The Romanian Corpus Annotated with Multiword Expressions. PARSEME-Ro Version 2.0

Verginica Barbu Mititelu[1], Mihaela Cristescu[2], Elena Irimia[1], Carmen Mîrzea Vasile[2,3]

[1]Romanian Academy Research Institute for Artificial Intelligence, [2]University of Bucharest,
[3]"Iorgu Iordan – Al. Rosetti" Institute of Linguistics, Romanian Academy
Bucharest, Romania
vergi@racai.ro, mihaela.cristescu@litere.unibuc.ro, elena@racai.ro, carmen.vasile@unibuc.ro

Abstract

The Romanian journalistic corpus previously annotated with verbal multiword expressions (PARSEME-Ro) has been extended recently with other journalistic texts and annotated with multiword expressions of all parts of speech closely observing version 2.0 of the PARSEME guidelines. The corpus size has been increased by about 40%, it underwent automatic morpho-syntactic annotation following the Universal Dependencies principles, as well as extensive semi-automatic annotation of multiword expressions of all morphological types (nominal, adjectival, adverbial, determiner, pronominal, prepositional, conjunction, interjection, and verbal for the newly added texts). We present here our work methodology, which involves an automatic annotation phase, but the manual work prevails in checking the annotation and its consistency. We also offer quantitative data about the new version of the corpus, the types of multiword expressions existing in Romanian and occurring therein, and characteristics thereof. The new version of the PARSEME-Ro corpus contributes to the field of developing multiword expressions resources per se, i.e. describing this language phenomenon, as well as resources for training, tuning and testing the performance of tools and large language models when dealing with this linguistic phenomenon.The paper also discusses some remarks on the MWE paraphrasing subtask in which a part of the corpus was used. The corpus is released with a permissive license.

Keywords: multiword expressions, Romanian, parts of speech, multiword expression identification, multiword expression paraphrasing

1. Introduction

Multiword expressions (MWEs) is a term that covers a variety of linguistic phenomena, from idioms like *kick the bucket* "to die" to collocations such as *neat and tidy* and phrasal verbs like *do away with*. Some of them are fixed, others are semi-fixed word combinations; most of them are non-compositional, others are partially compositional; some are continuous, others are discontinuous, etc. Such characteristics, among many others (Baldwin and Kim, 2010), alongside their frequency (rather frequent in lexicons, but most of them quite rare in language use) have made linguists consider them "a pain in the neck for natural language processing" (Sag et al., 2002). After more than two quite intense decades of research dedicated to them (Barbu Mititelu et al., 2025), to a great extent due to COST Actions such as PARSEME[1] and UniDive[2], the researchers are still collaborating towards developing or enriching existing resources containing or dedicated to MWEs, finding and testing new methodologies and technologies for dealing with MWEs, since they continue to raise problems to language technology downstream applications (Miletić and Schulte im Walde, 2024; Ramisch et al., 2023; Phelps et al., 2024; Mahajan et al., 2024).

Romanian is one of the languages for which not many resources are available for describing the characteristics of MWEs. Nevertheless, one corpus (PARSEME-Ro) annotated with verbal MWEs was created in the PARSEME COST Action (Barbu Mititelu et al., 2019) and further enriched a few years later to include one more type of verbal MWEs, i.e. adpositional verbs (Barbu Mititelu et al., 2022) (see below Section 3). A Romanian lexicon that describes verbal MWEs was created (Leseva et al., 2024): one major characteristic of this is the fact that it is interconnected with a lexicon of verbal MWEs for the Bulgarian language, both of them being aligned to the wordnets of the respective languages and to the Princeton WordNet (Fellbaum, 1998). Thus, equivalents of the verbal MWEs in Romanian and Bulgarian can be easily found in any other wordnet aligned to PWN, and, indirectly, to the Romanian and Bulgarian wordnets.

In this paper, we present a new version of the PARSEME-Ro corpus, which has been quantitatively and qualitatively enriched. The paper is structured as follows: Section 2 presents previous work related to the research on MWEs in Romanian theoretical and computational linguistics, Section 3 describes the corpus; the types of MWEs annotated in this new version are presented and exemplified in Section 4, while the next section expands on the methodology adopted for enhancing and annotating this new version; some statistics of the annotated corpus are given in Section 6; the corpus has already been used in two shared tasks that are presented in Section 7, where we also make some remarks on them, especially on

[1] https://typo.uni-konstanz.de/parseme/
[2] https://unidive.lisn.upsaclay.fr/

the paraphrasing subtask[3], before announcing the corpus availability and concluding the paper.

2. Related Work

Research on MWEs in Romanian has resulted in a diverse set of linguistic descriptions, lexical inventories, and annotated corpora, reflecting both theoretical and applied interest. From the perspective of language resource development, existing work can be broadly grouped into descriptive studies and educational inventories, computationally oriented lexical and corpus resources, and multilingual frameworks supporting cross-lingual comparability.

Early and subsequent studies in Romanian linguistics have documented MWEs primarily from a descriptive and pedagogical standpoint, with particular attention to verbal expressions (Ioanițescu, 1956; Dimitrescu, 1958). These contributions have clarified definitional criteria, structural variation, and graduality, and have resulted in representative inventories used in grammar teaching and linguistic analysis (Căpățână, 2007; Pană Dindelegan et al., 2025). While highly valuable as reference works, such resources are not designed for computational reuse and do not provide corpus-level annotation aligned with current NLP standards.

More recently, several resources have addressed Romanian MWEs from a computational perspective. These include a lexicon of Romanian verbal MWEs offering linguistically uniform descriptions across multiple levels (lexical, morphological, syntactic, semantic, and stylistic), also considering cross-linguistic comparisons with Bulgarian and English (Leseva et al., 2024). Treebank-based resources, such as the Romanian Reference Treebank (Barbu Mititelu, 2013), have further incorporated explicit annotation of functional MWEs, including conjunctions (Barbu Mititelu & Voicu, 2024) aligned with Penn Discourse Treebank relations (Webber et al., 2019). Together, these efforts have contributed to improved modelling of Romanian MWEs in parsing and discourse analysis, though they remain limited in scope, either focusing on a single MWE category (verbal ones) or covering only specific functional subclasses (multiword conjunctions).

In parallel, Romanian has been included in multilingual initiatives targeting the automatic identification of MWEs. The PARSEME shared tasks have provided corpora annotated with verbal MWEs for over 20 languages, used as training and evaluation data for supervised systems (Savary et al., 2018; Savary et al., 2023). Common annotation guidelines and typologies have ensured cross-lingual consistency, while allowing for language-specific distinctions, thereby enabling meaningful comparison across languages. However, until recently, these multilingual resources have largely concentrated on verbal MWEs, leaving other syntactic categories underrepresented.

Recent extensions of multilingual annotation frameworks have addressed this limitation by broadening the scope of MWE annotation to include nominal, modifier, and functional expressions. Updated guidelines[4] now support the annotation of MWEs across all major syntactic categories, providing a unified framework for corpus development and enabling richer linguistic coverage. This shift has opened new opportunities for creating language-specific resources that are both internally consistent and interoperable with multilingual datasets.

