Leaving No Stone Unturned When Identifying and Classifying Verbal Multiword Expressions in the Romanian Wordnet

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

Maria Mitrofan
RACAI
Bucharest, Romania
maria@racai.ro

Abstract

We present here the enhancement of the Romanian wordnet with a new type of information, very useful in language processing, namely types of verbal multiword expressions. All verb literals made of two or more words are attached a label specific to the type of verbal multiword expression they correspond to. These labels were created in the PARSEME Cost Action and were used in the version 1.1 of the shared task they organized. The results of this annotation are compared to those obtained in the annotation of a Romanian news corpus with the same labels. Given the alignment of the Romanian wordnet to the Princeton WordNet, this type of annotation can be further used for drawing comparisons between equivalent verbal literals in various languages, provided that such information is annotated in the wordnets of the respective languages and their wordnets are aligned to Princeton WordNet, and thus to the Romanian wordnet.

1 Introduction

The Romanian wordnet (RoWN) is a rich lexical and semantic resource. Its development followed the expand method (Vossen, 2002) and started within the BalkaNet project (Tufiș et al., 2004). Alignment with Princeton WordNet (PWN) (Miller, 1995; Fellbaum, 1998) was a consequence of this working method and has always been one of the objectives whenever new synsets were developed for enlarging the RoWN. Consequently, alignment with all the other wordnets aligned with PWN is obtained, which is a great asset for both interlingual lexical comparison or for applications working in a multilingual environment.

The expand model in wordnets development implies importing the structure of the PWN (that is, its semantic relations) and translating the source synsets (from PWN), so that the meaning encoded by the English synset is rendered in the target language (Romanian, here). As a consequence, a Romanian synset may have one of the following structures: (i) list of words; (ii) list of free word combinations; (iii) empty list. (i) A list of words is a list of simple words (ex. *zâmbi* ("smile")) and/or expressions (ex. *casă de bani* (house of money "strong box")). These expressions are what in lexicographic terms is called idioms, terms, etc. (ii) Whenever no word or expression could be found in Romanian for rendering the meaning of the English synset, a free word combination, when possible, was used for implementing the respective synset: ex.: *pune jos* is a literal in the Romanian synset equivalent to the PWN 3.1 {ground:10} (gloss: place or put on the ground). These are examples of Recurrent Free Phrases, as Bentivogli and Pianta (2004) call them. (iii) In case not even such a combination could be found, the synset was left empty and a special tag is used for keeping track of them (they are marked as NL, i.e. non lexicalized): ex.: the English synset {change state:1, turn:4} (gloss: undergo a transformation or a change of position or action) has a non-lexicalized corresponding synset in RoWN. However, as already pointed out (Vincze et al., 2012; Bentivogli and Pianta, 2004; Agirre et al., 2005), these lexical gaps should be reduced as much as possible when use of wordnets is envisaged for tasks in a multilingual environment (see machine translation), but also for word sense disambiguation (Bentivogli and Pianta, 2004).

As far as this structure of its synsets is concerned, RoWN looks as rendered in Table 1. One should bear in mind the fact that it is impossible to distinguish automatically between expressions and free word combinations. That is why, on rows

4 and 5 in Table 1 both types of literals, expressions and free combinations of words, are counted together. As one can see, almost 70% of all Romanian synsets are made up of only simple literals. Those made up of only multiword literals represent 21.2% of all synsets. Less than 5% of the Romanian synsets are made up of both simple and multiword literals, having almost the same distribution as non-lexicalized synsets.

Types of synsets	*Number*	*Percent*
all synsets	59,348	-
synsets containing only simple literals	41,188	69.5%
synsets containing simple literals, expressions and free word combinations	2,813	4.7%
synsets containing expressions and/or free word combination	12,590	21.2%
non-lexicalized synsets	2,757	4.6%

Table 1: Distribution of different types of synsets in RoWN.

As far as the distribution of simple literals and expressions in RoWN is concerned, Table 2 shows that, at the literal level, the situation is somehow different: almost 65% of the whole number of unique literals are simple ones, whereas 35% are multiword ones. When considering their all occurrences, we notice that the simple ones are more frequent (76.5%), given their polysemy which is bigger than that of multiword units (see also (Bentivogli and Pianta, 2004)), which account for only 23.5% of the number of all literals in RoWN.

