Two Birds with One Stone: Annotating Romanian Multiword Expressions with an Eye to the PARSEME 2.0 Guidelines Applicability

Verginica Barbu Mititelu[1], Mihaela Cristescu[2], Elena Irimia[1], Carmen Mîrzea Vasile[2]

[1]Romanian Academy Research Institute for Artificial Intelligence, [2]University of Bucharest
vergi@racai.ro,mihaela.cristescu@litere.unibuc.ro,elena@racai.ro,carmen_marzea@yahoo.fr

Abstract

This paper presents an enhanced version of the Romanian corpus previously annotated only for verbal multiword expressions. The new release extends the annotation to multiword expressions of other parts of speech, following version 2.0 of the PARSEME guidelines. The corpus has been expanded, its new part was automatically morpho-syntactically annotated based on the Universal Dependencies framework, followed by extensive semi-automatic annotation of multiword expressions across all morphological categories. The paper also reports quantitative data on the updated corpus and discusses the distribution and characteristics of Romanian multiword expressions. We also highlight language-specific annotation challenges and issues arising from the PARSEME 2.0 guidelines.

1 Introduction

Multiword expressions (MWEs) are everywhere, yet notoriously slippery to define and analyze. From idioms like *go bananas* to collocations such as *rancid butter* and phrasal verbs like *put up with*, these fixed or semi-fixed combinations of words play a key role in how meaning is packaged and conveyed. Understanding how MWEs function is easier said than done, as their meaning often goes beyond the sum of their parts. However, besides the semantic non-compositionality, MWEs also exhibit idiosyncrasies at other linguistic levels (Baldwin and Kim, 2010): lexical, syntactic, pragmatic and even statistical.

We present below the process of quantitatively and qualitatively enriching the Romanian component of the PARSEME corpus (Savary et al., 2023). Its initial version (Barbu Mititelu et al., 2019) contained annotations of only verbal MWEs, while now MWE of all parts of speech have been annotated. Another contribution of this paper is that of offering feedback regarding the PARSEME 2.0 guidelines[1] that were observed during the annotation.

The paper is structured as follows: Section 2 presents the current work concerning MWEs both within Romanian linguistics and in an international context. In Section 3 statistics of the enriched corpus is given. The types of MWEs annotated in the data are inventoried in Section 4, alongside the challenges their identification raises. We describe our work methodology in Section 5 and make some remarks on the frequency and variety of MWEs in this corpus in Section 6. The issues we had in the application of the decision trees of the PARSEME guidelines 2.0 are presented in Section 7, before concluding the paper.

2 Related Work

Related work concerns, on the one hand, the current situation of research on MWE in the Romanian language and, on the other hand, the larger international background against which our work has been carried out.

The systematic study of MWEs in Romanian linguistics dates back to the 1950s (Ioaniţescu, 1956), with early focus on verbal MWEs (Dimitrescu, 1958). Over time, scholars have used varied terminology and offered differing views on their definition, classification, and structure (Căpățână, 2007). A recent contribution to this field (Pană Dindelegan et al., 2025) provides a detailed overview of Romanian MWEs, covering theoretical issues (concepts, delimitation criteria, graduality, and terminology) and practical aspects (types, analysis methods, and exercises). Aimed primarily at educational and applied purposes, it offers representative inventories and solutions to common difficulties, making it a useful resource for students, teachers, and researchers of Romanian grammar.

[1] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 66–74

In recent years, there has been growing interest in the computational analysis of Romanian MWEs (for a larger context, see (Barbu Mititelu et al., 2025)). A corpus annotated with verbal MWEs is already mentioned above and was developed within PARSEME COST Action[2]. The lexicon of Romanian verbal MWEs (Leseva et al., 2024) provides uniform descriptions of such MWEs at several linguistic levels (lexical, morphologic, syntactic, semantic, stylistic) (and in comparison with a language historically in contact with Romanian, i.e. Bulgarian, and with English). Additionally, the Romanian Reference Treebank (Barbu Mititelu, 2013) has been annotated with multiword conjunctions (Barbu Mititelu and Voicu, 2024) which were assigned Penn Discourse Treebank relations (Webber et al., 2019), facilitating deeper insights into their syntactic and semantic roles.

