Annotation guidelines

Welcome to the official annotation guidelines of the PARSEME shared task on verbal MWE identification!

Here, you'll find detailed definitions, examples and linguistic tests to guide your decision as to whether a given combination in your language is a verbal multiword expression. Use the table of contents on the left to navigate between sections and the header buttons to show/hide examples.

In addition to these general guidelines, language teams may also provide extra documentation, like lists of borderline cases and decisions taken concerning them. They should all be compatible with these general guidelines.

If you spot errors or if something remains unclear after reading the guidelines, please contact us and we'll do our best to correct the problems.

Authors and contributors (alphabetical order)


Table of contents

- 1 Definitions and scope
  - 1.1 Notation
  - 1.2 Words and tokens
  - 1.3 Verbal multiword expressions
  - 1.4 Syntactic variants of VMWEs
  - 1.5 Lexicalization and open slots
  - 1.6 VMWEs versus collocations
  - 1.7 VMWEs versus metaphors
- 2 Textual annotation scope
- 3 Categories of VMWEs
- 4 Annotation process - decision tree
- 5 Generic tests - Identify VMWEs
- 6 Specific tests - categorize VMWEs
  - 6.1 Structural tests
  - 6.2 Light verb constructions
  - 6.3 Idioms
  - 6.4 Inherently reflexive verbs
  - 6.5 Verb-particle constructions
  - 6.6 Language-specific categories
  - 6.7 Other VMWEs
- 7 Language specific tests
  - 7.1 Particles, prepositions, prefixes
- 8 Annotation management
  - 8.1 Frequently asked questions
  - 8.2 Adding new examples
  - 8.3 Annotation platform FLAT
- 9 Glossary
- 10 Contact

Section 1

Definitions and scope

In this shared task, we aim at identifying verbal Multiword Expressions (VMWEs) in running texts in about 20 languages from several language families. VMWEs are of particular interest to the PARSEME COST action since they frequently introduce discontinuity and long-distance dependency issues, which are central to deep parsing and to other Natural Language Processing tasks.

This document defines the annotation scope and puts forward a classification of VMWEs together with linguistic tests for their identification and categorization.

Section 1.1

Notation

The notational convention used throughout the document is the following:

- **Italic** is used to display example sentences and expressions.
- **Bold** is used to highlight the lexicalized components of a candidate VMWE inside an example (positive or negative).
- **Underline** is used to focus the reader's attention on the important part of an example
Section 1.2
Words and tokens

While the definition of an MWE inherently relies on the notion of a word, manual annotation and automatic identification of VMWEs in our task is performed on texts which are automatically tokenized. It is therefore important to understand the distinction between words and tokens in the context of VMWEs.

A word is a linguistically (notably semantically) motivated unit. The detection of words is, thus, language-dependent and annotation experts should have a clear idea of how to define it for their own language (even if this definition proves hard in general).

A token is a technical and pragmatic notion, defined according to more or less linguistically motivated clues and depending on the particular tokenization tool at hand.

Tokens should ideally be as close as possible to words. However, in practice - due to the hardness of the (automatic) tokenization task - the relation between tokens and words is not always 1-to-1. The following cases occur:

- A token coincides with a word:
  - (BG) вземам, решение, наяве, бял, на, се, д-р
  - (CS) mít, hlad, se, úžas
  - (DE) einen, Spaziergang, machen, Überraschung
  - (EL) κάνω, άνω, κάτω, ποδήλατο, καλός
  - (EN) take, a, walk, astonishment
  - (ES) dar, un, paseo, sorpresa
  - (FA) من کتاب، دویست
  - (FR) faire, une, promenade, étonnement
  - (HR) napraviti/činiti, jedan, šetnja, začuđenost
  - (HU) tesz, egy, séta, meglepetés
  - (IT) fare, una, passeggiata, sorpresa
  - (MT) fekh, libes, sabih
  - (PL) robić 'to do', na 'on', dokładność 'precision'
  - (RO) face, o, plimbare
  - (SV) gå, på, promenad, förvåning
  - (SL) iti, na, en, sprehod, začuđenost

- Several tokens build up one word, like in abbreviations, possessive markers, words with "accidental" separators, inflected or derived forms of foreign names, etc. In this case we speak of a multitoken word (MTW):
  - (BG) т|.|н.| etc.
  - (DE) Wie geht|'|s lit. How goes it 'How are you'
  - (EL) κύριος lit. Mister'
  - (FA) آیت| |-الله، آیت| |-الکاب
  - (FR) aujourd|'|hui lit. 'today'
  - (HU) időjárás|-|jelentés 'weather forecast'
  - (IT) vice|-|presidente 'vice-president'
  - (MT) šebet 'she wore '
  - (ES) pl.| et |pl.| pages
  - (PT) p|. |ej.| lit. 'before Christ'
  - (SV) EU|:|s lit. 'EU's'

- One token can contain several words, like in contractions and compounds. In this case we speak of a multiword token (MWT):
  - (BG) m.|[м.| 'etc.'
  - (CS) na, ce, d-п
  - (DE) machen, Überraschung
  - (EL) ποδήλατο, καλός
  - (EN) take, a, walk, astonishment
  - (ES) dar, un, paseo, sorpresa
  - (FA) من کتاب، دویست
  - (FR) faire, une, promenade, étonnement
  - (HR) jedan, šetnja, začuđenost
  - (HU) séta, meglepetés
  - (IT) fare, una, passeggiata, sorpresa
  - (MT) libes, sabih
  - (PL) Chomsky|'|ego lit. 'of Chomsky'
  - (RO) vice|-|presidentu lit. 'vice-president'
  - (SV) EU|:|s lit. 'EU's'
  - (SL)溦, lit. 'pages'

Note: See also the representation of MWTs in Universal Dependencies
A VMWE contains several tokens, whether each of them coincides with a word or not:

- (BG) заем=решение 'make a decision' (2 words, 2 tokens)
- (DE) eine Rede halten (2 words, 2 tokens) lit. a speech hold 'to give a speech'
- (EL) παίζω στα δάχτυλα (2 words, 3 tokens) lit. play in-the fingers 'know very well'
- (IT) dar una caminhada 'to give a walk' (2 words, 2 tokens)
- (RO) de ar treabă 'to die' (3 words, 3 tokens) lit. do.3.SG.PRES+be.1.SG.AGL 'I did'
- (SL) nanj = na+njega 'on him'

While a VMWE always contains at least two words, the relation between VMWEs and tokens can be twofold:

- A VMWE contains several tokens, whether each of them coincides with a word or not:
  - (BG) заем=решение 'make a decision' (2 words, 2 tokens)
  - (DE) eine Rede halten (2 words, 2 tokens) lit. a speech hold 'to give a speech'
  - (EL) παίζω στα δάχτυλα (2 words, 3 tokens) lit. play in-the fingers 'know very well'
  - (IT) dar una caminhada 'to give a walk' (2 words, 2 tokens)
  - (RO) de ar treabă 'to die' (3 words, 3 tokens) lit. do.3.SG.PRES+be.1.SG.AGL 'I did'
  - (SL) nanj = na+njega 'on him'

- A VMWE contains one (multitoken) word:
  - (BG) 3 examples found for Bulgarian
  - (DE) vorbereiten lit. to pre-arrange 'to prepare'
  - (EL) εξετάσεις μικρά lit. give-up-1SG 'to manage'
  - (EN) to pretty-print
  - (ES) suicidarse lit. suicide self 'to commit suicide'
  - (FR) court-circuiter 'to short circuit'
  - (HE) pripremiti lit. unaprijed napraviti/ urediti 'to prepare'
  - (HU) kinyir lit. out.cut 'to kill'
  - (IT) corto-circuitere 'to short circuit' suicidarsi lit. suicide self 'to commit suicide'
  - (PL) no example found for Polish
  - (PT) queixar-se-la lit. complain-self would 'would complain' (2 words, 5 tokens)
  - (RO) și de ar treabă 'to die' (3 words, 3 tokens) lit. do.3.SG.PRES+be.1.SG.AGL 'I did'
  - (SV) klargöra lit. clear-make 'clarify'

Note finally that multitoken words are not considered verbal VMWEs since they contain one (multitoken) word only:

- (BG) no example found for Bulgarian
- (DE) ??
- (EL) n.a.
- (ES) maldicir lit. bad say 'curse' bien vivir lit. well live 'to live in comfort'
- (IT) ricercare 'to recharge'
- (PL) SMS-ować 'to write an SMS'
- (PT) pós-data 'to post-date'
- (RO) a re-măprălit/PREFIX-split' split again/with the aporiaesis of the sound 'i' in rapid speech; this is one word, multitoken
- (SL) SMS-jasi 'to write an SMS'

Note: The precise word forms cannot always be straightforwardly deduced from the MWT containing them and vice versa, as in don't, della, du, etc.
Section 1.3

Verbal Multiword expressions (MWEs) are (continuous or discontinuous) sequences of words with the following compulsory properties:

- They show some degree of orthographic, morphological, syntactic or semantic idiosyncrasy (see tests 1 to 5) with respect to what is considered general grammar rules of a language. Collocations, i.e. word co-occurrences whose idiosyncrasy is of statistical nature only (e.g. the graphic shows, drastically drop) are not annotated.
- Their component words include a head word and at least one other syntactically related word. Most often the relation they maintain is a syntactic (direct or indirect) dependence but it can also be e.g. a coordination. Depending on the category of the head word, the whole MWE can be nominal, adjectival, prepositional, verbal, sentential, etc.
- At least two components of such a word sequence have to be lexicalized. In this task we only annotate the lexicalized components and ignore open slots.

Probably the most salient property of MWEs is semantic non-compositionality. In other words, it is often impossible to deduce the meaning of the whole unit from the meanings of its parts and from its syntactic structure. For instance, while it is easy to interpret phrases like to kick the bucket or to spill some water from the words that compose them, it is almost impossible to guess, without knowing it beforehand, that to kick the bucket means 'to die' and to spill the beans actually means 'to reveal a secret'.

However, as non-compositionality is a subjective notion, we use inflexibility as a proxy in the tests. Our underlying hypothesis is that (verbal) MWEs have some degree of semantic non-compositionality that implies limited flexibility.

Verbal MWEs (VMWEs) are simply multiword expressions whose syntactic head in the prototypical form is a verb.

Section 1.4

Syntactic variants of VMWEs

VMWEs in this task include following syntactic structures:

1. Prototypical verb phrases: in most cases, the prototypical form of a VMWE is a verb in finite form – or a participle, infinitive or gerund with finite auxiliaries – in active voice, and whose other components depend directly or by transitivity on the verb. The VMWE can also contain coordinated verbs. These phrases can be:
   - Partly saturated, where only some of their arguments are lexicalized:
     - (BG) πυλή νενέρει πέτας lit. scrape someone's nerves 'to annoy someone'
     - (DE) traf eine Entscheidung 'made a decision'
     - (EL) παίρνει το μέτρα μου lit. take-1.SG a-FE.SG.AC decision-NE.PL.AC 'to make a difficult decision'
     - (EN) made a decision
     - (ES) tomó una decisión lit. took.he/she a decision 'he/she made a decision'
     - (FR) a eu du courage lit. has had courage 'had courage'
     - (HU) sétált tesz 'to take a walk'
     - (IT) prendere una decisione lit. take a decision 'take a decision'
     - (PL) podjął niejedną trudną decyzję lit. took.3.SG not-one hard decision 'he took several hard decisions'
     - (PT) eles deram uma caminhada lit. they gave a walk 'they took a walk'
     - (RO) a trece asta cu vederea lit. to pass this with sight.the.ACC 'to overlook'
     - (SV) fattade dit belsuf 'made a decision'
     - (SL) sprejeti odločitev 'to make a decision', zvezli komu srce lit. to break someone’s heart 'to upset someone by letting them know that you do not love them', vezdi si k srcu lit. take something to heart 'to think about something seriously'
   - Partly saturated, where the lexicalized arguments include the subject:
     - (BG) излиз си си време lit. appears.for me.DAT name 'a name sticks for/to me'
     - (DE) ein Vöglein hat mir gezischt lit. a little bird has to me twittered 'a little bird told me'
     - (EL) μου αφήνει ο τάκος lit. me.GEN left the chosk 'our patience runs out'
     - (EN) a little bird told someone, the problem lies in something
     - (ES) me lo ha dicho un pajarito lit. to.me it has said a little bird 'a little bird has told me'
     - (IT) me lo ha detto un uccellino lit. to.me it told a little bird 'a little bird told me'
     - (PL) mina komu zrzędła lit. the face someone.DAT thinned 'one lost one’s confidence'
     - (PT) a sua hora chegou lit. your time has arrived 'your time has come'
Partly saturated, where lexicalized head verbs are coordinated:

- (BG) Σφυνθω χωρα lit. to blossom and give fruit ‘(usually sarcastically) to prosper’
- (DE) leben und leben lassen lit. to live and let live ‘to live and let live’
- (EL) στορπολ και έξόριαμεν lit. wonder.1.SG and be-amazed.1.SG ‘to wonder’
- (EN) drink and drive
- (ES) coser y cantar lit. to_sew and to_sing ‘easy as pie, a piece of cake’
- (IT) vivi e lascia vivere lit. to live and let live ‘to live and let live’
- (PL) pijl i lapil lit. he spl and catched ‘(he) was lazy, (he) did nothing useful’
- (PT) pintar e bordar lit. paint and knit ‘to abuse’
- (RO) a tunat și-a adunat lit. it has thundered and CL.ACC.it has gathere.birds of a feather flock together ‘seamānā, dar nu răsarelit. sow.3SG (homonym of resemble), but not sprout.3SG’not to resemble’
- (SL) živi v pusni živetli lit. to live and let live ‘to live and let live’

Fully saturated:

- (BG) пипе не може да прекъсне lit. bird cannot fly in ‘it is very strictly guarded’
- (DE) der frühe Vogel füngt den Wurm lit. the early bird catches the worm’the early bird catches the worm’
- (EL) το έξυπνο πουλί από τη μύτη πιάνεται lit. the clever bird is-caught-3SG by the nose ‘people who consider themselves clever fail’
- (EN) the early birds the worm catches
- (ES) los ojos son el reflejo del alma lit. the eyes are the reflection from the soul ‘the eyes are the window of the soul’
- (FR) à quelque chose malheure est bon lit. for something bad luck is good ‘bad experiences may bring unexpected positive effects’
- (PL) nie od razu Kraków zbudowano lit. not at once Cracow was built ‘Rome was not built in a day’
- (PT) quem vê cara não vê coração lit. who sees face doesn’t see heart ‘a person can lie/or his/her feelings’
- (RO) se reversă zoriță lit. Refl.C1.3sg. Acc. row_out dawns ‘it is getting morning’
- (SL) kofor prej pride, prej mejlo lit. who came first mills first

2. Meaning-preserving variants belonging to the following syntactic categories.

- infinitives:
  - (DE) eine Entscheidung treffen lit. a decision meet ‘to make a decision’
  - (EL) έχω πάρει την απόφασή μου
  - (EN) to make a decision, to break one’s heart
  - (ES) tomar una decisión lit. to take a decision ‘to make a decision’
  - (FR) avoir du courage ‘to have courage’
  - (IT) prendere una decisione lit. take a decision ‘to make a decision’
  - (PL) podjąć niejedną trudną decyzję ‘to take several hard decisions’
  - (PT) tomar essa decisão ‘to take this decision’
  - (RO) lua o decizie ‘make a decision’
  - (SL) odlodičev je treba sprejeti’decisions has to be made’

- Nominal groups (headed by nominal complements from the prototypical VMWEs) with relative clauses:
  - (BG) решението, което аз се присъединявам lit. decision.DET which I.PRODROP made ‘the decision which I made’
  - (DE) Entscheidungen die wir treffen ‘decisions which we made’
  - (EL) η απόφαση που πήραμε lit. the decision which we took ‘decisions which we made’
  - (FR) les décisions que nous avons prises hier sont bonnes lit. the decisions that we have taken yesterday are good ‘the decisions that were made yesterday are good’
  - (IT) la decisione che prenderemo lit. the decision which we took ‘the decision which we made’
  - (PL) decyzja, która podjęła ‘decisions which he took’
  - (PT) a apresentação que Maria fez lit. the presentation that Mary made ‘the presentation which Mary made’
  - (RO) decizia pe care am luat-o ‘the decision that we have made’
  - (SL) odlodičev, ki jo je sprejel ‘decisions which he took’

- Gerunds:
  - (BG) емокийно трябва решението ‘(while) making this decision’
  - (DE) n.a. (?)
  - (EL) παίρνοντας απόφασης κάνοντας πλάκα lit. ‘one hasn’t enough courage to do something’, sekira pade v med komu

- Nominal and adjectival groups with participles:
  - (BG) акоято вече решението ‘the decisions already made’
  - (DE) früher getroffene Entscheidungen ‘earlier made decisions’
Like other VMWE occurrences, syntactic variants are also annotated if they contain one multi-word token only, e.g. particle verbs like (DE) *NVPC: nominal verb-particle constructions* so as to keep the contextualized nature of the annotation.

