
'Famous performers took part in the concert.'

b. В концерта **участие взеха** известни изпълнители. (bg)
V kontserta **uchastie vzeha** izvestni izpalniteli.
In concert-DEF part took famous performers.
'Famous performers took part in the concert.'

However, when (some) constraints exist with respect to the word order of components or only of some of them, they are clearly marked in the entry of the respective VMWE. Such examples include: (ro) *arunca praf în ochi* (lit. 'throw dust in eyes') 'pull the wool over one's eyes', in which the noun phrase (*praf*) and the prepositional phrase (*în ochi*) always occur in this order, and the verb can be moved after them, thus resulting in an emphatic construction. A relevant example is (bg) *mnogo mi znae ustata* (much my knows mouth-DEF, lit. 'my mouth knows a lot') 'have a big mouth'. The normal word order of the MWE is an emphatic one with the `advmod` first and the `nsubj` last instead of the neutral sentential order `nsubj + det + V + advmod`. Although even in this case different word order variants are possible, some of them such as the ones where the `advmod` follows the `V` or the `V` follows the `nsubj` are very rare and we mark them as such.

4.3.5 Intervening elements

Another syntactic characteristic of VMWEs in the two languages is the possibility for (sequences of) words that do not belong to the expression to occur between its components. This is a consequence of the relatively free word order characterising Bulgarian and Romanian. Such an example is: (ro) ***Învăț** adesea, cu drag, o poezie **pe de rost*** (lit. 'Learn often, with pleasure, a poem by heart'). A few words occur within the VID *învăța pe de rost* 'learn by heart': a frequency adverb (*adesea* 'often'), a manner prepositional phrase (*cu drag* 'with pleasure') and the direct object (*o poezie* 'a poem'). The first two are not part of the valence frame, whereas the last one is. We take the stance that by default the VWMEs obey the general rules of the language in question so that peculiarities resulting from the free word order need not be marked in any way.

However, there are also cases in which the possibility for intervening elements is blocked. Such an example is: compare (ro) *Stă cu mâinile adesea în sân* 'She often stays with her arms crossed' with ***Stă** adesea **cu mâinile în sân*** 'She often stays doing nothing'. The former example shows that it is not possible to insert the frequency adverb *adesea* 'often' between the two prepositional phrases of the VID and keep the non-compositional meaning (hence, the status of VID),

whereas the latter shows that this insertion is possible between the verb and the first prepositional phrase.

For Bulgarian, we note that in some cases external elements may be blocked between a dependent's modifier and its head, when both are part of the idiom (`V + [case + amod + obj]`), e.g. (bg) *stoya sas skrasteni ratse* (lit. 'sit with crossed arms') 'sit back, sit by'. In this case, the occurrence of such element signals that the phrase has a literal reading, as in (bg) *Toy* ***stoeshe sas skrasteni*** *otpred* ***ratse*** 'He stood with his arms crossed in front of his body'.

There are also cases where parts of the VMWE are themselves idiomatic and thus do not allow intervening elements. Consider the example (bg) *varvya v krak* 'keep in step' whose dependent *v krak* 'in step' functions as an idiomatic expression outside the VID, and therefore the noun cannot be modified.

Wherever we establish restrictions on the occurrence of intervening elements between the components of a VMWE, the lexicon entry clearly states the components between which such an insertion is blocked.

Theoretically, the intervening elements may belong to a particular part of speech, may be forms of a particular lexeme or lexemes, etc. At the current stage of the development of the MWE lexicon, we prefer to collect evidence of various types of idiosyncrasies whose tackling may be dealt with at present, or may be deferred to a later moment. One of the focuses of this part of our work are the cases that diverge from the regular syntactic and linearisation rules of the language under study. Currently, this description involves the specification of POS tags that are disallowed. In the above case, the VID (bg) *varvya v krak* 'keep in step' does not allow the modification of the noun, although the rules of Bulgarian license the adjectival modification of nominals in prepositional phrases.

4.4 Semantic description

The lexicon design proposed in this chapter falls in line with the trend of describing MWEs in dedicated lexicons that provide various types of linguistic information referring to the MWE and its components and may be employed in MWE recognition related tasks. Due to the fact that BulNet and RoWN are aligned to each other, to WordNet and to any other wordnet mapped to it (see §3.1), we make use of the rich semantic description provided in the WordNet, added from additional resources or supplied manually by the teams developing BulNet and RoWN. The use of WordNet further supports the multilingual dimension of the described resource through the possibility of directly deriving the relevant semantic description available for other languages.

The main components of the semantic description incorporated herein are: a definition (called gloss), a set of semantic relations to other WordNet concepts, usage examples, stylistic and connotation information.

4.4.1 Lexicographic definition

The lexicographic description of MWEs in the form of definitions was employed by various authors of MWE resources, including close non-MWE paraphrases (cf. §2). The use of definitions aims not only at documenting the meaning of a MWE, but also at distinguishing the particular sense from other senses of the same MWE lemma, thus accounting for polysemy.

The lexicographic definition adopted in WordNet and in the lexicon describes concepts regardless of the structure of the units that lexicalise them (single words or MWEs). Thus each MWE shares a definition with the remaining synonyms in the relevant synset in both languages, with the WordNet gloss serving as an intermediary.

