The Romanian Corpus Annotated with Verbal Multiword Expressions

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

Mihaela Cristescu
UB, Faculty of Letters
Bucharest, Romania
mihaella.ionescu@yahoo.com

Mihaela Onofrei
ICS, RA, Iasi branch
UAIC, FCS
Iaşi, Romania
mihaela.plamada.onofrei@gmail.com

Abstract

This paper reports on the Romanian journalistic corpus annotated with verbal multiword expressions following the PARSEME guidelines. The corpus is sentence split, tokenized, part-of-speech tagged, lemmatized, syntactically annotated and verbal multiword expressions are identified and classified. It offers insights into the frequency of such Romanian word combinations and allows for their characterization. We offer data about the types of verbal multiword expressions in the corpus and some of their characteristics, such as internal structure, diversity in the corpus, average length, productivity of the verbs. This is a language resource that is important per se, as well as for the task of automatic multiword expressions identification, which can be further used in other systems. It was already used as training and test material in the shared tasks for the automatic identification of verbal multiword expressions organized by PARSEME.

1 Introduction

Recent years marked an intense international preoccupation with the multiword expressions within a multilingual community of specialists with multiple interests: linguistic description (Iñurrieta et al., 2018), classification (Savary et al., 2018), specific resources identification (Losnegaard et al., 2016) or development (Savary et al., 2018), syntactic annotation practice (Rosén et al., 2015) and recommendations (Rosén et al., 2016), processing (Constant et al., 2017), crosslingual comparison (Koeva et al., 2018; Barbu Mititelu and Leseva, 2018), etc. They were made possible thanks to the PARSEME Cost Action (Savary et al., 2015). Such activities continue older preoccupations with such lexical units, given their problematic nature on several aspects (Sag et al., 2002; Baldwin and Kim, 2010): meaning, inflexion, discontinuity, ambiguity, translatability, etc. (Savary et al., 2018), which motivate further investigations in different languages.

Within this effervescent context, the aim of this paper is to describe the creation of the Romanian corpus annotated with verbal multiword expressions (VMWEs), its characteristics and availability, as well as the VMWEs occurring in it, from several perspectives. Section 2 presents, on the one hand, the state of the art of the work done in the Romanian linguistics with respect to VMWEs, and, on the other hand, other international initiatives of annotating VMWEs in corpora. In section 3 we present the types of VMWEs defined within the PARSEME guidelines and applicable to Romanian. The characteristics of the annotated corpus are identified in section 4. It is followed by a brief description of the annotation process (section 5). The largest part of the paper is dedicated to the presentation of the VMWEs annotated in the corpus (section 6). We start with some frequency remarks on various VMWEs types within the multilingual context of the annotation, then focus on the diversity of the VMWEs in the Romanian corpus, their average length. Their internal structure is presented in a detailed way in the same section and then conclusions are drawn (section 7) and future work is envisaged.

2 Related Work

In the Romanian linguistics the analysis of multiword expressions is an old concern, dating back to the '50s (Ioaniţescu, 1956). Even since then has there been a special interest in the Romanian verbal multiword expressions (Dimitrescu, 1958). However, as remarked by Căpăţână (2007), throughout time, the authors have used a lot of different terms for referring to such linguistic units, there have been divergent opinions with respect to their definition, to their classification and their structural description. Nowadays the lexicologists' interest in this linguistic phenomenon is still

Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 13–21
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

not so strong, while the computational aspects have also been poorly studied. Todiraşcu et al. (2009) identified verb-noun collocations in a multilingual context using lexico-grammatical constructions specific to them. Todiraşcu and Navlea (2015) used verb-noun collocations extracted from a parallel French-Romanian corpus to improve machine translation. Rizea et al. (2016) studied multiword expressions from their interest in negative polarity items. Within PARSEME, a template describing Romanian VMWEs syntactic structure, fixedness/flexibility of their parts, and idiomaticity (lexical, syntactic, semantic, pragmatic and statistical) was created. Within the same Cost Action, another initiative was the organization of a shared task for automatic identification and classification of MWEs in corpora. The focus was only on verbal MWEs and representatives of many languages, Romanian included, joined the effort of creating the resource necessary for training and testing the systems participating in the competition (Savary et al., 2018).