We also do not annotate MWEs containing verbs but functioning as adverbials or nominals (other than in case 2):

- (BG) *взято на решение* 'making a decision'
  - удар в грудь 'a stab in the back'
  - высоко вздыхаяет лёгкую 'highly raised bar 'high bar''
  - игра на карто 'card player'
- (DE) *Wortbruch* lit. word-break 'a promise which has not been hold'
- (EL) *αριθμός ημερολογίων* lit. *the-FE.SG.NOM taking-FE.SG.NOM decisions-GE.PL.GEN* 'deciding *'
- (EN) *a take-off*
- (ES) *toma de decisiones* lit. taking of decisions 'decision making'
  - puesta a punto lit. lay to point 'set-up'
- (FR) *la prise en compte* 'the fact of taking into account'
  - une mise à disposition 'the fact of making available''
- (IT) *la messa a disposizione* 'the made to availability 'the fact of making available''
- (PL) *zabawa czyimś kosztom* 'a play at someone else's expenses' derived from *bawić się czymś kosztom* 'to enjoy oneself at someone else's expenses'
- (PT) *a tomada de decisão* 'the making of decisions'
  - o tomador de decisão 'the decision maker'
- (RO) *luarea unei decizii* lit. take-noun.suffix a-genit decision 'making a decision'
- (SL) *sprejete odločitev* 'the decision'

If a candidate VMWE is a meaning-preserving variant of a prototypical verbal phrase, all linguistic tests to identify and categorize VMWEs must be applied to this prototypical phrase, rather than to the variant itself. The prototypical forms of a meaning-preserving variant are also known as its canonical forms.

Note that, for some VMWEs, the only possible form is not prototypical. For instance, some VMWEs appear in passive voice but never in active voice. In that case, the linguistic tests should be applied to the passive form with finite passive auxiliary.

Lexicalized verb complements can occur with all sorts of non-lexicalized simple and complex determiners, quantifiers and modifiers (the, some, half a dozen, an impressive number of, ...). These optional elements should not be included in the annotated VMWE. Moreover, depending on the syntactic theory used, the lexicalized elements will not depend directly on the verb, but only indirectly, by transitivity. In this case, the canonical form is the one that includes no complex determiners, quantifiers and modifiers, thus creating direct dependency links between the lexicalized components.

- (EN) *they had taken a significant number of steps* → *apply tests to they take steps*
- (PL) *dostać połówkę spadku* lit. to receive half of the heritage
- (EL) *ανάγκη αριθμοφόρων* lit. *the-FE.SG.NOM taking-FE.SG.NOM decisions-GE.PL.GEN* 'deciding *'
- (IT) *ele fez o restante do trabalho* 'to make the remainder of the work' → *apply tests to ele fez o trabalho* 'he did the work'

Expressions of the syntactic categories mentioned above are considered VMWEs only if they function as verb phrases (case 1) or nominal/participial phrases (case 2).

Other kinds of variants are not considered VMWEs. This concerns nominalizations morphologically derived from verbs and describing a process, result, state, agent, etc.
Section 1.5

Lexicalized components and open slots

Just like a regular verb, the head verb of a VMWE may have a varying number of compulsory arguments, that is, arguments that must be present in each occurrence of this VMWE. For instance, the direct object and the prepositional complement are compulsory in the VMWE to take someone by surprise.

Some components of such compulsory arguments may be lexicalized, that is, always realized by the same lexemes. Here, by surprise is lexicalized while someone is not. The head verb of a VMWE is always considered lexicalized. When it can be replaced by another verb, like in to make/take a decision, we consider that these are two different VMWEs, although possibly synonymous.

Conversely, a component of a compulsory argument which can be realized by a free lexeme taken from a relatively large semantic class is called an open slot. In the following VMWE examples (cited after Gross 1994), all having the same syntactic structure NP V NP Prep NP, the lexicalized arguments are highlighted in bold:

- Max took the bull by the horns.
- The news took John by surprise.
- Bob took part in the inquiry
- Money burns a hole in Bob's pocket.

Special cases

Prepositions have a special status with respect to the notion of lexicalization. In the first, second and fourth example above, the prepositions by and in are lexicalized since they introduce lexicalized complements (the horns, surprise and pocket). However, in the third case the preposition in introduces an open slot whose meaning compositionally combines with the meaning of the VMWE took part. We say in this case that the preposition is selected by the VMWE but it is not lexicalized and should not be annotated. Prepositions selected by the governing verb, noun, adjective or adverb are fixed in the sense that they cannot vary freely. However, this kind of fixedness belongs to the phenomenon of valency and is considered a regular property of the grammatical system, thus outside of our annotation scope.

Reflexive clitics in inherently reflexive verbs also have a special lexicalization status. In some languages, the same reflexive clitic is used regardless of the person and number, inflecting for case only:

- (BG) смеят се lit. laugh se.REFL 'to laugh'
- (DE) ??
- (EL) n.a.
- (ES) n.a.
- (PT) n.a.
- (SL) smeji se lit. laugh.2.SG self 'You laugh'

In other languages, reflexive clitics agree in person and number with the subject and the verb:

- (BG) No examples found for Bulgarian
- (DE) sie wundert sich lit. she wonders self.3.SG 'she wonders'
- (EL) n.a.
- (ES) yo me quejo lit. I self.1.SG complain 'I complain'
- (FR) je me trouve lit. I self.1.SG find 'I find myself'
- (IT) mi meraviglio lit. I self.1.SG wonder 'I wonder'
- (PT) eu me queixo lit. I self.1.SG complain 'I complain'
- (RO) eu ma gândeșc lit. I self.1.SG think 'I am thinking'

It this case, the clitic is realized by different lexemes, depending on the number and gender. Strictly speaking, it is not lexicalized. However, we admit that, regardless of the language, the reflexive clitic is a unique lexeme (with lemma sie, se, sich, etc.) inflecting for person and number. It is thus lexicalized in inherently reflexive verbs.

Section 1.6

Verbal multiword expressions versus collocations

Collocations are not considered VMWEs in this task and should not be annotated. However, the boundary between both categories is not always easy to define and should be handled with care.

We understand collocations as combinations of words whose idiosyncrasy is purely statistical. In other words, words in collocations tend to co-occur with each other more often than expected by chance, but they show no substantial orthographic, morphological, syntactic and (most notably) semantic idiosyncrasy.

Some combinations happen to be very frequent and are perceived as "frozen":

- (BG) kávaaμ əpahama 'raise the price'
- (DE) eine Frage beantworten 'to answer a question', die Graphik zeigt 'the graphic shows', einen Bus nehmen 'to take a bus'
- (EL) ξαδου βόλτα 'lit. take-1SG a walk
Section 2

Wordsworth, cited by Shutova: "However, verbal metaphors are not always VMWEs. Consider the newspaper title "eggs in one basket"

However, applying regular lexical alternations to them does not markedly impact their meaning.

However, the extent of the vocabulary concerned by this variability is different for collocations and VMWEs.

However, verbal metaphors are not always VMWEs. Consider the newspaper title "eggs in one basket". In other words it is based on similarity (presence of common characteristics) between two concepts.

However, the difficulty of distinguishing collocations from VMWEs lies in the fact that lexical variability is relevant to some VMWEs:

Section 1.7

Verbal multiword expressions versus metaphor

Another phenomenon closely related to VMWEs is metaphor. According to (Shutova 2010), "a metaphor occurs when one concept is viewed in terms of the properties of the other. In other words it is based on similarity (presence of common characteristics) between two concepts".

However, verbal metaphors are not always VMWEs. Consider the newspaper title "simple steps to lift your dark cloud of stress". The metaphorical expressions to lift dark cloud of stress to relax and my heart ... dances with the daffodils 'I am happy' are not semantically compositional. These expressions, however, were probably constructed for the needs of one article/poem only and are not sufficiently common in the statistical vocabulary to be considered VMWEs.

The distinction between VMWEs and metaphors is a relatively unstudied and open question. There are few precise tests, other than statistical, which would allow human annotators to resolve it reliably. Gross (1982) gives some clues on the reproducibility and predictability of metaphors. It remains to be seen how heavily this problem will impact the annotation of texts selected for our shared task. We suggest that the annotators take notes of such cases and discuss them within their communities, both local and international.
Section 2

Textual annotation scope

In this annotation task, all occurrences of all syntactic types of VMWEs are to be annotated in the text. We annotate, as integral parts of VMWEs, all lexicalized elements that can form a separate word. For instance, lexicalized particles are annotated at but case suffixes are not. Thus, in *put something up*, the verb and the particle are integral parts of the VMWE (see VPC tests), while in (H) *döntést hoz* valamivel lit. decision-ACC bring something-DEL 'make a decision', only *döntést hoz* is annotated, even if the delative case suffix is also lexically determined.

Both continuous and discontinuous sequences of lexicalized components of VMWEs are annotated. Verb+pronoun and verb+particle combinations are annotated essentially if they are inherently reflexive verbs or verb-particle combinations.

Note: In this version of the guidelines, verb+preposition combinations like *to rely on somebody* and *to come across something* are no longer considered VMWEs.

The annotation considers only flat, tokenized sentences whose tokens will be tagged by annotators as part of a VMWE or not. We do not annotate their internal syntactic structure. We do annotate, however, VMWEs embedded in other VMWEs. For instance, the VMWE *to let the cat out of the bag* contains the embedded VMWE *let out* and both are to be annotated as different VMWEs.

Once identified in a text, VMWEs are also to be assigned to exactly one of the categories described in the following sections.

Note: In this version of the guidelines, we no longer admit hesitation between two different categories. Hesitation can, however, be expressed in a comment and a particular value of the annotator’s confidence assigned to a particular VMWE occurrence.

Section 3

Categories of verbal MWEs

In this task we distinguish the following categories of verbal MWEs:

- Two *universal* categories, i.e. valid for all languages participating in the task:
  - light verb constructions (LVC):
    - (BG) *dържат под контрол* 'to keep under control'
    - (DE) *eine Rede halten* lit. a speech hold to give a speech'
    - (EL) *κάνω μια βόλτα* lit. make-1SG a walk 'to walk'
    - (EN) *to give a lecture*
    - (ES) *hacer una foto* lit. to make a picture 'to take a picture'
    - (FR) *avoir du courage* 'to have courage'
    - (IT) *fare un discorso* lit. give a speech' to give a speech'
    - (MT) *he deċiżjoni 'took a decision’
    - (PL) *podjąć decyzję* 'to take a decision’
    - (PT) *fazer uma promessa* 'to make a promise'
    - (RO) *o luca o decizie* lit. to take a decision*to make a decision’
    - (SL) *imeti predavanje, sprejet odločitev* 'to take a decision’
  - idioms (ID):
    - (BG) *превърнам се в дръжки шалката* lit. to behave myself as 'hold my hat' 'pretend to be naive and innocent'
    - (DE) *schwarz fahren* lit. to drive black 'to take a ride without a ticket'. In Kraft treten lit. into force step 'to come into effect', in die Waagschale werfen lit. in the weighing pan throw 'to bring to bear'
    - (EL) *χήναυ τα αυγά και τα καλάθια* lit. loose-1SG the eggs and the baskets 'to be at a complete and utter loss'
    - (EN) *to go bananas, fortune favors the bold*
    - (ES) *hacer de trampas corazón* lit. make of intestines heart 'to pluck up the courage' *entrar en vigor* lit. enter in vigor 'to come into force/effect'
    - (FR) *défendre son bifteck* lit. defend one’s beefsteak 'to defend one’s interests'
    - (IT) *entrare in vigore* lit. *to enter into force 'to come into effect', gettare le perle ai porci* lit. *to throw the pearls to the pigs 'to waste something good on someone who doesn’t care about it’
    - (MT) *ghasfur izbir qalī lit. a bird small told me ‘to hear something from the grapevine’
    - (PL) *rzucą grochem o ślane* lit. throw peas agains a wall 'to try to convince somebody in vain'
    - (PT) *fazer das trampas coração* lit. transform the tripes into heart 'to try everything possible'
    - (RO) *a trage pe sfoară* lit. to pull on rope 'to fool'
    - (SL) *ubiti dve muhi na en mah* 'to achieve two aims at once', *spati kot ubit* lit. *sleep like dead ‘sleep soundly’

- Two *quasi-universal* categories, valid for some language groups or languages but not all:
  - inherently reflexive verbs (IRRefV):
    - (BG) *усямхам се* 'to smile'
    - (DE) *sich bemühen* 'to endeavour', *sich enthalten* lit. himself contain 'to abstain'
    - (EL) *να*
    - (ES) *suicidarse* 'to suicide'
    - (FR) *se suicider* 'to suicide'
    - (IT) *suicidarsi* 'to suicide'
    - (PL) *ścieś się* lit. *to fear SELF* 'to be afraid'
    - (PT) *se queixar* 'to complain'
    - (RO) *a se gândi* 'to think'
    - (SL) *batiti se to be afraid’
  - *verb-particle combinations (VPC):*
    - (BG) *not applicable to Bulgarian*
    - (DE) *er gibt auf* 'he gives up', *er wirft ihr das vor lit. he throws her that against’ *he reproaches that to her’
    - (EL) *μπαίνω μέσα* lit. *get in’ to go bankrupt’
• (EN) to do in
• (ES) n.a.
• (IT) buttare giù lit. throw down 'to swallow'
• (PL) not applicable to Polish
• (PT) jogar fora This seems to be the only VPC in Portuguese. We annotate it as ID and do not use the VPC category.
• (RO) n.a.
• (SL) dati skozi lit. give through 'to go through', gre za lit. it goes about 'it is about'

language-specific categories, defined for a particular language in a separate documentation.

other verbal MWEs (OTH), which gather the types not belonging to any of the categories above:
• (BG) ухапна u супака lit. to blossom and give fruit (usually sarcastically) to prosper
• (DE) einen drauf setzen 'going one better'
• (EL) έτρωύω και έτρωραμα lit. wonder1SG.PST and be-amazed1SG.PST 'to wonder'
• (EN) to drink and drive
to voice act
to pretty-print
to short-circuit
to tumble dry
• (ES) coser y cantar lit. to sew and to sing 'easy as pie, a piece of cake'
• (FR) court-circuler 'to short-circuit'
• (IT) andare e venire lit. to come and go/back and forth'
• (MT) iqum u joghodh lit. jump and stay 'to fidget'
• (PL) plus i tapac lit. to spit and catch 'to be lazy, to do nothing useful'
• (PT) pintar e bordar lit. paint and knit 'to abuse'
• (RO) a tunat gî a adunatî lit. it has thundered and CL-ACC-it has gathered/birds of a feather flock together'

In practice, to identify and categorize verbal MWEs during manual annotation, one must use the rigorous generic and category-specific tests provided.

Section 4

Annotation process and decision tree

We propose the following methodology for VMWE annotation:

• Step 1 - identify a candidate, that is, a combination of a verb with at least one other word which could form a VMWE. If the candidate is a meaning-preserving variant of a prototypical verbal phrase, the following steps apply to this prototypical phrase, called the canonical form. This step is largely based on the annotators’ linguistic knowledge and intuition after reading this guide.
• Step 2 - determine which components of the candidate (or of its canonical form) are lexicalized, that is, if they are omitted, the VMWE does not occur any more. Corpus and web searches may be required to confirm intuitions about acceptable variants.
• Step 3 - formally check if the candidate (or of its canonical form) forms a VMWE and categorize it into one of the available categories, using the decision trees and detailed tests in the following sections.

We provide two decision trees that indicate the order in which tests should be applied in step 3. They determine the priority of different categories when several tests match. The decision trees are a useful summary to consult during annotation, but contain very short descriptions of the tests. Each test is detailed and explained with examples in the following sections.

Decision tree 1: Identification

Note: In this tree, one YES to one of the tests is sufficient to identify a VMWE

Apply test 1 - [CRAN: Candidate contains cranberry word?]

