Annotation guidelines

P A R S E M E shared task on automatic identification of verbal MWEs - edition 1.0 (2017)

New edition of the shared task: these are the guidelines for edition 1.0 (2017). For the most up-to-date version, please check the guidelines for edition 1.1 (2018).

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

Here, you'll find detailed definitons, 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)

Marie Candito, Fabienne Cap, Silvio Cordeiro, Vassiliki Foufi, Polona Gantar, Voula Giouli, Carlos Herrero, Mihaela Ionescu, Verginica Mititelu, Johanna Monti, Joakim Nivre, Mihaela Onofrei, Carla Parra Escartín, Manfred Sailer, Carlos Ramisch, Monica-Mihaela Rizea, Agata Savary, Ivelina Stonayova, Sara Stymne, Veronika Vincze.

Table of contents

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

- An asterisk (*) precedes ungrammatical examples.
- A hash (#) precedes examples where a standard modification yields unexpected meaning shifts with respect to the original expression.
- Different colors are used to display examples:
  - *Red* is used for counter-examples, that is, expressions which look like VMWEs but are not one, whatever the language.
  - According to the language, different colors are used for other examples, that is, positive examples of the phenomenon being discussed:
    - *Shades of green* are used for positive examples in Germanic languages.
    - *Shades of blue* are used for positive examples in Romance languages.
    - *Shades of orange* are used for positive examples in Slavic languages.
    - *Shades of pink* are used for positive examples in other language families.
- Examples are preceded by the 2-letter language code in parentheses
- Examples can be shown and hidden using the toggle buttons in the header.

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، دوست،
  - (FR) *faire, une, promenade, étonnement*
  - (HR) *napraviti/činiti, jedan, šetnja, začuđenost*
  - (HU) *tesz, egy, séta, meglepetés*
  - (IT) *fare, una, passeggiata, sorpresa*
  - (MT) *ferħ, libes, sabiħ*
  - (PL) *robić* 'to do', *na* 'on', *dokładność* 'precision'
  - (PT) *dar, uma, caminhada, supresa*
  - (RO) *face, o, plimbare*
  - (SV) *gå, på, promenad, förväning*
  - (SL) *iti, na, en, sprehod, začudenost*
- 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):
  Note: The pipe symbol '|' indicates token separation in these examples
  - (BG) *m|.|н|.* 'etc.'
    год|. 'year'
  - (DE) *z|.|B|.* 'for instance'
    Wie geht|'|s lit. *How goes it* 'How are you'
  - (EL) *κ.* lit. *κύριος* 'Mister'
    υπΔρ lit. *υποψήφιος διδάκτορας* ' PhD candidate'
  - (EN) *M|.* 'Mister'
    pp|. 'pages'
    Pandora|'|s
  - (ES) *a|. |C|.* lit. *antes de Cristo* 'before Christ'
    p|. |ej|. lit. *por ejemplo* 'for instance'
    Rte|. lit. *remitente* 'sender'
  - (FA) می|-|اروم، آیت|-|الله، کتاب|-|ها
  - (FR) *aujourd|'|hui* 'today'
  - (HR) *danas* 'today'
  - (HU) *időjárás|-|jelentés* 'weather forecast'
  - (IT) *vice|-|presidente* 'vice-president'
  - (MT) *libs|et* ' she wore '
  - (PL) *Chomsky|'|ego* 'of Chomsky'
    SMS|-|ować 'to write an SMS'
  - (PT) *vice|-|presidente* 'vice-president'
  - (RO) *prim|-|ministru* 'prime minister'
    d|-|voastră 'polite "you"'
  - (SV) *EU|:|s* 'EU's'
  - (SL) *g|.* 'Mister'
    str|. 'pages'
    le|-|to
- One token can contain several words, like in contractions and compounds. In this case we speak of a **multiword token** (MWT):
  Note: See also the representation of MWTs in Universal Dependencies

- (BG) *вагон-ресторант* lit. *train carriage+restaurant* 'train buffet'
- (DE) *Schulaufgabe = Schule+Aufgabe* lit. *school+exercise*'homework'
     Apfelbaum = Apfel+Baum lit. *apple tree*'apple tree'
- (EL) *στου = σε+του* lit. *at+the.GEN*
     στον = σε+τον lit. *at+the.ACC*
- (EN) *don't = do+not*
- (ES) *del = de+el* 'from the'
     pelirrojo = pelo + rojo lit. *hair+red* 'red-haired'
- (FA) کتابش=کتاب+ش
- (FR) *du = de+le* 'from the'
- (IT) *della = de+la* 'of the'
- (PL) *Białymstoku=Białym+stoku* lit. *white+slope* 'Białystok.INST (a city name)'
     robiłem=robi+łem lit. *do.3.SG.PRES+be.1.SG.PAST.AGL*'I did'
     żeśmy = że+śmy lit. *that+be.1.PL.AGL* 'that-we'
- (PT) *neles = em+eles* 'on them'
- (RO) *într-o = într-+o* 'in a'
- (SV) *arvsmassa = arv+massa* 'genetic stock'
- (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) *вземам решение* 'make a decision' (2 words, 2 tokens)
         прочитам от корица до корица 'to read from cover to cover' (5 words, 5 tokens)
    - (DE) *eine Rede halten* (2 words, 2 tokens) lit. *a speech hold* ' to give a speech'
         wie geht's (2 words, 4 tokens) lit. *how goes it* 'how are you'
    - (EL) *δίνω τον λόγο μου* (3 words, 3 tokens) lit. *give the speech* 'to promise'
         παίζω στα δάχτυλα (3 words, possibly 4 tokens) lit. *play in-the fingers* 'know very well '
    - (EN) *to take a walk* (2 words, 2 tokens)
         to open Pandora's box (3 words, possibly 5 tokens)
    - (ES) *dar un paseo* 2 words, 2 tokens lit. *to give a walk* 'to take a walk'
         dar por sentado 3 words, 3 tokens lit. *to give for seated* 'to take for granted' *irse de rositas* 3 words, 4 tokens lit. *to go_self of little_roses* 'to get off scot free'
    - (FA) 2) داد دستور words, 2 tokens)
    - (HR) *napraviti šetnju* (2 words, 2 tokens)
         otvoriti Pandorinu kutiju(3 words, 3 tokens)
    - (HU) *sétát tesz* 'to take a walk' (2 words, 2 tokens)
    - (IT) *tenere un discorso* (2 words, 2 tokens) lit. *hold a speech* ' to give a speech'
         cavalcare l'onda (3 words, 4 tokens) lit. *ride the wave* 'ride the wave'
    - (MT) *kien idur fuq il-fatt* 'turns on the fact'
    - (PL) *robi z igły widły* lit. *make.3.SG a pitchfork out of a needle* 'he makes a mountain out of a molehill' (4 words, 4 tokens)
         robiłem z igły widły lit. *made.3.SG.M1+be.1.SG.AGL a pitchfork out of a needle* 'I made a mountain out of a molehill' (4 words, 5 tokens)
    - (PT) *dar uma caminhada* 'to give a walk' (2 words, 2 tokens)
         cair de pára-quedas lit. *to fall with parachute* 'to arrive unprepared in the middle of a situation' (3 words, possibly 5 tokens) ⓘ
         queixar-se-ia lit. *complain-self-would* 'would complain' (2 words, possibly 5 tokens)
    - (RO) *a da ortul popii* 'to die' (3 words, 3 tokens)
    - (SV) *hålla ett tal* (2 words, 2 tokens) lit. *hold a speech* ' to give a speech'
    - (SL) *klicati jelene* lit. *to call cerfs* 'to vomit' (2 words, 2 tokens)
         vreči puško v koruzo lit. *throw a rifle in the corn* 'to give up' (4 words, 4 tokens)
- A VMWE contains one (multiword) token:
    - (BG) *no example found for Bulgarian*
    - (DE) *vorbereiten* lit. *to pre-arrange* 'to prepare'
         anfangen lit. *at-catch* 'to begin'
    - (EL) *έδωσα-πήρα* lit. *gave-1SG took-1SG* 'to manage '
    - (EN) *to pretty-print*
    - (ES) *suicidarse* lit. *suicide_self* 'to commit suicide'
    - (FR) *court-circuiter* 'to short circuit'
    - (HR) *pripremiti* lit. *unaprijed napraviti/ urediti* 'to prepare'
    - (HU) *kinyír* lit. *out.cut* 'to kill'
    - (IT) *corto-circuitare* 'to short circuit' *suicidarsi* lit. *suicide_self* 'to commit suicide'
    - (PL) *no example found for Polish*
    - (PT) *queixar-se-ia* lit. *compain-SELF-would* 'would complain'
    - (RO) *a se-ndura* lit. *RCLI.ACC-have.the.heart* 'to have the heart'
    - (SV) *klargöra* lit. *clear-make* 'clarify' *påpeka* lit. *on-point* 'point out'
    - (SL) *pripraviti* lit. *to pre-arrange* 'to prepare'