Besides developing language resources (a corpus and a lexicon), researchers have also been interested in developing systems dedicated to the task of identifying MWEs in Romanian corpora (Boros et al., 2017; Avram et al., 2023).

PARSEME has conducted a series of multilingual annotation campaigns and shared tasks dedicated to MWEs. Corpora for 20+ languages were annotated with verbal MWEs and further used as training, tuning and testing data for systems in three shared tasks for automatic MWE identification in corpora. One factor that made this effort even more valuable was the annotation of MWEs observing common guidelines developed with an eye to universality: a common typology of MWEs was created for and tested on all languages involved, at the same time making space for any particular language specificity, i.e. language specific MWE types were accepted and described.

The PARSEME annotation framework formalized a consistent procedure for identifying MWEs across languages, combining a decision-tree approach with cross-lingual validation and language-specific clarifications (Savary et al., 2018). This approach has been crucial for ensuring both comparability and adaptability across typologically diverse languages. With PARSEME 1.3 (Savary et al., 2023), the multilingual corpus expanded to 26 languages, was aligned with Universal Dependencies v.2, and further enhanced its linguistic coverage, consolidating its position as a major resource for multilingual MWE research.

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

3 The Corpus

The PARSEME-Ro corpus consists exclusively of journalistic texts published between 2003 and 2017, and includes 56,703 sentences totaling 1,015,623 tokens (i.e. syntactic words and punctuation). In this annotation campaign, we expanded the corpus with 14,517 sentences (407,801 tokens), also drawn from journalistic sources, in order to maintain genre homogeneity throughout the corpus. Table 1 shows that the newly added data contains longer sentences and represents about a third of the final corpus.

The number and proportion of MWEs of different parts of speech annotated in the corpus are rendered in Table 2. We can see that functional MWEs are the most frequent in the corpus. It is a closed morphological class but indispensable in rendering logical connection between syntactic units in sentences. The ratio token/MWE is 22 (i.e., there is an average frequency of one MWE per 22 tokens), which, correlated with the average sentence length (20 tokens, see Table 1), means that each sentence contains an average of 1.1 MWEs.

4 MWEs in Romanian

4.1 Types of MWEs occurring in Romanian

The types of MWEs currently represented in the Romanian corpus follow the PARSEME guidelines 2.0 and are as follows:

- Verbal
  - Verbal Idioms (VID): *avea de gând* (‘have of thought’ “intend”), *da viață* (‘bring life’ “bring to life”)
  - Light Verb Constructions (LVC)
    - full: *avea grijă* (‘have care’ “take care”), *da citire* (‘give reading’ “read”)
    - cause: *da asigurare* (‘give assurance’ “assure”), *pune la dispoziție* (‘put at disposal’ “provide”)
  - Reflexive Verbs (IRV): *se gândi* (“think”), *se abține* (“refrain”)
  - Inherently Adpositional Verbs (IAV): *conta pe* (“count on”), *depinde de* (“depend on”)
- Nominal
  - Nominal Idioms (NID): *bani gheață* (‘money ice’ “cash”), *bătaie de cap* (‘beating of head’ “trouble, nuisance”)

	Older data	New data	TOTAL data
sentences	56703	14517	71220
tokens	1015623	407801	1423424
tokens/sentence	18	28	20

Table 1: Statistics of the PARSEME-Ro 2.0.

MWE PoS	#	%
verbal	18084	28
nominal	8752	13
adjective/adverb	14672	23
functional	23093	36
TOTAL	64601	

Table 2: Statistics on MWEs in PARSEME-Ro 2.0.

  - Pronominal Idioms (PronID): *câte ceva* (“something”), *Excelența Sa* (“His Excellency”)
  - Deverbal Nominal MWEs (NV): *aducere aminte* (‘bringing to memory’ “memory, remembrance”), *băgare de seamă* (‘putting of notice’ “attention, observation”)

- Modifier
  - Adjectival Idioms (AdjID): *cât un purice* (‘as big as a flea’ “very small”), *de vină* (‘of guilt’ “guilty”)
  - Adverbial Idioms (AdvID): *de asemenea* (‘of the same’ “also, likewise”), *și așa mai departe* (‘and thus further’ “and so on”)
  - Deverbal adjectival / adverbial MWEs (AV): *cu luare aminte* (‘with taking notice’ “attentively”), *avut în vedere* (‘had in view’ “considered, took into account”)

