A Romanian Treebank Annotated with Verbal Multiword Expressions

Verginica Barbu Mititelu
Romanian Academy
RACAI
vergi@racai.ro

Mihaela Cristescu
University of Bucharest
mihaela.ionescu@litere.unibuc.ro

Maria Mitrofan
Romanian Academy
RACAI
maria@racai.ro

Bianca-Mădălina Zgreabăn
Utrecht University
madalinazgreaban0@gmail.com

Elena-Andreea Bărbulescu
University of Bucharest
adabarbulescu7@gmail.com

Abstract

In this paper we present a new version of the Romanian journalistic treebank annotated with verbal multiword expressions of four types: idioms, light verb constructions, reflexive verbs and inherently adpositional verbs, the last type being recently added to the corpus. These types have been defined and characterized in a multilingual setting (the PARSEME guidelines for annotating verbal multiword expressions). We present the annotation methodologies and offer quantitative data about the expressions occurring in the corpus. We discuss the characteristics of these expressions, with special reference to the difficulties they raise for the automatic processing of Romanian text, as well as for human usage. Special attention is paid to the challenges in the annotation of the inherently adpositional verbs. The corpus is freely available in two formats (CUPT and RDF), as well as queryable using a SPARQL endpoint.

Keywords: multiword expressions, Romanian, inherently adpositional verbs, idioms, light verb constructions.

1 Introduction

Language resources of the type electronic corpora annotated with syntactic information (most of the times on top of lexical and morphological annotations), i.e. treebanks, are now quite common for languages and even dialects. If a decade ago the number of treebanks for various languages was rather scarce, now we can find many such resources, though still of a modest size. The situation has greatly improved given the existence of two major multilingual initiatives: Universal Dependencies[1] (UD) (Nivre et al., 2016; de Marneffe et al., 2021) and PARSEME Cost Action (Savary et al., 2015), two open community efforts, active in improving and enhancing their results. UD is an initiative created with the aim of offering the instruments for a cross-lingual description at the morphologic and syntactic levels. Seventeen universal parts of speech (e.g., NOUN, VERB, AUX, PRON, ADJ, etc.) and a set of morphological features (e.g., Number, Gender, Tense, etc.) are used for the morphologic level, and 37 universal relations (e.g., `nsubj` for the nominal subject, `csubj` for the clausal subject, `obj` for the nominal direct object, `ccomp` for the clausal direct object, etc.) are defined for the syntactic description. These morphologic instruments are considered enough for the description of any language, while the inventory of syntactic relations is admittedly universal, but subtypes of the 37 universal relations are accepted for a more specific syntactic analysis: e.g., `nsubj:pass` for the nominal subject in passive constructions for the languages that do have passive; 26 such subtypes have been defined so far, which are specific to one or more languages. In its last release (May 2022), UD boasts 228 treebanks for 130 languages, all freely available.

The existence of treebanks for various languages released through UD has offered the premise for the development of automatic tools (Straka et al., 2016) that can be trained on these treebanks and further used to annotate new corpora. This paved the way to initiatives such as PARSEME, in which new corpora, collected according to certain requirements (such as text genre, size, license, etc.), were automatically morphosyntactically annotated with such tools and further enriched with a new level of annotation, i.e. semantic: verbal multiword expressions (VMWEs) were manually annotated following the same guidelines for all languages, that identify universal, quasi-universal and language-specific VMWE types. Within PARSEME, treebanks for 26 languages were annotated and one of them is for Romanian.

There are already several treebanks for Roma-

[1] universaldependencies.org

nian freely available: within UD, there is the Romanian Reference Treebank (RRT) (Barbu Mititelu, 2018) (containing sentences from various text genres), Romanian Non-Standard treebank (Colhon et al., 2017) (containing sentences from old texts or from folklore), the medical treebank SiMoNERo (Barbu Mititelu and Mitrofan, 2020) (which has an extra annotation level: medical entities of the types anatomical parts, chemicals, disorders and procedures) and the treebank of the Aromanian dialect of Romanian ArT (Barbu Mititelu et al., 2021). There is also another treebank, unavailable in UD, LegalNERO (Păiș et al., 2021), which has a further level containing gold annotations for five entity classes: organizations, locations, persons, time expressions and legal resources mentioned in legal documents.