The present work builds on these developments by contributing an expanded Romanian corpus annotated with MWEs across all syntactic categories. By focusing on systematic corpus annotation rather than isolated inventories or task-specific subsets, the resource aims to fill a gap in existing Romanian language resources (Boroș et al., 2017; Barbu Mititelu et al., 2019; Avram et al., 2023) and to support a wide range of applications, including MWE identification, syntactic and semantic analysis, and language learning tools. In this sense, the contribution is positioned primarily as a reusable and extensible language resource, aligned with current multilingual standards and designed to facilitate both language-specific and cross-lingual research.

3. The Corpus

The previous versions of the PARSEME-Ro corpus included 56,703 sentences totaling 1,015,623 tokens. In this annotation campaign, we added 17,168 sentences (486,141 tokens). The corpus had to be enriched so as to offer testing material for the shared task on MWE identification (see below Section 7). Given that the previous corpus had been in the public space since its creation and all Large Language Models had seen it, it was mandatory to create a new test set, from the same text domain (so that evaluation is made in-domain), and the shared task organizers imposed some constraints on the size of this new test set and on the number of MWEs to be included in it.

The PARSEME-Ro corpus (in all its previous releases[5] and the current version) contains

[3] Remarks on the MWE identification subtasks and on the challenges in the annotation of the MWEs are presented by Barbu Mititelu et al. (2026).

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

[5] Version 1.0: https://lindat.mff.cuni.cz/services/catalog/view/http:__hdl_handle_net_11372_LRT-2282, version 1.1: https://lindat.mff.cuni.cz/services/catalog/view/http:__hdl_handle_net_11372_LRT-2842, version 1.2: https://lindat.mff.cuni.cz/services/catalog/view/http:__h

exclusively journalistic texts. They are original texts that were retrieved from the daily issues of two newspapers from the period 2003-2017.

The choice of a journalistic corpus is motivated by the fact that this text type best represents the standard register of a language (see Crystal, 2008: 450, s.v. standard), offering the highest degree of representativeness (i.e., a stable mid-frequency lexicon, and also necessary lexical innovations).

Table 1 shows that the new corpus is about 40% larger than its previous versions and that the new data contains much longer sentences.

	Older data (v.1.3)	New data	TOTAL data
sentences	56,703	14,517	71,220
tokens	1,015,623	407,801	1,423,424
tokens/ sentence	18	28	20

Table 1: Statistics of the PARSEME-Ro 2.0.

The format of the corpus is the same as in previous versions (the CUPT format[6]), with 11 columns, and the span of a MWE is determined by assigning the same ID number to all its components in the last column (see Barbu Mititelu et al., 2019 for a thorough explanation of these aspects).

4. Types of MWEs Occurring in PARSEME-Ro

Earlier editions of the PARSEME-Ro corpus concentrated on the manual annotation of verbal MWEs, with their categorization and illustrative examples documented by Barbu Mititelu et al. (2019, 2022). The release of PARSEME-Ro version 2.0 extended the corpus to cover MWEs from all syntactic categories, following the PARSEME annotation guidelines v. 2.0. As a result, the corpus now includes the types of MWEs that are detailed in the Appendix of this paper. All types of nominal, modifier and functional MWEs defined in the PARSEME guidelines occur in Romanian, unlike the verbal ones where only five types (out of the eight defined) do occur in Romanian[7].

5. Work Methodology

To significantly decrease the amount of human work, we devised an annotation strategy that relies on automation as much as possible, given the expectedly large number of occurrences of MWEs: e.g., the functional MWE category is a closed class, but it is highly frequent. As they are rarely ambiguous, this makes the automatic annotation very efficient. Nevertheless, human validation and correction are essential.

The purpose was not to come up with technically flawless and complex heuristics for automation, but to find good enough solutions that reduce the manual processing time. In devising the automatic annotation strategy, some decisions had to be made to compromise between precision and recall:

1. Allowing intervening words between the different tokens in a MWE is prone to over-generation (low precision, good recall), predicting too many MWE candidates that have to be removed in the manual correction phase; on the other hand, eliminating completely the intervening words option leads to under-generation (good precision, low recall) and more manual annotation work; our decision was to allow for only two[8] intervening words between the tokens, with the aim of capturing the most distance dependencies while limiting noise.

2. Matching of the dictionary entries over the corpus could be done at lemma or at form level. The former is preferred when (some of) the MWE's components inflect, while the latter is appropriate for MWEs with no inflection of (some of) their components. Both matching strategies were used in our annotation work, in ways that will be clarified in the following sections.

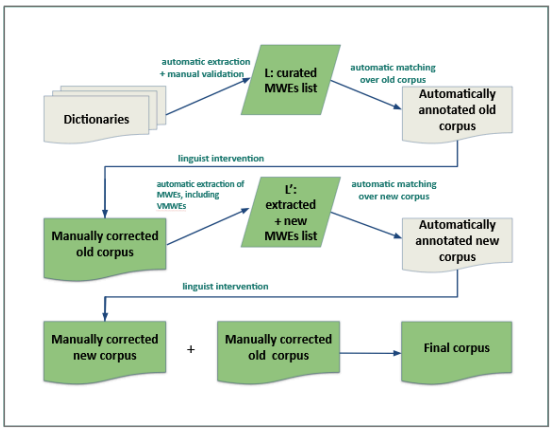

Figure 1. Workflow for corpus annotation

Figure 1 presents a diagram of the annotation workflow. The procedure has two steps: (i) Annotation of the older PARSEME-Ro corpus (version 1.3) with new, non-verbal MWE types (see Subsection 5.1 below). This includes a dictionary-based automatic annotation followed by manual validation. (ii) With the purpose of creating a new test set, we enriched the corpus with new texts and annotated them with all MWE types. This step involved automatic annotation of

dl_handle_net_11234_1-3367, version 1.3: https://lindat.mff.cuni.cz/services/catalog/view/http:__hdl_handle_net_11372_LRT-5124.

[6]https://gitlab.com/parseme/corpora/-/wikis/CUPT-format

[7] By occurring in Romanian, we mean that they are specific to the language, not necessarily present in this corpus.

[8] This number was established by observation of the data to be annotated.

MWEs of all parts of speech, out of which 300 are new ones, i.e. not occurring in the rest of the corpus (see Subsection 5.2 below), followed, again, by manual validation and correction.

5.1. Annotating older data with non-verbal MWEs

This subsection presents the way in which the old version of the PARSEME-Ro corpus (v.1.3) was enriched with annotation of MWEs of nominal, modifier and functional types.

5.1.1 Curating a dictionary-based list of non-verbal MWEs

We took advantage of the existence of Romanian MWE dictionaries in electronic format[9] (PDF): *DELS* (Mărănduc, 2010), *Dicționar de expresii românești în contexte*, Vol. 1-4 (Dictionary of Romanian Expressions in Context, Ilincan, 2015, 2016, 2021a, 2021b) and *Dicționar frazeologic al limbii române* (Phraseological Dictionary of the Romanian Language, Tomici, 2009). The text was extracted from these PDF files using dedicated scripts and iTextSharp .Net library. The TXT files that were obtained this way were further processed through a semi-manual curation workflow. The procedure comprised:

- content cleaning (removal of definitions, examples, usage notes, variants, and cross-references),
- correction of extraction-related errors (line-break segmentation and diacritic encoding issues),
- normalization of alternation-based entries (e.g. alternation marked by parentheses: *ca (la mama) acasă* results in two expressions, *ca acasă* and *ca la mama acasă;* alternation marked by *sau* 'or': *minciună cu coadă sau minciună cu coarne* results in *minciună cu coadă* and *minciună cu coarne*, having the similar meaning 'blatant lie'*),*
- formatting the candidate MWEs into single-entry lists.