PARSEME action is not the only initiative of annotating VMWEs in corpora: Kato et al. (2018) describe the annotation of VMWEs in an English journalistic corpus: it is rather the identification of a list of dictionary-based VMWEs and their labeling with a set of labels created on morphosyntactic grounds (verb-particle constructions, prepositional verbs, light verb constructions, verb-noun(-preposition), semi-fixed VMWEs). Vincze (2012) describes an English-Hungarian parallel corpus annotated with light verb constructions.

However, what makes the PARSEME action stand out is the multilingual perspective on VMWEs: the semantic, syntactic and morphological variations were considered in all the languages involved and unified annotation guidelines were created and used in the annotation of corpora for all these languages, allowing for interlingual comparison to a certain extent.

3 Romanian Verbal Multiword Expressions

Multiword Expressions (MWEs) are defined as "idiosyncratic interpretations that cross word boundaries" (Sag et al., 2002). They are considered "a pain in the neck for the NLP applications", due to their variation and discontinuities. Verbal MWEs are defined as "multiword expressions whose canonical form is such that their syntactic head is a verb V and their other lexical components form phrases directly dependent on V" (Savary et al., 2018).

Romanian participated in both preparatory phases of the PARSEME shared tasks. The results obtained, namely the identification of types applicable to Romanian and the corpus annotated with these types of VMWEs, got enhanced from version 1.0 to version 1.1 (the number of sentences in the corpus was increased with 5,203, which implied an increase with 1,351 of the number of annotated VMWEs) and the presentation that follows pertains to the latter. Out of the categories of VMWEs[1] defined in this edition, we present below only the ones applicable to Romanian:

1. universal categories - valid for all languages participating in the task. They are further divided into:
   (a) Light Verb Constructions (LVC), i.e. VMWEs consisting of a light verb and a noun denoting an event or state. Two subcategories are specified for them:
      i. LVC.full - in which the verb is semantically bleached: ex. *a avea acces* (to have access);
      ii. LVC.cause - in which the verb adds a causative meaning to the noun: ex. *a pune capăt* (to put end "to end");
   (b) Verbal Idioms (VID), including all VMWEs not belonging to other categories, and most often having a high degree of semantic non-compositionality: ex. *a o lua la goană* (to CLT3sgfemAcc take at rush "to run away");
2. quasi-universal categories - valid for some languages in the action. From this category only one type was annotated in Romanian:
   (a) Inherently Reflexive Verbs (IRV), which consist of a verb and a reflexive clitic. A VMWE is annotated as an IRV if (a) it never occurs without the clitic, or (b) the reflexive and non-reflexive versions of the verb have different meanings or subcategorization frames: ex. *a se face* (to SE make "to become"). The reflexive inflects for case (accusative and dative), for person and number in Romanian.

[1] See also http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/?page=030$_{C}ategories_{o}f_{V}MWEs$

Edition 1.1 of the PARSEME Shared Task also includes an experimental and optional VMWE category, called Inherently Adpositional Verbs (IAVs). Such an expression consists of a verb and a selected preposition or postposition. The annotation of this VMWE type is optional, since the overlapping with other categories can be quite frequent. Romanian has such verbs, for example *a conta pe* (to count on), but they were not annotated.

4 The Corpus

The Romanian corpus is a collection of articles from the concatenated editions of the Agenda newspaper. It was chosen because it raises no intellectual property rights problems, so that we could make it freely available in the PARSEME repository[2], edition 1.1, under a CC BY 4.0 license. There are several kinds of texts in it: columns, press releases, letters to the editor, news stories, feature stories, editorials, sports stories. Some repetitive constructions can be spotted, reflecting either the fixed style of the types of articles or the (permanent) authors' style. The average sentence length is about 18 tokens, which is very close to the average length of the sentences in CoRoLa, the representative corpus of contemporary Romanian (20.7 tokens/sentence) (Barbu Mititelu et al., 2018).