- Functional
  - Determiner Idioms (DetID): *tot felul de* (“all kinds of”), *ca atare* (“as such”)
  - Adposition Idioms (AdpID): *în legătură cu* (‘in connection with’ “regarding”), *cu excepția* (‘with the exception’ “except for”)
  - Conjunction Idioms (ConjID): *astfel încât* (“so that”), *pentru că* (‘for that’ “because”)
  - Interjection Idioms (IntjID): *așa să fie* (“so be it, amen”), *nici vorbă* (‘no word’ “no way, not a chance”)

4.2 Challenges in MWE Identification in Romanian

As mentioned above, the MWE status is decided based on the battery of tests organized as a decision tree in the PARSEME annotation guidelines. Inherent difficulties are detailed in the annotation guidelines (see Section 3 therein, e.g., the solutions provided for problematic reflexive expressions). Beyond these general challenges, MWE identification in Romanian faces additional difficulties stemming from the language’s specific lexico-grammatical features.

Romanian is a language with rich morphology (Pană Dindelegan, 2013), where some grammatical categories have analytical expression. There are specific word strings that warrant examination at the morphology-lexicon interface. This concerns primarily grammatical categories (e.g., the comparative of superiority) and morpho-lexical (sub)types (e.g., the supine, ordinal numerals, distributive numerals, etc.). Some of these analytic forms may acquire a specific, idiosyncratic meaning distinct from the compositional meaning specific to their paradigm. These sequences with a fixed idiosyncratic meaning have been treated as MWEs (not as converted analytical forms): e.g., the expression of comparison (Mîrzea Vasile, 2012, 32-33) and intensity (*mai ales* “especially”, *mai curând* “rather”, *mai mult* “more”, *mai puțin* “except for”, *cel mai tare* “especially”, etc.).

The supine is a non-finite verb form containing a grammaticalized functional preposition (*de* “of”) and a deverbal abstract noun (Pană Dindelegan, 2013, 233-243). Supine forms with an idiosyncratic meaning were considered MWEs in our annotation (AdvIDs or AdjIDs, e.g., *de împrumut* “borrowed, unfitting”, *de neuitat* “unforgettable”, *de neînchipuit* “unimaginable”), while those in free syntactic configurations were omitted (e.g., *termină de scris* “finishes writing”, *instrument de scris* “writing instrument”, *apă bună de băut* “water good for drinking”, *De băut, am băut.* “As for drinking, I drank.”).

Another characteristic of Romanian is the con-

version (or zero-derivation) of adjectives into adverbs. The few dozen adverbs suffixed with *-ește* in contemporary Romanian are most frequently used in fixed expressions: *a împărți frățește* "to share fraternally", *a fi răsplătit regește* "to be rewarded royally", etc. (Mîrzea Vasile, 2012, 91-128).[3] There are also many compositional adverbials with quite regular structures, which were not considered MWEs; e.g., *în mod* 'in manner' + adjective: *în mod special* "in a special manner", *în mod necinstit* "in a dishonest manner"; *cu/fără* "with/without" + abstract quality noun: *cu prietenie* "with friendship", *cu dragoste* "with love", *fără plăcere* "without pleasure", *fără frică* "without fear". The expressions that we retained are those which passed the PARSEME tests: *Cu plăcere!* "My pleasure!" (adverbial used as IntjID), *fără seamăn* "peerless, incomparably" (AdjID/AdvID containing the cranberry old noun *seamăn* "resemblance", cf. current equivalent *asemănare* "resemblance"), *fără stare* "restless, agitated" (AdjID, in which the noun *stare* has a special meaning, "calm, tranquility"), etc.

In Romanian, there are the variable elements *al* and *cel*, with semi-functional or functional status depending on the context: e.g., *al* can be an obligatory unbound possessive morpheme in contexts of non-adjacency with the definite enclitic article, and can function as a pronoun that cannot appear independently; *cel* is a morpheme of the relative superlative degree, but can also have a status similar to that of a pronoun, etc. (Pană Dindelegan, 2013, 265-267, 309-318). Compositional constructions with these elements were omitted, and those which have developed a special meaning were annotated as MWEs; e.g., *ai mei* "my folks" (but not *ai mei* "mine" from: *Pantofii tăi sunt curați, ai mei nu sunt.* "Your shoes are clean, mine are not."), *Cel de Sus* (please notice the capitalized words) "God" (but not *cel de sus* "the above one": *El culege mărul de jos, nu pe cel de sus.* "He picks the apple from below, not the one from above").