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

- (BG) *no example found for Bulgarian*
- (DE) *??*
- (EL) n.a.
- (ES) *maldecir* lit. *bad say* 'curse' *bienvivir* lit. *well live* 'to live in comfort'
- (IT) *ricaricare* 'to recharge'
- (PL) *SMS-ować* 'to write an SMS'
- (PT) *pós-datar* 'to post-date'
- (RO) *a re-mpărți*lit. *PREFIX-split*'to split again'with the aphaeresis of the sound 'î' in rapid speech; this is one word, multitoken
- (SL) *SMS-jati* 'to write an SMS'

Whenever the distinction between a word and a token is judged by a particular language team as hard to tackle, a possible option is to consider these two notions equivalent for the needs of this shared task.

Section 1.3

Verbal Multiword expressions

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 ball* 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) *пиля нечии нерви* lit. *scrape someone's nerves* 'to annoy someone'
       вземам трудно решение lit. *make a difficult decision* 'to make a difficult decision'
     - (DE) *traf eine Entscheidung* 'made a decision'
       nahm sich das zu Herzen 'took this to heart'
     - (EL) *παίρνω μία δύσκολη απόφαση* lit. *take-1SG a-FE.SG.AC difficult-FE.SG.AC decision-FE.SG.AC* ' to take a difficult decision'
       παίρνω τα μέτρα μου lit. *take-1.SG the-NE.PL.AC measures-NE.PL.AC my-1.SG.GE.POSS* ' to take precautions'
       γράφω κάποιον στα παλιά μου τα παπούτσια lit. *write-1.SG someone to-the-NE.PL.AC old-NE.PL.AC my-1.SG.GE.POSS shoes-NE.PL.AC* ' to ignore someone'
     - (EN) *made a decision*
       break her heart
       took this to heart
       could take this to heart
       would have been making a decision
       could have made a different decision
     - (ES) *tomó una decisión* lit. *took.he/she a decision* 'he/she made a decision'
       le hubiera roto el corazón lit. *him/her would_have broken.he/she the heart* 'he/she would have broken his/her heart'
     - (FR) *a eu du courage* lit. *has had courage* 'had courage'
     - (HU) *sétát tesz* 'to take a walk'
     - (IT) *prendere una decisione* lit. *make a decision* 'take a decision
       spezzare il cuore lit. *break the heart* 'break the heart'
       prendere a cuore
       lit. *take to heart* 'take to heart'
       '
     - (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'
       a trece ceva sub tăcere lit. *to keep something under silence.ACC* 'to keep quiet about something'
     - (SV) *fattade ett belsut* 'made a decision'
     - (SL) *sprejeti odločitev* 'to make a decision', *zlomiti komu srce* lit. *to break someone's heart* 'to upset someone by letting them know that you do not love them', *vzeti si k srcu* lit. *take something to heart* 'to think about something seriously'
   - Partly saturated, where the lexicalized arguments include the subject:
     - (BG) *излиза ми име* lit. *appears for-me.DAT name* 'a name sticks for/to me'
       чашата на търпението ми прелива lit. *glass.DET of patience my.POS overflows* 'my patience runs out'
     - (DE) *ein Vöglein hat mir gezwitschert* lit. *a little bird has to me twittered* 'a little bird told me'
     - (EL) *μου έφυγε ο τάκος* lit. *me.GEN left the chosk* 'I was very tired'
     - (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 komuś zrzedła* 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'

- (RO) *a mustra cugetul (pe cineva)* lit. *to chide consciousness-the (PE_Acc somebody)* 'the consciousness chides (PE somebody)'
- (SL) *srce pade v hlače* komu lit. *(someone's) hart drops into the pants* 'one hasn't enough courage to do somethiing', *sekira pade v med* komu
  - Partly saturated, where lexicalized head verbs are coordinated:
    - (BG) *цъфна и върже* 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) *απορώ και εξίσταμαι* lit. *wonder1.SG and be-amazed1.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) *pluł i łapał* lit. *he spit and catched* '(he) was lazy, (he) did nothing useful'
    - (PT) *pintar e bordar* lit. *paint and knit* 'to abuse'
    - (RO) *a tunat și i-a adunat*lit. *it.has thundered and CL.ACC-it.has gathered*'birds of a feather flock together'
      *seamănă, dar nu răsare*lit. *sow.3SG (homonym of resemble), but not sprout.3SG*'not to resemble'
    - (SL) *živi in pusti živeti* lit. *to live and let live* 'to live and let live'
  - Fully saturated:
    - (BG) *пиле не може да прехвръкне* lit. *bird cannot fly in* 'it is very strictly guarded'
    - (DE) *der frühe Vogel fängt 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 themelves clever fail'
    - (EN) *the early bird catches the worm*
    - (ES) *los ojos son el reflejo del alma* lit. *the eyes are the reflection from_the soul* 'the eyes are the window to the soul'
    - (FR) *à quelque chose malheur 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'
      *kości zostały rzucone*lit. *the dice have been thrown*'alea iacta est'
    - (PT) *quem vê cara não vê coração* lit. *who sees face doesn't see heart* 'a person can lie/omit his/her feelings'
    - (RO) *se revarsă zorile* lit. *Refl.Cl.3sg.Acc. flow_out dawns* 'it is getting morning'
    - (SL) *kdor prej pride, prej melje* lit. *who came first mills first*
2. **Meaning-preserving <u>variants</u>** 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'
       hacer ilusión lit. *to_make excitement* 'to look forward to, to be excited about'
     - (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) *odločitev je treba sprejeti* 'decisions has to be made'
   - Nominal groups (headed by nominal complements from the prototypical VMWEs) with relative clauses:
     - (BG) *решението, което взех* lit. *decision.DET which I.PRODROP made* 'the decision which I made'
     - (DE) *Entscheidungen die wir trafen* 'decisions which we made'
       Herzen die wir gebrochen haben lit. *hearts which we broke have* 'hearts which we have broken'
     - (EL) *η απόφαση που πήραμε*
       η πλάκα που κάναμε
     - (EN) *decisions which we made*
       heart which we broke
     - (ES) *la decisión que tomamos* lit. *the decision which made.we* 'the decision we took'
       la ilusión que le hizo lit. *the excitement which to_him/to_her made.it* 'he/she was very excited by it'
     - (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 prendemmo* lit. *the decision which we took* 'the decision which we made'
       i cuori che abbiamo spezzato lit. *the hearts which we have broken* 'hearts which we have broken'
     - (PL) *decyzje, które podjął* 'decisions which he took'
     - (PT) *a apresentação que Maria fez* 'the presentation that Mary made'
     - (RO) *decizia pe care am luat-o* 'the decision that we have made'
     - (SL) *odločitev, ki jo je sprejel* 'decisions which he took'
   - Gerunds:
     - (BG) *вземайки това решение* '(while) making this decision'
     - (DE) *n.a. (?)*
     - (EL) *παίρνοντας αποφάσεις*
       κάνοντας πλάκα
     - (EN) *decision making*
       heart breaking
     - (ES) n.a.
     - (FR) *En prenant cette décision* 'by making this decision'
       les personnes subissant plusieurs opérations sont fragiles lit. *the people undergoing several surgical operations are fragile*
     - (IT) *Prendendo decisioni importanti, imparerai* 'by making important decisions, you will learn'
     - (PL) *podejmowanie decyzji* 'decision making'
     - (PT) *Estou tomando uma decisão* lit. *I am taking a decision*
       A pessoa tomando banho sou eu lit. *The person taking shower am I*
     - (RO) *luând decizia* 'making the decision'
     - (SL) *sprejemanje odločitev* 'making the decision'
   - Nominal and adjectival groups with participles:
     - (BG) *взетите вече решения* 'the decisions already made'
       броящ звезди 'who counts stars'
       неразбитите все още от него сърца 'the hearts not yet broken by him'
     - (DE) *früher getroffene Entscheidungen* 'earlier made decisions'