In this paper we present a new version of the Romanian treebank, whose annotation started in PARSEME and which has recently been enriched with a new type of verbal expressions, i.e. inherently adpositional verbs. We call this corpus *PARSEME-Ro*. We will first outline the context of development of this corpus, namely the PARSEME shared tasks (Section 2), then present some idiosyncrasies displayed by the verbal expressions occurring in the corpus (Section 3). We describe the corpus itself: its levels of annotation (Section 4) and problems raised by annotating the new type of verbal expressions. Some general statistics about the corpus and statistics about the VMWE types in the corpus are given in Section 6, before concluding the paper.

2 Context of development

PARSEME is an international and multilingual community aiming at identifying MWEs in running texts. Although so far the interest has manifested only for verbal MWEs in a concerted way, MWEs of other morphological classes will also be approached in a multilingual perspective.The PARSEME shared tasks editions 1.0 (Savary et al., 2017), 1.1 (Ramisch et al., 2018) and 1.2 (Ramisch et al., 2020) focused on the identification of VMWEs because of their challenging features: complex structure, discontinuity, variability, ambiguity (Savary et al., 2017). The main aim of this initiative is to eventually automatically recognize VMWEs in corpora. The annotation guidelines are unified across languages and have been enhanced from edition 1.0 (Savary et al., 2017) to edition 1.1 (Ramisch et al., 2018).

Based on the experience gathered in the annotation for edition 1.0, as well as on the types of VMWEs identified in the corpora, starting with edition 1.1 of PARSEME, the following types of VMWEs have been annotated (Savary et al., 2018):

- *Universal categories*, that are valid for all languages participating in the task:
  - **Light verb constructions** (LVCs) with two subcategories:
    - **LVC.full**, in which the verb is semantically totally bleached: EN *to give a lecture*, RO *a lua o decizie* (to make a decision), *a face parte* (to be part);
    - **LVC.cause**, in which the verb adds a causative meaning to the noun: EN *to give a headache*, RO *a da bătăi de cap* (to give a bad time), *a pune capăt* (to pun an end);
  - **Verbal idioms** (VIDs), which have at least two lexicalized components including a head verb and at least one of its dependents and is characterised by a high degree of semantic non-compositionality: EN *to go bananas*, RO *a trage pe sfoară* (to double-cross), *a o lua la goană* (to start running);
- *Quasi-universal categories*, valid only for some languages:
  - **Inherently reflexive verbs** (IRVs), in which the reflexive clitic either always co-occurs with a given verb or changes its meaning or subcategorization frame: EN *to help oneself*, RO *a se gândi* (to think), *a se face* (to become);
  - **Verb-particle constructions** (VPC), which are made up of a verb and a particle: EN *to do in, to eat up*; this type is not applicable to Romanian;
  - **Multi-verb constructions** (MVC), which are made up of two verbs: EN *to let go, to make do*); neither is this type applicable to Romanian;
- *Language-specific categories*, valid only for the language for which they are defined, unless other languages claim them as well: so far, only one such type has been defined, namely **inherently clitic verbs** for Italian: it consists of a verb and one or more non-reflexive clitics

that represent the pronominalization of one or more complements: IT *infischiarsene* (not to worry about);

- *Experimental category*, annotated in the post-annotation step: **Inherently adpositional verbs** (IAVs), made up of a verb and a preposition: EN *to rely on*, RO *a conta pe* (to rely on).

For each language, a team of linguists was trained to apply the guidelines[2] for identifying VMWEs in a corpus and classifying them into one of the existing categories. Simultaneously, quality of the annotation was acquired by applying semi-automatic methods for ensuring full coverage of the VMWEs in the corpus, as well as for their consistent classification.

This is the context in which the creation of PARSEME-Ro took place, alongside corpora annotated with VMWEs for other languages.

The three editions of the PARSEME Cost Action (1.0, 1.1, 1.2) covered 18, 20, and 14 languages, respectively, from several language families: Romance languages (French, Italian, Portuguese, Romanian, Spanish), Balto-Slavic languages (Bulgarian, Czech, Croatian, Lithuanian, Polish, Slovene), Germanic languages (German, English, Swedish, Yiddish), and other languages (Arabic, Greek, Basque, Farsi, Maltese, Hebrew, Hindi, Hungarian, Turkish, Chinese, Irish).