5 Work Methodology

Our objective in this annotation campaign was to automate part of the workflow to optimize the overall process. The main motivations for introducing automation were the significantly larger number of non-verbal MWEs targeted for annotation and the inherent redundancy of functional MWEs, which makes them particularly amenable to automatic processing. All automatically generated annotations were manually validated and, when necessary, corrected, given that, as previously noted, the accurate identification and labeling of MWEs remains a challenging task. The following subsections outline the steps undertaken in the annotation process of the PARSEME-Ro corpus in its 2.0 version.

5.1 Automatic Retrieval of MWEs from Dictionaries

The automatic annotation approach was a resource-based one, involving Romanian idioms and expressions dictionaries. In selecting these linguistic resources, we restricted our focus to dictionaries that were already digitised, available online or in standard digital formats (e.g., .DOCX, .XLS, .PDF), enabling automatic processing and eliminating the need for manual scanning and subsequent OCR processing.

Using dedicated scripts, information was extracted from PDF files in the case of three dictionaries: DELS (Mărănduc, 2010), *Dicționar de expresii românești în contexte*, Vol. 1-4 (Dictionary of Romanian Expressions in Context, (Ilincan, 2015)) and *Dicționar frazeologic al limbii române* (Phraseological Dictionary of the Romanian Language, (Tomici, 2009)). Through a combination of automatic and manual processing, the resulting TXT files were parsed and curated to: (i) exclude verbal expressions (as already annotated in the previous versions of the corpus), (ii) remove definitions, examples, usage notes, variants, lexicographic cross-references, etc., (iii) expand expressions in case of variants rendered as alternations (e.g., the unique entry *Majestatea Ta/Sa/Voastră* was split into three different PronIDs, namely *Majestatea Ta* "Your Majesty" (2nd person singular), *Majestatea Sa* "His/Her Majesty" (3rd person singular) and *Majestatea Voastră* "Your Majesty" (2nd person plural)), (iv) correct errors arising from the automatic content extraction from PDFs (such as end-of-line word segmentation, diacritics misencoding), and (v) format candidate MWEs as lists with one entry per line.

The online dictionary *Dicționarul ortografic, ortoepic și morfologic al limbii române*[4] (DOOM, The Orthographic, Orthoepic, and Morphological Dictionary of the Romanian Language) offers the possibility to retrieve and download lists of idiomatic expressions through queries targeting a

[3] Such examples do not occur in the corpus.

[4] https://doom.lingv.ro/

specific part of speech. The online version of *Dicționarul explicativ al limbii române*[5] (DEX, The Explanatory Dictionary of the Romanian Language) was not downloaded, but it is a comprehensive resource that was manually consulted at all stages of human validation and annotation.

The MWEs extracted from the aforementioned dictionaries were consolidated into a single inventory after duplicate entries were removed. This inventory was then automatically matched to the corpus at the word-form level, and the resulting list of matched MWEs (2,034 unique occurrences) was carried forward to the next annotation stage.

5.2 Curation of the List of MWEs Extracted from Dictionaries

The automatically matched MWEs list was manually validated and labeled with PARSEME MWE categories by a team of six linguists, following a two-step procedure to allow cross-validation. Each expression was assigned one or more labels from the label set in the PARSEME guidelines version 2.0, applying the test battery therein. Some of these expressions were clearly erroneous, arising from errors in the automatic extraction process, while others were plausible MWEs but failed the PARSEME tests corresponding to their specific part of speech. All such expressions were subsequently labeled as NOT MWE.

Entries that could be assigned two different parts of speech given their possible distributions, and consequently two distinct MWE labels, were expanded into separate corresponding entries. For example, the entry *de mână* ('by hand') was split into *de mână* (AdjID), as in *scris de mână* ('writing of hand' "handwriting") (i.e., when having a noun as its syntactic head), and *de mână* (AdvID), as in *scrie de mână* ('writes by hand') (i.e., when having a verb as its syntactic head). This duplication procedure was not applied to MWEs exhibiting polysemy and, thus, occurring with the same part of speech (e.g., *în parte*, labeled as AdvID, has the meanings "partially" (*Ai dreptate, în parte.* "You are right, partially."), but also "separately, one by one" (*El răspunde la fiecare întrebare în parte.* "He answers each question separately.").