The corpus annotated with VMWEs is made up of 56,703 sentences containing 1,015,623 tokens, which makes it the biggest corpus in this action, however not the richest in VMWEs (Ramisch et al., 2018), as it contains only 5,891 VMWEs: their frequency in the corpus is 0.58% VMWEs (with 1 VMWE in 10 sentences).

The training and test files of the corpus were sentence split, tokenized, part-of-speech tagged and lemmatized with the TTL tool (Ion, 2007). The corpus was automatically syntactically annotated with UDPipe based on the romanian-ud-ro-2.0-170801.udpipe. The format of the corpus is *cupt* (Ramisch et al., 2018). Here is an example of a sentence from the corpus:

Ea se află acum în Timişoara.

She SE finds now in Timişoara.

"She is in Timişoara now."

The format of the file contains 11 columns: the first one identifies the position of each token in the sentence (here, from 1 to 7: six words and one punctuation mark). The second column contains the token, the third one - its lemma, the fourth column - its morphological category, the fifth - the morphosyntactic description of the token. For Romanian, the specifications for morphosyntactic description were created in the MULTEXT-East project (Dimitrova et al., 1998). The sixth column contains the same information in the format attribute=value, the seventh column identifies the syntactic head of the respective word by referring to its position in the sentence (column 1). The eighth column contains the name of the syntactic relation holding between the word and its head. The syntactic relations pertain to UD version 1.4. The ninth column contains no information (it is always_), the tenth one contains information only when the respective token is not followed by a blank space (usually when a punctuation mark follows or when words are hyphenated), while the last one contains the VMWEs annotation, when it is the case, otherwise it contains a star (*). Each occurrence of a VMWE in a sentence is counted starting from 1.

text = Ea se află acum în Timişoara.

```
1    Ea    el    PRON
Pp3fsr--------s    Case=Acc,Nom|
Gender=Fem|Number=Sing|Person=3|
PronType=Prs|Strength=Strong    3
nsubj    _    _    *

2    se    sine    PRON
Px3--a--------w    Case=Acc|
Person=3|PronType=Prs|Reflex=Yes|
Strength=Weak    3    expl:pv
_  _    1:IRV

3    află    afla    VERB
Vmip3s    Mood=Ind|Number=Sing|
Person=3|Tense=Pres|VerbForm=Fin
0    root    _    _    1

4    acum    acum    ADV    Rgp
Degree=Pos    3    advmod    _
_    *

5    în    în    ADP    Spsa
AdpType=Prep|Case=Acc    6    case
_    _    *

6    Timişoara    Timişoara
```

[2] https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2842

```
PROPN     Np       _      3      obl       _
SpaceAfter=No      *

7     .     .     PUNCT     PERIOD
_     3     punct      _      _     *
```

In order to determine the span of a VMWE all its components contain the same number on the last column. Only for the first element (in linear order) is this number followed by the VMWE type label. When one word belongs to two VMWEs (overlapping VMWEs), it bears two numbers: in the sentence beginning rendered below (*Când s-a lăsat întunericul...* When S-has left dark-the... "When it got dark...") we can see the VMWE *s- lăsat* (IRV) is part of the VMWE *s- lăsat întunericul* (VID). There are 53 such cases of overlapping VMWEs in the corpus, affecting not more than two words of each of the overlapping expressions (Savary et al., 2018).

```
1     Când     când     ADV     Rw
PronType=Int,Rel     4     advmod
_     *

2     s-     sine     PRON
Px3--a--y-----w     Case=Acc|
Person=3|PronType=Prs|Reflex=Yes|
Strength=Weak|Variant=Short     4
expl:pv     _     SpaceAfter=No
1:IRV;2:VID

3     a     avea     AUX     Va--3s
Number=Sing|Person=3     4     aux
_     _     *

4     lăsat     lăsa     VERB
Vmp--sm     Gender=Masc|Number=
Sing|VerbForm=Part     13     advcl
_     _     1;2

5     întunericul     întuneric
NOUN     Ncmsry     Case=Acc,Nom|
Definite=Def|Gender=Masc|Number=
Sing     4     nsubj     _
SpaceAfter=No     2
```

5 The Annotation Process

The annotation process was performed by a team of three native speaker linguists, according to the PARSEME guidelines[3], edition 1.1, using a dedicated web platform, FLAT[4].