All the annotated corpora from the editions 1.0[3], 1.1[4] and 1.2[5] are available for download, under the Creative Common license.

3 Characteristics of VMWEs Contributing to their (Automatic) Processing Difficulty

MWEs are defined as "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al., 2002). They are "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity" (Baldwin and Kim, 2010).

The identification of VMWEs in texts is a well-known challenge for NLP applications, because of their special characteristics, including discontinuity, overlaps, non-compositionality, heterogeneity, and syntactic variability. They are problematic not only for machines, but also for humans: on the one hand, for students learning a second language and, on the other, for native speakers who are exposed to rarer expressions.

One key characteristic of a VMWE is for it to be idiomatic. This property refers to "markedness or deviation from the basic properties of the component lexemes, and applies at the lexical, syntactic, semantic, pragmatic, and/or statistical levels" (Baldwin and Kim, 2010).

Lexical idiomaticity is displayed when one or more components of a VMWE are not *part of the conventional lexicon or are not used outside the respective VMWEs*: for Romanian, this is the case of the boldfaced words in the VMWEs *a-și aduce* ***aminte*** ('to remember') or *a avea* ***habar*** ('to have a clue'). Various Romanian VMWEs conserve lexical or semantic archaisms as the boldfaced words in the following expressions show: *aduce* ***aminte*** ('remind'), *nu da în* ***brânci*** ('have a soft job'), *da* ***ortul*** *popii* ('die'), *nu avea* ***habar*** ('have a clue'), *veni de* ***hac*** ('bear down'), *băga de* ***seamă*** ('notice'), *nu da inima* ***ghes*** (be reluctant to), *scoate la* ***iveală*** ('reveal'), *lua la* ***rost*** ('chide'), etc.

On the *morphological* level, there are VMWEs that display restrictions on the paradigmatic realization of the verbal head with respect to one or more morphosyntactic features, such as person, number, tense, mood, polarity, etc. or with respect to possible derived forms: e.g. RO *a nu privi cu ochi buni* (not watch with eyes good, 'to regard with disfavour') is always used with the negative marker *nu* 'not'. In addition, there are VMWEs that contain obsolete inflectional and derived forms, such as *a bate* ***câmpii*** (beat the fields 'to beat around the bush') (in which *câmpi*[6] is an old plural form of the word *câmp*, whose current plural form is *câmpuri*), *a pune pe* ***roate*** (pun on wheels 'to get on its feet') (in which *roate* is an old plural form of the word *roată*, whose current plural form is *roți*), *a lua cu* ***binișorul*** (take with wellness_DIMINUTIVE 'to let down easily') (the diminutive noun *binișorul* is not currently used outside expressions) (Căpățână, 2007).

Syntactic idiomaticity arises when the syntax of the VMWE is not derived directly from that of its components. The syntactic level of description

[2] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=home

[3] https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2282

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

[5] https://gitlab.com/parseme/sharedtask-data/-/tree/master/1.2

[6] The form *câmpii* is the definite one for *câmpi*.

would include any restrictions on the word order of the VMWE components and of the possible dependents. For example, in the RO VMWE *a da ortul popii* (give coin-the to priest-the 'to die') the object *ortul* always precedes the indirect object *popii*, though Romanian allows for any order of the direct and indirect object in case of their co-occurrence (though with different pragmatic salience in each case).

Semantic idiomaticity means non-compositionality of the expression, i.e. the meaning of a MWE is not explicitly derivable from the semantics of its parts. VMWEs displaying semantic idiomaticity have frequently components with metaphoric (*a lua taurul de coarne* take the bull of horns 'to take the bull by the horns'), hyperbolic (*a crăpa de frig* crack by cold 'to be very cold') or metonymic (*a nu ridica un deget* not lift a finger 'not to lift a finger') meaning in addition to their literal meaning. Semantic idiomaticity may imply either the fact that the expression's meaning is given rather by one of the components (see the descriptions for LVCs in Section 2) or the fact that the global sense of the expressions has nothing to do with the senses of the components: e.g. the words making up the VID *a tăia frunză la câini* (cut leaf for dogs 'to dally') have no semantic relation to the sense of the expression.