The online dictionary *Dicționarul ortografic, ortoepic și morfologic al limbii române*[10] (DOOM, The Orthographic, Orthoepic, and Morphological Dictionary of the Romanian Language) provides functionality for retrieving and downloading lists of idiomatic expressions based on part of speech (POS) queries, by means of which we extracted the MWEs therein, with their associated PoS.

A common MWEs inventory was compiled by concatenating all the entries extracted from the aforementioned dictionaries (duplicates were automatically removed). This extended inventory (L0) comprises over 30,000 MWEs (e.g., idioms, sayings, expressions, etc.). It only partially aligns with our journalistic corpus and contains many entries that are not relevant to it. Therefore, the inventory was matched to the corpus to restrict it to a shorter list of relevant entries.

As a matching strategy at this step, we opted for form matching, as the types of MWEs to be annotated display little inflexion (verbal MWEs, which have big morphological flexibility on their verbal head, are not the focus in this annotation phase while the other flexible category, NID, represents only around 10% of all the MWEs in the curated list): 2,034 unique MWEs were matched with the corpus.

They were manually labeled with PARSEME MWE categories by a team of six linguists, following a two-step procedure to allow cross-validation. The online version of *Dicționarul explicativ al limbii române*[11] (DEX, The Explanatory Dictionary of the Romanian Language), a comprehensive resource, was manually consulted in the process of human validation.

Step 1. For each expression, one label from the PARSEME Guidelines version 2.0 label set was assigned, following the prescribed test battery. Expressions that function as different POSes in different contexts, and therefore correspond to different MWE labels, were expanded in two separate corresponding entries. The most frequent such case is that of the AdjID/AdvID distinction, e.g.: *ca lumea* 'proper'/'properly' is an AdjID in *apartament ca lumea* 'proper apartment' and an AdvID in *să înveți ca lumea* 'to study properly'.

Some of the errors identified in the list originate in the automatic parsing of the dictionaries, while others were expressions that did not pass the MWE PARSEME tests, although considered as such by the dictionary authors. As is well known in the field, there is no universally accepted definition of MWEs (Baldwin & Kim, 2010). Erroneous entries were labeled as NOT MWE, as a marker for future elimination.

Step 2. Cross-validation involved independent annotation of each entry by two annotators. The inter-annotator agreement rate across all 2,034 entries was 57.9% (1,178 entries). Both the consensus and the disagreement sublists were subject to a third round of validation. Some of the homonymous entries mentioned before were identified by observing conflicting labelling. All other cases of disagreement were discussed in linguists team meetings, to select the correct label. Overall, 40% (816) of the matched MWEs list were classified as NOT MWE and excluded. The final dataset (L), after expanding homonymous entries, comprised 2,010 MWEs.

5.1.2 Automatic Annotation of the Old Corpus

The curated and labelled L inventory was again matched over PARSEME-Ro, automatically annotating in the corpus all matched occurrences of the inventory entries. As mentioned above, the strategy involved

[9] The MWE are described in a traditional fashion therein, i.e. they are mostly defined, with some remarks on their possible forms.

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

[11] https://dexonline.ro/

word-form matching and a 2 token window, which we considered a practical balance between precision and recall. Homonymous expressions were annotated with both associated labels, leaving the disambiguation of the label in context for the manual correction.

In devising the annotation script, we had to pay particular attention to the correct unique ID numbering of the new annotated expressions: since verbal MWE had already been annotated, we needed to identify the biggest ID number *n* used in a specific sentence and start the ID numbering of the new expressions from *n+1*. This is important both for subsequent proper rendering of the expressions in the human validation tool and for correct automatic processing in future corpus-based tasks such as MWE identification applications training.

5.1.3. Manual Correction of the Automatic Annotation of the Old Corpus

The correction and validation of the automatic annotation were made on the FLAT platform[12], in a PARSEME dedicated instance, under individual accounts. For this purpose, the corpus was partitioned in files limited to a maximum of 500 sentences, due to limitations of the FLAT platform. For details about the issues encountered in the manual validation, see Section 5.3.

5.2. Annotating the new data

In this subsection we present the way in which the newly added part of the corpus was annotated with MWEs of all parts of speech.

In this step, the automatic annotation employed a lemma-based strategy, which was necessary to cope with the morphological variability displayed by verbal MWEs. All annotated MWEs in PARSEME-Ro were automatically extracted from the corpus in their lemma form and an inventory (L1) with 4043 unique MWE lemmas was constructed.

The L1 inventory was used to select new journalistic data that contain at least 300 new MWEs (for evaluating the MWE identification systems in the shared task) by applying the following heuristic: (1) Compare L1 with the initial extended inventory (L0) extracted from dictionaries and create L2=L1-L0, a list that would contain new MWEs, un-seen in the old corpus. (2) Match L2 inventory over un-seen journalistic documents resulting in the inventory L3. (3) From L3, create a list of 300 curated, labelled MWEs (L4), as proceeded in 5.1.1; (4) Match L4 over the un-seen corpus and select documents until all MWEs in L4 have at least one match.

L1 and L4 inventories were combined in L' and used to automatically annotate the newly selected documents. Manual validation and correction then followed (see Section 5.1.3 and 5.3), producing a new dataset that was integrated into the corpus, resulting in the updated PARSEME-Ro 2.0.

5.3. Human validation in detail

Human validation of the automatic annotation involved such activities as:

(i) modifications, which include: correcting the MWE type label; removing one of the labels in case of homonymous expressions; adding and/or removing tokens from a MWE;

(ii) deletions of whole annotations in case of false positives, many of them stemming from the relaxed matching strategy, but some of them being occurrences with a compositional meaning: e.g. *în natură* "in kind", considered AdjID only in contexts such as *Firma nu plătește în natură.* ("The company does not pay in kind"), was automatically annotated as MWE in the syntagm *petreceți timp în natură* ("spend time in nature"), which is a false positive.

(iii) insertions: this were necessary when the corpus contained MWEs that were not recorded in the electronic dictionaries we relied on or when the distance between the components of a MWE was higher than 2 words (i.e., the limit we used in the automatic annotation): 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.").

As seen in Figure 2, the automation significantly reduced the number of required manual operations (35% of automatically annotated MWEs left unchanged after manual correction).