Cross-validation was conducted, with each entry being independently validated by two annotators. The overall inter-annotator agreement rate was 57.9% (1,178 out of 2,034 total entries). A third round of validation, carried out by a linguist, was performed on the consensus dataset, while the disagreement dataset was analysed in expert team meetings until a consensus was reached. In certain cases, conflicting annotations were both retained, particularly when one of the annotators had expanded an entry to account for two possible part of speech labels. This is the case of the expression *de mână* discussed above. In most cases, only one of the competing labels was selected as correct, the other(s) being annotation errors. In other cases, the expression was reclassified as NOT MWE for failing to pass the PARSEME tests battery, in spite of being considered MWEs by the authors of the dictionary from which they were automatically extracted, which once again shows that there is no universally accepted definition of MWEs. For example, *fără risipă* ("without waste"), initially labeled as AdjID, was ultimately retained as NOT MWE, while *cu judecată* ("with judgment, rationally"), initially tagged as AdvID and AdjID, was similarly reclassified as NOT MWE in the third validation round (see Subsection 4.2 above).

Overall, 46.9% of the agreed-upon and 30.8% of the disagreed-upon expressions (816 in total) were classified as NOT MWE for failing the PARSEME tests and were, consequently, excluded. The final dataset, after expanding homonymous entries, comprised 2,010 MWEs.

5.3 Automatic MWE Annotation

The manually validated resulting list was used to automatically annotate the PARSEME-Ro corpus, by performing word-form level matching. When identifying an expression for annotation, a window of up to two intervening tokens was permitted between any of its components, accommodating insertions typical of some MWEs. Although this approach does not ensure full recall, it offers a practical trade-off with precision, since the likelihood of false positives increases with the number of allowed intervening words.

5.4 Manual Correction of the Automatic Annotation

A dedicated FLAT platform instance, configured for the PARSEME project and providing individual accounts for each annotator, was used for manual validation of the automatically annotated corpus. The same team of six linguists participated in this stage, which involved: (i) removing one of the two annotations in cases of homonymous expressions

[5] https://dexonline.ro/

bearing two possible labels that could only be disambiguated in context: see the expression *de mână* explained in subsection 5.2; (ii) correct the MWE type label when a wrong one was automatically assigned; (iii) adding or removing components of an expression and (iv) deleting an annotation in case of false positives resulting either from the allowance of intervening tokens or from instances with a compositional meaning. For example, the construction *de preț* 'of price' "precious" is considered AdjID only in contexts such as *Am amintiri de preț din acel concediu.* ('Have-I memories of price from that vacation' "I have precious memories from that vacation."). However, the automatic annotation marked *de preț* as AdjID even in contexts such as *Marcatorul de preț este un aparat care...* ('Marker of price is a device that...' "A labelling system is a device that..."), in which case the annotator removed the label.

At the same time, expressions that were not automatically detected required manual annotation. This occurred primarily for three reasons:

- the list of manually validated MWEs was not exhaustive and therefore did not include some expressions that occur in our corpus. Such examples include expressions referring to meteorological warnings and alerts (*cod portocaliu* "orange code", *cod roșu* "red code"), which were classified as NIDs;
- the number of intervening words between the components of an expression exceed the predefined limit of two, e.g. *Ei **pun**, fără nicio îndoială, **bazele** statului modern.* ('They put, without any doubt, the foundations of the modern state.' "They are undoubtedly laying the foundations for the modern state."); a few functional MWEs also allow for this: e.g., *Au acționat **fără ca** măcăr atunci, în ultimul ceas, **să** le pese de ce simt ceilalți.* "They have acted without at least then, in the last hour, to care about what others feel.";
- the notion of MWE sometimes cover more than what is traditionally called expression (see the case of terms that observe the PARSEME definition of a NID), e.g., *date personale* "personal data".

As shown in Table 3, the automation of the procedure significantly reduced the number of required manual operations (see the great number of automatically annotated MWEs left unchanged after manual correction: 18,953). When counting unchanged MWEs, verbal expressions annotated in previous stages of the project were not taken into account, but any insertion, deletion or modification of verbal MWE was counted.