† YES ⇒ Annotate as a VMWE and go to test 6 - [HEAD]
† NO ⇒ Apply test 2 - [LEX: Regular replacement of a component ⇒ unexpected meaning shift?]
† YES ⇒ Annotate as a VMWE and go to test 6 - [HEAD]
† NO ⇒ Apply test 3 - [MORPH: Regular morphological change ⇒ unexpected meaning shift?]
† YES ⇒ Annotate as a VMWE and go to test 6 - [HEAD]
† NO ⇒ Apply test 4 - [MORPHSYNT: Regular morphosyntactic change ⇒ unexpected meaning shift?]
† YES ⇒ Annotate as a VMWE and go to test 6 - [HEAD]
† NO ⇒ Apply test 5 - [SYNT: Regular syntactic change ⇒ unexpected meaning shift?]
† YES ⇒ Annotate as a VMWE and go to test 6 - [HEAD]
† NO ⇒ Apply the LVC hypothesis - [Candidate has operator verb + activity or state noun?]
† YES ⇒ Assume a VMWE and go to test 6 - [HEAD]
† NO ⇒ It is not a VMWE, exit

Decision tree 2: Categorization

Apply test 6 - [HEAD: Unique verb as syntactic head of the whole?]

† NO ⇒ Annotate as a VMWE of category OTH
† YES ⇒ Apply test 7 - [DEP: verb v has exactly one dependent d’]
† NO ⇒ Annotate as a VMWE of category ID
† YES ⇒ Apply test 8 - [CATEG: What is the morphosyntactic category of d’?]
† Reflexive clitic ⇒ Apply iReflIV-specific tests ⇒ IReflIV tests positive?
† YES ⇒ Annotate as a VMWE of category iReflIV
† NO ⇒ It is not a VMWE, exit
† Particle ⇒ Apply VPC-specific tests ⇒ VPC tests positive?
† YES ⇒ Annotate as a VMWE of category VPC
† NO ⇒ It is not a VMWE, exit
Section 5

Generic tests for identifying VMWEs

In order to decide if a candidate is a VMWE, we apply the following generic idiosyncrasy tests. If a candidate expression passes at least one test from 1 to 5, we consider it to be a VMWE, and it can further be categorized by decision tree 2 based on category-specific tests. If tests 1 to 5 fail, the LVC hypothesis may apply but LVC-specific tests are needed to confirm the candidate’s VMWE status (at the same time as its LVC category).

Test 1 - [CRAN] - Cranberry word

Does the candidate expression contain a cranberry word?

- **YES** ⇒ it is a VMWE
  - (BG) хвасам имамено lit. catch in a tight place to coerce, to pressure → имамено is only used in MWEs
  - права на бъле и корана lit. to turn into elder and nettle to scold, to tell off → бъле is an old word, very rarely used independently
  - въземам предел и имам предел lit. to → предел (as adverb) is only used in MWEs
  - стоям напред и стой напред lit. to stay upright as in Osman council to stay ready to serve → напред is an old word, very rarely used independently
  - (DE) sich um etw. scharen 'to gather around something' → scharen is not a stand-alone word
  - (EL) μάλλιαν η γάλασα μου lit. in-full-of-hair-3SG the-SG.NOM tongue-SG.NOM.my-SG.GEN.Poss to repeat the same thing again and again' → μάλλιαν is not a stand-alone word
  - (EN) to go astray → astray is not a stand-alone word
  - (ES) sin decir ni chus ni mus → chus is not a stand-alone word lit. without to say neither chus nor mus 'without saying a word'
  - no decir ni chus ni mus → chus is not a stand-alone word lit. not to say neither chus nor mus 'not to say a word, to remain silent'
  - a troche y moche → troche is not a stand-alone word lit. to troche and dault 'in a nonsensical way, willy-nilly, haphazardly'
  - (FR) prendre la poudre d’escampette 'to escape' → escampette is not a stand-alone word
  - (IT) mangiare a ufo 'to eat without paying' → a ufo is not a stand-alone word
  - fare lo gnorri 'to play dumb' → lo gnorri is not a stand-alone word
  - scendere in lizza 'to enter the lists' → lizza is not a stand-alone word
  - (PL) odsadzić kogoś od cześci wiar lit. to refuse honor and faith to someone 'to drag sb’s name through the mire/mud, to damage someone’s reputation by saying insulting things about them’
  - (PT) ir para as cucúias 'to go wrong' → cucúias is not a stand-alone word
  - (RO) a nu avea habar 'to have no idea' → habar is not a stand-alone word
  - (SV) att komma ihåg 'to remember' → ihåg is not a stand-alone word
  - (SL) biti = kvit 'to pay up a debt, owe nothing to somebody' → kvit is not a stand-alone word

- **NO** ⇒ further tests are required
  - (BG) прая на сос → прая and сос are stand-alone words
  - (DE) sich um etw. herum stellen 'to stand around something' → all words are stand-alone words
  - (EN) to go away → go and away are stand-alone words
  - (ES) ir a la universidad 'to go to university' → ir, a, y universidad is stand-alone words
  - (IT) andare giù 'to go down' → andare and giù are stand-alone words
  - (PL) wyznac tajemnice 'to reveal a secret' → wyznac and tajemnice are stand-alone words
  - (PT) ir para a escola 'to go to school' → ir, para, and escola are stand-alone words
  - (RO) a nu avea idee 'to have no idea' → all words are stand-alone words
  - (SV) att komma på 'to figure out' → komma and på are stand-alone words

Test 2 - [LEX] - Lexical inflexibility

Does a regular replacement of one of the components by related words taken from a relatively large semantic class lead to ungrammaticality or to an unexpected change in meaning?

- **YES** ⇒ it is a VMWE
  - (BG) взимам змии и гущери lit. to spew snakes and lizards → йбълем елечуи (to spew reptiles)
  - всяка змия да си знае гълъба lit. every frog to know its own puddle → всяка змия да си знае ложата
  - (DE) die Katze aus dem Sack lassen 'to let the cat out of the bag' → Iden Hund aus dem Karton lassen 'to let the dog out of the box'
  - eine Entscheidung treffen lit. to meet a decision 'to make a decision' → meine Entscheidung machen/herstellen lit. a decision make/produce 'to make/produce a decision'
  - (EN) to let the cat out of the bag → bto allow the feline out of the container
    - to make a decision → 'to produce/build/create a decision'
    - to go on → 'to go upon'
    - to stand firm/fast → 'to stand hard/higid/solid'
  - (ES) meteoro en la boca del lobo lit. to get into self in the mouth from the wolf 'venture into the lion's den' → meteoro en el ojo del gato
  - tomar una decisión lit. to take a decision 'to make a decision' → hacere/coger/producir una decisión lit. to make/grab/produce a decision 'to make/produce a decision'
  - (IT) non dire gatto se non ce l’hai nel sacco lit. don’t say cat if it is not in the sack 'don’t count on something before it happens' → non dire cane se non ce l’hai nel sacco lit. don’t say dog if it is not in the sack
  - sputare il rosso lit. spit the red 'spit it out' → #sputare la rana lit. spit the frog
  - (PL) wiedź, co w trawie piszczy lit. to know what in grass squeals 'to be well informed' → wiedź, co w trawniku popiskuje
  - wzięć udział lit. to take participation 'to take part' → wto podjąć/pobrać/zabrać członkostwo/uczestniczenie
  - (PT) cavar um galho lit. to break a branch to help → cavar um ramo 'to damage a stem'
  - (RO) a da cu băţă în bală lit. to give with bat-the in pond 'to say sth embarrassing' → a da cu băţă în bală lit. to give with stick-the in pond, *a da cu băţă în lac* 'to give with bat-the in lake'
  - (SV) att Plocka rüssinen ur kakan lit. to pick the raisins out of the cake 'to choose only the best things' → att välja ut nötterna från kakan
Does a regular language’s morphology lead to unexpected change in meaning?

**Yes** → it is a VMWE

- **BG** хвърлих мяу. Lit. throw an eye 'to throw a glance' → хвърлях мяу.PLURAL PLURAL "to throw the bull by the horns" → хвърлях мяу.INDEF "to throw a bull by the horns" can not find a place for myself 'to be extremely nervous' → only exists in negative form
- **DE** ins Gras beißen lit. to bite into the grass 'to die' → *ein ein Gras beißen lit. to bite into a grass *in die Grüser beissen lit. to bite into the grasses. *in Kraft treten lit. into force step 'to come into effect' → *in Kräft treffen lit. into force step 'to come into effect'
- **IT** andare a letto con la gallinella, lit. to go to bed with the hen 'to go to bed early' → *andare a letto con la gallinella lit. to go to bed with the hen *in the hair in the egg 'to be pedantic' → *andare a letto con la gallinella lit. to go to bed with the hen *cercare il pelo nell'uovo 'to look for the hair in the egg'
- **PL** budować zamki na lodzie lit. to build castles on ice 'to rely on unstable foundations' → *budować zamki na lodzie lit. to build a castle on ice *mucha kogoś ugrzyła lit. a fly bit someone 'someone is in a bad temper' → *mucha kogoś ugrzyła lit. a fly bit someone *wyciągnąć nogi lit. to stretch.PERF leg(s) lit. 'to die'
- **PT** bater perna lit. hit leg 'to walk around' → bater a/uma/essas perna/pernas/perninhas/pernora 'to hit the one/these leg(s)/legs.SMALL/BIG'
- **RO** a da cofa lit. lit. to give corner.the. the 'to die' → *a da cofarelit. to give corners.the *ă vraide și la kraft lit. step in force 'to come into effect' → *ă vraide și krafter lit. step in forces
- **SL** klicati jelena lit. to call cerfs 'to vomit' → *klicati jelena lit. to call a cerf

**No** → further tests are required

**Yes** → it is a VMWE

- **BG** хвърлим мяу Lit. throw an eye 'to throw a glance' → хвърлим мяу.PLURAL PLURAL "to throw the bull by the horns" → хвърлим мяу.INDEF "to throw a bull by the horns" can not find a place for myself 'to be extremely nervous' → only exists in negative form
- **DE** einen Kuchen backen 'to bake a cake' → *einen Kuchen backen 'to bake a cake'
- **EN** to make a cake → *to make a/many/those/no cake/cakes
- **ES** mover el brazo 'to move the arm' 'to move the arm' → *mover/agitar/levantar/estirar el brazo las piernas/las piernas "to move the legs" *move/shake/raise/stretch the arm/the leg/the hands/the legs 'to move/stretch/shake/raise/stretch the arm/the leg/the hands/the legs'
- **IT** fare un dolce 'to make a sweet' → *fare un dolce 'to make a sweet'
- **PL** kształtować opinie 'to form an opinion' → *kształtować opinie 'to form opinions'

**No** → further tests are required

**Yes** → it is a VMWE

**Note** that more tests are required.
Does a regular syntactic change that would normally be allowed by general grammar rules lead to ungrammaticality or to an unexpected change in meaning?

**YES => it is a VMWE**

- (BG) *na star krastrapëhrna krastrapëhsta prodhvam* lit. an old cucumber seller cucumbers to sell "to try to cheat a more experienced person"
  
  #prodhvam krastrapëhsta na star krastrapëhrna. Krastrapëhstvëti sa prodhvadni

  (DE) Noun phrase (NP) or prepositional phrase (PP)

  (EN) *speak of the devil* 'the person one is talking about shows up' → *thee was speaking of the devil*

  (IT) *cantar y corder* 'to sing and to sew' **perder la cabeza** "to lose the head" to *I perder las cabezas* "to lose the heads"

  (PL) *kogoś krew zalewa* lit. blood foods someone 'someone gets furious' → *dzieci się niszczą* a *dog dies*!

  (PT) *pisar na bola* lit. step on the ball (regular modification blocked)*dobrze komus z oczu patrzy* lit. "do someone's eyes looks/someone looks like a good person" → *uprzejmość dobrze komus z oczu patrzy* lit. kindness well someone.DAT from eyes looks (subject prohibited)

  (SV) *del knallar och går* lit. it trots and walks *'it is OK/as usual"* → *det går och knallar* (a regular word order variability is blocked)

  (SL) *delaj se Francisco* lit. to pretend to be French → *delan Francoz* lit. made French

**NO -> further tests are required**

1. (BG) *konan eu apooba* 'to dig my grave' → *konan mukvyd/konan apooba* (to dig your/his/her/their grave)

2. (DE) or try mezhe seine Entscheidung 'he made his decision' → er traut meine/hue/unsere/uren Entscheidung 'he made my/her/our/your decision'

3. (ES) *Ha hecho eu trabajo* lit. Has he/she has done his/her/your work "He/She has done his/her/your work" → *Ha hecho mi/truo/nuestro trabajo* lit. Has he/she done your work/our work/has done your work/their work

4. (IT) *ha fatto il suo lavoro* → ha fatto il mio/uo/nuovo/voestro/lavoro

5. (PL) *Polish VMWEs do not seem to exhibit this kind of inflexibility*

**NG** 

Does a regular syntactic change that would normally be allowed by general grammar rules lead to ungrammaticality or to an unexpected change in meaning?

**YES => it is a VMWE**

- (BG) *na star krastrapëhrna krastrapëhsta prodhvam* lit. an old cucumber seller cucumbers to sell "to try to cheat a more experienced person"

  #prodhvam krastrapëhsta na star krastrapëhrna. Krastrapëhstvëti sa prodhvadni

- (DE) Noun phrase (NP) or prepositional phrase (PP)

- (EN) *speak of the devil* 'the person one is talking about shows up' → *thee was speaking of the devil*

  (IT) *cantar y corder* 'to sing and to sew' **perder la cabeza** "to lose the head" to *I perder las cabezas* "to lose the heads"

  (PL) *kogoś krew zalewa* lit. blood foods someone 'someone gets furious' → *dzieci się niszczą* a *dog dies*!

  (PT) *pisar na bola* lit. step on the ball (regular modification blocked)*dobrze komus z oczu patrzy* lit. "do someone's eyes looks/someone looks like a good person" → *uprzejmość dobrze komus z oczu patrzy* lit. kindness well someone.DAT from eyes looks (subject prohibited)

  (SV) *del knallar och går* lit. it trots and walks *'it is OK/as usual"* → *det går och knallar* (a regular word order variability is blocked)

  (SL) *delaj se Francisco* lit. to pretend to be French → *delan Francoz* lit. made French

**NO -> further tests are required**

1. (BG) *konan eu apooba* 'to dig my grave' → *konan mukvyd/konan apooba* (to dig your/his/her/their grave)

2. (DE) or try mezhe seine Entscheidung 'he made his decision' → er traut meine/hue/unsere/uren Entscheidung 'he made my/her/our/your decision'

3. (ES) *Ha hecho eu trabajo* lit. Has he/she has done his/her/your work "He/She has done his/her/your work" → *Ha hecho mi/truo/nuestro trabajo* lit. Has he/she done your work/our work/has done your work/their work

4. (IT) *ha fatto il suo lavoro* → ha fatto il mio/uo/nuovo/voestro/lavoro

5. (PL) *Polish VMWEs do not seem to exhibit this kind of inflexibility*
Specific tests for categorizing verbal MWEs

Once a candidate verbal MWE has been pre-identified according to one of the identification tests, the confirmation of its status as a MWE, as well as its categorization can be based on category-specific tests.

- Structural tests
- Light-verb constructions
- Idioms
- Inherently reflexive verbs
- Verb-particle constructions
- Language-specific categories
- Other verbal MWEs

Section 6.1

Structural tests

Structural tests are quite simple preliminary tests that help determine the syntactic structure of the VMWE. This is required in order to pursue categorization by pointing to the right category-specific tests in the last step. In practice, annotators will rarely need them since they will already have an intuition about the VMWE's category when they identify it.

Test 6 - [HEAD] - Syntactic head

Does the candidate contain a unique verb functioning as the syntactic head of the whole?

- Yes ⇒ continue to the next test
- No ⇒ the candidate is NOT a VMWE

The LVC hypothesis is not a real test, but its application is largely based on intuition and it may be hard to judge whether a verb is only performing the role of operator. This hypothesis accounts for LVCs that have otherwise no salient inflexibility but still correspond to multword predicates we want to annotate. If you are unsure, we advise you to assume that the combination is a VMWE and go to the LVC tests. If the expression fails the LVC tests, then you must change your mind and consider that the answer to the LVC hypothesis was actually NO.
Does the VMWE contain exactly one lexicalized syntactic dependent of the head verb?