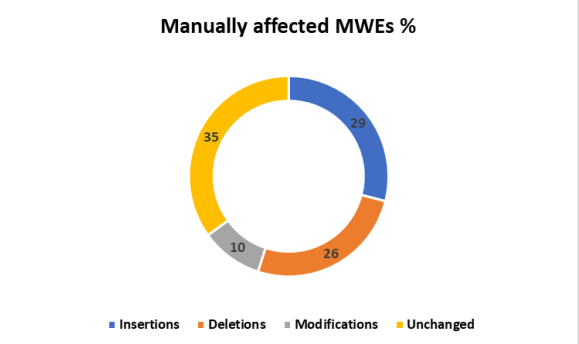

Figure 2: Operations done in the manual correction of the automatic annotation step

For time constraints and limited staffing reasons, each file underwent only one validation, equally distributing the files among the six linguists. Nevertheless, to evaluate the reliability of the annotation process and the degree of agreement within the annotation team, a subset of the corpus of 2,000 randomly sampled sentences underwent double annotation. The inter-annotator agreement score[13] of 0.78 indicates a strong level of consistency across

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

[13] This was computed using the evaluation scripts provided by the PARSEME team (Savary et al., 2017)

annotators and suggests that the annotation guidelines were applied in a largely uniform manner.

5.5. Ensuring Annotation Consistency

Annotation consistency was ensured using the methodology designed and implemented in the PARSEME project. By means of a collection of Python libraries, MWE annotations in PARSEME-Ro, together with instances of the same sequences that were omitted during the annotation stages, were extracted and grouped by the corresponding unique MWE. Human validators examined all occurrences of a specific MWE in context: visualising different annotations in similar contexts or identical annotations in divergent contexts made the errors easy to spot and correct. The F-measure between corpus versions before and after consistency check was 86, which suggests a fairly high level of corpus consistency.

6. Statistics on PARSEME-Ro v. 2.0

Figure 3 shows the distribution of the 64,601 MWEs annotated in the corpus according to their parts of speech, as annotated in version 2.0 of the PARSEME-Ro corpus. The data indicate that function MWEs represent the most frequent category within the corpus. Although they belong to a closed morphological class, they play a crucial role in establishing logical and syntactic relations between sentence constituents. The token-to-MWE ratio is 22; when considered alongside the average sentence length of 20 tokens (see Table 1), this suggests that, on average, each sentence contains approximately 1.1 MWEs.

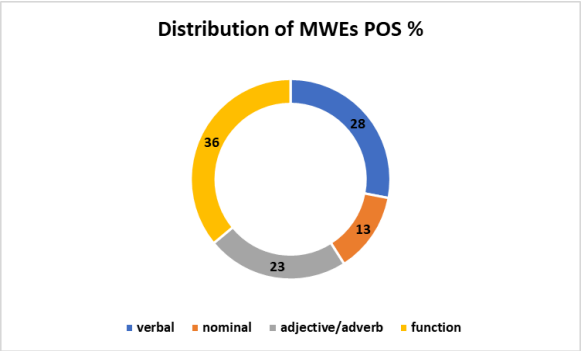

Figure 3: Statistics on MWEs in PARSEME-Ro 2.0.

As illustrated in Figure 3, functional MWEs represent the most frequent category, followed by verbal MWEs, adjectival and adverbial MWEs, and nominal MWEs. Within these broad classes, subtype distribution displays different degrees of asymmetry. Major imbalances can be observed in certain cases: for example, PronIDs occur considerably less frequently than fully lexical nominal MWEs, while DetIDs are markedly underrepresented among functional MWEs. Moderate asymmetries are also attested: although AdpIDs constitute the most frequent subtype, ConjIDs still show a substantial number of occurrences. By contrast, adjectival and adverbial MWEs (AdjIDs and AdvIDs) exhibit a relatively balanced distribution.

7. PARSEME-Ro v.2.0 in shared tasks

The new version of the corpus was used in one shared task with two subtasks recently organized by the PARSEME community: one on the automatic identification of MWEs in corpus and the other on paraphrasing a sentence containing exactly one MWE (Scholivet et al., 2026). The brief description of these tasks is made in the next subsections with some focus on the Romanian data and results.

Test data for both these subtasks was selected from among the new sentence additions of PARSEME 2.0 release, to assure that they are not “seen” by the participating systems. Different efforts were made in the PARSEME initiative to decrease the likelihood of data contamination, including privatising public corpus development Git repositories and secretising consistency checking pages.

7.1. Automatic MWE Identification

Edition 2.0[14] of the PARSEME Shared Task on MWE identification in corpora was based on version 2.0 of the PARSEME corpus, which comprises data for 17 languages, including Romanian. Ten systems were evaluated using two standard F-score variants: MWE-based and token-based.
The highest ranking 3 systems in this subtask for Romanian achieved 82.27, 83.60 and 85.65 MWE-based F-scores, the best one being a Romanian monolingual LLM based model.

7.2. Automatic MWE Paraphrasing

On the occasion of organizing the shared task on paraphrasing MWEs, a multilingual dataset containing paraphrases of 100 sentences in each of the 14 participating languages was compiled (by native speakers along unified guidelines).
Receiving as input a raw sentence containing exactly one unmarked verbal, nominal, or adjectival idiom, systems must produce a paraphrase that rewrites the original MWE while preserving the sentence’s meaning. A masked BERT-score is used to check if at least one of the original MWE components was removed and BERTscore (Zhang et al., 2020) is computed against two reference paraphrases: a minimal and a creative one. Supplementary, a manual score is devised and the final score is computed as an average of the BERTscore and the manual

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

score.

The manual scoring of the automatically generated paraphrases considered three distinct criteria: 1. the degree to which the meaning of the removed MWE is preserved; 2. the degree to which the overall sentence meaning is preserved; 3. grammaticality and naturalness. The final score was computed as a sum of the three criteria scores (with values ranging from 0 to 3); in cases where the MWE was not replaced at all, the only assigned value was 0.

Despite the granular criteria, the task encountered various challenges arising from the inherent heterogeneity of MWEs, from the peculiarities of the Romanian word-formation system and its interface areas with syntax and morphology (see Barbu Mititelu et al., 2026) and from the impossibility to account for every complex and unpredictable scenario produced by automatic paraphrasing. Furthermore, the conceptual scope of the term *paraphrase* introduces substantial challenges. While linguistics does not formally concern itself with a rigorous definition of paraphrase (utilizing the concept mainly for grammatical testing or as a definitional method in lexicography), the NLP field (Bhagat & Hovy, 2013; Pehlivanoğlu et al., 2024) adopts a broader interpretation; this scope encompasses what traditional grammar labels as 1. morpho-lexical synonyms (e.g., *durere dentară*, 'pain dental' and its perfect equivalent *durere de dinți*, 'lit. pain of teeth', both meaning "toothache") or 2. syntactic variations (e.g., *Firma construiește blocul* "The company is building the block", active voice, and its equivalent from another agency perspective *Blocul este construit de firmă* "The block is being built by the company", passive voice). So, although working definitions may appear straightforward ("paraphrases are sentences or phrases that convey the same meaning using different wording", Bhagat & Hovy, 2013), in practice, classifying paraphrase types and ensuring precise semantic equivalence remain difficult to manage. Bhagat & Hovy (2013) catalogue 25 operations that generate (correct) quasi-paraphrases and empirically validate them through corpus distribution and human judgment. Their core conclusion is that paraphrase is a gradient, heterogeneous phenomenon rather than a matter of strict semantic equivalence; this is a useful theoretical backdrop for the difficulties and the inherent inconsistencies encountered among human evaluators. Drawing upon these considerations, we will now outline the main challenges and shortcomings identified in automated paraphrase outputs.