At present, we are carrying out a bilateral (Romanian-Bulgarian) project of annotating the different types of multiword expressions in the Romanian wordnet. The first step is annotating the verbal multiword expressions (VMWEs). This follows naturally from our participation in the PARSEME Cost Action[1] and in the creation and annotation of the corpora used in the PARSEME

[1] https://typo.uni-konstanz.de/parseme/

Types of synsets	*Number*	*Percent*
all literals	85,277	-
simple literals	65,246	76.5%
expressions and/or free word combination	20,031	23.5%
unique literals	50,480	-
unique simple literals	32,664	64.7%
unique expressions and/or free word combination	17,816	35.3%

Table 2: Distribution of different types of literals in RoWN.

shared tasks 1.0 (Savary et al., 2017) and 1.1 (Ramisch et al., 2018). This paper focuses on the annotation of Romanian wordnet data. We present the PARSEME typology of VMWEs and the types applicable to Romanian (section 2), the process of annotating the verbal literals in RoWN with these types of VMWEs (section 3) and we discuss the obtained results, as well as a comparison with those from the annotation of a Romanian news corpus with the same types of VMWEs (section 4), before concluding the paper.

2 Typology of verbal multiword expressions

For the organization of a shared task on the automatic identification and classification of VMWEs, the existence of an annotated corpus was one of the prerequisites. The interest in this initiative manifested by representatives of quite a large number of languages lead to fruitful discussions and the creation of an annotation manual defining the scope of the task, the types of VMWEs to be annotated and their characteristics. The annotation guidelines capture the idiosyncrasies of all the languages involved.

According to the last version of these guidelines[2], VMWEs fall into universal, quasi-universal

[2] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/index.php?

and language specific categories, the first two having some subcategories, as follows:

- universal categories are types of VMWEs that exist in all natural languages (at least in those participating in the PARSEME corpus annotation action). Their subcategories are:
  - *light verb constructions* (LVC) - they are made up of a verb and a predicative noun (directly following the verb or being introduced by a preposition), the latter having semantic arguments. Depending on the semantics of the verb, two subtypes are identified:
    - LVC.full - these are expressions in which the verb's contribution to the expression's semantics is (almost) null (we call the verb "light"): example: *pay a visit*;
    - LVC.cause - in these expressions the verb has a causative meaning, i.e. it identifies the subject as the cause or source of the event or state expressed by the noun in the expression: example: *give a headache*;
  - verbal idioms (VID) - they are made up of a verb and at least one of its arguments and have a totally non-compositional meaning (Vincze et al., 2012): example: *kick the bucket* (die);
- quasi-universal categories exist only in some of the languages under study. They are:
  - *inherently reflexive verbs* (IRV) - these are verbs that are accompanied by a clitic pronoun with a reflexive meaning: example: *help oneself*;
  - *verb-particle construction* (VPC) - these are verbs accompanied by a particle which totally or partially changes the meaning of the verb: example: *put off*;
  - *multi-verb constructions* (MVC) - they are sequences of two adjacent verbs functioning together as a single predicate with the same subject; this type does not exist in English.

Romanian displays only the following types of VMWEs from the PARSEME classification: LVC.full: *lua o decizie* (make a decision), LVC.cause: *da bătăi de cap* (give headaches), VID: *trage pe sfoară* (pull on rope "cheat") and IRV: *se preface* ("pretend"). These labels were used for the annotation of the Romanian corpus used in the shared task version 1.1 (as in version 1.0 the VMWEs types were slightly different). No language specific categories were necessary in the corpus annotation.

3 Annotation of the Types of VMWEs in RoWN

The task of annotating the VMWEs in a wordnet is different in some respects from their annotation in a corpus. First, all components are present as one literal in the synset, whereas in a corpus they need to be identified, according to the specifications available for all languages (e.g., auxiliaries, clitics or negation are not annotated as parts of the expression). Second, whenever at least one element of the VMWE inflects for number, gender, etc., it has a unique form in the wordnet, the one considered lemma, while in the corpus all inflected forms may be found and need to be recognized. Third, no voice alternation is to be found in the wordnet, while this can be spotted in a corpus. Fourth, when the decision on whether a word combination is a VMWEs depends on the meaning of that combination, the gloss attached to the synset is useful for this and the decision is based on it.

The annotation of VMWEs in RoWN was done by one linguist, with experience in annotating VMWEs in a corpus, following the PARSEME guidelines. Thus, we cannot discuss here the difficulty of this annotation or any controversial cases. The data are stored in a standoff file[3]. The file contains the literals in each synset, their VMWE label and the unique identifier of each synset, which is taken from PWN 3.0.