YES → continue to next test

DE eine Fratze ziehen lit. a grimace pull ‘to make a face’ → the single dependent is the noun phrase, Fratze

EN to make a face → the single dependent is the noun phrase, Fratze

IT fare la linguacce lit. ‘to make grimaces’ ‘to make a face’ fare is the head and the NP depends on it

PL zbijać bąki lit. ‘to smash fart/too foot around, to do nothing useful’ zbijać is the head and the NP bąki depends on it
del kosm o popalici lit. ‘to let someone smoke/to make someone’s life hard’ del is the head and the infinitive popalici depends on it

PT cair vergonha na cara lit. ‘to cringe’ the head and the NP cara depends on it

RO a face baie lit. to make bath ‘to bath’ face is the head and the NP depends on it

SV get upp ‘to give up’ ge is the head and upp is the dependent on it

SL imeti krompir lit. ‘to have potatoes’ to be lucky → imeti is the head and the NP depends on it

NO annotate as OTH
The test covers only lexicalized dependents. There may be other, non-lexicalized dependents, which the test ignores. We explicitly call the non-verbal elements

Noun phrase (NP) or prepositional phrase (PP) headed by a preposition governing a noun

Particle - apply VPC tests. If the outcome is negative, discard the VMWE candidate.

(BG) Bulgarian does not have VPC expressions.

Test 8 - [CATEG] - Category of the dependent

What is the morphosyntactic category of the dependent $d$ that co-occurs with the head verb $v$?

**Reflexive cletic** - apply IrRefl tests. If the outcome is negative, discard the VMWE candidate.

(BG) съм нахален съ by me.REFL 'to be afraid'

(DE) sich wundern 'to wonder', sichschämen 'to be ashamed'

(EN) English does not have IrRefl expressions.

(FR) se suicider 'to suicide', s'vanourir 'to faint'

(IT) suicidarsi 'to suicide', seppurarsi 'to be ashamed'

(PL) bać się lit. 'to feel SELF'

(PT) suicidar-se 'to suicide', queixar-se 'to complain'

(RO) se sinunca 'to commit suicide' with obligatory ACC reflexive clitic

(SL) seboh 'to stare' with obligatory ACC reflexive cletic

**Particle** - apply VPC tests. If the outcome is negative, discard VMWE candidate.

(BG) Bulgarian does not have VPC expressions.

(DE) anfangen to begin, er fängt ar se 'he begins', er hat angefangen 'he has begun' in German, VPCs may occur separated or within one word, we annotate all occurrences!

(EL) τιμηθεί ομοιόμορφα, βάδι ομοιόμορφα, κάνω μια τους

(EN) to give up, to look forward to

(IT) far fuori lit. to make out 'to kill', lo fa fuori 'he kills him', lo ha fatto fuori 'he killed him'

(PL) Polish does not have VPC expressions.

(PT) Portuguese does not have VPC expressions.

(RO) Romanian does not have VPC expressions.

(SL) grij se 'it is about'

dali skoči 'to go through'

biti za 'to agree'

**Noun phrase (NP) or prepositional phrase (PP) headed by a preposition governing a noun - apply LVC tests. If the outcome is negative, categorize as ID.

(BG) търся към няколко лъвски 'I search for a lions'

(DE) die Nase rümpfen lit. the nose wrinkle 'turn up one's nose at sth.' → die Nase is a noun phrase composed of a determiner and a noun

(Kraft) trennen lit. into

(EN) to make a wish → a wish is a noun phrase composed of a determiner and a noun
take turns → turns is a noun phrase composed of a singular plural noun
mettere radici → radici is a noun phrase composed of a single plural noun

(IT) prendere in considerazione 'take into account' → in considerazione is a prepositional phrase composed of a preposition and a noun
rompere il silenzio 'to break the silence' → il silenzio is a noun phrase composed of an article and singular noun

(PL) podjąć decyzję 'to take a decision' → decyzję lit. decision is a nominal phrase composed of a single noun

dochodzi prostych drogą lit. to go (on) a straight road

(INST) to avoid complications → prosty drogą 'on a straight road' is a noun phrase composed of an adjective and a noun in (instrumental)

(RO) a rupe tăcerea lit. to break silence the 'to start talking' → tăcerea is a noun phrase composed of a singular plural noun

(a face bai lit. to do batfio 'to take a shower' → bai is a noun phrase composed of a single noun

(SL) biti v dvomih lit. 'doubt' → v dvomih lit. in doubt is a prepositional phrase composed of a preposition governing a noun

**Other** - categorize as ID.

**Adjective**

(BG) използвам син от една лъвка lit. to come out dry from the water 'to avoid taking responsibility'

(RO) to see red

(EL) τα βραδινά μυάλα lit. them-NE.PL.ACC paint-1.SG black-NE.PL.ACC 'be very sad'

(EN) to stand firm, to see red

(FR) voir rouge lit. to see red 'to be very angry'

(PT) pensar grande 'to think big'

The test covers only lexicalized dependents. There may be other, non-lexicalized dependents, which the test ignores. We explicitly call the non-verbal elements instead of arguments or complements because argument-adjunct distinction is irrelevant. The outcome of the test is positive if the verb has a single lexicalized dependent, which can be the subject, the direct or indirect object, but also an adverbial complement, adverb, particle, relative clause, etc.
Light verb constructions (LVC)

Light verb constructions (LVC) constitute a universal category. We retain the following key characteristics:

1. They are formed by a verb v and a noun n, which either directly depends on v or is introduced by a preposition.

   - (BG) ვედრა აბრუ ლ.: lit. to CL.ACC make wide/to party
   - (SL) narediti svoj: lit. to do one's own/to do what one is supposed to do

   **Verb:**

   - (BG) ვედრა აბრუ ლ.: lit. to keep under control
     - (DE) zum Einsatz kommen lit. to the use come 'to be called into action'
     - (IT) fare una passeggiata lit. to make a walk 'to have a walk'
     - (PL) osiąść sukces lit. carry-away success 'to be successful'
     - (PT) fazer um aborto lit. to make an abortion
     - (RO) a duce dornj lit. to carry yearning.the to miss somebody'
   
   **Adverb:**

   - (BG) изваждам намир lit. take out in the open 'to uncover'
   - (EL) να τρέξει στο λάθος lit. 'to fail'
   - (EN) to get well
   - (IT) andare l'aria lit. to throw down 'to hand over'
   - (PT) quiserem dire lit. want say 'to mean'
   - (SL) razoješi jo kom: lit. to play music to someone/play a joke on someone

   **Pronoun:**

   - (BG) мързи me.ACC lit. (it feels) lazy me.ACC 'to be lazy'
   - (EL) να μείνει πολλά lit. 'to make it'
   - (EN) to make it
   - (FR) faire une présentation lit. make a presentation
   - (IT) chiamare in causa lit. to call in cause 'to single out'
   - (PL) odnieść sukces lit. carry-away success 'to be successful'
   - (PT) fazer um aborto lit. to make an abortion
   - (RO) a duce dornj lit. to carry yearning/to miss somebody'
   - (SL) biti v dvoniti lit. 'to doubt'

2. The noun n is predicative, either referring to an event (e.g. decision, visit) or a state (e.g. fear, courage).

   - (BG) езмам решение lit. 'to make a decision'
   - (DE) eine Entscheidung treffen lit. 'to make a decision'

**Etc.:**

Section 6.2
(EL) μαζί με τη διάθεση, κάνε να δοθεί → noun refers to an event
(EL) συμφωνία, κάνε να καταληφθεί → noun refers to a state
(EN) to make a decision, to pay a visit → noun refers to an event
 make to have fear, to have courage → noun refers to a state
(FR) donner un conseil 'give advice' → noun refers to an event
(noir du courage 'to have courage' → noun refers to a state
(IT) fare una domanda → noun refers to an event
 avere paura, avere coraggio → noun refers to a state
(PL) prowadzić rozmowy lit. to lead conversations/lead negotiations' → the noun refers to an event
micię rację lit. to have right/to be right' → the noun refers to a state
(PT) fazer uma prece 'to make a prayer' → noun refers to an event
 ter simpatias 'to have symptoms' → noun refers to a state
(RO) a lua o decisie 'to make a decision', a face o vizita 'to pay a visit' → noun refers to an event
 a avea curaj → noun refers to a state
(SL) imeti poizaj 'to have courage' → noun refers to a state
sprejeti odločitev 'to make a decision' → noun refers to an event
3. The verb v is "light", i.e. it contributes to the meaning of the whole only by bearing tense and mode. This may be implied by "light" either per se, or when used in the specific context of the noun. This implies that v's syntactic subject is v's semantic argument.

Many authors distinguish support verbs from light verbs, still others differentiate between true light verbs and vague action verbs. On the one hand, we take a narrower scope by ignoring aspectual or causative support verbs, since they do contribute an additional (change of state) meaning to the expression. For instance, for the predicative noun walk, we will consider the light verb to have, but not the aspectual verbs to start, to pursue, to stop a walk. For the noun bloom, which is in itself inchoative, we do consider come into bloom as LVC (both the verb and the noun are inchoative, so the verb does not add any semantics to the noun). In the same vein, we do not consider constructions with causative support verbs as LVCs (as in give a headache compared to have a headache).

On the other hand we do take in cases in which the verb has per se a light semantics (it only bears the tense and mood in any case), which hence cannot be described as "bleached" as is usually said of support verbs. For instance, whereas to pay does not have its usual meaning in to pay a visit, it cannot really be said that commit does not have one of its meanings in commit a crime (note that commit can be used with any negatively charged achievement noun, e.g. suicide, crime, fraud, felony...). These are borderline cases in which they do not fulfill the tests 1 to 5, but we take them as LVCs.

The noun n functions as a regular syntactic dependent, so LVCs exhibit regular syntactic variants.

- (BG) эзма решение → решението, което през протекшния етап 'the decision that the president made'
- (DE) eine Entscheidung treffen → die Entscheidung die der Direktor zu treffen hatte.
- (EL) τα προσωπικά "νόμισμα" α προσωπικά "που πρέπει κάθες να υπάρξει"
- (EN) make a decision → the decision that the director has to make.
- (FR) prendre une décision → la décision prise par la directrice.
- (IT) prendere una decisione → la decisione che il direttore ha dovuto prendere.
- (PL) wziąć udział lit. to take participation.ACC to take part → wzięcie udziału lit. taking.GER participation.GEN taking part', biorący udział lit. taking.PART participation.ACC taking part'
- (PT) tomar banho lit. to take shower → o banho que eu tomei estava bom lit. the shower which I took was good
- (RO) a lua o decisie 'to make a decision' → decizia ce care directorul trebuie să o ia 'the decision that the director has to make'.
- (SL) sprejeti odločitev 'to make a decision' → odločitev, ki jo je moral sprejeti direktor 'the decision that the director has to make'

In many cases of LVCs, it can be said that there is some degree of selection of the verb by the noun.

- (BG) эзмам решение 'to make a decision' vs "эзмам ометовачното" 'to take responsibility'
- (DE) eine Entscheidung treffen lit. a decision meet 'make a decision's vs. 'eine Entscheidung machen lit. a decision make vs. "einen Beschluss treffen lit. a resolution meet
- (EL) κάνε να διαλέξει vs. "κάνε να διαλέξει μαζί με τη διάθεση"
- (EN) have a walk vs "have a race
 run a race vs "run a walk
- (FR) faire une promenade lit. make a walk "take a walk" vs "procéder à une promenade lit. perform a walk but faire/procéder à une enquête lit. make/perform an inquiry
- (IT) prendere una decisione lit. take a decision 'make a decision's vs. "fare una decisione lit. make a decision vs. "prendere una conclusione lit. take a conclusion
- (PL) wziąć udział lit. to take participation vs. "wziąć uczestnictwo
miec rację lit. to have right/to be right' vs. "posiadać rację lit. to possess right
- (PT) fazer uma prece 'to make a prayer' vs. "dar uma prece 'to give a prayer' but fazer/ridar uma caminhada lit. to make/give a walk
- (RO) a da divorț lit. to make divorce 'to divorce' vs. "a oferi divorț
- (SL) postaviti vprašanje lit. to put a question 'to pose a question' vs "postaviti odgovor

Yet some regularities exist. For example, large classes of nouns function with have (e.g. +property) or commit (+negative) charged achievement noun. Therefore, we chose not to retain the selection of the verb as a criterion for LVC categorization. Instead, the following decision tree should be applied to decide whether a candidate should be annotated as an LVC.

LVC-specific decision tree:

Note: In this tree, a single NO to one of the tests is sufficient to decide that a candidate is not an LVC.

- Apply test 9: [N-EVENT: The noun describes an event/state]
  - NO: It is not an LVC, exit
  - YES: Apply test 10: [N-SEM: The noun keeps its usual sense?]
    - NO: It is not an LVC, exit
    - YES: Apply test 11: [V-LIGHT: The verb adds zero semantics?]
      - NO: It is not an LVC, exit
      - YES: Apply test 13: [N-PROHIB-ARG: Noun prohibits a regular argument?]
        - NO: It is not an LVC, exit
        - YES: It is an LVC, exit
Test 9 - [N-EVENT] Noun denoting an event/state

Does the noun \( n \) refer to an event or state (including permanent or non permanent properties, relations) with at least one semantic \( 0 \) argument?

» YES ⇒ continue to next test

- (BG) nocemasan azevha 'to emphasize' → event, with two arguments: the agent and the object being emphasized
- (DE) einen Besuch abhalten 'to pay a visit' → event, with two arguments: the visitor and the visitee
- (EI) këvës përu influencë lit. 'to make a visit' → event, with two arguments: the entity glancing and the entity glanced at
- (PT) ter pame lit. 'to have hunger' 'to be hungry' → property, with one argument: the entity that is hungry
- (FR) avoir du courage 'to have courage' → event, with two arguments: the entity having courage and the entity glancing at
- (IT) fare una visita → event, with two arguments: the visitor and the visitee
- (PL) złożyć lit. 'to submit a complaint' 'to make a complaint' → event, with two arguments: the complaining person and the one he/she complains about
- (RO) a face o vizita lit. 'to make a visit' 'to pay a visit' → event, with one argument: the entity that visits
- (SL) imeti predvajanje 'to give a lecture' → event, with two arguments: a lecturer and the people who are attending the lecture
- (BG) meri khomu fokanye 'van threw out the garbage' → physical entity (not event/state)
- (DE) Joe macht einen Kuchen — physical entity (not event/state), even though Joe could be considered its semantic argument
- (EN) Joe makes a cake → physical entity (not event/state), even though Joe could be considered its semantic argument
- (FA) Anna a un vélo → noun not an event, nor a state
- (IT) Joe fa un dolce → physical entity (not event/state), even though Joe could be considered its semantic argument
- (PL) złożyć kartkę lit. 'to fold a sheet' → physical entity (not event/state)
- (PT) quebrar a cabeça lit. 'to break one's head' 'to rack one's brain' → physical entity (not event/state)

We exclude from the test abstract nouns represeting: informational content that do not require agents (informações, notícias), natural phenomena (chuva, neve, tornado).

- (RO) Joe a făcut o grăitură 'Joe made a cake' → physical entity (not event/state), even though Joe could be considered its semantic argument
- (SL) Janez ima avto → the person that has a car could be considered as a semantic argument, but the car is not an event or a state

Test 10 - [N-SEM] Noun keeping its sense

Is the noun \( n \) used in one of its original senses?

» YES ⇒ continue to next test

- (BG) noezematsa meemana 'to take responsibility' → the noun is literally understood
- (DE) einen Besuch abhalten 'to pay a visit' → event, with two arguments: the visitor and the visitee
- (EN) pay a visit → noun is literally understood
- (IT) fare una visita → noun is literally understood
- (PL) podjąć decyzję 'to take a decision' → the noun decyzja is literally understood
- (PT) tomar banho → noun is literally understood
- (RO) face o vizită lit. 'to make a visit' 'to pay a visit' → noun is literally understood
- (SL) imeti pokajni 'to have courage' → noun is literally understood
- (BG) meri khomu fokanye lit. 'van threw out the garbage' → physical entity (not event/state)
- (DE) Herzklopfen haben lit. 'heartbeating' 'to be in love' → noun is not used in one of its normal senses
- (EN) have tingles 'to be worried or angry' → noun is not used in one of its normal senses (it's an ID)
- (FR) jeter l'éponge 'to give up' → noun is not used in one of its normal senses (it's an ID)
- (IT) avere il batticuore lit. 'to have heartbeating' 'to be in love' → noun is not used in one of its normal senses
- (PL) nasadziwiać karkuł lit. to expose neck:GEN'to take personal risks' → the noun karkuł, neck is not literally understood
- (PT) quebrar um galho lit. 'to break a branch' 'to do a favor' → noun is not used in one of its normal senses (it's an ID)
- (RO) face fațălită 'to make face' 'to face to succeed' → noun is not used in one of its normal senses (it's an ID)
- (SL) imeti krompir lit. 'to have potatoes' 'to be lucky' → noun is not used in one of its normal senses (it's an ID)

Test 11 - [V-LIGHT] Verb with light/void semantics

Does \( v \) only bear tense and mood, and add no semantic \( 1 \) that is not already present in \( n \), other than pointing to which semantic role is played by \( v \)'s subject with respect to \( n \)'s predicate?