Criterion 1, which assessed how well the paraphrase preserves the meaning of the removed MWE, was straightforward and presented no particular challenges: e.g., for the MWE *a face față* "to cope, to manage", the paraphrase *a rezista* "to resist" received the maximum score, while *a se adapta* "to adapt" received a lower score. However, a broader limitation emerged: the difficulty of capturing the explanatory power of a paraphrase within the existing criteria. Paraphrasing involves lexical substitution to varying degrees. Many paraphrases, however, were based on inflectional and derivational changes within the same paradigm and have very limited explanatory power: e.g., *sunt arestați preventiv* "are being held on remand", for the nominal MWE *arestare preventivă* "preventive detention", where an abstract noun was replaced by an adjectivized participle of the same verb *a aresta* "to arrest"; *bugetul statului* "the state's budget", for *bugetul de stat* "the state budget", where a prepositional phrase is replaced by a genitive construction; *statele europene estice* "Eastern European states", for *Europa de Est* "Eastern Europe", where a noun was replaced by an adjectival form within the same lexical family (cf. Bhagat & Hovy, 2013, for the low scores assigned for paraphrase types involving verb-to-adjective/adverb conversion or verb-to-deverbal-agent-noun substitution, i.e., the paraphrase of a verbal MWE by a deverbal agent derivative).

Paraphrases based on relative clauses also have limited explanatory power (e.g., *membrii care au fondat* "the members who founded", for the MWE *membrii fondatori* "founding members"), just like those based on changes in functional elements (e.g., *gândul de pe urmă*, for *gândul din urmă* "last thought/afterthought"; cf. paraphrases based on function word variations, which score 85 out of 100 in Bhagat & Hovy, 2013).

A special case is represented by zero paraphrases, where the MWE was fully omitted without loss of the overall meaning of the sentence, reflecting contextual or lexicalized semantic condensation (e.g., *independență ø* "independence", for *independență de stat* "state independence"; *comunicat ø* "release" for *comunicat de presă* "press release"; *partide ø* "parties" for *partide politice* "political parties").

With respect to criterion 2, some lexical, grammatical, semantic, and instruction-following hallucinations were identified (cf. the hallucination typology in Gogoulou et al., 2025). Accordingly, depending on the situation, a low score was assigned in cases of:

(i) lexical fabrication, i.e., the generation of non-existent word forms (such as the verb *să *imponheze* in the sentence *Am solicitat Partidului Mișcarea Populară să imponheze sau să restricționeze o asemenea formulare* "I have requested the People's Movement Party to ?imponheze and to restrict such a wording");

(ii) morphological and morphosyntactic hallucinations, a type particularly prominent in Romanian given its rich morphology (e.g., instead of the correct verbal form *să ne adaptăm* "(we) adapt (ourselves)", generated as a paraphrase for the MWE *a face față* "to cope, to manage", the model produced *să ne *adaptezăm*, a form containing the morpheme *-ză-*, absent from the first person plural of the present indicative paradigm, resulting from overgeneralization of this morpheme across all persons of the verb; as a paraphrase for the same MWE, the model also generated the reflexive construction **să ne supraviețuim* "to survive (*ourselves)", which is ungrammatical, as the verb *supraviețui* "survive" does not take a reflexive pronoun in Romanian);
(iii) incoherent over-paraphrasing, whereby the model targeted not only the MWE itself but the entire context in which the MWE was embedded, and introduced unauthorized external, factual-sounding information not present in the source text (e.g., the sentence[15] *Cuplul R. și A. M. M., nașii de cununie ai E. B. și R. I., au decis să* [[*pună capăt*]] *căsătoriei* "The couple R. and A. M. M., the godparents of E. B. and R. I., decided to [[put an end to]] the marriage" was transformed into *Cuplul R. și A. M. M., făcuți de la început unii dintre cei mai influenți membri ai familiei B., a anulat căsătoria* "The couple R. and A. M. M., made from the beginning some of the most influential members of the B. family, cancelled the marriage");
(iv) failure to adhere to negative constraints, which occurs when the model fails to suppress the original MWE as instructed, instead embedding it unchanged within what is effectively a dictionary definition (e.g., the MWE *certificatul de naștere* "birth certificate" was not paraphrased, but rather incorporated into a lengthy definitional entry: *Certificatul de naștere este un document care...* "The birth certificate is a document that...");
(v) instruction-following failure, manifesting as paraphrase overproduction, i.e., the generation of an excessively large number of paraphrases (9-10, instead of 1) for the same MWE.
Analysis of the paraphrase assessment process suggests that criterion 3, covering grammaticality and naturalness/register, quite often involves conflicting signals and warrants decoupling. In a hierarchical evaluation, grammaticality must be assessed first, as naturalness becomes moot if the utterance is ill-formed. Annotators exhibited a binary bias, frequently assigning 0 to any malformed paraphrase despite the availability of a '1' score for minor grammatical/syntactic errors (e.g., when deciding whether the ungrammatical reflexive use of *a supraviețui* "to survive" as a paraphrase for *a face față* "to cope, to manage" is a major or minor error, as in: *Vom fi în măsură să ne supraviețuim unei noi recesiuni economice?* "Will we be able to survive [ourselves] a new economic recession?"). Regardless, any perceived grammatical error tended to preclude further evaluation of naturalness.
The best system for this subtask for Romanian (89.25 global masked BERT-score and 77.31 manual score) relies on prior MWE identification with the highest ranking system of subtask 1 followed by GPT-4o queries using POS category-oriented prompts.

8. Corpus Availability

The PARSEME-Ro corpus has been released with a permissive license since its first version: CC-BY 4.0[16]. This new, enlarged and enhanced version is made available[17] with such a license, in a common release with all the other languages that participated in PARSEME v. 2.0. The paraphrasing subset will also be released alongside data for the other languages.

9. Conclusions and Future Work

In this paper, we have presented the most recent efforts to expand and enhance the Romanian corpus annotated for multiword expressions, namely PARSEME-Ro. This resource can be integrated with an electronic lexicon of Romanian MWEs (Leseva et al., 2024), enabling a more detailed linguistic description of these expressions.

Beyond its relevance for shared tasks and evaluation campaigns, the corpus can also serve as a valuable resource for language learning, particularly in the context of second language acquisition. It provides learners with numerous authentic contexts illustrating the use of a wide range of MWEs.

The corpus is likewise useful for linguists and language specialists, offering a contemporary snapshot of the inventory, frequency, and distribution of MWEs in the journalistic genre. When compared with existing dictionaries, such data may reveal tendencies toward the obsolescence of certain expressions as well as the emergence of new ones. These observations should, however, be interpreted with caution, as the corpus is not representative of the entirety of journalistic writing, let alone of the Romanian language as a whole. Future work will focus on extending the analysis of MWEs by incorporating additional text genres into the corpus, which will make it possible to identify similarities and

[15] The names of the public figures were abbreviated for GDPR reasons.