4.4.2 Stylistic and/or register information

The inventories for encoding stylistic/register information in MWE resources are usually subsets of those adopted in standard dictionaries (§2.7). Note that while stylistic remarks are usually assigned to an entry, which means that they characterise all the occurrences of the respective lexical unit, Fellbaum & Geyken (2005) assign the labels to usages, thus accounting for the fact that the same idiom may have different stylistic features depending on the context.

In the model adopted, we assign stylistic/register information as a permanent value attached to a MWE, using one or more labels, established in the lexicographic practice and adopted in the BulNet for both single and MWE lexemes: "colloquial", "slang", "literary", "figurative", "dialect", "obsolete", "pejorative". The values were assigned to the RoWN counterparts and reviewed manually, as lexical items describing the same concept may differ stylistically. Thus, the corresponding VIDs (bg) *davam pet pari* (lit. 'give five paras') and *davam puknata para* (lit. 'give broken para') and (ro) *da doi bani* (lit. 'give two coins') 'give a hang' are marked as "colloquial", whereas (ro) *da două parale* (lit. 'give two paras') having the same meaning but pertaining to a different register is marked as "literary".

4.4.3 Connotation

We include connotative information which is automatically assigned to BulNet and RoWN from SentiWordNet (Baccianella et al. 2010). This is an open lexical

resource designed for supporting sentiment classification and opinion mining applications which resulted from the automatic annotation of all the synsets in WordNet with one of three possible values: positive (between 0.00 and 1.00), negative (between −1.00 and 0.00) and neutral (0.00). The sentiment values were assigned to BulNet and RoWN as part of previously implemented tasks.

In our current work, we undertook a check of the values at the level of individual VMWEs (not the level of the synset), as different literals may have different connotation. For instance, the colloquial (bg) *hvarlyam prah v ochite* and (ro) *arunca praf în ochi* ('throw dust in the eyes') have negative connotation, but the synset was assigned a positive value of 0.5. We marked where the connotation value assigned from WordNet were reconsidered in our resource.

4.4.4 Semantic relations

Another trend in MWE lexicon crafting was to integrate MWEs into the lexical system of the language as individual entities, while accounting for their morphological, syntactic and semantic properties. This integration may involve the encoding of various relations to other single and MWE lexemes (§2).

By virtue of their integration in the WordNet's structure, the VMWE in the devised lexicon are explicitly associated to their synonyms (i.e., the remaining synset members, both single words and MWEs), see Figure 2. Through their membership in synsets, VMWEs are also connected to other synsets in WordNet via a number of conceptual-semantic relations – hypernymy (and its inverse hyponymy), holonymy (and its inverse meronymy), etc. – and/or lexical relations, e.g., antonymy (Miller 1995, Fellbaum 1998b).

The Bulgarian and Romanian MWEs in the target synset are connected to their hypernym (also containing MWE literals in Bulgarian). In addition, WordNet includes derivational relations (marked as `eng_derivative`), part of which are assigned semantic values that denote various roles in the situation described, eventualities or properties, i.e., the so-called morphosemantic links (Fellbaum et al. 2009). Derivational relations require validation as they might not be true across languages, e.g., (bg) magazin:1 and (ro) magazin:1; prăvălie:1 ('shop') are not derivationally related to the target synset. Their semantic values, however, are considered to be language-independent. In Figure 2, such relations are: `has_location` that connects the target synset to the location where it takes place; `has_agent` – pointing to the invariant agent (a person who shops); `has_event` – the act of doing shopping. Another relation, `category_domain`, describes the domain to which a synset pertains (if relevant). In this case it relates the target synset to the domain of commerce.

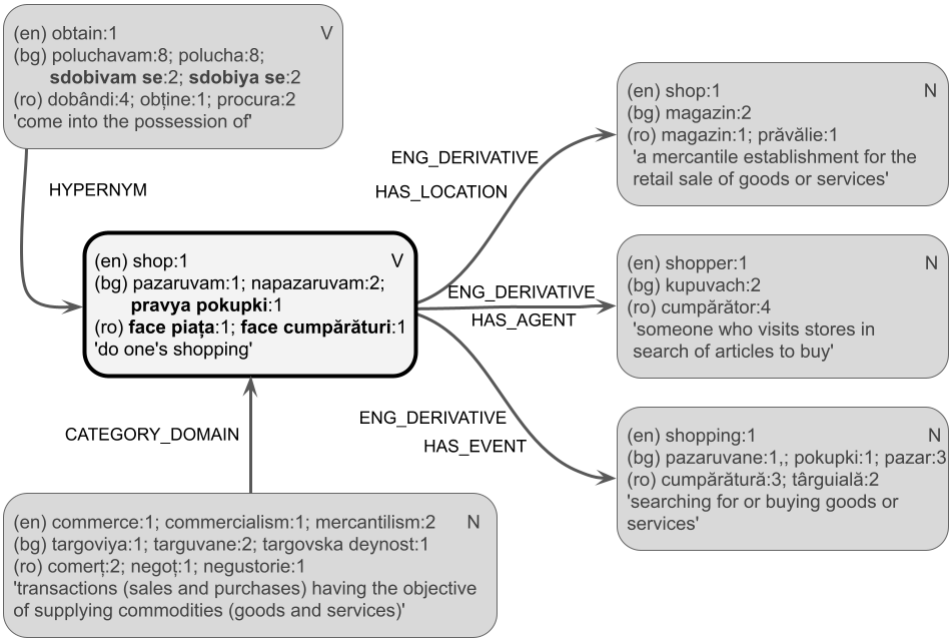

Figure 2: Synset relations within WordNet.