All VMWEs in RoWN were identified, extracted and were assigned to one of the types of VMWEs applicable to Romanian (LVC.full, LVC.cause, IRV and VID). However, these types proved not enough for this task. The free word combinations with a verb as head could not be annotated with any of these labels, as expected, in fact. Consequently, we marked them with a new label, NONE: they have a literal, compositional meaning, they do not display the characteristics of the VMWE classes: such an example is *culege nuci* (pick nuts).

page=home

[3] http://www.racai.ro/en/tools/text/

This type of annotation is done at the literal, not at the synset level (see also the discussion about the distribution of different types of VMWEs within a synset, in the next section).

Although the vast majority of VMWEs belong to only one type, there are literals which are annotated differently when belonging to different synsets, i.e. when having different meanings. Out of only a handful of such cases, here is one example: the expression *scoate fum* (give out smoke) is annotated as NONE when being in the synset corresponding to the English {fume:4; smoke:4} (gloss: emit a cloud of fine particles) and it is annotated as VID when belonging to the synsets corresponding to the English {steam:3} (gloss: get very angry).

4 Annotation Results

The distribution of the types of VMWEs in the RoWN is presented in Table 3. As one can see, there is a great number (1,211) of artificial verbal expressions (the label NONE). The most frequent type of expressions is IRV (989), followed by VID (614). The numbers of LVC.full and LVC.cause are quite low: 102 and 42, respectively.

Type	***No.***	***%***	***%*** *ignoring NONE*
LVC.full	102	3.4	5.8
LVC.cause	42	1.4	2.4
VID	614	20.9	35
IRV	989	33.3	56.5
NONE	1,211	40.8	
double ann.	5	0.2	0.3
TOTAL	2,963		

Table 3: The distribution of VMWEs types in the RoWN.

As far as the correlation of these figures with those found in the corpus annotated in PARSEME (see Table 4) is concerned, we notice that the frequency distribution is roughly the same, with IRV the most frequent type, followed by VID, while the subtypes of LVC are both rare.

We can conclude that the IRV type is the most frequent both at the lexicographic level and in language use for Romanian.

Figure 1 shows the presence of VMWEs in synsets of different lengths. We notice their greatest presence in shorter synsets (especially of lengths 1 or 2).

Type	***No.***	***Freq.***	***Rel. freq.***
LVC.full	39	312	5.31
LVC.cause	8	181	3.08
VID	171	1,602	27.28
IRV	268	3,777	64.32
TOTAL	486	5,872	-

Table 4: The distribution of VMWEs types in a Romanian news corpus.

		No. of VMWEs per synset					
		1	**2**	**3**	**4**	**5**	**6**
No. of literals per synset	**1**	867					
	2	246	220				
	3	79	54	41			
	4	30	18	15	12		
	5	11	7	6	3	4	
	6	3	0	2	0	0	1
	7	1	0	0	0	0	0
	8	1	1	0	0	0	0

Figure 1: The distribution of VMWEs in synsets of different lengths.

Figure 2 shows the distribution of RoWN synsets made up only of VMWEs by the number of literals in the synset. This is relevant for the productivity of the synonymy relation between VMWEs. As one can see, most of these expressions (867) do not have synonyms. It is noteworthy that this is the case mainly with those annotated as NONE, which is further proof of their artificial nature. There are 220 literals in which there are pairs of synonymous VMWEs. Synonymy among three VMWEs is displayed by 41 synsets, among four VMWEs by 12 synsets, among five VMWEs by 4 synsets, among six VMWEs by 1 synset, and among twelve VMWEs by 1 synset. This very rich synset is {fi de gardă, fi de pază, fi de strajă, fi de santinelă, face de gardă, face de strajă, face de pază, face de santinelă, sta de pază}, which is the equivalent of the PWN synset {stand guard:1, stand watch:1, keep guard:1, stand sentinel:1} (gloss: watch over so as to protect). This Romanian synset is based, on the one hand, on the synonymy among the nouns in the VMWEs

structure (*gardă, pază, strajă, santinelă*) and, on the other hand, on their collocation with three different verbs (*fi, sta, face*) for rendering the same meaning.

We analyzed the (277) synsets in which all literals are VMWEs in order to identify the synsets for which all types of MWEs occurring in the respective synsets are the same. After excluding those synsets containing only strings annotated as NONE (129), we counted 37 synsets in which the literals are all VID, 3 in which they are all LVC.full, 2 in which they are all LVC.cause and other 2 in which they are all IRV.