Operations	**#**
Insertions	4579
Deletions	5177
Modifications	1474
Unchanged	24844
MWEs after manual correction phase	36074
Existing verbal MWEs	5891
Unchanged minus verbal MWEs	18953

Table 3: Operations done in the manual correction of the automatic annotation step.

Unfortunately, due to time constraints and reduced number of staff, the files of the corpus were manually checked only by one linguist. However, in order to understand the extent to which the team of annotators agree in their evaluation, a part of the corpus was doubly annotated: 2000 sentences were randomly selected from the automatically annotated ones and four pairs of annotators manually checked them. We calculated the inter-annotator agreement score using the scripts made available by the PARSEME team (see (Savary et al., 2017)). Its value is 0.78, which shows high consistency among the annotators in our team.

5.5 Ensuring Annotation Consistency

A methodology to ensure consistency in the PARSEME annotation, implemented at the project level through a suite of Python libraries, reflects the initiative's commitment to producing high-quality datasets. All annotations in the corpus are automatically extracted and grouped according to the unique MWE they pertain to, alongside occurrences of the same sequences that were skipped during the automatic or manual annotation stages. This setup allows human validators to examine all contexts of a given word sequence and assess the MWE status and assign a label for each occurrence. By presenting the user with MWEs in contexts, both similar and divergent ones together, the process facilitates more accurate and consistent annotation of MWEs.

When the manual consistency check was over, the F-measure between the manual annotations and the outcomes of the consistency check was calculated and its value is 86, which is indicative of a fairly consistent corpus.

6 Remarks on the Frequency and Variety of MWEs in the PARSEME-Ro Corpus

As shown in Section 3, Table 2, functional MWEs constitute the most frequent category (36% of total MWEs), followed by verbal MWEs (28%), adjectival and adverbial MWEs (23%), and nominal MWEs (13%). Within these main categories, the distribution of subtypes exhibits varying degrees of asymmetry: major imbalances (e.g., PronIDs are significantly less frequent than fully lexical nominal MWEs; DetIDs are extremely underrepresented compared to other functional MWEs), moderate imbalances (AdpIDs show the highest frequency, yet ConjIDs still register a substantial number of occurrences), or balanced distributions (AdjIDs and AdvIDs).

As expected, the content categories (nominal, verbal, adjectival, and adverbial expressions) exhibit greater variety in the corpus compared to functional ones (prepositional and conjunctional): e.g., the expressions (NIDs) *amor propriu* "self-esteem", *cotă de piață* "market share", (AdjIDs) *în putere* "in force, powerful", *la cheie* "turnkey", (AdvIDs) *an de an* "year after year", *cu zâmbetul pe buze* "with a smile on one's face", (VIDs) *a sări în ochi* "to catch the eye", *a aduce pe lume* "to bring into the world, to give birth" have fewer than 5 occurrences each, whereas functional ones (ConjIDs) *pentru că* "because", *după ce* "after", *după cum* "as", (AdpIDs) *în cazul* "in the case of", *față de* "compared to" have between 200 and 500 occurrences each.

Also as expected, the list of MWEs from dictionaries is only partially found in the corpus (see Section 5.1); conversely, new MWEs not recorded in dictionaries were identified. Thus, of the 94 forms of polite pronominal expressions listed in DOOM (grouped into 70 separate dictionary entries), the corpus contains only scattered occurrences of fewer than 10 distinct PronIDs (*Majestatea Sa* "His/Her Majesty", *Sanctitatea Sa* "His/Her Holiness", *Preasfinția Sa* "His/Her Grace", etc.). Therefore, the vast majority of polite pronominal expressions included in the contemporary prescriptive dictionary DOOM (*Luminarea Voastră* "Your Reverence", *Panevghenia Ta* "Your Eminence", *Preacucernicia Sa* "Your Piety", etc.) seem not to be in current use. Manually identified MWEs that are not included in consulted dictionaries (i.e. 53% of the all unique MWEs annotated in the corpus) include AdjIDs (e.g., *de tristă amintire* "of bitter memory" (from the communist period in Romania), NIDs (e.g., *cod galben* "yellow code", *chestiune arzătoare* "a burning issue"), AdpIDs (e.g., *funcție de* "depending on"), etc.