[16] https://creativecommons.org/licenses/by/4.0/deed.en
[17] http://hdl.handle.net/11372/LRT-6123

differences between journalistic discourse and other types of texts.

10. Acknowledgements

The work presented here was carried out with support from the following projects: (i) *Large Language Models for the European Union (LLMs4EU)*, project no. 101198470, call DIGITAL-2024-AI-B-06-LANGUAGE, funded by the European Union; (ii) a grant of the Ministry of Research, Innovation and Digitalization - UEFISCDI, PNCDI IV, project number PN-IV-P8-8.2-EUD-2025-0061; (iii) CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology); (iv) a grant of the Ministry of Education and Research, CCCDI – UEFISCDI, project number PN-IV-PCB-RO-MD-2024-0142, within PNCDI IV. Views and opinions expressed are however those of the authors 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.

11. Bibliographical References

Avram, A., Barbu Mititelu, V. and Cercel, D.-C. (2023). Romanian Multiword Expression Detection Using Multilingual Adversarial Training and Lateral Inhibition. In Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), pages 7–13, Dubrovnik, Croatia. Association for Computational Linguistics.

Bhagat, R., Hovy, E. (2013). What is a paraphrase? Computational Linguistics, 39(3), p. 463-472.

Baldwin T., and Kim, S.N. (2010). Multiword expressions. In Nitin Indurkhya and Fred J. Damerau (Eds.), Handbook of Natural Language Processing. Boca Raton, FL: CRC Press.

Barbu Mititelu, V. (2013). Sistemul all-inclusive în reprezentarea cunoştinţelor lexicale. In Ofelia Ichim (ed.), Tradiţie/inovaţie - identitate/ alteritate: paradigme în evoluţia limbii şi culturii române, Iaşi, Editura Universităţii „Alexandru Ioan Cuza”, 2013, p. 9-18.

Barbu Mititelu, V., Cristescu, M. and Onofrei, M. (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.

Barbu Mititelu, V., Cristescu, M., Mitrofan, M., Zgreabăn, B.-M., & Bărbulescu, E.-A. (2022). A Romanian Treebank Annotated with Verbal Multiword Expressions. In *Proceedings of the Fifth International Conference on Computational Linguistics in Bulgaria* (CLIB 2022), pages 137–145.

Barbu Mititelu, V. and Voicu, T. (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.

Barbu Mititelu, V., Giouli, V., Korvel, G., Liebeskind, C., Lobzhanidze, I., Makhachashvili, R., Markantonatou, S., Markovic, A. and Stoyanova, I. (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)*, p. 41–57, Albuquerque, New Mexico, U.S.A.. Association for Computational Linguistics.

Barbu Mititelu, V., Cristescu, M., Irimia E., Vasile Mîrzea, C. (2026). Two Birds with One Stone: Annotating Romanian Multiword Expressions with an Eye to the PARSEME 2.0 Guidelines Applicability. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 66–74, Rabat, Morocco. Association for Computational Linguistics.

Boroș, T., Pipa, S., Barbu Mititelu, V. and Tufiș, D. (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.

Căpățână, C. (2007). *Elemente de frazeologie*. Editura Universitaria, Craiova.

Crystal, D. (2008). *A dictionary of linguistics and phonetics*, 6th ed. Blackwell Publishing.

Dimitrescu, F. (1958). *Locuțiunile verbale în limba română* (The verbal locutions in Romanian). EA, București.

Ilincan, V. (2015). *Dicționar de expresii românești în contexte* [DERC]. A-C. Volume I; 2016, D-N. Volume II; 2021a, O-R. Volume III; 2021b, S-Z. Volume IV, Presa Universitară Clujeană, Cluj-Napoca.

Ioanițescu, E. (1956). Locuțiunile. *Limba română*, 6:48–54.

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

Gogoulou, E., Zahra, S., Guillou, L., Dürlich, L., & Nivre, J. (2025). Can LLMs Detect Intrinsic Hallucinations in Paraphrasing and Machine Translation? *arXiv:*2504.19857.

Leseva, S., Barbu Mititelu, V., Stoyanova, I. and Cristescu, M. (2024). A uniform multilingual approach to the description of multiword expressions. In Voula Giouli and Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 73–116. Berlin: Language Science Press.

Mahajan, U. P., Patil, A. S., and Patil, N. P. (2024). A survey of tools and techniques for multiword expression detection. International *Journal of Computer Applications*, 186(32):11–18.

Mărănduc, C. (2010). *Dicționar de expresii, sintagme și locuțiuni ale limbii române*, DELS, Corint, Bucharest.
Miletić, F. and Schulte im Walde, S. (2024). Semantics of Multiword Expressions in Transformer-Based Models: A Survey. *Transactions of the Association for Computational Linguistics*, 12:593–612.
Pană Dindelegan, G., Brăescu, R., Aranghelovici, C. (2025). *Locuțiunile limbii române*, Univers Enciclopedic, Bucharest.
Pehlivanoğlu, M., Gobosho, R., Syakura, M., Shanmuganathan, V., and De La Fuente Valentín, L. (2024). Comparative analysis of paraphrasing performance of ChatGPT, GPT-3, and T5 language models using a new ChatGPT generated dataset: ParaGPT. *Expert Systems, 41*.
Phelps, D., Pickard, T., Mi, M., Gow-Smith, E. and Villavicencio, A. (2024). Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.
Ramisch, C., Walsh, A., Blanchard, T. and Taslimipoor, S. (2023). A Survey of MWE Identification Experiments: The Devil is in the Details. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 106–120, Dubrovnik, Croatia. Association for Computational Linguistics.
Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15.
Savary, A., Ramisch, C., Cordeiro, S.R., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., Doucet, A. (2017). *The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions*. In the Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), 4 April 2017, Valencia, Spain.
Savary, A., Candito, M., Barbu Mititelu, V., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S. R., Eryiğit, G., Giouli, V., van Gompel, M., HaCohen-Kerner, Y., Kovalevskaitė, J., Krek, S., Liebeskind, C., Monti, J., Parra Escartín, C., van der Plas, L., QasemiZadeh, B., Ramisch, C., Sangati, F., Stoyanova, I., and Vincze, V. (2018). PARSEME multilingual corpus of verbal multiword expressions. In Markantonatou, S., Ramisch, C., Savary, A., and Vincze, V. (eds.), *Multiword Expressions at Length and in Depth: Extended Papers from the MWE 2017 Workshop*, pages 87–147, Language Science Press, Berlin.
Savary, A., Ben Khelil, C., Ramisch, C., Giouli, V., Barbu Mititelu, V., Mohamed, N. H., Krstev, C., Liebeskind, C., Xu, H., Jiang, M., Stymne, S., Güngör, T., Pickard, T., Guillaume, B., Bhatia, A., Butler, A., Candito, M., Gantar, A., Iñurrieta, U., Gatt, A., Kovalevskaite, J., Krek, S., Lichte, T., Ljubešic, N., Monti, J., Escartín, C.P., Shamsfard, M., Stoyanova, I., Vincze, V., Walsh, A. (2023). *PARSEME Corpus Release 1.3*. In the Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), 6 May 2023, Dubrovnik, Croatia.
Scholivet, M. Savary, A., Ramisch, C., Bilinski, E., Nakamura, T., Mitrofan, M., Păiș, V. (2026). *Edition 2.0 of the PARSEME Shared Task on Multilingual Identification and Paraphrasing of Multiword Expressions*. In Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026).
Tomici, M. (2009). *Dicţionar frazeologic al limbii române*. Editura Saeculum Vizua, Bucharest.
Webber, B., Prasad, R., Lee, A. and Joshi, A. (2019). *The Penn Discourse Treebank 3.0 Annotation Manual*. Philadelphia, University of Pennsylvania, 35:108.
Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A. (2023). *Multimodal chain-of-thought reasoning in language models*. arXiv preprint arXiv:2302.00923.