4.5 Derivational information

MWEs can be bases for derivation in both Romanian and Bulgarian, but this property was not consistently accounted for in WordNet.[22] Barbu Mititelu & Leseva (2018) showed that derivation of MWEs can result into both other MWEs and one-word compounds; the authors also analysed some syntactic patterns identified in the derivation of VMWEs extracted from two lexicons of MWEs in Bulgarian and Romanian. However, the VMWEs in our lexicon display only the derivational relations between two MWEs.

We also investigated the derivational potential of the VMWEs included in the lexicon. Our datasets do not coincide with those used by Barbu Mititelu & Leseva (2018), although a certain overlap is naturally possible. However, after the manual investigation of the derivational possibilities of the VMWEs in BulNet

[22]Note that we cannot claim that the discussed patterns are indeed resulting from a process of derivation that occurred in the language history. Rather, we mean that there are multiword formations that are semantically and structurally related to VMWEs and that those formations involve the employment of some mechanism of derivation (or even inflection) concerning one or more of the elements of the respective VMWE, as well as internal syntactic restructuring.

and RoWN, we could confirm the patterns enumerated there.[23] Table 4 shows the syntactic patterns involved in the VMWEs derivation, alongside examples for each language in which they are found. Derivational patterns with the same syntactic transformation, but involving different semantics, are presented as distinct patterns (e.g., `V + obj` > `N_V-derived + case + nmod` for deriving Event or Agent). The head of the derivation is marked by boldface.

The data shows a vast number of nouns designating events, which is in line with the findings by Barbu Mititelu & Leseva (2018), while derivation involving a result pertaining to other semantic types is less numerous. The semantic labels provided in Table 4 are mostly based on the inventory analysed by Barbu Mititelu & Leseva (2018).

In light of the most represented syntactic patterns in the datasets, the primary bases for forming the most productive type – event deverbal – are VMWEs exhibiting the relations `V + obj` and `V + [case + obl]`. In addition to expressing the VMWE complement as a prepositional modifier in the resulting nominalisation, Romanian exhibits a pattern where the VMWE complement is turned into a genitive modifier, which the Bulgarian language does not allow for.

From the same syntactic patterns, but involving a different (e.g., agentive) suffix in the derivation of the deverbal noun, we obtain Agents (bg) *perach na pari* 'brainwasher', Patients (ro) *muritor de foame* 'very poor person', etc.

VMWEs exhibiting the `V + obj` relation allow the formation of noun expressions (NMWEs) whose head is the object of the VMWE modified by a participial (adjective) – a past (passive) participle, cf. the examples in Table 4: (bg) *promiya mozaka* > *promit mozak* and (ro) *trage sfori* > *sfori trase*. The meaning is resultative and aligns with such examples in English: (en) *close the door* > *closed door*, *break the heart* > *broken heart*.

VMWEs exhibiting both the relations `V + obj` and `V + case + obl` regularly correspond to formations headed by a participle of the head verb in the VMWE. In Bulgarian different participles take part in this process: present (active) participles, e.g. (bg) *smrazyavam kravta* > *smrazyavasht kravta* 'curdle the blood' > 'curdling the blood', 'blood-curdling'; past active participles: (bg) *umiram ot glad* > *umryal ot glad* 'die of hunger' > 'dead of hunger', 'starved'; past passive participles: (bg) *vlyubya se do ushi* > *vlyuben do ushi* 'fall in love to the ears' > 'fallen in love to the ears', 'be head over heels in love' > 'head over heels in love'. Some of them become established in the language and are converted to adjectives,

[23] The patterns presented by Barbu Mititelu & Leseva (2018) are described in terms of dependency grammar, but using syntactic functions such as subject, complements, adjuncts. The confirmation of those patterns was possible by converting them into the UD format.

Table 4: The most frequent syntactic patterns involved in the VMWE-to-OtherPOS-MWE derivation.

Romanian examples	Bulgarian examples	
`V + case + obl` > **`N_V-derived`** `+ case + nmod`		Event
ieși la iveală > *ieșirea la iveală* 'exit at apparition' > 'exit (N) at apparition' 'come to light' > 'coming to light'	*umiram ot glad* > *umirane ot glad* 'die of hunger' > 'an act of dying of hunger' 'starve' > 'starving, starvation'	
`V + obj` > **`N_V-derived`** `+ case + nmod`		Event
spăla bani > *spălare de bani* 'launder money' > 'laundering of money', 'money laundering'	*pera pari* > *prane na pari* 'launder money' > 'laundering of money', 'money laundering'	
`V + obj` > **`N_V-derived`** `+ nmod`		Event
spăla creierul > *spălarea creierului* 'brainwash' > 'brainwashing'	–	
`V + obj` > **`N_V-derived`** `+ case + nmod`		Agent
spăla creierul > *spălător de creiere* 'brainwash' > 'brainwasher'	*promivam mozaka* > *promivach na mozatsi* 'brainwash' > 'brainwasher'	
`V + obj` > **`ADJ_V-derived`** `+` **$\mathtt{N_{obj}}$**		Result
trage sfori > *sfori trase* 'pull strings' > 'pulled strings'	*promiya mozaka* > *promit mozak* 'brainwash' > 'a brainwashed brain'	
`V + case + obl` > **`ADJ_V-derived`** `+ case + obl`		Characteristic
muri de foame > *mort de foame* 'die of hunger' > 'dead of hunger' 'starve' > 'starving' *spăla creierul* > *spălat pe creier* 'brainwash' > 'brainwashed	*umra ot glad* > *umryal ot glad* 'die of hunger' > 'dead of hunger' 'starve' > 'starving'	
`V + case + obl` > **`ADJ_V-derived`** `+ case + obl`		Characteristic
scoate din minți > *scoatere din minți* 'take-out from minds' > 'taking-out from minds' 'madden' > 'maddenning'	*umiram ot glad* > *umirasht ot glad* 'die of hunger' > 'dying of hunger' 'starve' > 'starving'	