The annotation process consists of two stages: the identification of a VMWE and its classification into one of the aforementioned categories. A number of Structural Tests have been defined, in order to help the annotators determine the type of a VMWE. The annotation was followed by consistency check and homogenization with the help of a tool developed and made available by the shared task organizers (Savary et al., 2018), improving the results: inconsistency among annotators were eliminated, skipped VMWEs were found and annotated, incorrectly identified VMWEs were unannotated.

A set of 2,503 sentences was double-annotated and it was used by the organizers of the shared task for calculating the inter-annotator agreement scores (Ramisch et al., 2018) in order to assess the quality of the annotation, as well as the task difficulty. Two aspects were considered: VMWE span and their categorization. For the former, the F_{span} score, i.e. the MWE-based F-measure when considering that one annotator tries to predict the other one's annotation, is 0.533, while K_{span}, i.e. the agreement between annotators on the VMWE span, is 0.491.

Table 1 provides statistics of the Romanian corpus annotated for the edition 1.1 of the PARSEME Shared Task.

Entity	Number
Sentences	56,703
Tokens	1,015,623
VID	1,611
LVC.full	313
LVC.cause	183
IRV	3,784
TOTAL VMWEs	5,891

Table 1: Statistical data about the Romanian corpus.

6 Characteristics of the Annotated Romanian VMWEs

As compared to other languages. As seen in Table 1, the category IRV is the best represented in the Romanian corpus. Reflexive verbs are the

[3] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/?page=home

[4] https://flat.readthedocs.io/en/latest/

most frequent type of VMWEs also in Bulgarian, Spanish and Polish, according to the data provided by the shared task v. 1.1 organizers[5]. However, it is interesting that the Romance languages (Romanian among them) participating in the task display differences both with respect to the types of VMWEs they contain and to their distribution in the corpora. On the other hand, we have to keep in mind the fact that these corpora do not contain the same kind of texts (Ramisch et al., 2018) or their type is even unknown (Savary et al., 2018). Even so, we can say that Romanian stands alone among Romance languages and displays characteristics of some Slavic languages in this respect.

Diversity in the corpus. The 5,891 occurrences of VMWEs are forms of 486 unique VMWEs, as seen in Table 2: the second column presents the total number of occurrences of each type in the corpus, the third column – the number of unique VMWEs of each type, the fourth one – the relative frequency of each type, while the last column contains the number of VMWEs occurring only once in the corpus (hapax legomena). We can see the high frequency of each VMWE type. This correlates with the repetitive nature of the texts in the corpus, as mentioned in section 4. Moreover, 122 (about a quarter) of all VMWEs are hapax legomena. This implies an even higher real relative frequency of the other VMWEs. This distribution is suggestive of the low diversity of VMWEs in the corpus (see also the discussion about verbs productivity in VMWEs in section 6).

Type	#occ.	unique	% occ.	#hapax
VID	1,611	171	9	65
LVC. full	313	39	8	8
LVC. cause	183	8	22	3
IRV	3,784	268	14	46
TOTAL	5,891	486	12	122

Table 2: Distribution of VMWEs in the corpus.

Here is a list with the most frequent 5 VMWEs of each category: between brackets we noted the frequency of each VMWE. One can notice that there are several very frequent ones, especially IRVs, and the frequency drops drastically with the second (in case of LVC.full), the third (in case of LVC.cause) or of the fifth (in case of VID) expression in the series:

VID: *avea loc* (have place "take place") (683), *avea dreptul* (have right-the "have the right") (104), *avea în vedere* (have in sight "have in mind") (81), *fi vorba* (be speech "be about") (79), *trimite în judecată* (send in judgement "send to court")(28);

IRV: *se desfăşura* (SE unfold "take place") (432), *se afla* (SE found "exist") (296), *se adresa* (SE address "address") (201), *se putea* (SE can "be possible") (190), *se prezenta* (SE present "go") (112);

LVC.full: *face parte* (make part "be part") (127), *lua parte* (take part) (27), *lua decizie* (take decision "make a decision") (19), *avea acces* (have access) (19), *lua hotărâre* (take decision "make a decision") (13);

LVC.cause: *pune la dispoziţie* (put at disposal "make available") (92), *pune în vânzare* (put in sale "put up for sale") (68), *pune capăt* (put end "put an end") (9), *pune în circulaţie* (put in circulation "circulate") (6), *pune în pericol* (put in danger) (3).