» YES ⇒ continue to next test
• (BG) ζητώμενο ρώσιμο 'make a decision' → ζητώμενο adds no meaning to ρώσιμο 'decision' besides that of performing an act
• (DE) eine Entscheidung treffen lit. a decision meet 'to make a decision' → treffen adds no meaning to Entscheidung besides that of performing an activity
• (EN) take a walk → take adds no meaning to walk besides that of performing an activity

Note that this light semantics of the verb is either usual for that verb (i.e. the verb is a pure syntactic operator, like

YES 'order to the soldiers to leave'

have an opinion

pay a visit → the verb in its usual sense means 'to spend some money on a visit', but here it is not used in this sense and does not add any semantics to the "visiting" event

deliver a speech → the verb in its usual sense means 'to move from one placeto another', but here it is not used in this sense and does not add any semantics to the "speech" event

undergo a surgery → undergo adds no meaning to surgery besides indicating that the subject is the patient of the surgery

(R) έλθε τους στρατιώτες έναν το άλλον 'order to leave' → adds no meaning to τους στρατιώτες besides that of having a property

Peter's threats to Paul

Jean makes a search

I have a hypersensitivity

Peter's threats to Paul

Ce médicamente présente un risque lit. This medicine presents a risk → le risque de ce médicament'this medicine's risk'
The noun test 13 - [N-PROHIB-ARG] Noun's prohibited argument

Does the noun n, in the presence of v, prohibit at least one syntactic argument a which it normally licensed in the absence of v (except when a is in the whole–part relation with v's subject)?

Note: An alternative formulation for this test is the following: Let s be the subject of v, and let t be the semantic role that s plays with respect to the noun n. Is it prohibited for t to be realized both by s and by a syntactic argument a of n, except when a is in the whole–part relation with s? The rationale for this test is that a semantic argument a cannot be realized as its syntactic dependent, since it is already realized as its syntactic dependent instead (usually as v's subject). For instance the noun visit takes two semantic arguments, the visitor and the visited entity, as in

Paul made a conquer → Paul made a good impression → #Impresia lui Paul despre soția sa 'Paul's impression on his wife' — different semantics

Test 13 - [N-PROHIB-ARG] Noun's prohibited argument

Yes: It is an LVC

(BG) Петър Стоянов взе решение да подпише договора 'Ivan made the decision to sign the contract' vs 'решение на президента да подпише договора' — the noun cannot be modified by the person performing the act/event (which is the subject)

(DE) Die Königin hat dem Premierminister einen Besuch abgestattet 'The Queen has paid a visit to the Prime Minister' vs ein Besuch der Dame beim Premierminister 'a visit of the Lady to the Prime Minister' — the discussing entity cannot be a modifier of visit

The street carries the forename of a famous poet. 'The street carries the name of a famous poet.'

Paul leads the discussion vs Peter's discussion — Peter leads Peter's discussion but Paul leads the discussion of the committee — the discussing entity can only modify discussion when the subject is part of the committee

La ministre a rendu une visite aux victimes vs la visite de la ministre aux victimes — the visitor cannot be a modifier of visit

Bjarnar a marqué un but vs le but d'Arnason — Paul a marqué le but d'Arnason but Paul a marqué le but de l'Islande — the scoring entity can only modify goal in the last case, when they are part of the Iceland team

Il primo ministro ha preso la decisione di dimettersi vs the Prime Minister decided to resign vs le dimissioni del governo — the resigner cannot be a modifier of resignation

Paweł złożył wymówienie vs Paweł made a good impression — #Paul's Eindruck auf seine Freunde 'Paul's impression on his friends' vs #погледът на Иван върху вестника

The director position -> the resignation cannot be modified by the resigning person
Section 6.3

Idioms (ID)

Idioms constitute a universal category. An idiom (ID) has at least two lexicalized components including a head verb and at least one of its arguments. The argument can be of different types. Here are some examples:

Subject

* (BG) Спом емъ ци пебама lit. 'Ivam transmitted the decision to his sister' + решението на комисията → Иван предаде решение на комисията 'Ivan predareshenieto na komisiata'

* (DE) Paul hat seiner Schwester die Entscheidung überbracht 'Paul has transmitted the decision to his sister' + Peter's Entscheidung 'Peter's decision' → Paul hat seiner Schwester Peter's Entscheidung überbracht 'Paul transmitted Peter's decision to his sister' — the decision can be modified by its author

* (PL) Paweł podważył niedawną decyzję 'Paweł questioned the recent decision' + decyzja Piotra's decision 'Piotra's decision' → Paweł podważył niedawną decyzję Piotra's decision — the decision can be modified by its author

* (PT) Paulo tem noticias sobre o conflito 'Paul has news about the conflict' + as noticias de Pedro 'Peter's news' → Paulo tem as noticias de Pedro sobre o conflito 'Paul has Peter's news about the conflict' — the news can be complemented by its source

Direct object

* (BG) същата същност 'the same entity' + същност on the movie 'the movie' → винаги сам да си говори 'lit. always (someone) to say or do something stupid'

* (PL) Paweł transmitte ordre aux soldats 'Pawel transmitted the order to the soldiers' + ordre de Paul aux soldats 'Paul's order to the soldiers' → Paul transmitte ordre aux soldats 'Paul transmitted the order to the soldiers' — the order can have as a complement the person who gave it

* (RO) Paweł a transmis sfatul surorii sale 'Pawel transmitted the advice to his sister' + sfatul lui Peter 'Peter's advice' → Pawel a transmis sfatul lui Peter surorii sale 'Pawel transmitted Peter's advice to his sister' — the advice can be modified by its author

* (SL) Janec mi je povetdal mnenje o filmu 'Janez told me his opinion on the movie' + mnenje o filmu 'Nina's opinion on the movie' → Janez mi je povetdal Ninino mnenje o filmu 'Janez told me Nina's opinion on the movie' — the opinion can be modified by the person who has this opinion

Structural tests

* (DE) er hat den Schuss nicht gehört 'lit. he did not hear it' + ein Schuss 'a shot' + lit. 'lit.' + to keicken the bucket + lit. 'lit.'

* (RO) o gopiti o pâsăricălit 'lit. whispered a bird little' + lit. 'lit.'

* (SL) srce me je padlo v hlace 'lit. His heart fell into his pants' + someone has lost courage

* (BG) нямам пълен урок 'lit. I have a full class' + урок 'class' + lit. 'lit.'

* (PL) wiercić dziurę w brzuchu 'lit. to drill a hole in one's belly' + lit. 'lit.' + to intrusively solicit someone, to insist too much

* (PT) levar em conta 'lit. to bring in account' + to take into account + lit. 'lit.'

* (RO) o luz in considerare 'lit. to bring in account' + to take into account + lit. 'lit.'

* (SL) spati kot ubar 'lit. to sleep like dead' + to sleep soundly

It is often challenging to distinguish IDs from other VMWEs if only one argument of the head verb is lexicalized. The VMWE categorisation depends on the category of this argument:

* Noun or preposition governing a noun: fine-grained tests need to be applied in order to discriminate between an LVC and an ID. See the section on Structural tests.

* Particle or reflexive pronoun: the VMWE is either a VPC (particle) or an IrEfV (reflexive pronoun), never an ID.

With an argument of any other category, the VMWE is always an ID, including the following:

* (PT) João está tomando banho 'lit. John is taking shower' + o banho do Pedro lit. 'the bath cannot be modified by a bath taker'

* (DE) Pedro hat seiner Schwester die Entscheidung überbracht 'Paul has transmitted the decision to his sister' + Peter's Entscheidung 'Peter's decision' → Paul hat seiner Schwester Peter's Entscheidung überbracht 'Paul transmitted Peter's decision to his sister' — the decision can be modified by its author

* (PL) Paweł podważył niedawną decyzję 'Paweł questioned the recent decision' + decyzja Piotra's decision 'Piotra's decision' → Paweł podważył niedawną decyzję Piotra's decision — the decision can be modified by its author

* (RO) Pawel a transmis sfatul surorii sale 'Pawel transmitted the advice to his sister' + sfatul lui Peter 'Peter's advice' → Paul a transmis sfatul lui Petere surorii sale 'Paul transmitted Peter's advice to his sister' — the advice can be modified by its author

* (SL) Janec mi je povetdal mnenje o filmu 'Janez told me his opinion on the movie' + mnenje o filmu 'Nina's opinion on the movie' → Janez mi je povetdal Ninino mnenje o filmu 'Janez told me Nina's opinion on the movie' — the opinion can be modified by the person who has this opinion
• Preposition governing a complex noun phrase
  - (BG) удрём немного в юбку lit. hit someone in the back 'to stab someone in the back'
  - (DE) und springt im Dreieck lit. s.o. jumps in the triangle 's.o. can soon no more control his anger'
  - (EN) to take something with a pinch of salt
  - (ES) coger algo con pinzas lit. to hang something with pegs 'to take something with a pinch of salt'
  - (PL) dopiąć coś na ostatni zukit lit. to button something up to the last button 'to complete something'
  - (PT) bater na mesma tecla lit. to hit the same key 'to insist on something'
  - (RO) a da cu pictură lit. to give with leg-the 'to give up the chance'
  - (SL) skrivati glavo v pesek lit. to hide head in the sand 'to pretend not to see a problem'

• Adjectival phrase
  - (DE) schwarm fahren lit. to drive black 'to take a ride without a ticket'
  - (EN) to come clean, to stand firm
  - (ES) jugar sucio lit. to play dirty 'to play dirty'
  - (PL) zrobić swoje lit. to do one's own 'to do what one is supposed to do'

• Relative clause
  - (DE) wissen wo es langgeht lit. to know where things are heading 'to know on which side one's bread is buttered'
  - (EN) to know on which side the bread is buttered
  - (PL) wiedzieć, skąd wieje wiatr lit. to know where wind blows from 'to know on which side your bread is buttered, to know how to take advantage of the situation'
  - (PT) saber onde pisar lit. to know where to-step 'to know the way to succeed in something'

• Verbal phrase
  - (DE) will sagen lit. want to say 'that is to say'
  - (EN) to make do
  - (ES) n.a.
  - (FR) laisser tomber lit. to let fall 'to give up'

• Non-reflexive pronoun
  - (DE) es gibt lit. it gives 'there is'
  - (EL) το κυρίεσπρου, τη μαμάκι lit. (EN) to make it
  - (FR) l'emporter lit. to take it away 'to win'
  - (IT) prendere le lit. to take it 'to be beaten'

• Adjective
  - (PT) a o intinde lit. to her extend 'to fly the coop'
  - (RO) o a o întinde lit. to her extend 'to fly the coop'

• Sentential expressions with no open slots, such as proverbs and conventionalized sentences, are included in the scope of IDs.

  - (DE) Rom wurde nicht an einem Tag erbaut 'Rome was not build in a day' wer A sagt muss auch B sagen lit. who says A must also say B 'you must finish what you start'
  - (EN) Rome was not built in a day
  - (RO) Un補au nu merge de multe ori la spațiu lit. 'The pitcher goes so often to the well that it is broken at last'
  - (SL) vedeti koliko je ura lit. to know what time it is 'to realize the truth'

If more than one argument of the head verb is lexicalized, then the candidate VMWE it is always classified as an ID.

  - (DE) die Katze aus dem Sack lassen to let the cat out of the bag'
  - (EN) to let the cat out of the bag 'to draw a long story short, to call it a day'
  - (FR) se faire des idées lit. to make SELF ideas 'to imagine something false' s'en aller lit. to go SELF from there 'to leave! It y a lit. it has there 'there is'

In case of severallexicalized arguments, special care must be taken to identify and also annotate embedded VMWEs.

  - (DE) einen Plan aufstellen lit. to set up a plan 'to draw up a plan' → contains the VPC
  - (EN) to let the cat out of the bag 'to give up the chance'
  - (RO) a da cu pictură lit. to give with leg-the 'to give up the chance'
  - (SL) skrivati glavo v pesek lit. to hide head in the sand 'to pretend not to see a problem'
We start by listing the various categories of meaning. REFLV can be categorized into different classes, some of which should be annotated as verbal MWEs.

- Reflexive verbs (REFLV)
- Reflexive clitics (RCLI)
- Inherently reflexive verbs (IReflV)

Section 6.4

Inherently reflexive verbs (IReflV)

Reflexive clitics (RCLI) are clitic pronouns that refer to the subject of the verb, like #self in English. They are very common in many languages and play several semantic roles depending on the context, as detailed below.

Reflexive verbs (REFLV), sometimes also called pronominal verbs, are formed by a full verb combined with a RCLI, although the clitic does not always have a reflexive role. IReflV can be categorized into different classes, some of which should be annotated as verbal MWEs.

Namely, we will only annotate a REFLV as an Inherently reflexive verb (IReflV) when (a) it never occurs without the clitic, or (b) the REFLV and non-reflexive versions have clearly different senses or subcategorization frames. Inherently reflexive verb constitute a quasi-universal category.

We start by listing the various categories of REFLV before providing tests to decide whether to annotate a given occurrence as IReflV.

- **Inherently reflexive** ➔ **ANNOTATE** as IReflV
  - The verb without the RCLI does not exist
    - (CS) stydět se 'to be ashamed', divít se 'to wonder'
    - (DE) sich schämen 'to be ashamed', sich wundern 'to wonder'
    - (ES) suicidarse 'to suicide', abstenere 'to abstain'
    - (FR) s'évanouir 'to faint', se suicider 'to suicide'
    - (IT) suicidersi 'to suicide', arrabbiarsi 'to get angry'
    - (PL) dowiedzieć się 'to find out', bać się 'to be afraid'
    - (PT) queixar-se 'to complain', abster-se 'to abstain'
    - (RO) a se teme 'to be afraid' with obligatory ACC reflexive clitic
      - a fiu de bisericii 'to be honest'
    - (SV) att föröva sig 'to sleep in'
    - (SL) sramovati se 'to be ashamed', bati se 'to be afraid'

  - The verb without the RCLI does exist, but has a very different meaning
    - (DE) sich enthalten 'to contain', sich (um etw.) handeln 'to be # handle to'
    - (ES) recoger # recogerse 'to gather # to go home', empeñar # empeñarse 'to pawn # to insist'
    - (FR) s'apercevoir # s'apercevoir 'to realize # to see', s'agir # agir 'to be # to act'
    - (IT) riferire # riferirsi 'to report, tell # to refer'
    - (PL) znajdować # znajdować się 'to find # to be', radzić # radzić sobie 'to advise # to manage'
    - (PT) encontrar-se # encontrar 'to be # to meet', referir-se # referir 'to concern # to refer'
    - (RO) a se indura 'to have the heart # to suffer'
    - (SV) att föröva sig 'to feel sad/angry # to touch'
    - (SL) tkati se česa 'to tick' 'to concern # to refer', pobrati se 'to pick up'

  - **Reciprocal** ➔ **NOT ANNOTATED**

  - The RCLI has a sense of mutually:
    - (CS) ibat se 'to kiss each other', potkávat se 'to meet each other'
    - (DE) sich küszen 'to kiss each other', sich treff'en 'to meet each other'
    - (ES) besarse 'to kiss each other', verse 'to see each other'
    - (FR) s'embrasser 'to kiss each other', se rencontrer 'to meet each other'

Idioms whose head verb is the copula (to be) can pose special challenges because their complements may be (nominal, adjectival, etc.) MWEs themselves. In this task, we consider constructions with a copula to be VMWEs only if the complement does not retain the idiomatic meaning when used without the verb.

- (DE) sein Froesch lit. 'be no frog' 'be no chicken' 'idiom because kein Froesch 'no frog' loses the meaning
- (EN) to be no chicken → idiom because ño chicken loses the meaning
  - to be somebody → idiom because ño somebody loses the meaning
  - it is double Dutch to me → non-VMWE because the copula can be omitted, as in he seems to speak double Dutch
- (ES) Ser un pelota lit. a ball 'to suck' 'idiom because'
- (PL) być jedną nogą na tamtym świecie lit. 'to be with one leg in the other world' loses the meaning
  - być do rzeczy lit. 'to be the thing/to be relevant' → non-VMWE because the copula can be omitted, as in dat pare argumentów całkiem do rzeczy he gave a couple of quite relevant arguments'
- (PT) ser alguém na vida lit. 'to be somebody' → idiom because alguém na vida loses the meaning
  - não ser flor que se cheire lit. 'to not be a flower that one may smell' loses the meaning
  - isso é grego pra mim lit. 'that's greek to me' → non-VMWE because the copula can be omitted, as in você está falando grego
- (RO) a fi ușă de bisericii lit. 'to be in church' 'idiom because ușă de bisericii loses the meaning
  - a fi un papă-lapte lit. 'to be a papa-latte' 'idiom because un papă-lapte preserves the meaning'
- (SL) biti trn v peti komu lit. 'to be in somebody's heel' 'idiom because trn v peti loses the meaning

Note that special care must be taken in languages in which the copula omission is a regular or even a compulsory phenomenon (e.g. in Russian). In those cases, language-specific tests are required to distinguish a copula-based idiom from a non-verbal MWE.