- (EL) n.a.
- (EN) *decisions* previously *made*
  - all *hearts* *broken* by him
  - *heart-breaking*
  - *breaking* her *heart*
- (ES) *las decisiones tomadas* ayer lit. *the decisions taken yesterday* 'the decisions made yesterday'
  - *el trato hecho* previamente lit. *the agreement made previously* 'the previously made agreement'
- (FR) *les décisions prises* hier sont bonnes lit. *the decisions taken yesterday are good* 'the decisions that were made yesterday are good'
  - *les personnes subissant* plusieurs *opérations* sont fragiles lit. *the people undergoing several surgical operations are fragile*
- (IT) *decisioni prese* 'made decisions'
- (PL) *podejmujący* trudne *decyzje* 'making hard decisions'
- (PT) a *decisão tomada* ontem 'the decision made yesterday'
  - a mulher *tomando* um *banho* ' a woman taking a bath'
- (RO) *decizia* recent *luată* 'the decision recently made'
- (SL) včeraj *sprejete odločitve* 'the decision made yesterday'
  - Diathesis alternation (passive, middle) ❶:
    - (EN) *important decisions* were *made* (passive alternation)
    - (FR) *les décisions* importantes se *prennent* en groupe lit. *the decisions important SELF take in group* 'important decisions are taken collectively' (middle alternation)

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* połowę *spadku* lit. *to receive half of the heritage*
  - nie *mieć* cienia *wątpliwości* lit. *not to have a shadow of a doubt* 'to have no doubt'
- (PT) *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 <u>not</u> considered VMWEs. This concerns nominalizations morphologically derived from verbs and describing a process, result, state, agent, etc.

- (BG) *вземане* на *решение* 'making a decision'
  - *удар* в *гърба* 'a stab in the back'
  - *високо вдигната летва* lit. *highly raised bar* 'high bar'
  - *играч* на *карти* 'card player'
- (DE) *Wortbruch* lit. *word-break* 'a promise which has not been hold'
- (EL) *η λήψη αποφάσεων* 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*lit. *the made to availability* 'the fact of making available'
- (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'
- (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) *sprejeta odločitev* 'the decision'

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

- (BG) *може би* lit. *(it) may be* 'maybe'
  - *разбира се* lit. *(it) is understood* 'of course'
- (DE) *Vergiss-mein-nicht*lit. *forget-me-not*'forget-me-not'
- (EL) *τα πάρε-δώσε* lit. *the.NE.PL.ACC give-2SG.IMP take-2SG.IMP* ' relationship of some type'
- (EN) *forget-me-not*
- (ES) *sacacorchos* lit. *take_out corks* 'corkscrew'
  - *portalápices* lit. *carry pencils* 'pencil case'
  - *nomeolvides*lit. *not to_me forget_you-*'forget-me-not'
- (FR) *peut-être* lit. *may-be* 'maybe'
  - *porte-feuille* lit. *carry-sheets* 'wallet'
- (IT) *non-ti-scordar-di-me*lit. *not you forget-me-*'forget-me-not'
- (PL) *zrobić* coś za *Bóg-zapłać* lit. *do something for a God-pay* 'to do something for free'
- (PT) um *saca-rolhas* lit. *a pull-corks* 'a corkscrew'
  - um *faz-de-conta* lit. *a make-as-story* 'a make-believe'
- (RO) *treacă-meargă* lit. *pass-go*'let it be'

Particular language teams may decide to extend the annotation scope to these variants. It is recommended in this case to introduce a new category for them (e.g. NVPC: nominal verb-particle constructions) so as to keep the (quasi-)universal categories intact.

Like other VMWE occurrences, syntactic variants are also annotated if they contain one multi-word token only, e.g. particle verbs like (DE) *aus|machen*.

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 <u>must</u> 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.❶ 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 <u>not</u> 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) ***смея се*** lit. *laugh se.REFL* 'to laugh'
    - ***намирам се*** lit. *find se.REFL* 'to be (somewhere)'
- (DE) *??*
- (EL) n.a.
- (ES) n.a.
- (PL) ***znajduję <u>się</u>*** lit. *find.1.SG.PRES self* 'I find myself'
    - ***znajdujesz <u>się</u>*** lit. *find.2.SG.PRES self* 'you find yourself'
    - ***znajdują <u>się</u>*** lit. *find.3.PL.PRES self* 'they find themselves'
- (PT) n.a.
- (RO) n.a.
- (SL) ***smejim <u>se</u>*** lit. *laugh.1.SG self* 'I laugh'
    - ***smejiš <u>se</u>*** lit. *laugh.2.SG self* 'You laugh'
    - ***smejijo <u>se</u>*** lit. *laugh.3.PL self* 'they 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 <u>sich</u>*** lit. *she wonders self.3.SG* 'she wonders'
    - *ihr **wundert <u>euch</u>*** lit. *you.PL wonder.2.PL self.2.PL* 'you wonder'
- (EL) n.a.
- (ES) *yo **me quejo*** lit. *I self.1.SG complain* 'I complain'
    - *tu **te quejas*** lit. *you self.2.SG complain* 'You complain'
- (FR) *je **me trouve*** lit. *I self.1.SG find* 'I find myslef'
    - *tu **te trouves*** lit. *you self.2.SG find* 'you find yourself'
- (IT) *io **mi meraviglio*** lit. *I self.1.SG wonder* 'I wonder'
    - *tu **ti meravigli*** lit. *you self.2.SG woder* 'you wonder'
- (PT) *eu **me queixo*** lit. *I self.1.SG complain* 'I complain'
    - *tu **te queixas*** lit. *you self.2.SG complain* 'You complain'
- (RO) *eu **mă gândesc*** lit. *I Refl.Cl.1sg.Acc. think* 'I am thinking'
    - *tu **te gândești*** lit. *you Refl.Cl.2sg.Acc. think* 'you are 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 *się*, *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 <u>not</u> considered VMWEs in this task and should <u>not</u> 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) ***качвам цената*** 'raise the price'
- (DE) *eine **Frage beantworten*** 'to answer a question', *die **Graphik zeigt*** 'the grahpic shows', *einen **Bus nehmen*** ' to take a bus'
- (EL) *κάνω βόλτα* lit. *take-1SG a walk*

- (EN) *drastically drop*
  - *the graphic shows*
  - *to take a bus*
- (ES) *responder a una pregunta* 'to answer a question'
  - *el gráfico muestra* 'the graphic shows'
  - *coger el autobús* 'to take the bus'
- (IT) *rispondere a una domanda* 'to answer a question'
  - *il grafico mostra* 'the graphic shows'
  - *prendere un bus* ' to take a bus'
- (PL) *zalać rynek* lit. *to flood the market* 'to dominate the market'
- (PT) *bater um recorde* 'to break a record' (*bater* 'to beat' has a regular sense of *to overcome* in addition to the litteral sense)
  - *entrar em cartaz* lit. *enter into poster* 'arrive in theaters (for a movie)' (the MWE is *em cartaz* lit. *in poster* 'in theaters', the verb just usually collocates with this MWE)
- (RO) *lua un autobuz* 'take a bus'
- (SL) *drastičen upad* 'drastically drop', *graf prikazuje* 'graphic shows', *vzeti taksi* 'to take a taxi'