Pragmatic idiomaticity occurs when a VMWE is associated with a fixed set of situations or a particular context of use: see the case of greetings that are specific to the different parts of the day: e.g. EN *good morning*, RO *noapte bună* (night good 'good night'), etc.

Statistical idiomaticity is triggered by the high frequency a particular combination of words occurs with: e.g. EN *black and white* is semantically equivalent to RO *alb-negru* (white-black), in spite of the lack of the conjunction and of the reversed order of the two components.

All these characteristics of expressions may prevent their correct automatic interpretation, but also their understanding in inter-human communication, needless to say their grammatically correct and semantically adequate usage by second language learners. These justify the necessity for (computational) linguists' focusing more on phraseology. The insufficient attention paid to them leads to inconsistent terminology, inconsistent treatment of such units in lexicography, partial descriptions in grammars and dictionaries and little focus on it in textbooks, though, admittedly, expressions are a touchstone of language command.

4 Annotation Levels

The PARSEME corpus for Romanian (PARSEME-Ro) is journalistic and was automatically tokenized, part-of-speech tagged, lemmatized and syntactically parsed using UDPipe (Straka et al., 2016) trained on RRT. In a first step (consisting of all three annotation phases pertaining to the participation in the three editions of the shared tasks), the annotation of the different types of VMWEs was manual: the annotators identified and classified the VMWEs belonging to the LVC.full, LVC.cause, VID and IRV types. In the first edition four annotators were involved, in the second one there were three, and two participated to the last edition. Each annotator had a portion of the morpho-syntactically processed corpus to annotate: using the FLAT platform[7] (Savary et al., 2017), their task was to read the text, to spot a potential VMWE and, using the decision tree and the battery of tests from the PARSEME guidelines, to decide if the respective string was indeed a VMWE and specify its type. Only for a small portion of the data (2500 sentences) was the annotation double, so as to measure the agreement between annotators (Savary et al., 2017; Ramisch et al., 2018).

In a second, recent, step, IAVs were annotated in PARSEME-Ro. This time, the annotation was automatic, followed by manual validation and correction, in two phases. Starting from the list of 1,725 adpositional verbs created by Geană (2013), all their occurrences in the corpus were identified and annotated as IAVs. This was done automatically by using a Python script that performed a global search of the IAVs tokens within the corpus text. This search was enhanced to include a span window in order to capture situations where other tokens were interleaved with the actual IAV in the corpus text. In cases where several matches were found for one of the tokens of the IAV (this applies mostly to prepositions) the principle of the minimum distance length between the tokens was used. Finally, based on these matches, the corpus tokens found to correspond to an IAV were automatically annotated. Then the first phase of the manual validation and correction step followed: two annota-

[7] github.com/proycon/flat, flat.science.ru.nl

tors, students in linguistics, were presented with all automatically annotated instances and, using an annotation platform, they could modify the annotations in the sense of deleting expressions or adjusting their size (i.e. adding or removing parts), using the BRAT tool (Stenetorp et al., 2012), integrated in the RELATE platform (Păiș et al., 2020).

Several sources of errors could be identified in the automatic annotation of IAVs:

- homonyms that had been erroneously part-of-speech-tagged as verbs: adjectives with participle origin (***scutite*** *de la plată* 'exempted from payment'), nouns zero-derived from participles (*în* ***trecut*** *la* 'in the past at'), etc.;

- ambiguity: the structure verb + adposition is ambiguous between an IAV and a mere combination with a different meaning from that specific to the IAV construction: the combination *a se lovi de* (REFL.CL hit of 'to bump into') is an IAV in a sentence like *Copilul s-a lovit de perete.* (Child-the REFL.CL-has hit of wall 'The child bumped into the wall.'), but not in the sentence *Copilul s-a lovit de dimineață* (Child-the REFL.CL-has hit of morning 'The child got hit in the morning.'), where the same preposition introduces a time adverbial. A particular example of this type is represented by constructions that are structurally similar to prepositionally marked direct objects: e.g. *a* ***lăsa pe*** (leave on 'to bend on') (as in *Ion s-a lăsat pe spate.* 'John leaned back.') as opposed to *lăsa pe cineva* (leave/let someone): as in ***lăsând****-o orfană* ***pe*** *micuța Ornella* (leaving-CL3SgFem orphan PE little Ornella 'leaving little Ornella orphan'), where PE is a marker of the direct object;