As in other languages, some MWEs have been borrowed or partially or totally loan-translated. During the decision tree application process, we observed that certain diagnostic features of these expressions are due to the source language. For example, the plural form *forțe* "forces" in *forțele armate* "armed forces" (NID) is borrowed through translation from Fr. *forces armées*; the definite articulated form of the noun in *cu forța* "by force" can be similarly explained, cf. Fr. *avec la force*; in contrast, the noun in *cu putere* "forcefully" lacks an article. Some MWE elements have meanings difficult to grasp outside these expressions; e.g., the terminological senses of expressions containing the noun *fond* (e.g., *instanță de fond* "court of first instance" (cf. Fr. *juridiction de fond*), *zgomot de fond* "background noise" (cf. Fr. *bruit de fond*), *a judeca pe fond* "to judge on the merits" (cf. Fr. *juger sur le fond*)); in the NID *fond de rulment* "working capital" (cf. Fr. *fonds de roulement*), *rulment* is a cranberry word — it does not occur independently in Romanian nor in other expressions (it may be mistaken for its homonym denoting a ball bearing).

Some MWEs, particularly those fixed in specialized registers, exhibit pleonastic features; e.g., *funcționar public* (NID) "civil servant" (by definition, see DEX, such employees work exclusively in the public sector), *drept pentru care* (ConjID) "wherefore, because" (both prepositions *drept* and *pentru* express causality, the former being archaic).

7 Issues in Tests Application

In assigning the MWE status, as defined in the project, the application of the proposed tests may encounter several issues, a situation potentially applicable to other languages, not only Romanian. Among these issues, three are particularly noteworthy:

- The cranberry status of a component within an expression, a strong diagnostic test for all MWE types, is sometimes difficult to determine, as it involves relatively subjective evaluations of (i) whether an element is archaic or obsolete (no strict delimitation exists: some outdated lexemes remain familiar to certain speakers through literature, historical texts, etc.), and (ii) the size of the closed set of contexts in which it still occurs. For example, the

conjunction *necum* 'not-how' is perceived as archaic both when used independently (*Ea nu a văzut arma nici măcar în poze, necum în realitate.* "She hasn't seen the weapon even in pictures, let alone in reality.") and when it appears in two other MWEs: *necum să* "so much the less" (ConjID) and *cum-necum* "one way or another" (AdvID); the stand-alone noun *preajmă* is out of use and appears in two AdpIDs: *în preajma* "in the vicinity of" and *din preajma* "from the vicinity of".

- The morphological inflexibility test is inoperative for expressions containing generic nouns or other types of defective forms, as morphological non-prototypicality in such cases has internal semantic constraints. For example, abstract nouns such as *plac* "liking", *libertate* "freedom", *dispoziție* "mood", *pace* "peace", *siguranță* "safety" are inherently morphologically restricted as singularia tantum, and this restriction carries over into expressions: *pe plac* "to one's liking" (AdvID), *în libertate* "free(ly)" (AdvID/AdjID), *și pace!* "and that's it!" (IntjID), etc. Furthermore, in Romanian, expressive or intensifying shifts between number values sometimes occur, as in *la început* (sg.) → *la începuturi* (pl.) "at the beginning" or *fără margini* "boundless" (pl.) → *fără margine* (sg.) "without limit(s)".

- In Romanian, the prohibited modification test (available for NIDs, AdvIDs, AdjIDs, DetIDs), like the morpho-syntactic inflexibility test, fails to determine the MWE status for specialized terms when relative adjectives are present (e.g., *persoană fizică* "natural person", *placă turnantă* "turntable", *politică externă* "foreign policy", *relații publice* "public relations"), as this class systematically blocks gradation (**persoană foarte fizică* 'person very physical') and anteposition (**fizică persoană* 'physical person', (Pană Dindelegan, 2013, 418-419)).

The annotation team discussed these issues and established consistent guidelines.

8 Conclusions and Future Work

We presented here the most recent work for enlarging and enhancing the Romanian corpus annotated with MWEs, namely PARSEME-Ro. It is a resource that will be coupled with an electronic lexicon of MWEs for Romanian (Leseva et al., 2024), so that the linguistic description of the MWEs gets more detailed.