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

- (BG) *вдигам цената* 'raise the price', *увеличавам цената* 'raise the price', *качвам залога* 'raise the bet', *качвам температурата* 'raise the temperature'
- (DE) *eine Anfrage beantworten* 'to answer a request', *das Diagramm zeigt* 'the diagram shows', *mit einem Bus fahren* 'to take a bus'
- (EL) *πάω βόλτα* lit. *go for a walk*
- (EN) *significantly drop*, *drastically decrease*, *the diagram shows*, *the graphic illustrates*, *to take a coach*
- (ES) *responder a una petición* 'to answer a request'
  - *el diagrama muestra* 'the diagram shows'
  - *coger un tren* 'take a train'
- (IT) *rispondere a una richiesta* 'to answer a request'
  - *il diagramma mostra* 'the diagram shows'
- (PL) *zdominować/zarzucić/zapełnić/nasycić rynek* 'to dominate/overwhelm/fill/saturate the market'
- (PT) *quebrar/bater/ultrapassar/estabelecer um recorde* lit. *to break/beat/overcome/establish a record*
  - *o recorde foi quebrado* lit. *the record was broken*
  - *entrar/estar/permanecer/ficar/continuar/ter em cartaz* lit. *enter/be/remain/stay/continue/have in poster*
- (RO) *lua o mașină*
- (SL) *občuten upad* 'significantly drop', *drastično zmanjšanje* 'drastically decrease', *diagram prikazuje* 'diagram shows', *slika prikazuje* 'picture shows'

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

- (BG) *нямам пукната пара/пукнат грош* 'to not have a single penny, be very poor'
  - *имам твърда/дебела глава* 'to have a thick head, to be stubborn and not listen to advice'
- (DE) *einen Willen/Menschen brechen* 'to break a will/person'
- (EN) *to come in handy/useful*, *to stand firm/fast*, *to break someone's spirit/will*, *to take the cake/biscuit*
- (ES) *dar un paseo/ una vuelta* lit. *give a walk / a turn* 'to go for a walk'
  - *darse/tomar una ducha* lit. *give.self/take a shower* 'take a shower'
- (IT) *cogliere/prendere di sorpresa*, *dare/fornire un contributo*
- (PL) *zapisać się złotymi literami/zgłoskami* lit. *to record iteself with golden letters/syllables* 'to be remebered and commemorated for a merit'
  - *zamarznąć na kość/lód/sopel* lit. *to freeze to bone/ice/icicle* 'to freeze strongly'
- (PT) *levar em conta/consideração* 'take into account/consideration'
  - *chutar o balde/pau da barraca* lit. *to kick the bucket/the tent's stick* 'to act irresponsibly'
- (RO) *lua o decizie/hotărâre* 'make a decision'
- (SL) *sprejeti odločitev/sklep* 'make a decision'

However, the extent of the vocabulary concerned by this variability is different for collocations and VMWEs. Namely, a head verb in a collocation usually selects a <u>whole semantic class</u> for each of its required arguments. For instance, the verb *to take* 'to use a vehicle to travel' selects a whole semantic class of means of transport. Similarly, the verb *to drop* can select a large set of adverbs describing the degree: *drastically/significantly/remarkably/slightly/reasonably drop*. Conversely, lexical variability in a VMWE is limited to a closed list of lexemes, sometimes only loosely semantically related. For instance, the VMWEs *to take a cake/biscuit* and *to stand firm/fast* do not keep their idiomatic readings with semantically close complements: #*to take a cookie/wafer*, **to stand hard/rigid/solid* etc. See also test 2.

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*".

Many VMWEs, especially idioms, are based on metaphors. For instance, *to take the bull by the horns* means to address a problem (the bull) starting with its most challenging aspect (the horns). *To set the world on fire* is to do something extraordinary and get the admiration (set on fire) of other people (the world), *to put all one's eggs in one basket* means to rely on one particular course of action (a basket) for success rather than giving oneself several possibilities.

However, verbal metaphors are not always VMWEs. Consider the newspaper title "*simple steps to lift your dark cloud of stress*", and the extract of a poem by Wordsworth, cited by Shutova: "*and then my heart with pleasure fills, and dances with the daffodils*". 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 established in the common vocabulary to be considered VMWEs.

The distinction between MWEs 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.

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 *to **put** something **up***, the verb and the particle are integral parts of the VMWE (see VPC tests), while in (HU) **döntést hoz** *valamiről* 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.

Reflexive pronouns, **particles** and **prepositions** need to be handled with special care, given their particular lexicalization status. 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.

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) *държа под контрол* 'to keep under control'
    - (DE) *eine **Rede halten***lit. *a speech hold*'to give a speech'
    - (EL) *κάνω μία 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) *ħa deċizjoni* 'took a decision'
    - (PL) ***podjąć decyzję*** 'to take a decision'
    - (PT) ***fazer** uma **promessa*** 'to make a promise'
    - (RO) *a **lua** o **decizie*** lit. *to take a decision*'to make a decision'
    - (SL) ***imeti predavanje**, **sprejeti odločitev*** 'to take a decision'
  - idioms (**ID**):
    - (BG) *правя се на дръж ми шапката* lit. *to behave myself as 'hold my hat'* 'pretend to be naive and innocent'
    - (DE) ***schwarz fahren*** lit. *to drive black* '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) *χάνω 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 tripas 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) ***għasfur żgħir qalli*** lit. *a bird small told me* ' **to hear something from the grapevine'**
    - (PL) ***rzucać grochem o ścianę*** lit. *throw peas agains a wall* 'to try to convince somebody in vain'
    - (PT) ***fazer das tripas 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 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 (**IRefIV**):
    - (BG) *усмихвам се* 'to smile'
    - (DE) ***sich bemühen*** 'to endeavour', ***sich enthalten*** lit. *himself contain* 'to abstain'
    - (EL) n.a.
    - (ES) ***suicidarse*** 'to suicide'
    - (FR) ***se suicider*** 'to suicide'
      quejarse 'to complain'
    - (IT) ***suicidarsi*** 'to suicide'
    - (PL) ***bać się*** lit. *to fear SELF*'to be afraid'
    - (PT) ***se queixar*** 'to complain'
    - (RO) *a **se gândi*** 'to think'
    - (SL) ***bati 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 reproches 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 <u>only</u> 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) *цъфна и върже* lit. *to blossom and give fruit* '(usually sarcastically) to prosper'
    река и отсека lit. *to say and cut* 'to say firmly, decisively'
  - (DE) *einen drauf setzen* 'going one better'
  - (EL) *απορώ και εξίσταμαι* 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-circuiter* 'to short-circuit'
  - (IT) *andare e venire* lit. *to come and go* 'back and forth'
    corto-circuitare
    'to short-circuit '
  - (MT) *iqum u joqghod* lit. *jump and stay* ' to fidget '
  - (PL) *pluć i łapać* 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 și i-a adunat* 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 ❶ 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 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