Average length. The examples given above also show the reduced length of Romanian VMWEs. Their average length is 2.15 words per VMWE. The longest VMWE is a VID: *bea până la ultima picătură paharul amar* (drink to at last drop glass-the bitter "suffer to the very end"). However, as shown by Savary et al. (2018), this is the case with almost all languages in the initiative. The discussion below about the internal structure of VMWEs sheds more light on the understanding of the Romanian VMWEs length.

Verbs productivity. There are two verbs that occur in three types of VMWEs: *da* (give) and *pune* (put). Their productivity in each VMWE types is rendered in Table 3. Noteworthy, they are the only verbs creating LVC.cause expressions in this corpus.

Verb	VID	LVC.full	LVC.cause
da	23	7	2
pune	22	4	6

Table 3: Productivity of two verbs

With respect to LVC.full, there is one verb more

[5] The distribution of the types and frequency of VMWEs in these languages are available at http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_04_LAW-MWE-CxG_2018___lb__COLING__rb__&subpage=CONF_40_Shared_Task

productive than them: *face* ("do/make") occurs in 12 different LVC.full expressions. The verb *lua* ("take") is as productive as *da*: it heads 7 expressions. Other verbs in LVC.full VMWEs are: *avea* ("have") – productivity: 6, *aduce* ("bring") – productivity: 2, and *intra* ("enter") – productivity: 1. With respect to *aduce* we remark the fact that the two expressions it heads are synonymous: *aduce contribuţia* and *aduce aportul* ("bring contribution").

As far as VID MWEs are concerned, they are, on the one hand, the most numerous among the VMWEs if we exclude IRVs, and, on the other hand, they display a large variety of head verbs: there are 47 different verbs heading VIDs, 22 of them occurring in only one expression. Besides *da* and *pune*, which are the most productive for this type, the next five most productive ones are: *lua* ("take") – 17 VIDs, *avea* ("have") – 13 VIDs,*face* ("do/make") – 13 VIDs, *ţine* ("hold") – 6 VIDs, and *aduce* ("bring") – 5 VIDs.

We cannot discuss of verbal productivity in case of IRVs.

Internal syntactic structure. With respect to **IRV**s, we can mention that in Romanian they may take either an Accusative or a Dative clitic. In this corpus, most of them take an accusative clitic, which reflects, in fact, their general occurrence in language.

However, we can identify several internal structures in case of LVC.cause, LVC.full and VID expressions.

LVC.cause. Although neither frequent nor numerous, the expressions of this type display one of the two internal structures:

1. verb + indefinite noun (functioning as a direct object): e.g. *da foc* (give fire, "put on fire"). This structure is displayed by 3 VMWEs;

2. verb + preposition + indefinite noun: e.g.: *pune în circulaţie* (put in circulation "circulate"). This structure is displayed by 5 VMWEs.

LVC.full. They display the same two types of structures as LVC.cause. However, what distinguishes them is the fact that these structures show some variation in the case of LVC.full.

1. The structure verb + noun presents the following subtypes:
   (a) verb + definite singular noun: e.g. *face apariţia* (make appearance-the "appear") – there are 8 such VMWEs;
   (b) verb + indefinite singular noun: e.g.: *da citire* (give reading "read") – there are 17 such VMWEs;
   (c) verb + noun (without restriction on its form): e.g.: *da declaraţie* (give declaration "declare") – there are 8 such VMWEs;
   (d) verb + indefinite plural noun: e.g.: *da asigurări* (give assurances "assure") – one such VMWE was found;

2. The structure verb + preposition + indefinite singular noun does not have any subtypes: e.g.: *intra în coliziune* (enter in collision "collide") – 5 expressions display this structure.

The structure without preposition is more frequent than the one with preposition in the case of LVC.full, whereas in the case of LVC.cause the one with preposition is more frequent.