12. Language Resources References

Savary, Agata, Ramisch, Carlos, Cordeiro, Silvio Ricardo, Sangati, Federico, Vincze, Veronika, QasemiZadeh, Behrang, Candito, Marie, Cap, Fabienne, Giouli, Voula, Stoyanova, Ivelina, Doucet, Antoine, Adalı, Kübra, Barbu Mititelu, Verginica, Bejček, Eduard, El Maarouf, Ismail, Eryiğit, Gülşen, Galea, Luke, Ha-Cohen Kerner, Yaakov, Liebeskind, Chaya, Monti, Johanna, Parra Escartín, Carla, Kovalevskaitė, Jolanta, Krek, Simon, van der Plas, Lonneke, Aceta, Cristina, Aduriz, Itziar, Antoine, Jean-Yves, Attard, Greta, Azzopardi, Kirsty, Boizou, Loic, Bonnici, Janice, Boz, Mert, Bumbulienė, Ieva, Busuttil, Jael, Caruso, Valeria, Cherchi, Manuela, Constant, Matthieu, Czerepowicka, Monika, De Santis, Anna, Dimitrova, Tsvetana, Dinç, Tutkum, Elyovich, Hevi, Fabri, Ray, Farrugia, Alison, Findlay, Jamie, Fotopoulou, Aggeliki, Foufi, Vassiliki, Galea, Sara Anne, Gantar, Polona, Gatt, Albert, Gatt, Anabelle, Herrero, Carlos, Iñurrieta, Uxoa, Jagfeld, Glorianna, Hnátková, Milena, Ionescu, Mihaela, Klyueva, Natalia, Koeva, Svetla, Kovács, Viktória, Kuzman, Taja, Leseva, Svetlozara, Louisou, Sevi, Lynn, Teresa, Malka, Ruth, Martínez Alonso, Héctor, McCrae, John, de Medeiros Caseli, Helena, Miral, Ayşenur, Muscat, Amanda, Nivre, Joakim, Oakes, Michael, Onofrei, Mihaela, Parmentier, Yannick, Pasquer, Caroline, Pia di Buono, Maria, Priego Sanchez, Belem, Raffone, Annalisa, Ramisch, Renata, Rimkutė, Erika, Rizea, Monica-Mihaela, Simkó, Katalin, Spagnol, Michael, Stefanova, Valentina, Stymne, Sara, Sulubacak, Umut, Tabone, Nicole, Tanti, Marc, Todorova, Maria, Urešová, Zdenka, Villavicencio, Aline, and Zilio, Leonardo, Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions

(edition 1.0), distributed via LINDAT-CLARIAH.CZ

Ramisch, Carlos, Cordeiro, Silvio Ricardo, Savary, Agata, Vincze, Veronika, Barbu Mititelu, Verginica, Bhatia, Archna, Buljan, Maja, Candito, Marie, Gantar, Polona, Giouli, Voula, Güngör, Tunga, Hawwari, Abdelati, Iñurrieta, Uxoa, Kovalevskaitė, Jolanta, Krek, Simon, Lichte, Timm, Liebeskind, Chaya, Monti, Johanna, Parra Escartín, Carla, QasemiZadeh, Behrang, Ramisch, Renata, Schneider, Nathan, Stoyanova, Ivelina, Vaidya, Ashwini, Walsh, Abigail, Aceta, Cristina, Aduriz, Itziar, Antoine, Jean-Yves, Arhar Holdt, Špela, Berk, Gözde, Bielinskienė, Agnė, Blagus, Goranka, Boizou, Loic, Bonial, Claire, Caruso, Valeria, Čibej, Jaka, Constant, Matthieu, Cook, Paul, Diab, Mona, Dimitrova, Tsvetana, Ehren, Rafael, Elbadrashiny, Mohamed, Elyovich, Hevi, Erden, Berna, Estarrona, Ainara, Fotopoulou, Aggeliki, Foufi, Vassiliki, Geeraert, Kristina, van Gompel, Maarten, Gonzalez, Itziar, Gurrutxaga, Antton, Ha-Cohen Kerner, Yaakov, Ibrahim, Rehab, Ionescu, Mihaela, Jain, Kanishka, Jazbec, Ivo-Pavao, Kavčič, Teja, Klyueva, Natalia, Kocijan, Kristina, Kovács, Viktória, Kuzman, Taja, Leseva, Svetlozara, Ljubešić, Nikola, Malka, Ruth, Markantonatou, Stella, Martínez Alonso, Héctor, Matas, Ivana, McCrae, John, de Medeiros Caseli, Helena, Onofrei, Mihaela, Palka-Binkiewicz, Emilia, Papadelli, Stella, Parmentier, Yannick, Pascucci, Antonio, Pasquer, Caroline, Pia di Buono, Maria, Puri, Vandana, Raffone, Annalisa, Ratori, Shraddha, Riccio, Anna, Sangati, Federico, Shukla, Vishakha, Simkó, Katalin, Šnajder, Jan, Somers, Clarissa, Srivastava, Shubham, Stefanova, Valentina, Taslimipoor, Shiva, Theoxari, Natasa, Todorova, Maria, Urizar, Ruben, Villavicencio, Aline, and Zilio, Leonardo, Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1), distributed via LINDAT-CLARIAH.CZ.

Guillaume, Bruno, Ramisch, Carlos, Waszczuk, Jakub, Monti, Johanna, Di Buono, Maria Pia, Sangati, Federico, Speranza, Giulia, Carlino, Carola, Güngör, Tunga, Yirmibeşoğlu, Zeynep, Sak, Haşim, Saraçlar, Murat, Giouli, Voula, Foufi, Vassiliki, Ramisch, Renata, Rademaker, Alexandre, Vale, Oto, Wilkens, Rodrigo, Candito, Marie, Crabbé, Benoît, Segonne, Vincent, Liebeskind, Chaya, Stymne, Sara, Hajič, Jan, Ginter, Filip, Luotolahti, Juhani, Straka, Milan, Zeman, Daniel, Barbu Mititelu, Verginica, Cristescu, Mihaela, Vaidya, Ashwini, Bhatia, Archna, Lichte, Timm, Ehren, Rafael, Jiang, Menghan, Xu, Hongzhi, Walsh, Abigail, Irimia, Elena, and Dowling, Meghan, 2020. *Morpho-syntactically annotated corpora provided for the PARSEME Shared Task on Semi-Supervised Identification of Verbal Multiword Expressions (edition 1.2)*, distributed via LINDAT-CLARIAH.CZ