whereas others are used in context but are not established as lexicographic units. Nevertheless, such constructions need to be described both from the perspective of generation, as they are formed on the basis of VMWEs having a certain syntactic structure and morphological properties according to certain rules, and recognition (being able to associate a relevant string of forms as related to the source VMWE).

With respect to derivation, the Romanian dataset contains a large number of VMWEs which are bases for derived nominal MWEs by means of conversion applied to the supine verb of the base VMWE. For example, (ro) *da socoteală* lit. 'give payoff' 'answer for' is the base for *datul socotelii*: the derived nominal MWE is obtained from the base MWE by converting the supine of the verb *da*, namely *dat*, into a noun, shown here by adding the definite article *-(u)l* to it. Equally often we find cases when the participle of the verb allows for the derivation of an adjectival MWE from the verbal one, also by means of conversion: e.g., (ro) *trage pe sfoară* lit. 'pull on rope' 'play a trick on' is the base for *tras pe sfoară*: the derived adjectival MWE is obtained from the base MWE by conversion of the supine of the verb *trage*, namely *tras*, into an adjective, which is a frequent phenomenon in Romanian.

4.6 Visualisation and basic query interface

Figure 3 shows the basic visual interface that allows access to and queries on the dataset. There are several filtering parameters: (i) the type of the VMWE (All, VID, LVC); (ii) word order variability; (iii) syntactic flexibility – whether the VMWE allows its components to be modified; (iv) stylistic register of the MWE; (v) structure of the VMWE – syntactic patterns according to the UD scheme; (vi) search terms in either Bulgarian and/or Romanian VMWEs or abstract lemmas. The result of the filtering is a list of all VMWE pairs that match the filtering criteria. Each VMWE pair is first identified by its synset ID and WordNet definition. If more than one VMWE pairs are available for a given synset, the user can select among possible Bulgarian-Romanian literal pairs to align and compare. Upon selection, the pair of VMWEs is presented in parallel for Bulgarian and Romanian (see Figure 4) with the features outlined in §4.1–§4.5.

Search instructions

Type: ☑ All ☐ VID ☐ LVC

Word order: ☑ All ☐ Frozen ☐ Limited ☐ Free

Allowing modifiers: ☑ All ☐ No modifiers ☐ Accepting Modifiers

Register: ☑ All ☐ Marked as colloquial ☐ Other marked

Syntactic pattern:

☑ All	☐ V + [amod + case + obl]	☐ V + [amod + obj]	☐ V + [case + advmod]
☐ V + [case + amod + obl]	☐ V + [case + case + obl]	☐ V + [case + obl]	☐ V + [nummod + obj]
☐ V + [obj (pronoun) + amod]	☐ V + advmod	☐ V + expl:poss + [case + obl]	☐ V + expl:poss + obj
☐ V + expl:poss + obj + xcomp	☐ V + expl:pv + [case + obl]	☐ V + expl:pv + [obl + case + obl]	☐ V + obj
☐ V + obj + [case + obl]	☐ V + obl (short dative pron) + [case + obl]	☐ cop + [case + ROOT]	☐ cop + [case + nummod + ROOT]
☐ cop + [case+ ROOT]	☐ cop + advmod	☐ neg + V + xcomp	

Search by component in the MWE or abstract lemma of the MWE:

Word (bg) Word (ro)

Filter

Figure 3: Search interface to filter MWE data.

- eng-30-00839194-v

conceal one's true motives from especially by elaborately feigning good intentions so as to gain an end

bg	ro
○ хвърля прах в очите	◉ arunca praf în ochi
◉ хвърлям прах в очите	○ duce de nas
	○ trage pe sfoară

Feature	BG	RO
WORDNET ID	eng-30-00839194-v	eng-30-00839194-v
Literal	хвърлям прах в очите	arunca praf în ochi
PARSEME TYPE	VID	VID
Definition	карам някого да мисли или прави нещо в моя угода, прикривайки реални факти, мотиви или цели	A induce pe cineva în eroare, printr-o viclenie sau printr-o minciună, pentru a trage un folos sau pentru a se amuza
MWE Lemma	хвърлям прах в очите	arunca praf în ochi
MWE Abstract Lemma	хвърлям прах в око	arunca praf în ochi
Aspect (BG only)	IMPERF	-
Regular morphosyntactic representation		NO
Restrictions on the verbal head	NO	NO
Restrictions on dependents	2_fixed: N = s; D = 0 & 3_fixed & 4_fixed: N = p; D = d	praf - only singular, only without article; ochi - only without article
Internal structure (in UD format)	V + obj + [case + obl]	V + obj + [case + obl]
Valences	nsubj, nmod:poss	nsubj, iobj
Word order restrictions		praf în ochi - as is
Intervening words blocked		praf în ochi
Register	colloquial	colloquial
Sentiment - pos	0	0
Sentiment - neg	0.5	0.5
Derivation	хвърляне на прах в очите	aruncatul prafului în ochi; aruncarea prafului în ochi

Figure 4: Visualisation of aligned bilingual VMWEs.