VID. This type of VMWEs is characterized by internal structural variation: most VIDs are short, containing 2 words, one of them being the verb. The first two structures below are the most frequent, while the others are attested by several VMWEs:

1. verb + noun: 81 VIDs. Several subtypes can be distinguished:
   - the noun is the subject of the verb: 4 VIDs: e.g.: *fura somnul* (steal sleep-the "fall asleep");
   - the noun is the direct object of the verb: 77 cases. The noun can be:
     - syntactically unmodified: 65 cases: e.g: *prinde viaţă* (catch life "come to life");
     - modified by an adjective (in the canonical word order in Romanian, i.e. noun + adjective): 4 cases: *da undă verde* (give wave green "give the go-ahead");
     - modified by preposition + noun: 4 cases: *aduce o rază de lumină* (bring a ray of light "bring hope");
     - modified by a genitive: 2 cases: *vedea lumina zilei* (see light-the day-of-the "be born");

    – modified by a defining relative clause: 2 VIDs, which are, in fact, synonyms: *face tot ce stă în putere* and *face tot ce stă în putinţă* (make all that stay in power "do one's best");

2. verb + prepositional phrase (PP): 72 cases with the following subtypes:

    - the PP is made up of a preposition and a noun: 65 cases: *înceta din viaţă* (cease from life "die"); in 2 of these cases the non-anaphoric feminine accusative personal clitic *o* functions as an expletive: *o lua de la capăt* (CL3SgFemAcc take from end "start again");
    - the PP is made up of a preposition and a modified noun: it can be modified by an adjective, by a genitive noun or by a prepositional phrase - 4 cases: e.g., *nu privi cu ochi buni* (not watch with eyes good "disfavour"). Notice here the negative form of the VMWE, which is mandatory;
    - the PP is made up of a preposition and an adjective: 2 cases: *trece la cele veşnice* (pass to the eternal "die"). The demonstrative determiner is obligatory in this VMWE, but this is not the case with all PPs of this kind;
    - the PP is made up of a preposition and a participle: 1 case: *lăsa de dorit* (leave of desired "fall short");

3. verb + two syntactic arguments: 10 cases. Several subtypes exist here as well:

    - direct object and indirect object: only one such VID could be found: *pune capăt vieţii* (put end life-to "commit suicide");
    - subject and a PP functioning as a place adverbial: 3 cases: *îngheţa sângele în vine* (freeze blood-the in veins "get cold feet");
    - direct object and a PP: 6 cases: e.g. *găsi drumul în viaţă* (find road-the in life "find one's way in life"). We found one VID in which: the PP precedes the direct object, the PP contains a compound preposition, the noun in the PP is preceded by a prenominal adjective, the direct object noun is modified by an adjective: *bea până la ultima picătură paharul amar* (drink up to last drop glass-the bitter "suffer to the very end");

4. verb + adverb: 4 cases: *da afară* (give outside "remove from job, eliminate");

5. varia - there are 4 VIDs that have various structures that do not fall under any of the previous types and we will not detail them here.

7 Conclusions and Future Work

In this paper we presented the Romanian PARSEME corpus annotated with VMWEs in the edition 1.1 of the shared task. The corpus offers insights into the use of VMWEs in a journalistic corpus made up of concatenated editions of the same newspaper. The characteristics identified for the VMWEs are not meant to be a general characterization of Romanian VMWEs, they pertain only to the expressions occurring in this corpus.

Such a corpus-based study completes the lexicon-based ones (Căpăţână, 2007) or the general, descriptive ones. In a multilingual context, we offer not only descriptions of Romanian VMWEs of preestablished types, but we also notice frequencies of types and productivity of head verbs. The analysis could be extended with morphological or syntactic remarks on the behaviour of these verbs: how grammatical categories are blocked by the participation to such expressions, how selectional restrictions are affected by this, what syntactic alternations, such as voice, are also blocked, etc.

Given the universal annotation guidelines, the corpus can be used in comparative linguistic studies, from various perspectives revealed by the data.

8 Acknowledgements

This work was supported by a grant of the Romanian Ministry of Research and Innovation, PCCDI - UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0818/73, within PNCDI III.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions.