Savary, Agata, Ramisch, Carlos, Guillaume, Bruno, Hawwari, Abdelati, Walsh, Abigail, Fotopoulou, Aggeliki, Bielinskienė, Agnė, Estarrona, Ainara, Gatt, Albert, Butler, Alexandra, Rademaker, Alexandre, Maldonado, Alfredo, Villavicencio, Aline, Farrugia, Alison, Muscat, Amanda, Gatt, Anabelle, Antić, Anđela, De Santis, Anna, Raffone, Annalisa, Riccio, Anna, Pascucci, Antonio, Gurrutxaga, Antton, Bhatia, Archna, Vaidya, Ashwini, Miral, Ayşenur, QasemiZadeh, Behrang, Priego Sanchez, Belem, Griciūtė, Bernadeta, Erden, Berna, Parra Escartín, Carla, Herrero, Carlos, Carlino, Carola, Pasquer, Caroline, Liebeskind, Chaya, Wang, Chenweng, Ben Khelil, Chérifa, Bonial, Claire, Somers, Clarissa, Aceta, Cristina, Krstev, Cvetana, Bejček, Eduard, Lindqvist, Ellinor, Erenmalm, Elsa, Palka-Binkiewicz, Emilia, Rimkute, Erika, Petterson, Eva, Cap, Fabienne, Hu, Fangyuan, Sangati, Federico, Wick Pedro, Gabriela, Speranza, Giulia, Jagfeld, Glorianna, Blagus, Goranka, Berk, Gözde, Attard, Greta, Eryiğit, Gülşen, Finnveden, Gustav, Martínez Alonso, Héctor, de Medeiros Caseli, Helena, Elyovich, Hevi, Xu, Hongzhi, Xiao, Huangyang, Miranda, Isaac, Jaknić, Isidora, El Maarouf, Ismail, Aduriz, Itziar, Gonzalez, Itziar, Matas, Ivana, Stoyanova, Ivelina, Jazbec, Ivo-Pavao, Busuttil, Jael, Waszczuk, Jakub, Findlay, Jamie, Bonnici, Janice, Šnajder, Jan, Antoine, Jean-Yves, Foster, Jennifer, Chen, Jia, Nivre, Joakim, Monti, Johanna, McCrae, John, Kovalevskaitė, Jolanta, Jain, Kanishka, Simkó, Katalin, Yu, Ke, Azzopardi, Kirsty, Adalı, Kübra, Uria, Larraitz, Zilio, Leonardo, Boizou, Loïc, van der Plas, Lonneke, Galea, Luke, Sarlak, Mahtab, Buljan, Maja, Cherchi, Manuela, Tanti, Marc, Di Buono, Maria Pia, Todorova, Maria, Candito, Marie, Constant, Matthieu, Shamsfard, Mehrnoush, Jiang, Menghan, Boz, Mert, Spagnol, Michael, Onofrei, Mihaela, Li, Minli, Elbadrashiny, Mohamed, Diab, Mona, Rizea, Monica-Mihaela, Hadj Mohamed, Najet, Theoxari, Natasa, Schneider, Nathan, Tabone, Nicole, Ljubešić, Nikola, Vale, Oto, Cook, Paul, Yan, Peiyi, Gantar, Polona, Ehren, Rafael, Fabri, Ray, Ibrahim, Rehab, Ramisch, Renata, Walles, Rinat, Wilkens, Rodrigo, Urizar, Ruben, Sun, Ruilong, Malka, Ruth, Galea, Sara Anne, Stymne, Sara, Louizou, Sevasti, Hu, Sha, Taslimipoor, Shiva, Ratori, Shraddha, Srivastava, Shubham, Cordeiro, Silvio Ricardo, Krek, Simon, Liu, Siyuan, Zeng, Si, Yu, Songping, Arhar Holdt, Špela, Markantonatou, Stella, Papadelli, Stella, Leseva, Svetlozara, Kuzman, Taja, Kavčič, Teja, Lynn, Teresa, Lichte, Timm, Pickard, Thomas, Dimitrova, Tsvetana, Yih, Tsy, Güngör, Tunga, Dinç, Tutkum, Iñurrieta, Uxoa, Tajalli, Vahide, Stefanova, Valentina, Caruso, Valeria, Puri, Vandana, Foufi, Vassiliki, Barbu Mititelu,

Verginica, Vincze, Veronika, Kovács, Viktória, Shukla, Vishakha, Giouli, Voula, Ge, Xiaomin, Ha-Cohen Kerner, Yaakov, Öztürk, Yağmur, Yarandi, Yalda, Parmentier, Yannick, Zhang, Yongchen, Zhao, Yun, Urešová, Zdeňka, Yirmibeşoğlu, Zeynep, Qin, Zhenzhen, Stank, Cristescu, Mihaela, Zgreabăn, Bianca-Mădălina, Bărbulescu, Elena-Andreea, and Stanković, Ranka, PARSEME corpora annotated for verbal multiword expressions (version 1.3), distributed via LINDAT-CLARIAH.CZ

Appendix

Types of MWEs occurring in PARSEME-Ro 2.0.

PoS type	MWE type	Example	Gloss	Translation
Verbal	Verbal Idiom (VID)	*avea în vizor*	'have in sight'	"keep in view, target"
		da nas în nas	'give nose in nose'	"run into, bump into"
	Light Verb Construction (LVC) - LVC.full	*avea întâlnire*	'have meeting'	"have a meeting"
	Light Verb Construction (LVC) - LVC.cause	*pune în aplicare*	'put in application'	"implement, carry out"
	Inherently Reflexive Verb (IRV)	*-și permite*	'3SG.DAT.REFL allow'	"afford"
	Inherently Adpositional Verb (IAV)	*se baza pe*	'3SG.ACC.REFL base on'	"rely on, be based on"
Nominal	Nominal Idiom (NID)	*act de identitate*	'act of identity'	"identity card"
		an de grație	'year of grace'	"year of Our Lord, A.D."
	Pronominal Idiom (PronID)	*Domnia Sa*	'His/Her Lordship'	"His Excellency"
	Deverbal Nominal (NV)	*punere în funcțiune*	'putting in function'	"commissioning, putting into operation"
Modifier	Adjectival Idiom (AdjID)	*cu cântec*	'with song'	"with flair, extravagantly"
		cu o falcă în cer și cu una în pământ	'with one jaw in sky and one in earth'	"very furious"
	Adverbial Idiom (AdvID)	*așa cum se cuvine*	'so as 3SG.ACC.REFL befit'	"properly, as it should be"
		ca oamenii	'like people.THE.PL'	"like decent people, properly"
	Deverbal adjectival / adverbial (AV)	*luat în seamă*	'taken in notice'	"taken into account, considered"
Functional	Determiner Idiom (DetID)	*picior de*	'foot of'	"not a single"
	Adposition Idiom (AdpID)	*de dragul*	'of dear.DEF.SG.M'	"for the sake of"
	Conjunction Idiom (ConjID)	*pe motiv că*	'on reason that'	"because, on the grounds that"
	Interjection Idiom (IntjID)	*ce folos*	'what use'	"what's the point?"