- overgeneration: the presence of the adposition in the context of the verb, although syntactically belonging to a phrase without direct dependence on the verb, is misinterpreted as being part of an IAV: *a lua două șunci* ***de*** *porc* (take two hams of pork): here, *de* is a preposition linking the noun *pork* to its nominal head *șunci*, not to the verb;

- the combination verb – adposition is already part of another VMWE: *a da în judecată* (give in trial 'to sue') is already classified as VID, thus no IAV is annotated in this case;

		correctly annotated	
#	Total IAVS	#	%
AUTO annot.	4,686	3,128	66.75
annot. 1	3,462	3,085	89.11
annot. 2	3,519	3,185	90.5
both annots	-	2,981	
gold IAVs	-	**3,311**	

Table 1: General statistics of the IAV annotation process

- using the span window to match IAVs that have interleaved tokens has made the algorithm match false-positives to a high degree (34% of all automatically annotated IAVs).

Consistency of annotation was ensured differently for each step: for the annotation in the context of the shared tasks, we used the consistency checking tools made available by the organizers (Savary et al., 2018), helping to spot the skipped occurrences of VMWEs, as well as inconsistent type assignment of the same VMWE.

For the step involving the annotation of IAVs, we envisaged a second phase of the validation and correction step: all cases of agreement between the two student annotators were considered correct decisions (see the 2,981 cases marked as "both annots" in Table 1). All cases of disagreement between them were further checked by two linguists experienced in the PARSEME annotation. Table 1 shows that two thirds of the automatically annotated IAV are actually correct IAVs and that the decision to automatically annotate them was a time saving one. They represent 94.47% of the IAVs that should have been annotated, i.e. of the cases called "gold IAVs" in the table and which are the result of the experienced annotators' validation and correction of the two student annotators' validations. Each individual initial manual validation covers almost 90% of all correct cases: see the last column of lines two and three in Table 1.

5 Defining and refining the class of IAVs annotated in PARSEME-Ro

PARSEME guidelines 1.2 define an IAV as follows: "It consists of a verb or VMWE and an idiomatic selected preposition or postposition that is either always required or, if absent, changes the meaning of the verb of VMWE significantly."[8]. Their annotation is done after the annotation of other VMWEs,

[8] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=050_

because (i) adpositional verbs occurring within other VMWEs should not be annotated as IAVs, and (ii) VMWEs can also be adpositional, just like verbs. The annotation of IAVs in the PARSEME-Ro corpus aimed at marking only the adpositional verbs for the time being, so as to serve as an exercise that would reveal the challenges this type of constructions raises.

PARSEME guidelines offer only one test for IAVs, which is meant to show the semantic difference between the verb occurring with the adposition and its use without it: if, in response to a declarative statement containing the potential IAV, a question cannot be asked about the circumstances of the verbal event using the verb, but not the adposition, then the combination verb – adposition is annotated as IAV.

Geană (2013: p. 46) defines adpositional verbs as constructions in which the verb is "capable of getting a prepositional complement", where the complement is defined as an obligatory valence of the verb, irrespective of its semantics. This means that in the class of adpositional verbs we can have examples in which the adposition is part of an adverbial, e.g. a place adverbial: I **live *in*** *London*. Geană (2013: p. 118) further distinguishes between adpositional verbs using as criterion the type of adposition, namely:

- merely **functional adpositions**: their sole role is to case-mark the nominal which is a thematic argument of the verb: e.g. *Noi ne bazăm **pe** ajutorul vostru*. (En. "We count on your help") – the adposition *pe* imposes the accusative case on the noun *ajutorul*;
- **semi-lexical adpositions** (Corver and van Riemsdijk, 2013): in the case of verbs requiring a semantic argument, the adposition carries the specific semantic content and, at the same time, case-marks the nominal with that role: e.g. *Ne plimbăm pe alee*. (En. We walk on the alley.) – the adposition has a locative meaning and imposes the accusative case to the noun *alee*.