Idioms typically have both a literal and an idiomatic reading. Thus, they are closely connected to the phenomenon of a metaphor (see also the section on VMWEs versus metaphors). This often makes them semantically totally non-compositional, i.e. none of their lexicalized components retains any of their original meanings.
Inchoative Middle with preverbal subject, also called synthetic passive

Reflexive ⇒ NOT ANNOTATED
- The RCLI marks the **reflexive or reciprocal** construction, that is, the clitic plays the role of subject in English
  - (CS) myt si nohy lit. wash RCLI.DAT the feet 'wash one's feet'
  - (DE) sich waschen lit. 'wash oneself', sich kratzen lit. 'to scratch oneself'
  - (ES) lavarse lit. 'to wash oneself', vestirse lit. 'to dress oneself'
  - (FR) se laver lit. 'wash oneself', se parler lit. 'to talk to oneself'
  - (IT) lavarsi lit. 'to wash oneself', vestirsi lit. 'to dress oneself'
  - (PL) myć się lit. 'to wash oneself', drapać się po głowie lit. 'to scratch oneself on the head'
  - (PT) apressar-se lit. 'hurry oneself', vestir-se lit. 'to dress oneself'
  - (RO) a se spăla lit. 'wash oneself'
  - (SV) sittvåta sig lit. 'wash oneself'

- (SL) umivati se lit. 'wash oneself', praskati se lit. 'to scratch oneself'

Body part, also called possessive reflexive ⇒ NOT ANNOTATED
- Specific type of reflexive use in which the direct object is a body part or, more generally, an inalienable part of the subject
  - (CS) myt si nohy lit. wash RCLI.DAT the feet 'wash one's feet'
  - (DE) sich das Bein brechen lit. RCLI the leg break 'break one's leg'
  - (ES) rascarse el brazo lit. scratch.RCLI the arm 'scratch one's arm'
  - (FR) se gratter la tête lit. RCLI scratch the head 'scratch one's head'
  - (IT) grattarisi la testa lit. RCLI scratch the head 'scratch one's head'
  - (PL) myć sobie nogi lit. wash RCLI.DAT the feet 'wash one's feet'
  - (PT) impossível, uses possessive instead
  - (RO) a-și rupe mâna lit. RCLI break arm 'break one's arm'
  - (SL) umivati noge lit. wash RCLI.DAT the feet 'wash one's feet', zlomiti roko lit. RCLI.DAT break arm 'break one's arm'

Middle with preverbal subject, also called synthetic passive ⇒ NOT ANNOTATED
- The clitic marks a regular syntactic alternation for transitive verbs. Just like in regular passive alternation, the direct object of the transitive version appears as the subject of the REF LV version, and thus the verb agrees with the subject.
- Differently from inchoative (see below), the subject of the transitive version is absent in the REF LV version but it exists necessarily, though it is underspecified
  - (DE) die Häuser verkaufen sich gut lit. the houses sell RCLI well 'the houses sell well'
  - (ES) las casas se venden bien lit. the houses RCLI sell well 'the houses sell well'
  - (FR) les pots se vendent bien lit. the pots RCLI sell well 'the pots sell well'
  - (IT) le case si affittano lit. the houses RCLI rent 'the houses are rented'
  - (PL) domy dobrze się sprzedają lit. houses sell.PL RCLI well 'houses sell well'
  - (PT) as casas se vendem bem lit. the houses RCLI sell well 'the houses sell well'
  - (RO) casele se vând bine lit. houses-the RCLI sell well 'houses sell well'
  - (SL) hiše se dobro prodajajo lit. the houses sell.RCLI well 'the houses sell well'

Middle with postverbal subject, also called syntactic passive ⇒ NOT ANNOTATED
- In some languages, middle alternative with preverbal subject sounds unnatural and middle alternative with postverbal subject is preferred. Depending on the languages, it is viewed as a postverbal subject (ES, PL, PT, RO) or as an object which agrees with the unaccusative verb form (IT). Middle alternation with postverbal subject is impossible in FR and DE.
  - (ES) se aqlitian casas lit. RCLI rent houses 'people rent houses'
  - (IT) si affittano case lit. RCLI rent houses 'people rent houses'
  - (PL) dobrze sprzedają się te domy lit. well sell.RCLI these houses 'these houses sell well'
  - (PT) alugam-se casas lit. rent-RCLI houses 'people rent houses'
  - (RO) se vând bine apartamentele din blocurile noii lit. RCLI sell well apartments-the from blocks-the new 'Apartments from new blocks sell well'
  - (RO) casele se vând bine lit. RCLI sell well 'houses sell well'
  - (SL) novo hiše se gradijo lit. new houses RCLI build 'new houses are built'

Impersonal ⇒ NOT ANNOTATED
- The RCLI marks an impersonal verb alternation possible for various transitivity classes, depending on the language: only transitive verbs (FR), only intransitive verbs with manner adjuncts (DE), preferably intransitive but tolerated for transitive verbs (PT), either transitive or intransitive verbs (IT, ES, RO, PL)

- There is no noun phrase before the verb (empty subject slot), the presence of the RCLI indicates a verb interpreted with a generic and underspecified subject
  - The verb is in third person singular, even when the object is plural
    - (DE) hier tanzt es sich gut lit. here dances it RCLI well 'people dance well here'
    - (ES) se busca a actores lit. RCLI searches to actors 'people look for actors'
    - (FR) se dit des bêtises lit. it RCLI says silly things 'people say silly things'
    - (IT) si lavora troppo lit. RCLI works too much 'people work too much'
    - (PL) za dużo się praca lit. too much RCLI works 'people work too much'
    - (PT) domne-se muito lit. sleeps-RCLI much 'people sleep a lot'
    - (RO) se lucrează până târziu lit. RCLI works until late 'people work until late'
    - (SL) se lucreza dobro lit. works-the RCLI well 'people work well'

- Inchoative ⇒ NOT ANNOTATED
- Similar to middle, but the RCLI marks a less productive syntactic alternation:
  - the direct object of the transitive version appears as subject of the REF LV
  - the subject of the transitive version is not only absent, it is also semantically unclear or nonexistent
    - (CS) dvele se olvírají lit. 'the door opens'
    - (DE) die Tür öffnet sich lit. 'the door opens'
IReflV-specific decision tree

Test 14 - [INHERENT] Inherent clitic

Does the verb only exist with the RCLI and never occurs without it?

- YES ⇒ annotate as IReflV
- NO ⇒ Apply test 15 - [DIFF-SENSE]
- NO or UNSURE ⇒ Apply test 16 - [DIFF-SUBCAT]
- YES ⇒ Annotate as IReflV

Test 15 - [DIFF-SENSE] - Different sense

Given the same verb without the RCLI, are all of its meanings clearly different from the REFLV form?

- YES ⇒ annotate as IReflV
- NO ⇒ next test

Test 16 - [DIFF-SUBCAT] - Different subcategorization frame

Is the subcategorization frame of the simple verb without the RCLI different from the subcategorization frame of the REFLV, except for the addition of a direct or indirect object corresponding to the same argument as the RCLI in the REFLV version?

- YES ⇒ annotate as IReflV
When you replace the RCLI by the REFLV version? In other words, when you move the subject to the object position, remove the RCLI and add a generic subject (people, somebody), thus building a transitive version, does it imply the REFLV form without changing the REFLV's meaning?

When you move the subject to the object position, remove the RCLI and add a generic subject (people, somebody), thus building a transitive version, does it imply the REFLV form without changing the REFLV's meaning?
Problematic cases and remarks

Polysemy

Keep in mind that both simple and reflexive verbs can have several senses. In test 15, we ask that ALL senses you can think of are different from the REFLV form in the given context. For example, French verb trouver can mean to find something, to have an opinion about something, discover something, etc. But it has a totally different and unrelated meaning of to be (located at) in the sentence L'église se trouve à Paris 'the church is located in Paris'. It should thus be annotated as a MWE. As the REFLV is polysemous itself, it should NOT be annotated as IReflV in sentences like Elle se trouve grosse 'she finds herself fat' where it means have an opinion about (herself), equivalent to the non-reflexive version.

Clitics position and concatenation

In some languages the clitics are joint with the verb, sometimes using a hyphen but not always. When there is no hyphen, the REFLV will probably be tokenized as a single token in the corpus.

- In French, orthography and pronunciation rules require the clitic to be concatenated with the verb and its last vowel to be replaced by an apostrophe (liaison):
  - (FR) s'abstenir 'to abstain'
- In Spanish and Italian, the clitic can appear concatenated after the verb in some verbal forms (e.g. infinitives, gerunds):
  - (ES) enamorarse 'to fall in love'
  - (IT) alzarsi 'to get up'
- In Portuguese, there are always hyphens for postponed clitics (enclisis), but in conditional tense the clitic is in the middle of the verb (mesocnosis), separating the root from the suffix:
  - (PT) queixar-se-ia 'would complain'
- In Romanian the clitic and the verb are either separate or have a hyphen between them:
  - (RO) se aude un clopot lit. RCLI hears a bell 'a bell is heard'
  - s-aude un clopot lit. RCLI-hears a bell 'a bell is heard'

The current annotation format allows annotating a single token as a MWE if it is a multiword token. Therefore, it should be annotated as an MWE.

Overlap ID - IReflV

Some idiomatic constructions include reflexive clitics. Two cases are possible:

- If a syntactically comparable literal construction is impossible or the REFLV would not be annotated in syntactically comparable literal constructions, annotate only the ID:
  - (DE) sich über etwas im Klaren sein dass S lit. RCLI about s.th. in the clear be 'to be aware of s.th./that S' ⇒ "sich in N sein, dass for any noun N meterse en lios 'to get in trouble' ⇒ REFLV not annotated in literal equivalents like meterse en una tienda 'to get in a store'
that should be annotated as ID rather than IReflV include:

- Non-reflexive clitics

Some expressions can have double clitics. Only the first two words belong to the IReflV:

- (PT) dar-se mal 'to fall' ⇒ dar-se ADV intransitive is acceptable only for antonym bem 'well'
- (RO) meter-se numa fria 'lit. to get-RCLI in-a cold' ⇒ REFLV not annotated in literal equivalent meter-se numa cabine 'to get into a cabin'

- (SL) puliti si lase lit. tear RCLI the hair 'to worry' ⇒ REFLV not annotated in literal equivalents like puliti si obrvi 'to pluck one's eyebrows'

If the REFLV would be annotated as IReflV in syntactically comparable literal constructions, annotate both the IReflV and the ID as embedded MWEs (rare):

| (FR) | se rendre compte de 'to realize' ⇒ se rendre N de for any noun N |
| (IT) | rendersi con to 'realize' ⇒ ai rendere N di for any noun N |
| (PL) | zdawać sobie sprawę z 'to realize' ⇒ zdawać sobie N z for any noun N |
| (PT) | dar-se mal 'to fall' ⇒ dar-se ADV intransitive is acceptable only for antonym bem 'well'
| (RO) | meter-se numa fria 'lit. to get-RCLI in-a cold' ⇒ REFLV not annotated in literal equivalent meter-se numa cabine 'to get into a cabin' |

Overlap LVC - IReflV

It is rare, although possible, to find light verb constructions in which a reflexive clitic changes the original meaning significantly, thus characterizing an IReflV:

| (FR) | il y avoir |
| (DE) | en avoir marre |
| (IT) | s'en aller |
| (RO) | a se duce pe apa Bistriței he goes on the river Bistrița |
| (PL) | radzić sobie |
| (CS) | bati se |

It is rare, although possible, to find light verb constructions in which a reflexive clitic changes the original meaning significantly, thus characterizing an IReflV:

Dative clitics and double clitics

In some languages, e.g. Polish, clitics inflect for case. Most cases of IReflV seem to be restricted to the accusative case:

| (CS) | bát se 'to be afraid' |
| (PL) | bać się 'to be afraid' |
| (RO) | a se sinchiliš. lit. to RCLI.ACC care 'to care' |
| | a se sfiliț. to RCLI.ACC be.shy 'to be shy'
| | a se căiță. to RCLI.ACC care 'to care'
| (SL) | batí se 'to be afraid' |

However, other cases can appear in IReflV:

| (CS) | poradit si lit. to advise oneself.DAT 'to manage' |
| (PL) | radzić sobie lit. to advise oneself.DAT 'to manage' |
| (RO) | a-gi înșisi lit. to-RCLI.DAT appropriate/to appropriate' - with a Dative clitic |
| | a-gi apropiatil. to-RCLI.DAT appropriate/to appropriate' - with a Dative clitic |
| (SL) | drzniti si lit. to dare oneself.DAT 'to dare' |

Some expressions can have double clitics. Only the first two words belong to the IReflV:

| (PL) | przyglądać się sobie lit. to observe RCLI.ACC RCLI.DAT 'to observe each other' |
| (RO) | n.a. |
| (SL) | nasmehnuti se sobi 'to smile at oneself' |

Non-reflexive clitics

This category does not cover other types of pronouns and clitics. They are covered by regular ID tests and should be annotated as such. Examples of constructions that should be annotated as ID rather than IReflV include:

| (DE) | es gibt lit. it gives 'there is' |
| (FR) | l'emporter lit. to take it away 'to win' |
| | s'en aller lit. to self from-it go 'to leave' |
| | en avoir marre lit. to have from-it enough 'to be fed up' |
| | il y avoir lit. it at-it have 'to exist' |
Section 6.5

Verb-particle constructions (VPCs)

Verb-particle constructions (VPCs), sometimes called phrasal verbs or phrasal-prepositional verbs, like

- (DE) umfahren lit. overdrive 'to run over', mitkommen lit. with|come 'to join', vorbereiten lit. before|prepare 'to prepare'
- (EN) to put off to blow up to do in
- (PL) n.a.
- (RO) n.a.
- (SL) biti za lit. to be for 'to agree', šteti za lit. to count for 'to consider as' constitute another quasi-universal category. They have the following general characteristics:

1. They are formed by a lexicalized head verb $v$ and a lexicalized particle $p$ dependent on $v$.
2. The meaning of the VPC is non-compositional. Notably, the change in the meaning of $v$ goes significantly beyond adding the meaning of $p$:

   - (DE) die Fische sind eingegangen lit. the fish went in 'the fish died'
   - (EN) to do in 'to die'
   - (ES) n.a.
   - (PL) n.a.
   - (RO) n.a.
   - (SL) biti za lit. to be for 'to agree', šteti za lit. to count for 'to consider as' VPCs are pervasive in English, German, Swedish, Hungarian and possibly some other languages but irrelevant to or very rare in Romance and Slavic languages or in Farsi and Greek for instance.

In some Germanic languages and also in Hungarian, verb-particle constructions can be spelled either as one (multiword) token or separated. Both types of occurrences are to be annotated:

- (DE) Die Kinder sollen in der Schule aufpassen 'The children must pay attention at school'
  Herr Müller, passen Sie auf 'Mr. Müller, be careful'
- (ES) n.a.
- (PL) n.a.
- (RO) n.a.
- (SL) n.a.

The first challenge in identifying a VPC is to properly distinguish the particle from a possibly homographic preposition, e.g.:

- (EN) to get up a petition vs to get up a hill
- (ES) n.a.
- (PL) n.a.
- (RO) n.a.
- (SL) biti za njeno idejo 'He agrees with her idea' vs biti za zaveso 'He is behind the curtain'

or a verbal prefix:

- (DE) um- in umfahren vs umfahren
  (ES) n.a.
  (PL) n.a.
  (RO) n.a.

Namely, a particle, contrary to a preposition, cannot introduce a complement

- (EN) to do sb in, 'to do in sb!
- (ES) n.a.
- (PL) n.a.
- (RO) n.a.

and prefixes can never be spelled separately from the verb, nor can the past tense of prefixed verbs be formed with the infix -ge-

- (DE) 'er fuhr den See um
  'er hat den See umfahren, instead: er hat den See umfahren 'he drove around the lake' but: er hat das Schild umfahren 'he run over the sign'

See the language-specific tests for more details on distinguishing particles from prepositions and verbal prefixes.