5 Discussion, conclusions, and future work

We consider the important aspects of our work to be (i) its focus on languages other than English, and (ii) the use of a common framework for an in-depth linguistic description of VMWEs. Bulgarian and Romanian are morphologically richer languages than English and belong to different families (Slavic and Romance, respectively). The description of VMWEs in these two languages is made in a multilingual landscape offered by aligned wordnets. Using of a common framework for an in-depth linguistic description of VMWEs allows for highlighting both similarities and differences between the MWEs in the two languages. Moreover, this framework is encoded in a transparent, flexible, expressively capable, versatile and friendly way (Lichte et al. 2019).

Our lexicon is rooted in WordNet: the organisation principles therein explain the work methodology and the representation of information. Thus, for a MWE, we do not encode a list of lexemes that can substitute components in an expression, as is the case with some other such lexicons (see §2). Whenever such substitutions are possible, the whole expression is encoded as a different literal occurring in the same synset as its synonyms (thus, labelled with different literal IDs, see §4.1). One such example is the pair (ro) *da doi bani – da două parale* 'give a hang', which differ in their last component: *ban* is a current unit of money, while *para* is an older one, not used anymore. An argument in favor of the distinct treatment of lexical variants is that, other differences aside, as we showed earlier, the two MWEs belong to different lexical registers – one is colloquial and the other is literary.

There are also cases when two expressions vary by means of one component that is added to offer emphasis to the expression in use: see the pair (ro) *își da silința – își da toată silința* 'do one's best', which differ only in the determiner *toată* 'all' added to the direct object of the verb, thus making it more emphatic. This affects the communicative status of the different variants and may determine the choice of one over the other in a context, the preference of different equivalents or translations in other languages, etc.

Another consequence of including in the lexicon MWEs from wordnets is that no relationship is encoded between an expression and the entries for its components, i.e., the synset(s) to which the MWE belongs and the synsets to which its components belong (unlike traditional thesauri where MWEs often appear under one or more of its components). Each word sense and each MWE sense are separately encoded. However, by means of the relations in the networks, when any semantic relation exists between one meaning of a component and the meaning of a MWE of which it is a component, then this (close or distant) relation can

be retrieved by traversing the edges starting from one synset and reaching the other one.

The multilingual dimension of the resource presented here springs from the fact that the Bulgarian-Romanian lexicon exploits the alignment between the two wordnets, thus being a resource on top of two linked monolingual ones. The alignment was possible via Princeton WordNet and this actually opens the way to alignment to any other such lexicon either built on top of other wordnets or linked to them. A possible future development towards the multilingual extension would be to employ a large-scale densely populated resource providing access to aligned MWE entities such as BabelNet (Navigli & Ponzetto 2012).

Lichte et al. (2019) discuss what they call general virtues of MWE encoding, namely *transparency, flexibility, power to generalise, implementation friendliness, electronic versatility*, as prerequisites for a lexical resource. *Transparency* concerns the ability of the human user to map the encoding back to the source set of lexical properties, i.e. the simplicity of the encoding of linguistic features and the straightforwardness of their interpretation by novices or non-expert users. *Flexibility* is the adaptability of a format to dealing with unforeseen properties or changes in properties. The *power to generalise* allows the user to group properties and assign them collectively, thus avoiding redundancies and errors. The *implementation friendliness* relates to the existence of tools that assist a human user with encoding or its validation. *Electronic versatility* describes the ease of converting the lexical encoding into a lexical resource, in particular, the existence of conversion tools or the possibility to produce them.

To ensure the *transparency* of the encoding, we adopted a straightforward link between the linguistic properties and their values. The field names serving to encode the properties are both easy to encode manually and to interpret. The basic tabular format of the template used to describe each MWE component facilitates the adaptation to new or unforeseen properties, thus ensuring the *flexibility* of the data encoding. New (categories of) fields and values may be defined as appropriate when needed and added to the predefined VMWE description template. This is especially relevant with respect to language-specific features (e.g., verb aspect in Bulgarian), as it allows the two teams to work independently. The unified description of the data for each language enabled us to consider two aspects of the *power to generalise*: (i) the possibility to identify and extract linguistic regularities, including groups of relevant properties in the VMWEs that share them, thus identifying possible classes of VMWEs with similar characteristics (from a certain perspective); and (ii) the possibility to look into linguistic regularities or shared features between the languages as well as to extract semantic, structural, etc. correspondences between VMWEs in them. *Implementation friendliness* and *electronic versatility* stem from the simple form in which the data are described.

Currently, we did not use a particular tool, but the explicitness of the format and the encoding of features makes it easy to convert to various formats according to the relevant requirements of the existing tools.