Testing the two types of examples against the PARSEME criterion, we notice that in the case of functional adpositions the test holds, as one cannot ask about the circumstances of the verbal event using the verb only, not also the adposition: **Când ne bazăm noi?* (En. "When do we count?") is not a grammatically complete question in Romanian. However, in the case of semi-lexical adpositions, the test does not hold: asking a question like *Când vă plimbați?* (En. When do you walk?) is grammatically complete.

Given these remarks on the types of IAVs annotated in PARSEME-Ro, we consider that the annotated data will need some further refinement: adpositional verbs will need to be further classified into two subtypes: IAV.functional and IAV.semi-lexical. The existence of subclasses inside a class is not new for PARSEME: see the two subtypes of LVCs, namely LVC.full and LVC.cuase (Section 2). However, continuing the PARSEME custom of testing classes and subclasses against data in more languages before coining them officially, the next step we envisage is collaborating with teams working on IAVs for other languages, so as to share findings.

no. of sentences	56,664
no. of tokens	1,014,908
no. of words	806,540
no. of verbal lemmas	61,323
no. of unique verbal lemma	3,815

Table 2: General statistics of the PARSEME-Ro corpus

6 Corpus Statistics

General information about the corpus size is available in Table 2, whereas information about the VMWEs types and their frequency in the corpus is provided in Table 3, which shows that the majority (2 thirds) of the VMWEs in the corpus are reflexive verbs or adpositional ones. Such distribution of the types in the corpus should not be taken as general in the language, but should be interpreted with respect to the corpus texts genre, as well as their characteristics inherent to their source: being issues of the same daily newspaper, written by the same journalists, on more or less similar topics.

The most frequent (usually ten[9]) verbs in each type of VMWEs are enumerated below, and, between brackets, their frequency with the respective type of VMWEs; for verbs that are among the 20 most frequent ones in the corpus, we also indicate between brackets the relative frequency with which they are used in that type of VMWEs:

Cross-lingual_tests/070_Inherently_adpositional_verbs__LB_IAV_RB_

[9]We give less than 10 verbs when they have more than 1 occurrence.

Type	Number
IRV	3.826
LVC.cause	182
LVC.full	516
VID	1.644
IAV	3.311
TOTAL	9479

Table 3: Number of VMWEs of each type

- IRV: *desfășura* (unfold) (303, i.e. 47% of all its occurrences in the corpus), *afla* (find) (294, i.e. 42% of all its occurrences in the corpus), *adresa* (address) (203), *putea* (can) (190, i.e. 8% of all its occurrences in the corpus), *prezenta* (present) (117, i.e. 19% of all its occurrences in the corpus), *derula* (unreel) (93), *încheia* (finish) (91), *naște* (give birth) (87), *număra* (count) (63), *deplasa* (travel) (61);

- LVC.cause: *pune* (put) (179), *da* (give) (6);

- LVC.full: *avea* (have) (192, i.e. 7% of all its occurrences in the corpus), *face* (make, do) (173, i.e. 17% of all its occurrences in the corpus), *lua* (take) (108), *da* (give) (26), *aduce* (bring) (10), *pune* (put) (7);

- VID: *avea* (have) (804, i.e. 31% of all its occurrences in the corpus), *pune* (put) (108), *lua* (take) (102), *da* (give) (85), *fi* (be) (76, i.e. 9% of all its occurrences in the corpus), *intra* (enter) (65), *ține* (hold) (51), *trimite* (send) (50), *face* (make, do) (43, i.e. 4% of all its occurrences in the corpus), *sta* (stay) (41);

- IAV: *beneficia* (benefit) (185), *participa* (participate) (149), *desfășura* (unfold) (130, i.e. 20% of all its occurrences in the corpus), *intra* (enter) (120), *ajunge* (reach) (116), *pune* (put) (100), *trece* (pass) (98), *duce* (take to) (81), *lua* (take) (63), *ridica* (lift) (59).