<u>Note</u>: 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 - [**1DEP**: *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 IReflV-specific tests ⇒ IReflV tests positive?
        ↳ **YES** ⇒ Annotate as a VMWE of category **IReflV**
        ↳ **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 <u>at least one test</u> 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) *хващам* <u>натясно</u> lit. *catch in a tight place* 'to coerce, to pressure' → *натясно* is only used in MWEs
   правя на <u>бъзе</u> *и* <u>коприва</u> lit. *to turn into elder and nettle* 'to scold, to tell off' → *бъзе* is an old word, very rarely used independently
   вземам <u>предвид</u>, *имам* <u>предвид</u> lit. *to* → *предвид* (as adverb) is only used in MWEs
   стоя диван <u>чапраз</u> lit. *to stay upright as in Osman council* 'to stay ready to serve' → *чапраз* is an old word, very rarely used independently
- (DE) *sich um etw.* <u>scharen</u> 'to gather around something' → *scharen* is not a stand-alone word
- (EL) *μάλλιασε η γλώσσα μου* lit. *is-full-of-hair-3SG the-SG.NOM tongue-SG.NOM my-SG.GEN.POSS* 'to repeat the same thing again and again' → *μάλλιασε* is not a stand-alone word
- (EN) *to go* <u>astray</u> → *astray* is not a stand-alone word
- (ES) *sin decir ni* <u>chus</u> *ni mus* → *chus* is not a stand-alone word lit. *without to_say neither chus nor mus* 'without saying a word'
   no decir ni <u>chus</u> *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 <u>troche</u> *y moche* → *troche* is not a stand-alone word lit. *to troche and dulled* 'in a nonsensical way, willy-nilly, haphazardly'
- (FR) *prendre la poudre d'*<u>escampette</u> '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' → *gnorri* is not a stand-alone word
   scendere in lizza 'to enter the lists' → *lizza* is not a stand-alone word
- (PL) <u>odsądzić</u> *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'
- (PT) *ir para as* <u>cucuias</u> 'to go wrong' → *cucuias* is not a stand-alone word
- (RO) *a nu avea* <u>habar</u> 'to have no idea' → *habar* is not a stand-alone word
- (SV) *att komma* <u>ihåg</u> 'to remember' → *ihåg* is not a stand-alone word
- (SL) <u>biti</u> *si* <u>kvit</u> 'to pay up a debt, owe nothing to somebody' → *kvit* is not a stand-alone word

↳ **NO** ⇒ further tests are required
- (BG) *правя на сос* → *правя* 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, la* and *universidad* are stand-alone words
- (IT) *andare giù* 'to go down' → *andare* and *giù* are stand-alone words
- (PL) *wyznać tajemnicę* 'to reveal a secret' → *wyznać* and *tajemnica* are standalone words
- (PT) *ir para a escola* 'to go to school' → *ir, para, a* 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 <u>replacement</u> of one of the components by related words taken from a <u>relatively large semantic class</u> lead to <u>ungrammaticality</u> or to an <u>unexpected change in meaning</u>?

↳ **YES** ⇒ it is a VMWE
- (BG) *бълвам змии и гущери* lit. *to spew snakes and lizards* → #*бълвам влечуги (to spew reptiles)*
   всяка жаба да си знае гьола lit. *every frog to know its own puddle* → #*всяка жаба да си знае локвата*
- (DE) *die Katze aus dem Sack lassen* 'to let the cat out of the bag' → #*den 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' → #*eine Entscheidung machen/herstellen* lit. *a decision make/produce* '#to make/produce a decision'
- (EN) *to let the cat out of the bag* → #*to 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/rigid/solid*
- (ES) *meterse en la boca del lobo* lit. *to_get_into.self in the mouth from_the wolf* 'venture into the lion's den' → #*meterse en el ojo del gato*
   tomar una decisión lit. *to_take a decision* 'to make a decision' → #*hacer/coger/producir una decisión* lit. *to_make/grab/produce a decision* '#to make/grab/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 rospo lit. *spit the toad* 'spit it out' → #*sputare la rana* lit. #*spit the frog*
- (PL) *wiedzieć, co w trawie piszczy* lit. *to know what in grass squeals* 'to be well informed' → #*wiedzieć, co w trawniku popiskuje*
   wziąć udział lit. *to take participation* 'to take part' → #*to podjąć/pobrać/zabrać członkostwo/uczestniczenie*
- (PT) *quebrar um galho* lit. *break a branch* 'to help' → #*danificar um ramo* 'to damage a stem'
- (RO) *a da cu bâta în baltă* lit. *to give with bat-the in pond* 'to say sth embarrassing' → *a da cu băţul în baltă* 'to give with stick-the in pond', *a da cu bâta în lac* 'to give with bat-the in lake'
- (SV) *att Plocka russinen ur kakan* lit. *to pick the raisins out of the cake* 'to choose only the best things' → #*att välja ut nötterna från kakan*

- (SL) *imeti mačka* lit. *to have a cat* 'to have a hangover' → #*imeti psa* lit. *to have a dog*
  - *iti rakom žvižgat* lit. *to go whistling to cancers* 'to fail, to die' → #*iti jastogom pet* lit. *to go singing to the lobsters*
- ↳ **NO** ⇒ further tests are required
  - (BG) *изнасям доклад* 'present a report' → *изнасям урок/лекция/презентация и т.н.*
  - (DE) *den Bus nehmen* 'to take the bus' → *den Zug/ das Flugzeug, etc nehmen* 'to take the train/plain/etc'
  - (EN) *to take a plane* → *to take a bus/car/boat, etc.*
  - (ES) *coger el autobús* lit. *to_take the bus* 'to take the bus' → *coger el avión/tren, etc.* 'to take the plain/train/etc.'
  - (IT) *prendere il treno* 'to take the bus' → *prendere il bus/aereo/etc* 'to take the bus/plain/etc'
  - (MT) *jqum u joqgħod* 'always moving about'
  - (PL) *sprawić kłopot* 'to make a trouble' → *sprawić przykrość/trudność/niedogodność/problem/zawikłanie/nieprzyjemność* 'to make a(n) nuisance/difficulty/inconvenience/problem/complication'
  - (PT) *quebrar um braço* 'to break an arm' → *quebrar uma perna/costela/falange* 'to break a leg/rib/phalanx'
  - (RO) *a lua o decizie* lit. *to take a decision* 'to make a decision' → *a lua o hotărâre* lit. *to take a decree* 'to make a decision'
  - (SV) *att ta bussen* 'to take the bus' → *att ta tåget/flyget, etc* 'to take the train/plain/etc'
  - (SL) *delati težave* 'to make a trouble' → *delati preglavice/probleme/* 'to make a(n) nuisance/problem'

Usual modifications for [LEX] include replacing content words in the candidate by synonyms, hypernyms, hyponyms, antonyms, troponyms, meronyms, and related words in general.

Test 3 - [MORPH] - Morphological inflexibility

Does a regular <u>morphological change</u> that would normally be allowed by general grammar rules lead to <u>ungrammaticality</u> or to an <u>unexpected change in meaning</u>?