Besides its importance for shared tasks and evaluation campaigns (the annotated corpus was used in the PARSEME 2.0 shared task[6]), the corpus can also serve as a resource for language learning, especially second language learning, as foreigners can find here various contexts for a large number of expressions.

The corpus is also relevant for the language specialists, who can find here a current snapshot of the number, frequency, and inventory of MWEs in the journalistic genre. When compared with existing dictionaries, this can show the tendency of some MWEs to become obsolete, as well as the appearance of new ones. Of course, they need to take this with a grain of salt, given that our corpus is not representative of the whole journalistic writing, let alone of the language in general. We are also interested in extending the analysis of MWEs by adding new text genres to the corpus and then notice similarities and differences between them and the journalistic one.

9 Acknowledgment

Part of the work presented here was was carried out within the "Large Language Models for the European Union (LLMs4EU)", project no. 101198470, call DIGITAL-2024-AI-B-06-LANGUAGE, funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. Another part of this work was supported by a grant of the Ministry of Research, Innovation and Digitalization - UEFISCDI, project number PN-IV-P8-8.2-EUD-2025-0061, within PNCDI IV. Another part of the work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

References

Andrei-Marius Avram, Verginica Barbu Mititelu, and Dumitru-Clementin Cercel. 2023. Romanian multiword expression detection using multilingual adver-

[6] https://unidive.lisn.upsaclay.fr/doku.php?id=other-events:parseme-st

sarial training and lateral inhibition. *arXiv preprint arXiv:2304.11350*, no volume:no pages.

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2 edition, pages 267–292. CRC Press, Boca Raton, USA.

Verginica Barbu Mititelu. 2013. istemul all-inclusive în reprezentarea cunoştinţelor lexicale. In Ofelia Ichim, editor, *Tradiţie/inovaţie - identitate/ alteritate: paradigme în evoluţia limbii şi culturii române*, pages 9–18. Editura Universităţii „Alexandru Ioan Cuza", Iași.

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

Verginica Barbu Mititelu, Voula Giouli, Gražina Korvel, Chaya Liebeskind, Irina Lobzhanidze, Rusudan Makhachashvili, Stella Markantonatou, Aleksandra Markovic, and Ivelina Stoyanova. 2025. Survey on lexical resources focused on multiword expressions for the purposes of NLP. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 41–57, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Verginica Barbu Mititelu and Tudor Voicu. 2024. Function multiword expressions annotated with discourse relations in the Romanian reference treebank. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 90–97, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Tiberiu Boros, Sonia Pipa, Verginica Barbu Mititelu, and Dan Tufiș. 2017. A data-driven approach to verbal multiword expression detection. PARSEME shared task system description paper. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 121–126, Valencia, Spain. Association for Computational Linguistics.

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

Florica Dimitrescu. 1958. *Locuțiunile verbale în limba română*. Editura Academiei, Bucharest.

Vasile Ilincan. 2015. *Dicționar de expresii românești în contexte [DERC]*. Presa Universitară Clujeană, Cluj-Napoca.

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

Svetlozara Leseva, Verginica Barbu Mititelu, Ivelina Stoyanova, and Mihaela Cristescu. 2024. A uniform multilingual approach to the description of multiword expressions. In Voula Giouli and Verginica Barbu Mititelu, editors, *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, pages 73–116. Language Science Press, Berlin.

Carmen Mîrzea Vasile. 2012. *Eterogenitatea adverbului românesc. Tipologie și descriere*. Editura Universității din București, Bucharest.

Cătălina Mărănduc. 2010. *Dicționar de expresii, sintagme și locuțiuni ale limbii române, DELS*. Corint, Bucharest.

Gabriela Pană Dindelegan, editor. 2013. *The Grammar of Romanian*. Oxford University Press, Oxford.

Gabriela Pană Dindelegan, Raluca Brăescu, and Cristiana Aranghelovici. 2025. *Locuțiunile limbii române*. Univers Enciclopedic, Bucharest.

Agata Savary, 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, and 9 others. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomir 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, and 3 others. 2018. *PARSEME multilingual corpus of verbal multiword expressions*. Number 2 in Phraseology and Multiword Expressions. 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.

Mile Tomici. 2009. *Dicţionar frazeologic al limbii române*. Editura Saeculum Vizua, Bucharest.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. Technical report, University of Pennsylvania.