Note that in this shared task we do not account for compositional verb-particle combinations, i.e. those whose meaning can be deduced from the meaning of the preposition and of the verb:
Language-specific categories

Language-specific categories can be proposed for annotation in this task provided that they are carefully defined and accompanied by linguistic tests that allow to distinguish them from other categories. We recommend not redefining the universal and quasi-universal categories described here, but introducing new names and abbreviations in order to answer such needs.

Section 6.7

Other verbal MWEs (OTH)

This category is meant to contain VMWEs which do not fit to the preceding categories, that is, whose lexicalized components do not include a single head verb and at least one of its arguments. VMWEs in this category fail the structural test 6 [HEAD]. They include:

- coordinations of verbs
  - (DE) *leben und leben lassen* 'live and let live'
- (EN) *to drink and drive*
- (ES) *coser y cantar*
- (PL) *pluć i łapać* lit. *spit and catch* 'to be lazy, to do nothing useful'
- (RO) *tunat* și *i-a adunat* lit. 'it has thundered and CL.ACC-it.has gathered'
- (SL) *živi in pusti* živeti lit. 'to be so good as to do something'
- (DE) *ich halte es nicht länger aus mit ihm* 'I can no longer put up with him'
- (EN) *I can no longer put up with him*  
  (RO) n.a.  
  (PL) n.a.  
  (RO) n.a.

Section 7
Language-specific tests

Language-specific tests may be necessary in one of 3 cases. Firstly, a VMWE category may be universal or quasi-universal but it may require different tests in different languages. Secondly, any category specific to a language must be associated with appropriate test in the same language. Thirdly, universal categorisation tests can build upon more elementary language-specific tests (e.g. to distinguish a particle from a preposition).

Section 7.1
Particles versus prepositions and prefixes

The following tests allow to properly identify prepositional verb particles in cases where they might be homographic with prepositions in prepositional phrases (PPs) or with verbal prefixes. The word to be discriminated is referred to as a *candidate word*. The tests are language-specific and concern English and German.

English-specific test for distinguishing particles from preposition

The following tests concern English words which can be either a proposition or a particle depending on the context, e.g. *up*, *on*, *through*, etc. If a candidate word passes any of the three following tests it can be categorized as a particle.

Test 7.1.EN - [PART+PREP] - Particle followed by a preposition

Is the candidate word followed by a preposition?

- **YES** ⇒ it is a particle
  - (DE) *Ich halte es nicht länger aus mit ihm* 'I can no longer put up with him'
  - (EN) *I can no longer put up with him*
  - (PL) n.a.
  - (RO) n.a.

- **NO** ⇒ other tests are needed
  - (DE) *Ich klettere den Berg hinauf* 'I climb up the hill'
  - (EN) *I got up the hill*
  - (PL) n.a.
  - (RO) n.a.

Test 7.2.EN - [FIN-PART] - Sentence-final particle

Does the candidate word w occur at the end of a clause which is: (i) affirmative or imperative, (ii) headed by the verb governing w, and (iii) not a relative clause? Or can the sentence be reformulated so as to put the candidate word at the end of such a clause?

- **YES** ⇒ the candidate word is a particle
  - (DE) n.a.
  - (EN) *They got up a petition on Monday. They got it up.*
  - (PL) n.a.
  - (RO) n.a.

- **NO** ⇒ other tests are needed
  - (DE) n.a.
  - (EN) *I got up the hill. I got it up.*
  - (PL) n.a.
  - (RO) n.a.
Test 7.3.EN - [AD-INS] - Adjunct insertion

Is an insertion of a circumstantial adjunct prohibited between the governing verb and the candidate word?

↳ YES ⇒ the candidate word is a particle
- (DE) n.a.
- (EN) I took off my clothes at once. *I took off my clothes at once.* She always tries to take in her clients. *She tries to take in her clients.*
  - (PL) n.a.
  - (RO) n.a.

↳ NO ⇒ other tests are needed
- (DE) n.a.
- (EN) He has been off alcohol recently. He has been off alcohol.
  - (PL) n.a.
  - (RO) n.a.

German-specific tests for distinguishing particles from prepositions and verbal prefixes

The following tests concern German words which can be both a particle and either a preposition or a verbal prefix, depending on the context, e.g. mit, um, vor, etc. If a candidate word passes any of the three following tests it can be categorized as a particle.

Test 7.1.DE - [FIN-PART] - Sentence-final particle

Does the candidate word occur at the end of the sentence or can the sentence be reformulated so as to put the candidate word at the end?

↳ YES ⇒ it is a particle
- (DE) Kommst Du mit? lit. come you with? *Are you coming?* Ich schlage vor allen zu verzeihen. *I propose to forgive everyone* Ich schlage es vor *I propose it*
  - (PL) n.a.
  - (RO) n.a.

↳ NO ⇒ other tests are needed
- (DE) n.a.
- (EN) Kommt Du mit jemandem? *Are you coming with someone?* *Kommt Du jemandem mit?*
  - (PL) n.a.
  - (RO) n.a.

Test 7.2.DE - [SEP-PART] - Separable particle

Can the verb and the candidate word be spelled both separately and together?

↳ YES ⇒ it is a particle
- (DE) Passen Sie auf die Autos auf! *Be careful with the cars!* Sie müssen auf die Autos aufpassen! *You must be careful with the cars!*
  - (PL) n.a.
  - (RO) n.a.

↳ NO ⇒ other tests are needed
- (DE) Er fuhr den ganzen See mit dem Fahrrad. *He drove around the whole lake with a bike* *Er fuhr ihn um*
  - (PL) n.a.
  - (RO) n.a.

Section 8

Annotation management

This section groups the documentation on practical aspects of the annotation campaign management. Some of these aspects are specific to this shared task, such as the edition of examples by language leaders and the use of the annotation platform FLAT. Others are more generic and concern the guidelines in general, such as the FAQ section.

Section 8.1

Frequently Asked Questions (FAQ)

Annotators often face questions and challenging examples. When several annotators ask the same question, we will update the list of frequently asked questions.

However, we suggest that language teams set up another communication platform to deal with questions that are specific to a language. This can take the form of a shared online document, a wiki, a dedicated bug tracking system or mailing list. We also suggest keeping track of decisions taken considering borderline examples (with a list of expressions to which the decision applies). These should be kept in a centralized document or page that all annotators can access.

Whenever you think that a question can also be interesting to other languages, please notify the organizers and we will try to update this page.

1. How to define an unexpected change in meaning?
2. How to annotate lexicalized words which belong to contractions and compounds?
3. How to annotate coordinated VMWEs sharing some components?
4. How to annotate elliptical occurrences of VMWEs?
5. How to annotate VMWEs that seem to belong to more than one category?
6. How to annotate embedded VMWEs?
7. Are existential expressions with there is/are considered VMWEs?
8. How to categorize VMWEs which seem LVCs but do not pass all LVC tests?
9. Why are verb+noun constructions with pure operator verbs (to commit, to make, to have etc.) considered LVCs?
10. Does the IRefV category include verbs with non-reflexive clitics?
11. Should nominalizations of VMWEs be annotated?
12. How to express hesitation between different VMWE categories?
13. In test 9, how can one decide whether an abstract noun is an event or a state?
14. How does one decide if a more or less frozen determiner is a lexicalized VMWE component?
15. Should I annotate compound and serial verbs as VMWEs? Of which category?
16. If an LVC contains a complex (fixed) NP as a dependent, should I include the whole NP or just the head?

1. How to define an unexpected change in meaning?

Check the glossary entry that defines unexpected change in meaning

2. How to annotate lexicalized words which belong to contractions and compounds?

In some languages prepositions, clitics and determiners are subject to contractions (i.e. they yield multilword tokens, MWTs). Tokenizers might not handle contraction splitting properly. In this case, a lexicalized component of a VMWE can be merged with an external word:

- (DE) n.a.
- (ES) haberse suicidado lit. have+REFL suicided ‘committed suicide’
- (PL) n.a.
- (SL) n.a.

A similar problem occurs in languages with productive compounding, where a lexicalized component of a VMWE and a free modifier can build up a multistoken word (since compound splitting might not be a standard feature of a tokenizer):

- (DE) unter Drogen einfluss stehen ‘to be under the influence of drugs’
  Heiss Hunger haben lit. to have hot hunger ‘to be ravenously hungry’
- (SL) n.a.

Since the current annotation format is token-based, we prohibit correcting tokenization errors and compound splitting by the annotators for the sake of coherence. Therefore the annotation of such contractions and compounds finds no fully satisfactory solution in our schema. We propose to annotate a whole MWT each time it contains a word which is part of a VMWE. Annotators should add a textual comment about the mixed status of this MWT:

- (DE) Drogen einfluss → MWT containing a lexicalized VMWE component Einfluss and an external word Drogen
  Heiss Hunger → MWT containing a lexicalized VMWE Hunger and an additional modifier heiss
- (ES) haberse → MWT containing a lexicalized VMWE component se and an external word haber
- (PL) n.a.

3. How to annotate coordinated VMWEs sharing some components?

A component shared by two or more coordinated VMWEs should be annotated as belonging to both of them.

- (DE) Regeln und Richtlinien aufstellen lit. to set up rules and guidelines ‘to draw up rules and guidelines’ → aufstellen must be annotated both as part of to Regeln aufstellen ‘to lay down rules’ and of Richtlinien aufstellen ‘to draw up guidelines’
- (EN) to have a walk or a ride → have must be annotated both as part of to have a walk and of to have a ride
- (PL) odprowadzić mszę i pokutę lit. celebrate a mass and a penance → odprowadzić should be annotated both as part of odprowadzić mszę lit. to celebrate a mass and of odprowadzić pokutę lit. to celebrate a penance
- (SL) imeti dober želodec in dobre živce lit. to have a good stomach ‘to bear something well’ lit. and good nerves ‘to be mentally strong’ → imeti ‘have’ must be annotated both as part of imeti dober želodec and of imeti dobre živce

4. How to annotate elliptical occurrences of VMWEs?

Instances of a VMWE in which all but one lexicalized component were omitted or pronominalized should not be annotated. This concerns in particular the cases where a nominal component is concerned by anaphora. For instance, in this decision was hard but he took it, we should not annotate take and decision or it as an instance of a VMWE. We annotate only the transformations in which the syntactic dependency link between the head verb and the lexicalized complement is preserved, e.g. the decision which he took.

5. How to annotate VMWEs that seem to belong to more than one category?

Such hesitation issues should normally be solved by the decision trees 1 and 2. For instance, consider the German expression sich eine Frage stellen lit. SELF a question put ‘to doubt’. It may seem to belong to both IRefV, since sich is required only if stellen co-occurs with Frage, and LVC, since Frage keeps its original meaning and stellen brings no additional meaning. However, test 7 [1DEP] indicates that an expression like this should be annotated as ID, since the verb has more than one lexicalized syntactic dependent.

Similarly, the French expression avoir peur lit. have fear ‘to be afraid’ seems to have features of an ID. Unlike most LVCs, it does not allow a determiner *avoir une peur lit. have a fear, except when the noun is modified avoir une grande peur lit. have a great fear. However, test 8 [CATEG] in decision tree 2, and the LVC-
6. How to annotate embedded VMWEs?

Candidate VMWEs embedded in other VMWEs should be annotated only if they have a VMWE status also outside the particular context. For instance, the VMWE to let the cat out of the bag should be annotated as ID, and its embedded VMWE to let out as a VPC.

On the other hand, the French expression se faire des idées lit. SELF make DET.PL ideas 'to imagine things which are not true', se faire should not be annotated as IReflV, since it is not inherently reflexive as a standalone verb+clitic combination.

7. Are existential expressions with there is/are considered VMWEs?

Hesitations about a possible LVC status can arise with respect to existential constructions with nouns introducing events or properties (see test 9 [N-EVENT]) as in:

- (DE) es aigl Beschwerden *there are complaints*
- (EN) there are complaints
- (FR) il existe des plaintes lit. if there has complaints *there are complaints*
- [PL] n.a.
- (PT) ha queixas lit. has complaints 'there are complaints'
- (SL) imepi priprombë lit. have complaints 'there are complaints'

Namely, the noun keeps its original sense and the existential verb to be or to have brings no additional meaning. However, a candidate LVC must also pass test 12 [V-REDC]. This requires the modification of the noun by the verb's subject, which is impossible with impersonal and empty subjects like there. Therefore, such candidates cannot be LVCs.

Note, however, that existential expressions themselves can be VMWEs of type ID. For instance, in the French example il y a des plaintes lit. if there has complaints *there are complaints*, two dependents of the verb a lit. has are lexicalized: if lit. it and y lit. there, therefore it is an ID (see test 7 [1DEP]).

8. How to categorize VMWEs which seem LVCs but do not pass all LVC tests?

If at least one of the five LVC tests (9 to 13) is not passed, the candidate is not considered an LVC. For the sake of a deterministic VMWE categorization and higher inter-annotator agreement, we admit a definition of an LVC which might seem more restrictive than some linguistic studies usually assume. Thus, we exclude from the LVC scope:

- expressions in which the verb's syntactic subject is not necessarily the noun's semantic subject, like to give courage or to make an impression. These candidates do not pass test 12 [V-REDC].
- expressions where the lexicalized nominal dependent of the verb is its subject, as in the problem lies in something; these candidates do not pass test 12 [V-REDC].
- expressions with aspectual verbs, as in to start, to pursue, to stop a walk. These do not pass test 11 [V-LIGHT] since they add (aspectual) semantics to the noun. The only exception is when the noun itself is already aspectual, as in to come into bloom.

9. Why are verb+noun constructions with pure operator verbs (to commit, to make, to have etc.) considered LVCs?

Pure operator verbs, i.e. such verbs which never have any semantics per se but only carry the grammatical (tense, mood etc.) information, seem to contradict the intuition behind a VMWE. Namely, they usually select a whole semantic class of nouns. For instance to commit selects any negative act (a crime, a suicide, a theft) and to perform selects any activity (a task, an experiment, a miracle). In this sense, their complements resemble open slots and the whole combinations resemble collocations. However, for the sake of a deterministic VMWE categorization and higher inter-annotator agreement, we do include verb+noun combinations with pure operator verbs, such as to commit a crime and to perform a task, into the LVC category. This is because such combinations pass all 5 LVC-specific test (9 through 13).

We could have organized decision tree 1 differently and exclude such cases from the VMWE scope by eliminating the LVC hypothesis. Then, to commit a crime and to perform a task would pass none of the tests from 1 to 5 and would be eliminated. However, we would also have to eliminate prototypical LVCs like to make a decision (it passes none of the tests from 1 to 5 either), which we do wish to take in as an LVC.

10. Does the IReflV category include verbs with non-reflexive clitics?

No, the IReflV category only includes (some) combinations of a head verb with a reflexive clitic. As indicated in the borderline cases page of IReflV category, other pronouns, whenever lexicalized, trigger the ID category. Recall that whenever more than one dependent of the verb is lexicalized (including or not a reflexive clitic), the VMWE is always categorized as an ID

- (DE) sich Fragen stellen lit. SELF questions put 'to doubt'
- (FR) s’en aller lit. SELF of-there go 'to leave'
- (PL) n.a.
- (SL) ucvreti jo lit. to escape her 'to escape something/someone by running'

11. Should nominalizations of VMWEs be annotated?

specific decision tree indicate that it belongs to the LVC category.
The only nominal VMWE variants within our annotation scope are those:

- headed by the gerund stemming from the head verb of the VMWE - taking of the decision, and
- in which a noun stemming from a VMWE is modified by a participle or a relative clause headed by the verb stemming from the same VMWE - the decisions taken yesterday, the decision which he took.

Other nominalizations are excluded:

- (DE) Wortbruch lit. word-break 'a promise which has not been hold'
- (EN) a break-down, a forget-me-not
- (FR) la prise en compte lit. the taking into account 'the fact of taking something into account'. peut-être lit. may-be. porte-feuilles lit. carry-sheets 'wallet'
- (PL) zabawa czyimś kosztem 'a play at someone else's expenses' derived from bawić się czyimś kosztem 'to enjoy oneself at someone else's expenses'
- (SL) šala na tuj račun lit. a joke at someone else's expenses derived from šaliti se na tuj račun 'to play a joke on someone'

For practical reasons (e.g. compatibility with an existing annotation, or usefulness for a particular application) they can be considered language-specific VMWEs but then a new category should be defined for them, so as to keep the universal and the quasi-universal categories intact.