Adopting the same work methodology made it possible for the teams to work independently from each other using a predefined template that includes the relevant linguistic features (on the basis of previous data analysis) and expanding it to new features when the need arises. The model is thus adaptable to languages that share similar linguistic properties, possibly to genetically and/or typologically related ones.

Future work will aim at the enrichment of the monolingual lexicons with descriptions of the VMWEs that are in the individual wordnets, as for now we created entries only for those that are mutually equivalent VMWEs in the two languages. The further development of the two wordnets will allow for the identification of other (V)MWEs equivalents, thus enriching the bilingual lexicon and extending it to MWEs of other parts of speech.

Syntactic transformations have not been tackled yet in our resource. As most of them show the regular syntactic behavior of free phrases, we have decided, during the next stage of our work, to start marking the cases where a certain transformation is impossible and proceed to describe the conditions for blocking. This will be implemented in a manner similar to the encoding of morphosyntactic restrictions, i.e. by defining a relevant field 'Syntactic transformations' and listing the restrictions using a predefined list of the names of the transformations as values. A further, more in-depth treatment of syntactic transformations will depend on the analysis of the data after we have collected them.

As corpora annotated with VMWEs exist for both Romanian (Barbu Mititelu et al. 2019a) and Bulgarian (Koeva et al. 2012), associating the lexicon entries with relevant corpora occurrences is a natural next step that would contribute a syntagmatic dimension to the resource.

Abbreviations

BulNet	Bulgarian WordNet
HPSG	Head-driven Phrase Structure Grammar
IAV	inherently adpositional verb
IRV	inherently reflexive verb
LFG	Lexical-Functional Grammar
LVC	light verb constructions
MWE	multiword expression
N	noun
NLP	Natural Language Processing
NMWE	noun multiword expressions
RoWN	Romanian WordNet
UD	Universal Dependencies
V	verb
VID	verbal idiom
VMWE	verbal multiword expression

Acknowledgements

The authors are grateful to the anonymous reviewers and to the editors of this volume for their remarks on the previous versions of this chapter, which helped improving it.

References

Autelli, Erica. 2020. Phrasemes in Genoese and Genoese-Italian lexicography. In Joanna Szerszunowicz & Eva Gorlewska (eds.), *Applied linguistics perspectives on reproducible multiword units: Foreign language teaching and lexicography* (Intercontinental Dialogue on Phraseology 8), 101–127. Białystok: University of Białystok Publishing House. DOI: 10.1007/978-3-642-30910-6_12.

Baccianella, Stefano, Andrea Esuli & Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the seventh international Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf.

Barbu Mititelu, Verginica, Mihaela Cristescu & Mihaela Onofrei. 2019a. The Romanian corpus annotated with verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 13–21. Florence, Italy: Association for Computational Linguistics. DOI: 10.18653/v1/W19-5103.

Barbu Mititelu, Verginica & Svetlozara Leseva. 2018. Derivation in the domain of multiword expressions. In Manfred Sailer & Stella Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective*, 215–246. Language Science Press. DOI: 10.5281/zenodo.1182601.

Barbu Mititelu, Verginica, Ivelina Stoyanova, Svetlozara Leseva, Maria Mitrofan, Tsvetana Dimitrova & Maria Todorova. 2019b. 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)*, 2–12. Florence, Italy: Association for Computational Linguistics. DOI: 10.18653/v1/W19-5102.

Bresnan, Joan. 1978. A realistic transformational grammar. In Morris Halle, Joan Bresnan & George A. Miller (eds.), *Linguistic Theory and Psychological Reality*, 1–59. MIT Press.

Chiarcos, Christian, Maxim Ionov, Elena-Simona Apostol, Katerina Gkirtzou, Besim Kabashi, Anas Fahad Khan & Ciprian-Octavian Truică. 2024. Multiword expressions, collocations and the OntoLex vocabulary. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources:*

Linguistic, lexicographic, and computational perspectives, 187–227. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998641.

Dalrymple, Mary (ed.). 2023. *Handbook of Lexical Functional Grammar* (Empirically Oriented Theoretical Morphology and Syntax 13). Berlin: Language Science Press. DOI: 10.5281/zenodo.10037797.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47(2). 255–308. DOI: 10.1162/coli_a_00402.

Dyvik, Helge, Gyri Smørdal Losnegaard & Victoria Rosén. 2019. Multiword expressions in an LFG grammar for Norwegian. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions*, 69–108. Language Science Press. DOI: 10.5281/zenodo.2579037.

Fellbaum, Christiane. 1998a. Towards a representation of idioms in WordNet. In *Usage of WordNet in natural language processing systems*, 52–57. https://aclanthology.org/W98-0707.

Fellbaum, Christiane (ed.). 1998b. *WordNet: An electronic lexical database*. Cambridge, MA: The MIT Press.

Fellbaum, Christiane & Alexander Geyken. 2005. Transforming a corpus into a lexical resource: The Berlin Idiom Project. *Revue française de linguistique appliquée* 10(2). 49–62. DOI: 10.3917/rfla.102.62.

Fellbaum, Christiane, Anne Osherson & Peter E. Clark. 2009. Putting semantics into WordNet's "morphosemantic" links. In Zygmunt Vetulani & Hans Uszkoreit (eds.), *Lecture notes in computer science*, 350–358. Berlin, Heidelberg: Springer.