Verginica Barbu Mititelu and Svetlozara Leseva. 2018. *Derivation in the domain of multiword expressions*, pages 215–246. Language Science Press, Berlin.

Verginica Barbu Mititelu, Dan Tufiş, and Elena Irimia. 2018. The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mathieu Constant, Glen Eryiit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Cecilia Căpăţână. 2007. *Elemente de frazeologie*. Editura Universitaria, Craiova.

Florica Dimitrescu. 1958. *Locuţiunile verbale în limba română (The verbal locutions in Romanian)*. EA, Bucureşti.

Ludmila Dimitrova, Tomaz Erjavec, Nancy Ide, Heiki Jaan Kaalep, Vladimir Petkevic, and Dan Tufis. 1998. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern European languages. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 315–319, Montreal, Quebec, Canada. Association for Computational Linguistics.

Uxoa Iñurrieta, Itziar Aduriz, Ainara Estarrona, Itziar Gonzalez-Dios, Antton Gurrutxaga, Ruben Urizar, and Iñaki Alegria. 2018. Verbal multiword expressions in basque corpora. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 86–95, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Eugen Ioaniţescu. 1956. Locuţiunile. *Limba română*, 6:48–54.

Radu Ion. 2007. *Word Sense Disambiguation Methods Applied to English and Romanian*. PhD Thesis, Romanian Academy.

Akihiko Kato, Hiroyuki Shindo, and Yuji Matsumoto. 2018. Construction of large-scale english verbal multiword expression annotated corpus. In *Proceedings of the 11th Language Resources and Evaluation Conference*, pages 2495–2499, Miyazaki, Japan. European Language Resource Association.

Svetla Koeva, Cvetana Krstev, Duko Vitas, Tita Kyriacopoulou, Claude Martineau, and Tsvetana Dimitrova. 2018. *Semantic and syntactic patterns of multiword names: A cross-language study*, pages 31–62. Language Science Press, Berlin.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartn, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. PARSEME Survey on MWE Resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2299–2306.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Monica-Mihaela Rizea, Gianina Iordăchioaia, and Frank Richter. 2016. A collocational approach to romanian strong negative polarity items. In *Proceedings of the 12th International Conference Linguistic Resources and Tools for Processing the Romanian Language*, pages 173–185.

Victoria Rosén, Koenraad De Smedt, Gyri Smørdal Losnegaard, Eduard Bejcek, Agata Savary, and Petya Osenova. 2016. MWEs in Treebanks: From Survey to Guidelines. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2323–2330, Paris, France. European Language Resources Association (ELRA).

Victoria Rosén, Gyri Smørdal Losnegaard, Koenraad De Smedt, Eduard Bejček, Agata Savary, Adam Przepiórkowski, Petya Osenova, and Verginica Barbu Mititelu. 2015. A survey of multiword expressions in treebanks. In *Proceedings of the 14th International Workshop on Treebanks & Linguistic Theories conference*, pages 179–193, Warsaw, Poland.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejek, Fabienne Cap, Slavomr pl, Silvio Ricardo Cordeiro, Glen Eryiit, Voula Giouli, Maarten van Gomple, Yaakov HaCohen-Kerner, Jolanta Kovalevskaite, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartn, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and

Veronika Vincze. 2018. *PARSEME multilingual corpus of verbal multiword expressions*, pages 87–147. Language Science Press, Berlin.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Mathieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland.

Amalia Todiraşcu, Christopher Gledhill, and Dan Stefanescu. 2009. Extracting collocations in contexts. In *Human Language Technology. Challenges of the Information Society*, pages 336–349, Berlin, Heidelberg. Springer Berlin Heidelberg.

Amalia Todirascu and Mirabela Navlea. 2015. Aligning Verb+Noun Collocations to Improve a French - Romanian FSMT System. In *MUMTTT workshop*, pages 82–99, Malaga, Spain.

Veronika Vincze. 2012. Light verb constructions in the SzegedParalellFX English–Hungarian parallel corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2381–2388, Istanbul, Turkey. European Language Resources Association (ELRA).