- ↳ **YES** ⇒ it is a VMWE
  - (BG) *хвърлям око* lit. *throw an eye* 'to throw a glance' → #*хвърлям очи.PLURAL*
    - *хващам бика.DEF за рогата* 'take the bull by the horns' → #*хващам бик.INDEF за рогата*
    - *не мога <u>да си намеря място</u>* lit. *cannot 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' → #*in ein Gras beißen* lit. *to bite into a grass* #*in die Gräser beißen* lit. *to bite into the grasses*, *in Kraft treten* lit. *into force step* 'to come into effect' → #*in Kräfte treten* lit. *into forces step*
  - (EN) *to kick the bucket* → #*to kick the buckets*
    - *to pretty-print* → **to prettier-print*
    - *to take turns* → #*to take a turn*
  - (ES) *coger el toro por los cuernos* lit. *to_take the bull by the horns* 'to take the bull by the horns' → #*coger el toro por el cuerno* lit. *to_take the bull by the horn* #*to take the bulls by the horns* lit. *to_take the bulls by the horns* '#to take the bulls by the horns'
    - *entrar en vigor* lit. *to_enter in vigor* 'to come into effect' → #*entrar en vigores* lit. *to_enter in vigors* '#to come into effects'
  - (FR) *prendre le taureau par les cornes* 'to_take the bull by the horns' → #*prendre le taureau par une corne* lit. *to_take the bull by a horn*
  - (IT) *andare a letto con le galline* lit. *to go to bed with the hens* ' to go to bed early' → #*andare a letto con la gallina* lit. *to_go to bed with the hen*
    - *cercare il pelo nell'uovo* lit. *to look for the hair in the egg* ' to be pedantic ' → #*cercare i peli nell'uovo*
  - (PL) *budować zamki na lodzie* lit. *to build castles on ice* 'to rely on unstable foundations' → #*budować zamek na lodzie* 'to build a castle on ice'
    - *mucha kogoś ugryzła* lit. *a fly bit someone* 'someone is in a bad temper' → #*mucha kogoś ugryzie* lit. *a fly will bite someone*
    - *wyciągnąć nogi* lit. *to stretch.PERF legs* 'to die' → #*wiciągać nogi* lit. *to stretch.IMPERF legs* (imperfective aspectual variant prohibited)
  - (PT) *bater perna* lit. *hit leg* 'to walk around' → *bater a/uma/essas perna/pernas/perninha/pernona* 'to hit the/one/these leg/legs/leg.SMALL/leg.BIG'
  - (RO) *a da colțul* lit. *to give corner.the* 'to die' → *a da colțurile lit. *to give corners.the*
  - (SV) *träda i kraft* lit. *step in force* 'to come into effect' → #*träda i krafter* lit. *step into forces*
  - (SL) *klicati jelene* lit. *to call cerfs* 'to vomit' → #*klicati jelena* 'to call a cerf'
- ↳ **NO** ⇒ further tests are required
  - (BG) *хвърлям топка* 'to throw a ball' → *хвърлям топка/топката/топки/топките*
  - (DE) *einen Kuchen backen* 'to bake a cake' → *viele/keine/den Kuchen backen/machen* 'many/no/the cake bake/make'
  - (EN) *to make a cake* → *to make a/many/those/no cake/cakes*
  - (ES) *mover el brazo* lit. *to_move the arm* 'to move the arm' → *mover/agitar/levantar/estirar el brazo/la pierna/las manos/las piernas* lit. *to_move/shake/raise/stretch the arm/the leg/the hands/the legs* 'to move/shake/raise/stretch the arm/the leg/the hands/the legs'
  - (IT) *fare un dolce* → *fare un/molti/dei/quei/nessun dolce/dolci*
  - (PL) *kształtować opinię* 'to form an opinion' → *kształtować opinie* 'to form opinions'
  - (PT) *bater o braço* 'to hit the arm' → *bater o/os/um/esse braço/braços/bracinho* 'hit the/the.PL/a/this arm/arms/arm.SMALL'
  - (RO) *a face o prăjitură* 'to make a cake' → *a face multe/aceste prăjituri* 'to make many/these cakes'
  - (SV) *att baka en kaka* 'to bake a cake' → *att baka flera/den där/några/ingen kaka/kakor* ' to bake several/that/some/no cake(s)'
  - (SL) *vzeti taksi* 'to take a cab' → *ne vzeti nobenega taksija/en taksi/dva taksija* 'to take no/one/two/… cab(s)'

Usual modifications for [MORPH] include inflecting content words in the candidate for gender, number, case, tense, mood, aspect, etc. - depending on the target language's morphology.

Test 4 - [MORPHSYNT] - Morpho-syntactic inflexibility

Does a regular <u>morpho-syntactic change</u> that would normally be allowed by general grammar rules lead to <u>ungrammaticality</u> or to an <u>unexpected change in meaning</u>?

- ↳ **YES** ⇒ it is a VMWE
  - (BG) <u>*аз ти давам думата си*</u> 'I give you my word' → #<u>*аз ти давам неговата дума*</u> *(I give you HIS word)*
    - <u>*аз си продавам душата*</u> 'I sell my soul' → #<u>*аз продавам неговата душа*</u> *(I sell his soul)*
  - (DE) <u>*Ich* werde *mein Bestes* tun</u> lit. *I will my best do* 'I will do my best' → **Ich werde dein Bestes tun* 'I will do your best', *Ich gebe dir mein Wort* 'I give you my word' → **Ich gebe dir ihr Wort* 'I give you her word'
  - (EN) <u>*I*</u> will **do** <u>*my*</u> **best** → *<u>*I*</u> will do <u>your</u> best
    - <u>*I*</u> **give** you <u>my</u> **word** for that → #<u>*I*</u> give you <u>his</u> word for that
    - *he* was **pulling** <u>my</u> **leg** → #<u>*I*</u> was pulling <u>my</u> leg
  - (ES) <u>*te*</u> **doy** <u>mi</u> **palabra** lit. *to_you give_I my word* 'I give you my word' → #<u>*te*</u> **doy** <u>su</u> **palabra** lit. *to_you give_I his/her word* 'I give you his/her word'
  - (FR) <u>*il*</u> **vide** <u>son</u> **sac** lit. *he empties his bag* 'he reveals his secret thoughts' → #<u>*il*</u> vide <u>mon</u> sac 'he empties my bag'
  - (IT) <u>*Io*</u> farò del <u>mio</u> meglio → *<u>*Io*</u> farò del <u>tuo</u> meglio
    - <u>*Io*</u> ti **do** la <u>mia</u> **parola** → #<u>*Io*</u> ti do la <u>sua</u> parola
  - (PL) Polish VMWEs do not seem to exhibit this kind of inflexibility

- (PT) *ele se suicidou* lit. *he self.3P.SG suicided* → **ele me suicidou*
  eu perdi meu tempo 'I wasted my time' → *eu perdi teu/seu/nosso tempo* ⓘ
- (RO) *Îți dau cuvântul meu* lit. *CL.DAT give.1SG word.the my* 'I give you my word' → *#Îți dau cuvântul lui* lit. *CL.DAT give.1SG word.the his* 'I give you his word'
- (SV) *Jag gör mitt bästa* lit. *I do my best* 'I do my best' → **Jag gör ditt bästa* 'I do your best'
- (SL) *Vlečeš me za nos* lit. *you are pulling my nose* 'you're pulling my leg ' → **Vlečeš se za nos* lit. *you're pulling your nose*
  Pojdi se solit! lit. *to go salt oneself* 'Get lost!' → **Pojdi ga solit* lit. *go salt him*

↳ **NO** ⇒ further tests are required
- (BG) *копая си гроба* 'to dig my grave' → *копая ти/му/ѝ/им гроба (to dig your/his/her/their grave)*
- (DE) *er traf seine Entscheidung* 'he made his decision' → *er traf meine/ihre/unsere/eure Entscheidung* 'he made my/her/our/your decision'
- (EN) *he did his job* → *he did my/her/our/your job*
- (ES) *Ha hecho su trabajo* lit. *Has_he/she done his/her work* 'He/She has done his/her work' → *Ha hecho mi/tu/nuestro trabajo* lit. *Has_he/she done my/your/our work* 'He/She has done my/your/our work'
- (IT) *hafatto il suo lavoro* → *ha fatto il mio/tuo/nostro/vostro/loro lavoro*
- (PL) Polish VMWEs do not seem to exhibit this kind of inflexibility
- (PT) *Eu fiz meu trabalho* 'I did my job' → *Tu/ele/nós fizeste/fez/fizemos meu trabalho* 'You/he/we made my job'
- (RO) *el își face tema* lit. *he his does homework.the* 'he does his homework' → *el îmi/ne/le face tema* lit. *he my/our/their does homework.the* 'he does my/our/their homework'
- (SV) *han gör sitt jobb* 'he does his job' → *han gör mitt/hennes/vårt jobb* 'he does my/her/our job'
- (SL) *opravil je svojo nalogo* 'he did his job' → *opravil je mojo/njeno/našo/tvojo nalogo* 'he did my/her/our/your job'

Usual modifications for [MORPHSYNT] involve agreement or loss of agreement between some components in the candidate.