Fotopoulou, Aggeliki, Stella Markantonatou & Voula Giouli. 2014. Encoding MWEs in a conceptual lexicon. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, 43–47. Gothenburg, Sweden: Association for Computational Linguistics.

Fried, Mirjam & Jan-Ola Östman. 2004. Construction Grammar: A thumbnail sketch. In Mirjam Fried & Jan-Ola Östman (eds.), *Construction Grammar in a cross-language perspective*, 11–86. Amsterdam: John Benjamins.

Giouli, Voula, Vera Pilitsidou & Hephestion Christopoulos. 2024. A FrameNet approach to deep semantics for MWEs. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 147–186. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998639.

Grégoire, Nicole. 2007. Design and implementation of a lexicon of Dutch multiword expressions. In Nicole Gregoire, Stefan Evert & Su Nam Kim (eds.), *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, 17–

24. Prague: Association for Computational Linguistics. https://aclanthology.org/W07-1103.

Gross, Gaston. 1996. *Les expressions figées en français: Noms composés et autres locutions*. Paris: Ophrys.

Gross, Maurice. 1975. *Méthodes en syntaxe: Régime des constructions complétives*. Paris: Hermann.

Gross, Maurice. 1982. Une classification des phrases «figées» du français. *Revue québécoise de linguistique* 11(2). 151. DOI: 10.7202/602492ar.

Al-Haj, Hassan, Alon Itai & Shuly Wintner. 2013. Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography* 27. 130–170. DOI: 10.1093/ijl/ect036.

Hnátková, Milena, Tomáš Jelínek, Marie Kopřivová, Vladimír Petkevič, Alexandr Rosen, Hana Skoumalová & Pavel Vondřička. 2019. Lexical database of multiword expressions in Czech. In *Trudy meždunarodnoj konferencii Korpusnaja Lingvistika*, 9–16. Saint Petersburg, Russian Federation: Saint Petersburg University Press.

Iñurrieta, Uxoa, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka & Kepa Sarasola. 2018. Konbitzul: An MWE-specific database for Spanish-Basque. In *Proceedings of the eleventh international Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). https://aclanthology.org/L18-1397.

Koeva, Svetla. 2010. Bulgarian WordNet: Current state, applications and prospects. In *Bulgarian-American dialogues*, 120–132. Sofia: Prof. M. Drinov Academic Publishing House.

Koeva, Svetla. 2021. The Bulgarian WordNet: Structure and specific features. *Papers of the Bulgarian Academy of Sciences* 8(1). 47–70.

Koeva, Svetla, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova & Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and practice in corpus design. *Journal of Language Modelling* (1). 65–110. DOI: 10.15398/jlm.v0i1.33.

Koeva, Svetla, Ivelina Stoyanova, Maria Todorova & Svetlozara Leseva. 2016. Semi-automatic compilation of the dictionary of Bulgarian multiword expressions. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the GLOBALEX 2016 workshop: Lexicographic Resources for Human Language Technology, LREC*, 86–95. Paris: European Language Resources Association (ELRA).

Leseva, Svetlozara, Verginica Barbu Mititelu & Ivelina Stoyanova. 2020. It takes two to tango: Towards a multilingual MWE resource. In *Proceedings of the 4th international conference on Computational Linguistics in Bulgaria (CLIB 2020)*, 101–111. Sofia, Bulgaria: Department of Computational Linguistics, IBL – BAS. https://aclanthology.org/2020.clib-1.11.

Lichte, Timm, Simon Petitjean, Agata Savary & Jakub Waszczuk. 2019. Lexical encoding formats for multi-word expressions: The challenge of "irregular" regularities. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 1–33. Berlin: Language Science Press. DOI: 10.5281/zenodo.2579033.

Markantonatou, Stella, Nikolaos T. Kokkas, Panagiotis G. Krimpas, Ana O. Chiril, Dimitrios Karamatskos, Nicolaos Valeontis & George Pavlidis. 2024. Description of Pomak within IDION: Challenges in the representation of verb multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 39–72. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998633.

Markantonatou, Stella, Panagiotis Minos, George Zakis, Erasmia Koletti, Elpiniki Margariti & Emilia Stripeli. 2020. Idion (ιδιον): A lexicographic environment for the documentation of Greek idioms. In Stella Markantonatou & Anastasia Christofidou (eds.), *Multiword expressions in Greek: Deltio epistimonikis orologias ke neologismon.*

Markantonatou, Stella, Panagiotis Minos, George Zakis, Vassiliki Moutzouri & Maria Chantou. 2019. IDION: A database for Modern Greek multiword expressions. In *Proceedings of the joint workshop on multiword expressions and WordNet (MWE-WN 2019) at ACL 2019*, 130–134. Florence. DOI: 10.18653/v1/W19-5115.

Mel'čuk, Igor. 1981. Meaning-text models: A recent trend in Soviet linguistics. *Annual Review of Anthropology* 10. 27–62.

Mel'čuk, Igor. 2006. Explanatory combinatorial dictionary. In Giandomenico Sica (ed.), *Open problems in linguistics and lexicography*, 225–355. Monza, Italy: Polimetrica.

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

Monti, Johanna. 2014. An English-Italian MWE dictionary. In *Proceedings of the first Italian conference on Computational Linguistics CLiC-it 2014 and of the fourth international workshop EVALITA*, 265–269. Pisa: Pisa University Press.