12. How to express hesitation between different VMWE categories?

Once identified in a text, each VMWE is to be assigned to exactly one category. Note that in this version of the guidelines we no longer admit "hesitation labels" (e.g. LVC/ID) used in the pilot annotation. Hesitation can, however, be expressed in a comment and a particular value of the annotator's confidence assigned to a particular VMWE occurrence.

13. In test 9, how can one decide whether an abstract noun is an event or a state?

The goal of test 9 is to identify whether a noun is predicative, that is, whether it requires some semantic arguments. We talk about events and states to circumvent the question of whether a noun is predicative. Here, they are understood very largely as roughly corresponding to binary and unary predicates. For instance, we consider that an event is something that happens, and can be related to an action, activity, process or phenomenon. A state is understood as a property that may or may not change over time, including feelings, sensations, permanent and temporary properties and relations between entities. These are a very generic definitions that go far beyond the scope of what is commonly understood as an event or state.

While it is hard to define required tests to identify a predicative noun, there are some useful clues that can be used for abstract nouns (sufficient criteria).

Verb paraphrase: Is the abstract noun derivationally related to a verb with the same semantics?

- (EN) John makes a decision = John decides
  John has a walk = John walks

Adjective paraphrase: Is the abstract noun derivationally related to an adjective with the same semantics?

- (EN) John has courage = John is courageous → and, more generally, characteristics and attributes
  John has hunger/thirst = John is hungry/thirsty → and, more generally, physical sensations
  John has passion/fear/anger = John is passionate/afraid/angry → and, more generally, feelings and emotions
  John has problems/difficulties = Something is problematic/difficult for John → and, more generally, states

Synonym verb or adjective paraphrase: Does the abstract noun have a synonym/hypernym derivationally related to a verb or adjective with the same semantics?

- (EN) John and Mary reach a consensus = John and Mary agree → consensus has no corresponding verb or adjective, but agreement is a synonym
  John has a chance to do something = John is likely to do something → chance has no corresponding verb or adjective, but likelihood is a synonym

For many classes of abstract nouns, it can be tricky to apply the tests above. We advise listing in a separate document those classes of nouns that pass test 9 in your language. We suggest considering that the following categories pass test 9:

- (EN) Illnesses, symptoms and health conditions:
  John has a flu = John is ill (illness is a hypernym of flu)
  Relations:
  John has contact with somebody = John contacts somebody
  John has an affair with somebody = John is involved with somebody (involvement is a synonym of affair)
  Mental content (internal to a cognizer):
  John has a worry = John worries
  John has an idea = John thinks (thought is a synonym of idea)
  John has an opinion = John believes (belief is a synonym of opinion)

Please notice that events and states that have no semantic arguments do not pass test 9, even if they have verbal/adjectival paraphrases:

- (EN) Natural phenomena: rain, snow, tornado, flood, earthquake

Informational content (external to a cognizer): information, news

Finally, notice that not any verb + predicative noun combination forms an LVC. Additionally, the verb needs to be "light", not adding semantics to the noun. The remaining LVC tests (tests 10 to 13) guarantee this.

14. How does one decide if a more or less frozen determiner is a lexicalized VMWE component?
Most of the time, it is easy to test whether a determiner is lexicalized by searching alternatives in corpora (or on the web). For instance, the is lexicalized in to kick the bucket because searches for other determiners (this, a, some, three, many, etc.) either do not return any result or return only literal uses of this verb phrase.

However, borderline cases do exist, in which alternatives are rare but possible, specially for LVCs and decomposable IDs. For instance, while the standard form of the idiom spill the beans forbids some determiners (it spill these/many/any beans), it is possible to find some variation (spill these/more/many beans).

We argue that the selection of some determiners (but not all) by a VMWE is comparable to selected prepositions for verbs. Thus, it can be seen as a regular grammatical phenomenon, suggesting that when the determiner varies, then it should not be included. In some VMWEs, though, determiner variation may be considered as marginal and/or incorrect, which means that it should be included in the scope of the annotated VMWE.

In short, determiners can exhibit limited variability. As a consequence, each language should document their decisions as to whether to include them or not for particular VMWE classes, to ensure consistency.

- (FR) avoir la pêche lit. have the peach ‘to have much energy’
- avoir de la chance lit. have some luck ‘to be lucky’
- avoir l’occasion ‘to have the opportunity’

After annotation, we suggest that LLs use the provided analysis scripts to detect inconsistencies in the annotation of the same VMWE (e.g., including or not a determiner). They can then take an arbitrary decision and homogenise all annotated occurrences.

### 15. Should I annotate compound and serial verbs as VMWEs? Of which category?

It depends. Most of the languages covered by the shared task for the moment do not have this kind of verb. The guidelines were written having these languages in mind, so they are not clear about compound verbs.

In many Indo-European languages (including Germanic, Romance and Balto-Slavic families), verbal chains using auxiliary and modal verbs are used to express tense, modality and aspect. This is a regular linguistic phenomenon that can be applied to any verb and should not be annotated.

On the other hand, some languages like Maltese have many compound verbs that do not necessarily express tense, mood and modality. We suggest that, when the verb combinations regularly combine with any other verb adding a given meaning, they should not be annotated. Future versions of these guidelines should study the need for a new category for compound verbs, in order to cover this phenomenon.

In short, verbal chains should only be annotated as ID when they are idiomatic:

- (FR) laisser tomber lit. let fall ‘to give up’
- vouloir dire lit. want say ‘to mean’
- faire tomber lit. make fall ‘to drop’
- vouloir changer lit. want change ‘to want to change’
- (MT) dak x’mar jghid itibern li. that (person) what’he-went he-says yesterday ‘what the hell did he say yesterday’
- (PT) querer dizer lit. want say ‘to mean’
- querer falar lit. want speak ‘to want to speak’

### 16. If an LVC contains a complex (fixed) NP as a dependent, should I include the whole NP or just the head?

The guidelines determine that only lexicalized components should be annotated. Therefore, we suggest that, in such cases, if the NP is compositional, only the head of the NP is included in the scope of the LVC. This may lead to the annotation of odd LVCs that actually never occur by themselves without a modifier. This is not a problem and is already the case for other VMWEs, e.g. the ones that only occur with a determiner, but the determiner is not lexicalized. The only cases where the NP should be included as a whole is if the complement is a non-compositional MWE, so that it would not make any sense to annotate only the head.

- (EL) το παιχνίδι του ευρωσκεπτικισμού lit. to-play the paper the.SG.GEN euroscepticism.SG.GEN ‘to use the asset of euroscepticism, to use euro-scepticism as an asset’
- κάνω στάση εργασίας lit. to make stop work.SG.GEN ‘to go on strike, to strike ‘→ the expression στάση εργασίας is non-compositional (term)
- (FR) présenter un Syndrome Coronaires Aigu ‘to present an acute coronary syndrome’
- mener une vie de débauche ‘to have a life of pleasures’
- faire un faux pas lit. make a false step ‘to commit a faux pas’ → the expression faux pas is non-compositional
- (PL) mieć wyrzuty sumienia lit. to have reproaches of the conscience ‘to feel guilty’
- (PT) fazer roleta russa lit. to make russian roulette ‘to play russian roulette’ → the expression roleta russa is non-compositional
- ter uma situação financeira/profissional/estável ‘to have a financial/professional/stable situation’

Notice that these suggestions also apply to LVCs whose nominal complements are introduced by prepositions (i.e. verb+PP LVCs). As usual, the preposition should be included if it is lexicalized and then the NP introduced by the preposition is analyzed exactly as described above.

If the complex dependent is an acronym, you may want to add the textual comment "PART" to indicate that only part of the full version is lexicalized (generally, the head), just like for contractions and compounds.

### Section 8.2

**Adding new examples in your language**

It is often useful to have examples of a phenomenon shown in your own language. We collect these examples for each language using an online shared spreadsheet, and we present these examples as in the template below:

- (BG) MWEs with their lexicalized components in Bulgarian are indicated like this.
Examples with tags

If you have not done it yet, open the spreadsheet and look for the entry 8.2_A_template-mwe. Let us analyse the English example (look for EN in the fourth column). The fifth column should read as follows:

MWEs with <lex>their lexicalized components</lex> in English are indicated like this.

As you can see, this is exactly the same text that was shown in the template above, except that the lexicalized components are surrounded by the tags <lex> and </lex>. When writing an example, you will often have to use XML tags. We describe below the most important ones.

Bold: you should surround lexicalized components with the tags <lex> and </lex>. For example, consider the code "He will <lex>take</lex> a <lex>shower</lex>". This code is presented as follows:

- (EN) He will take a shower

Red: By default, all examples are typeset using the language's color. Sometimes, examples contain counter-examples, that is, something that looks like a VMWE but that should not be annotated. The <nmwe> and </nmwe> tags can be used to represent these non-MWEs, which will be shown in red. For example, the code "This is not an MWE" yields the following:

- (EN) This is not an MWE

Underlining: Some examples use underlining to focus on some of the words. This can be done with the tags <u> and </u>. For example, the code "This is <u>not</u> an MWE" yields the following:

- (EN) This is not an MWE

Gloss icon: You should also provide English glosses and translation for your examples. Glosses and translations should always be provided in English, and never in another language. Glosses must be surrounded by the tags <gl> and </gl>. Translations must be surrounded by <trans> and </trans>. English examples can also use the tag <trans> to indicate the meaning of an idiomatic expression. For example, the code "<lex>defend one's beefsteak</lex>" yields the following:

- (EN) This is not an MWE
Section 8.3

Annotation platform FLAT

The annotation will be performed using the online annotation platform FLAT.

- The FLAT user manual for the PARSEME shared task
- Link to the PARSEME shared task FLAT platform

Section 9

Glossary

Cranberry word

A cranberry word is a token that does not have the status of a stand-alone word, has no proper distribution, and no stand-alone meaning, but it may have a syntactic category and an inflection paradigm. It only occurs in a particular expression (or a closed list of expressions) and can never be found in different contexts, as the underlined words below:

- (DE) jemandem Angst einjagen lit. to-someone chase-in fear 'to frighten someone'
  jemanden einen Besuch abstatten
- (FR) se mettre martel en tête lit. SELF put a hammer in head 'to worry a lot'
- (PL) odsadzić kogoś od czci i wiary lit. to refuse honor and faith to someone 'to drag sb's name through the mire/mud, to damage someone's reputation by saying insulting things about them'
- (SL) biti si kril 'owe nothing to somebody; each party got what it deserved/asked for'

Candidate VMWE

A candidate VMWE is group of tokens that seems to have some idiosyncrasy of the type listed in the MWE definition. However, further tests are required to decide whether it is to be annotated as a true VMWE or, instead, it was a false alarm. The lexicalized elements of candidate VMWEs are highlighted in **bold**.

Syntactic operator

A syntactic operator is a verb that only bears the grammatical features (person, number, tense and mood) but adds no semantics to the complement. This definition is more restricted than the traditional notion of a light verb. Notably, aspectual light verbs (which adds aspectual semantics to the complement), as in to *start a walk, to give courage*, are not considered operators. Operators are typical head verbs of light-verb constructions:

- (DE) eine Entscheidung treffen 'to make a decision'
  Angst haben 'to have fear'
  ein Verbrechen begehen 'to commit a crime'
- (EN) to make a decision
  to have fear
  to commit a crime
- (PL) oddać hold lit. to give-back tribute 'to pay tribute'
- (SL) priti v poštev lit. to come into consideration 'to consider'

Collocation

A collocation is a word co-occurrence whose idiosyncrasy is of statistical nature only. Collocations are not considered VMWEs in this task:

- (DE) eine Anfrage beantworten 'to answer a request', das Diagramme zeigt 'the diagram shows', mit einem Bus fahren 'to take a bus'
- (EN) the graphic shows
  drastically drop
- (PL) załać rynek lit. to flood the market 'to dominate the market'
- (SL) občutno zmanjšati 'significantly reduce',
  drastično zmanjšati 'drastically reduce'

Canonical form
The canonical form of a candidate VMWE is a prototypical verbal phrase preserving the same meaning.

- **(DE)** the canonical form of *das Herz welches er bricht* 'the heart which he breaks' is *er bricht ihr das Herz* 'he breaks her heart'
  - the canonical form of *Wortbruch* lit. word-break 'a promise which has not been kept' is *Wort brechen* lit. to break the word 'not to hold a promise'
- **(EN)** the canonical form of *the heart which he broke is he broke (her) heart*
  - the canonical form of *making an impression on him is (she) makes an impression on him*
- **(PL)** the canonical form of *decyzyja. Mieć podjąć 'decisions which he took' is podjąć decyzyj 'he took a decision'
  - the canonical form of *odłóżcie, ki jo je sprejel 'decisions which he took is sprejeli odločitev 'he took a decision'*

**Reflexive clitics**

Reflexive clitics are a special type of object pronoun that refers to the subject of the verb. See the guidelines of IReflV category for more details. In English, the reflexive is expressed as a suffix -self appended to object pronouns. However, many languages have special reflexive pronouns, which are a relatively small closed class of words:

- **(DE)** mich, dich, sich, uns, euch
- **(FR)** me, te, se, nous, vous
- **(IT)** mi, ti, si, ci, vi
- **(PL)** się, sobie
- **(PT)** me, te, se, nos, vos
- **(SL)** se, si

**Particles**

Particles are hard to distinguish from homographic prepositions:

- **(DE)** ich schlage vor allen zu verzeihen 'I propose to forgive everyone'
  - *ich schlage vor allen Dingen die Sahne* 'I mix prior to anything the cream'
- **(EN)** to get up a petition
  - *to get up a hill*
- **(PL)** n.a.
- **(PT)** n.a.
- **(SL)** *sem za njen predlog* 'I support her proposal'
  - *sem za hišo* 'I am standing behind the house'

The fundamental property to capture is that a preposition governs a prepositional group, while a particle functions as an adverbial. In some languages particles can also be homographic with verbal prefixes:

- **(DE)** das Schäd umfahren 'to drive over the sign'
  - *den See umfahren* 'to drive around the lake'

Most tests discriminating particles from prepositions and prefixes are language-specific and should be proposed by the individual language team. See the guidelines on particles for more details.

**Unexpected change in meaning**

An unexpected change in meaning, signaled by the # (hash) sign, is a phenomenon referred to in generic and category-specific tests, based on the notion of inflexibility. Inflexibility is verified by attempting a regular modification which yields an unexpected acceptability or meaning shift, that is, beyond what would be expected by the initial modification. In order to judge whether a shift in acceptability or meaning is unexpected, one can try to apply the same modification to a similar compositional construction, using analogy. For example, book and word have synonyms including notebook/novel/volume/publication and term/expression/headword, respectively. However, while the slight shift in the meaning of book is compositionally reflected in:

  - **(EN)** *I give you my book* → *I give you my notebook/novel/volume/publication*
- **(PL)** *daję ci książkę* 'I give you a book' → *daję Ci zeszyt/powieść/tom/publikację* 'I give you a notebook/novel/volume/publication'
  - **(SL)** *daj ti knjigo* 'I give you a book' → *dam ti zvezek/roman/publikacijo* 'I give you a notebook/novel/publication'

the same does not hold for:

- **(DE)** *Ich gebe Dir mein Wort* 'I give you my word, i.e. I promise' → *Ich gebe Dir mein(e) Publikation/Doktorarbeit/Kapitel/Novelle/Ausgabe* 'I give you my publication/thesis/chapter/novel/edition'
  - **(EN)** *I give you my word* → *I give you my notebook/novel/volume/publication*
- **(PL)** *daję ci słowo* lit. *I give you a word* → *daję Ci wyraz/sylabę/czasownik* 'I give you a word/syllable/verb'
  - **(SL)** *daj ti besedo* lit. *I give you a word* → *dam ti izraz/zlog/glago* 'I give you a word/syllable/verb'

I.e. the latter replacement produces an unexpected change of meaning that goes beyond the semantic difference between the original and the replaced word. Thus, Test 2 [LEX] applies and:

- **(DE)** *jm. sein Wort geben 'to give one's word to s.o.*'
  - **(EN)** *to give one's word to someone*
- **(PL)** *dać komuś słowo* lit. *to give someone a word* 'I give one's word to someone'
  - **(SL)** n.a.

is a VMWE.

Similarly, Test 22 [V+PART-DIFF-SENSE] refers to an unexpected change in meaning of the verb stemming from the addition of the particle. We do so by checking if the situation described by the verb with the particle implies the one described without the particle:

- **(DE)** *ich fange das Buch an 'I begin to read the book' does not imply ich fange das Buch 'I catch the book'*
  - *Ich lege das Buch auf den Tisch* ab 'I put down the book on the table' implies *Ich lege das Buch auf den Tisch* 'I put the book on the table'*