Test 5 - [SYNT] - Syntactic inflexibility

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) *на стар краставичар краставици продавам* lit. *to an old cucumber seller cucumbers to sell* 'to try to cheat a more experienced person' → *#продавам краставици на стар краставичар*, *#краставиците са продадени*
  бълвам змии и гущери → *#бълвам гущери и змии*
- (DE) Noun phrase (NP) or prepositional phrase (PP)
- (EN) *speak of the devil* 'the person one is talking about shows up' → *#he was speaking of the devil*
  to go bananas 'to get crazy' → *#bananas are gone*
  to drink and drive → *#drive and drink*
  to kick the bucket → *#the bucket was kicked*
- (ES) *coser y cantar* lit. *to_sew and to_sing* 'easy as pie, a piece of cake' → *#cantar y coser* 'to sing and to sew'
  perder la cabeza lit. *To_loose the head* 'to go bananas' → *#Perder las cabezas* 'To_loose the heads'
- (IT) *alzare la cresta* lit. *to lift the crest* 'become cocky' → *#la cresta è stata alzata* 'the crest has been lifted'
  andare in malora lit. *go to ruin* 'go to ruin ' → *#nella malora è andata* lit. *in ruin was gone*
  vivi e lascia vivere 'live and let live' → *#lascia vivere e vivi* 'let live and live'
- (PL) *kogoś krew zalewa* lit. *blood foods someone* 'someone gets furious'→ *#ktoś jest zalewany przez krew* lit. *someone is flooded by blood* (passive blocked)
  robić bokami lit. *to do with-sides* 'to have serious financial problems'→*#robić swoją robotę bokami* lit. *to do one's job with sides* (regular modification blocked)
  dobrze komuś z oczu patrzy lit. *well someone.DAT from eyes looks* 'someone looks like a good person' → *#uprzejmość dobrze komuś z oczu patrzy* lit. *kindness well someone.DAT from eyes looks* (subject prohibited)
  nie zagrzać miejsca w pracy lit. *not to warm a place at work* 'not to stay long at one work' → *#zagrzać miejsce w pracy* lit. *to warm a place at work* (negation is compulsory)
  zdechł pies! lit. *died the dog!* 'it is a lost cause'→ *#pies zdechł* lit. *the dog died* (a regular word order variability is blocked)
  wziąć w łeb lit. *to take into head* 'to fail' → *#wziąć porażkę w łeb* lit. *to take failure into head* (direct object prohibited for the normally transitive verb *wziąć* 'to take')
- (PT) *pisar na bola* lit. *step on the ball* 'make a mistake' → *#a bola na qual ele pisou* 'the ball on which he stepped'
- (RO) *a da colțul* lit. *to give corner.the* 'to die' → **colțul a fost dat* lit. *corner.the has been given*
- (SV) *det knallar och går* lit. *it trots and walks* 'it is OK/as usual' → *#det går och knallar*
- (SL) *delati se Francoza* lit. *to pretend to be French* 'to pretend to be indifferent' → **delan Francoz* lit. *made French*

↳ **NO** ⇒ further tests are required
- (BG) *продавам неговата кола* 'I sell his car' → *колата му беше продадена (his car was sold)*, *неговата кола, която тя продаде (his car which she sold)*, *т.н.*
- (DE) *jemandes Auto waschen* 'to wash one's car' → *ihr Auto wurde gewaschen* 'her car was washed', *das Auto, welches sie wusch* 'the car that she washed', *Autowaschen* 'car-washing', *etc*
- (EN) *to wash one's car* → *her car was washed, the car that she washed, car washing, etc.*
- (ES) *pisar la arena* 'to step on the sand' → *la arena que pisaste* 'The sand on which you stepped'
- (IT) *lavare la macchina* →*la sua macchina è stata lavata, la macchina che ha lavato, il lavaggio della macchina, etc.*
- (PL) *kształtować opinię* 'to form an opinion' → *opinia jest kształtowana* lit. *the opinion is formed*
- (PT) *pisar na areia* 'to step on the sand' → *a areia na qual você pisou* 'the sand on which you stepped'
  jogar futebol 'to play football' → *?futebol é jogado* 'football is played' ⓘ
- (RO) *a spăla mașina* 'to wash the car'→ *mașina a fost spălată, mașina pe care a spălat-o, spălarea mașinii etc.* 'the car was washed, the car that he/she washed, car washing'
- (SV) *att tvätta bilen* 'to wash one's car' → *min bil tvättades* 'my car was washed', *bilen som hon tvättade* 'the car that she washed', *biltvätt* 'car-wash' *etc.*
- (SL) *narediti film* 'to make a movie' → *Film, narejen po knjigi* 'a movie based on a book'

LVC Hypothesis

Does the candidate consist of a verb and a nominal complement, where the verb has a purely syntactic operator function (performing an activity or being in a state) and the noun expresses this activity or state?

↳ **YES** ⇒ assume that it is a VMWE
- (BG) *вземам решение* 'to make a decision' → the semantics is expressed by the noun *решение*, while the verb *вземам* expresses tense, aspect and mood, and assigns semantic role 'agent' to the subject

- (DE) *eine **Entscheidung treffen*** lit. *a decision meet* 'to make a decision' → *treffen* only expresses that an activity (*Entscheidung*) happened
  Mut haben 'to have courage' → *haben* only expresses that the subject has a property (*Mut*)
- (EN) *to **make a decision*** → *make* only expresses that an activity (*decision*) happened
  *to **commit suicide*** → *commit* only expresses that an activity (*suicide*) happened
  *to **have courage*** → *have* only expresses that the subject has a property (*courage*)
- (ES) ***tomar** una **decisión*** 'to make a decision' → *tomar* only expresses that an activity (*decisión*) happened
  cometer** un **crimen 'to commit a crime' → *cometer* only expresses that an activity (*crimen*) happened
  tener coraje 'to have courage' → *tener* only expresses that the subject has a property (*corage*)
- (IT) ***dare** un **suggerimento*** lit. *give a suggestion* 'give a suggestion' → *dare* only expresses that an activity (*suggerimento*) happened
  avere coraggio 'to have courage' → *avere* only expresses that the subject has a property (*coraggio*)
- (PL) ***wziąć udział*** lit. *to take a participation* 'to take part' → *wziąć* only expresses an activity (participation)
- (PT) ***tomar** uma **decisão*** 'to make a decision' → *tomar* only expresses that an activity (*decisão*) happened
  cometer** um **crime 'to commit a crime' → *cometer* only expresses that an activity (*crime*) happened
  ter coragem 'to have courage' → *ter* only expresses that the subject has a property (*coragem*)
- (RO) *a **lua** o **decizie*** lit. *to take a decision* 'to make a decision' → *lua* only expresses that an activity (*decizie*) happened
- (SV) ***ta** ett **beslut*** lit. *take a decision* 'to make a decision' → *ta* only expresses that an activity (*beslut*) happened
- (SL) ***sprejeti odločitev*** → *sprejeti* only expresses that an activity (*odločitev*) happened
  narediti samomor → *narediti* only expresses that an activity (*samomor*) happened
  imeti pogum → *imeti* only expresses that the subject has a property (*pogum*)

↳ **NO** ⇒ the candidate is **NOT** a VMWE
- (BG) ***вземам пари*** 'to take money' → *вземам* has a concrete meaning and *пари* (money) is not an activity or state
- (DE) *einen **Kuchen machen*** 'to make a cake' → *machen* has a concrete meaning and the thing being made (*Kuchen*) is not an activity or state.
  Nachbarn haben 'to have neighbours' → *haben* might be an operator verb but *Nachbarn* are not activities or properties***Hoffnung geben*** 'to give hope' → *Hoffnung* is a state/property, but *geben* adds inchoative (i.e. change-of-state) semantics to it
- (EN) *to **make a cake*** → *make* has a concrete meaning and the thing being made (*cake*) is not an activity or state
  *to **have neighbors*** → *have* might be an operator verb but *neighbors* are not activities or properties
  *to **give hope*** → *hope* is a state/property, but *give* adds inchoative (i.e. change-of-state) semantics to it
- (ES) ***tomar** un **zumo*** lit. *to_take a juice* 'to have a juice' → *tomar* has a concrete meaning and the thing being taken (*zumo*) is not an activity or state
  tener vecinos 'to_have neighbors' → *tener* might be an operator verb but *vecinos* are not activities or properties
  dar esperanza 'to give hope' → *esperanza* is a state/property, but *dar* adds inchoative (i.e. change-of-state) semantics to it
- (IT) ***fare** un **quadro*** lit. *to make a painting* ' to make a painting ' → *fare* has a concrete meaning and the thing being made (*quadro*) is not an activity or state
  avere fratelli → *avere* might be an operator verb but *fratelli* are not activities or properties
  dare speranza → *speranza* is a state/property, but *dare* adds inchoative (i.e. change-of-state) semantics to it
- (PL) ***odmówić udziału*** lit. *to refuse the participation* 'to refuse to participate' → *udział* is a state/activity but *odmówić* adds semantics to it (refusal to perform the activity)
- (PT) ***tomar** um **suco*** lit. *take a juice* 'have a juice' → *tomar* has a concrete meaning and the thing being taken (*suco*) is not an activity or state
  ter vizinhos 'to have neighbors' → *ter* might be an operator verb but *vizinhos* are not activities or properties
  dar esperança 'to give hope' → *esperança* is a state/property, but *dar* adds inchoative (i.e. change-of-state) semantics to it
- (RO) *a **face** o **prăjitură*** 'to make a cake' → *face* has a concrete meaning and the thing being made (*prăjitură*) is not an activity or state
- (SL) ***narediti palačinke*** 'to make pancakes' → *narediti* 'make' has a concrete meaning and the thing being made (*palačinke*) is not an activity or state.