Müller, Stefan, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.). 2021. *Head-Driven Phrase Structure Grammar: The handbook* (Empirically Oriented Theoretical Morphology and Syntax 9). Berlin: Language Science Press. DOI: 10.5281/zenodo.5543318.

Navigli, Roberto & Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193. 217–250. DOI: 10.1016/j.artint.2012.07.001.

Odijk, Jan. 2013. Identification and lexical representation of multiword expressions. In Peter Spyns & Jan Odijk (eds.), *Essential speech and language technology for Dutch: results by the STEVIN programme*, 201–217. Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-30910-6_12.

Odijk, Jan, Martin Kroon, Sheean Spoel, Ben Bonfil & Tijmen Baarda. 2024. MWE-Finder: Querying for multiword expressions in large Dutch text corpora. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 229–267. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998643.

Osenova, Petya & Kiril Simov. 2024. Representation of multiword expressions in the Bulgarian integrated lexicon for language technology. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 117–146. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998637.

Pollard, Carl & Ivan A. Sag. 1987. *Information-based syntax and semantics*, vol. 1: Fundamentals. Stanford: CSLI Publications.

Pollard, Carl & Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Przepiórkowski, Adam, Elżbieta Hajnicz, Agnieszka Patejuk & Marcin Woliński. 2014a. Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of workshop on lexical and grammatical resources for language processing*, 83–91. Dublin, Ireland: Association for Computational Linguistics & Dublin City University. DOI: 10.3115/v1/W14-5811.

Przepiórkowski, Adam, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski & Marek Świdziński. 2014b. Walenty: Towards a comprehensive valence dictionary of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the ninth international conference on Language Resources and Evaluation (LREC'14)*, 2785–2792. Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/279_Paper.pdf.

Savary, Agata. 2008. Computational inflection of multi-word units: A contrastive study of lexical approaches. *Linguistic Issues in Language Technology* 1. DOI: 10.33011/lilt.v1i.1195.

Savary, Agata, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze & Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, 24–35. Dubrovnik, Croatia: Association for Computational Linguistics. https://aclanthology.org/2023.mwe-1.6.

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Cebiroğlu Eryiğit, Voula Giouli, Maarten Van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaite, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke Van Der Plas, Behrang Qasemizadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI: 10.5281/zenodo.1471590.

Savary, Agata, Silvio Cordeiro & Carlos Ramisch. 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)*, 79–91. Florence. DOI: 10.18653/v1/W19-5110.

Savary, Agata, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Matthieu Constant, Petya Osenova & Federico Sangati. 2015. PARSEME: PARSing and Multiword Expressions within a European multilingual network. In *7th Language and Technology Conference: Human language technologies as a challenge for computer science and linguistics (LTC 2015)*. Poznań, Poland. https://hal.archives-ouvertes.fr/hal-01223349.

Schafroth, Elmar. 2015. Italian phrasemes as constructions: How to understand and use them. *Journal of Social Sciences* 3(11). 317–337. DOI: 10.3844/jssp.2015.317.337.

Shudo, Kosho, Akira Kurahone & Toshifumi Tanabe. 2011. A comprehensive dictionary of multiword expressions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 161–170. Portland, OR: Association for Computational Linguistics.

Skoumalová, Hana, Marie Kopřivová, Vladimír Petkevič, Tomáš Jelínek, Alexandr Rosen, Pavel Vondřička & Milena Hnátková. 2024. LEMUR: A lexicon of Czech multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 1–37. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998631.

Straka, Milan. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In Daniel Zeman & Jan Hajič (eds.), *Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies*, 197–207. Brussels, Belgium: Association for Computational Linguistics. DOI: 10.18653/v1/K18-2020.

Tufiș, Dan & Verginica Barbu Mititelu. 2014. The lexical ontology for Romanian. In Nuria Gala, Reinhard Rapp & Nuria Bel-Enguix (eds.), *Language Production, Cognition, and the Lexicon*, 491–504. Cham: Springer.

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

Vietri, Simonetta. 2014a. *Idiomatic constructions in Italian: A lexicon-grammar approach.* Amsterdam/Philadelphia: John Benjamins.

Vietri, Simonetta. 2014b. The lexicon-grammar of Italian idioms. In Jorge Baptista, Pushpak Bhattacharyya, Christiane Fellbaum, Mikel Forcada, Chu-Ren Huang, Svetla Koeva, Cvetana Krstev & Éric Laporte (eds.), *Proceedings of the workshop on lexical and grammatical resources for language processing (LG-LP 2014)*, 137–146. Dublin, Ireland: Association for Computational Linguistics & Dublin City University. DOI: 10.3115/v1/W14-5817.

Villavicencio, Aline, Timothy Baldwin & Benjamin Waldron. 2004a. A multilingual database of idioms. In *Proceedings of the fourth international conference on Language Resources and Evaluation (LREC'04)*, 1127–1130. Lisbon. http://www.lrec-conf.org/proceedings/lrec2004/pdf/760.pdf.

Villavicencio, Aline, Ann Copestake, Benjamin Waldron & Fabre Lambeau. 2004b. Lexical encoding of MWEs. In Takaaki Tanaka, Aline Villavicencio, Francis Bond & Anna Korhonen (eds.), *Proceedings of the second ACL workshop on multiword expressions: Integrating processing*, 80–87. Barcelona: ACL.