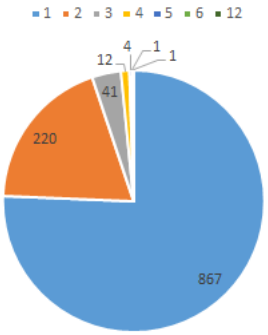

Figure 2: Distribution of synsets containing only VMWEs.

5 Conclusions

We have presented here the enhancement of the RoWN with a new type of syntagmatic information, namely labels for VMWEs. The importance of and, at the same time, the challenges raised by these lexical units for processing natural languages have been previously discussed (see, among many others, (Sag et al., 2002), (Baldwin and Kim, 2010)). Moreover, the impact of MWEs resources on the MWEs recognition in texts was proven by RiedlBiemann, : "In the case that high quality MWE resources exist, these should be used. If not, it is possible to replace them with unsupervised extraction methods". Savary et al. (2019) are also in favour of the creation of language resources containing MWEs, as many and diverse as possible; their presence in resources available for training systems for MWE identification being more important than their frequency (in annotated corpora). The results obtained in the annotation of the VMWEs in the RoWN are presented, as well as a a comparison with those obtained by annotating a news corpus with these types of VMWEs is drawn, showing that the distribution of types and their frequencies at the lexicon level are different from those at the corpus level. As further work, we envisage adding information about prepositional restrictions of the verbs in RoWN. This was another type of VMWEs in PARSEME, but annotating it was optional and we neglected it. The data annotated as presented here have been compared and discussed with the Bulgarian data, as the wordnets for both these languages have been annotated with VMWEs (Barbu Mititelu et al., 2019).

6 Acknowledgements

Part of the work reported here has been carried out within the Multilingual Resources for CEF.AT in the legal domain – MARCELL Action (http://marcell-project.eu/). Another part has been undertaken under the bilateral project *Enhancing Multilingual Language Resources with Derivationally Linked Multiword Expressions* (2018–2020) between the Institute for Bulgarian Language at the Bulgarian Academy of Sciences and the Research Institute for Artificial Intelligence at the Romanian Academy. The authors are grateful to the three anonymous reviewers for their valuable comments meant to improve the quality of the initially submitted form of this paper.

References

Eneko Agirre, Izaskun Aldezabal and Eli Pociello. 2005. *Lexicalization and Multiword Expressions in the Basque WordNet*. In Proceedings of the 3rd Global WordNet Conference, Jeju Island.

Timothy Baldwin and Su Nam Kim. 2010. *Multiword expressions.* Handbook of Natural Language Processing, Second Edition, 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Verginica Barbu Mititelu, Ivelina Stoyanova, Svetlozara Leseva, Maria Mitrofan, Tsvetana Dimitrova and Maria Todorova. 2019. *Hear about Verbal Multiword Expressions in the Bulgarian and the Romanian Wordnets Straight from the Horse's Mouth.* Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), ACL, 2–12.

Luisa Bentivogli and Emanuele Pianta. 2004. *Extending WordNet with Syntagmatic Information*. In Proceedings of the 2nd Global Wordnet Conference (GWC 04), Czech Republic, 47-–53.

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

George A. Miller. 1995. *WordNet: A Lexical Database for English.* Communications of the ACM Vol. 38, No. 11: 39–41.

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

Martin Riedl and Chris Biemann. 2016. *Impact of MWE Resources on Multiword Recognition.* Proceedings of the 12th Workshop on Multiword Expressions (MWE 2016), ACL, 107–111

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. *Multiword Expressions: A Pain in the Neck for NLP.* In Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), 1–15, Mexico City, Mexico.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova and Antoine Doucet. 2017. *The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions.* Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), ACL, 31-47.

Agata Savary, Silvio Cordeiro and Carlos Ramisch. 2019. *Without lexicons, multiword expression identication will never fly: A position statement.* Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), ACL, 79–91.

Dan Tufiș, Dan Cristea and Sofia Stamou. 2004. *BalkaNet: Aims, Methods, Results and Perspectives. A General Overview.* Journal on Information Science and Technology, Special Issue on BalkaNet, Romanian Academy, 7 (1-2), 7–41.

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

Piek Vossen. 2002. *EuroWordNet general document version 3*. Report, University of Amsterdam.