We notice that verbs may tend to occur in one type of VMWEs, but there are many exceptions, with the verb *pune* (put) occurring with LVC.cause, LVC.full, VID and IAV expressions, and the verb *lua* (take) occurring with three types: LVC.full, VID and IAV. There are others occurring only with LVC.full and VID: *avea* (have), *face* (make, do), *da* (give). All are verbs with rich polisemy, sometimes even bleached in frozen combinations.

7 Conclusions

The new version of the PARSEME-Ro corpus comes with a new type of VMWEs: IAV. Such expressions are widely spread in the corpus: they represent a third of the total number of VMWEs occurring therein. This makes them an important phenomenon to be made explicit in a corpus. A comparative analysis of the cases when the same combination verb + adposition is either annotated as an IAV or not will be carried out, coupled with grammatical and semantic characteristics of the context, to better understand what the specific contexts for IAVs are.

So far, only verbal IAVs have been annotated in PARSEME-Ro, while VMWEs IAVs (IAVMWEs) are left for further investigations. Prior to this, we consider that the status of IAVs needs to get clarified, as our analysis of such expressions has shown that the type could be further classified into two subtypes: IAV.functional and IAV.semi-lexical.

The new version of PARSEME-Ro will be made fully and freely available in the first annual release within PARSEME, scheduled for mid 2022, in a format that will be agreed upon within the community. It is also available on our website of language resources in Linked Data format[10] and can be queried using the SPARQL endpoint[11].

Acknowledgments

Part of the work reported here has been carried out within Action 2020-EU-IA-0088 which has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278547.

References

Timothy Baldwin and Su Nam Kim. 2010. *Handbook of Natural Language Processing, 2nd edition*, chapter Multiword Expressions. CRC Press, Boca Raton, FL, USA.

Verginica Barbu Mititelu. 2018. Modern Syntactic Analysis of Romanian. In Daniela Butnaru Marius-Radu Clim Veronica Olariu Ofelia Ichim, Luminiţa Botoşineanu, editor, *Clasic şi modern în cercetarea filologică românească actuală*, pages 67–78. Publishing House of "Alexandru Ioan Cuza" University.

[10] https://www.racai.ro/p/llod/index.html

[11] https://relate.racai.ro/datasets/dataset.html

Verginica Barbu Mititelu, Mihaela Cristescu, and Manuela Nevaci. 2021. Un instrument modern de studiu al dialectului aromân: corpus adnotat morfosintactic. In Ioan-Mircea Farcaș Manuela Nevaci, Irina Floarea, editor, *Ex Oriente lux. In honorem Nicolae Saramandu*, pages 143–162. Edizioni dell'Orso, Alessandria.

Verginica Barbu Mititelu and Maria Mitrofan. 2020. The Romanian Medical Treebank - SiMoNERo. In *Proceedings of the 15th International Conference "Linguistic Resources and Tools for Natural Language Processing"*, pages 7–16.

Mihaela Colhon, Cătălina Mărănduc, and Cătălin Mititelu. 2017. Multiform Balanced Dependency Treebank for Romanian. In *Proceedings of Knowledge Resources for the Socio-Economic Sciences and Humanities, (KnowRSH)*, pages 9–18, Varna, Bulgaria.

Norbert Corver and Henk van Riemsdijk, editors. 2013. *Semi-lexical Categories: The Function of Content Words and the Content of Function Words*. De Gruyter Mouton.

Cecilia Căpățână. 2007. *Elemente de frazeologie*. Editura Universitaria Craiova.

Ionuț Geană. 2013. *Construcții verbale prepoziționale în limba română*. Editura Universității din București.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: a multilingual treebank collection. In *Proc. of LREC*, pages 1659–1666, Portorož, Slovenia.

V. Păiș, M. Mitrofan, V. Gasan, C.L.and Coneschi, and A. Ianov. 2021. Named Entity Recognition in the Romanian Legal Domain. In *Proceedings of the Natural Legal Language Processing Workshop*, pages 9–18.

Vasile Păiș, Radu Ion, and Dan Tufiș. 2020. A processing platform relating data and tools for romanian language. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 81–88.

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.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLing 2002*, pages 1–15.

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

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. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

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, Matthieu 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.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text

annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, page 4290–4297, Portorož, Slovenia.