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 multiword 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.

Section 6

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 determining 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
- (BG) ***гушна букета*** lit. *to hug the bunch of flowers* 'to die' → *гушна* is the head and the NP depends on it
  правя на салата lit. *to make into salad* 'to scold' → *правя* is the head and the PP depends on it

- (DE) *eine **Fratze** <u>ziehen</u>* lit. *a grimace pull* 'to make a face' → *ziehen* is the head and the NP depends on it
  - *er **gibt** <u>auf</u>* 'he gives up' → *gibt* is the head and *auf* is the particle depending on it
- (EL) *κάνω γκριμάτσα* lit. *to make grimace* 'to make a face' *κάνω* is the head and the NP depends on it
- (EN) *to <u>make</u> a **face*** → *make* is the head and the NP depends on it
  - *to <u>give</u> **up*** → *give* is the head and *up* is a particle depending on it
- (IT) *to <u>fare</u> le **linguacce*** lit. *to make the grimaces* → *fare* is the head and the NP depends on it
  - *to <u>far</u> **fuori*** lit. *to make out* ' to kill '→ *far* is the head and *fuori* is a particle depending on it
- (PL) *<u>zbijać</u> **bąki*** lit. *to smash farts*'to fool around, to do nothing useful'→ *zbijać* is the head and the NP *bąki* depends on it
  - *<u>dać</u> komuś **popalić***lit. *to let someone smoke*'to make someone's life hard' → *dać* is the head and the infinitive *popalić* depends on it
- (PT) *<u>bater</u> as **botas*** → *bater* is the head and the NP depends on it
  - *<u>criar</u> **vergonha na cara*** → *criar* is the head and the two NPs depend on it
- (RO) *a <u>face</u> **baie*** lit. *to make bath* 'to bath' → *face* is the head and the NP depends on it
  - *a <u>ieşi</u> **înainte*** lit. *to go forth* 'to greet' → *ieşi* is the head and *înainte* is a particle depending on it
- (SV) *att <u>ge</u> **upp*** 'to give up' → *ge* is the head and *upp* is the particle depending on it
- (SL) *<u>imeti</u> **krompir*** lit. *to have potatoes* 'to be lucky' → *imeti* is the head and the NP depends on it

↳ **NO** ⇒ annotate as OTH
- (BG) *цъфна и върржа* → none of the verbs is clearly the head, as there is no universally accepted syntactic representation of coordination
- (DE) *leben und leben lassen* 'live and let live' → none of the verbs is clearly the head, as there is no universally accepted syntactic representation of coordination
- (EN) *to **pretty-print*** → there is an unusual case of an adjective modifying a verb
  - *to **drink** and **drive*** → none of the verbs is clearly the head, as there is no universally accepted syntactic representation of coordination
- (IT) *vivi e lascia vivere* 'live and let live' → none of the verbs is clearly the head, as there is no universally accepted syntactic representation of coordination
- (PL) *pluć i łapać* lit. *to spit and catch*'to be lazy, to do nothing useful' → none of the verbs is clearly the head, as there is no universally accepted syntactic representation of coordination
- (PT) *pintar e bordar* lit. *paint and knit* 'to abuse'
- (SV) *det knallar och går* lit. *it trots and walks* 'it is OK/as usual' → none of the verbs is clearly the head, as there is no universally accepted syntactic representation of coordination

The aim of this test is to distinguish VMWEs of category OTH from those that require further tests. For the special case of nominal, participle and gerund variants of VMWEs, the test should be applied to the canonical verbal form instead. This is required because there may be no verb or the verb may not be the syntactic head

- (BG) *вземам решение* passes the test → variants like *решението, което беше взето* pass the test as well
- (DE) *eine **Entscheidung** treffen* 'make a decision' passes the test → variants like *die **Entscheidung** wurde **getroffen*** 'the decision was made', *die **Entscheidung**, welche **getroffen** wurde* 'the decision which was made', *das **Treffen** der **Entscheidung*** 'the making of the decision' pass the test as well
- (EN) *to **make** a **decision*** passes the test → variants like *the **decision** which was **made**, **decision-making**, the **making** of the **decision*** pass the test as well
- (IT) *prendere una decisione* lit. *to take a decision* 'make a decision' passes the test → variants like *la decisione è stata presa* 'the decision was made', *la decisione, che è stata presa* 'the decision which was made', *la presa di decisione* 'the making of the decision' pass the test as well
- (PL) *<u>zbijać</u> **bąki*** lit. *to smash farts*'to fool around, to do nothing useful' passes the test → variants like *zbijanie bąków* lit. *farts smashing*'fooling around, doing nothing useful', *zbijający bąki* lit. *smashing farts* pass the test as well
- (PT) *tomar uma decisão* 'make a decision' passes the test → variants like *a decisão que foi tomada* 'the decision which was made', *decisão tomada* 'decision made' pass the test as well
- (RO) *a lua o decizie* 'make a decision' passes the test → variants like *decizia care a fost luată* 'the decision which was made', *luarea deciziei* 'decision-making' pass the test as well
- (SL) *sprejeti odločitev* 'make a decision' passes the test → variants like *odločitev, sprejeta dne ...* 'the decision was made' pass the test as well

Test 7 - [1DEP] - Single dependent

Does the VMWE contain exactly one lexicalized syntactic dependent *d* of the head verb *v*?

↳ **YES** ⇒ continue to next test
- (BG) *ритам **камбаната*** lit. *kick the bell* 'to die' → the single dependent is a noun phrase, *камбаната*
  - *ставам **на кайма*** lit. *turn into mince* 'to be destroyed' → the single dependent is a prepositional phrase, *на кайма*
  - *одирам **жив*** lit. *skin alive* 'to make someone suffer' → the single dependent is an small clause (adjective), *жив*
- (DE) *eine **Fratze** <u>ziehen</u>* lit. *a grimace pull* 'to make a face' → the single dependent is a noun phrase, *Fratze*
  - *, in **Betracht** <u>ziehen</u>* 'to take into consideration' → the single dependent is a prepositional phrase, *in Betracht*
  - *er **gibt** <u>auf</u>* 'he gives up' → the single dependent is a particle *auf*
- (EN) *to **make** a <u>face</u>* → the single dependent is a noun phrase, *face*
  - *to **take** <u>into account</u>* → the single dependent is a prepositional phrase, *into account*
  - *to **take** <u>turns</u>* → the single dependent is a noun, *turns*
  - *to **give** <u>up</u>* → the single dependent is a particle, *up*
- (IT) *fare le <u>linguacce</u>* lit. *to make the grimaces* 'to make a face' → the single dependent is a noun phrase, *linguacce*
  - *, prendere <u>in considerazione</u>* 'to take into consideration' → the single dependent is a prepositional phrase, *in considerazione*
  - *egli lo fa <u>fuori</u>* 'he kills him ' → the single dependent is a particle *fuori*
- (PL) *bić <u>na alarm</u>* lit. *to strike on alarm*'to raise the alarm' → the single dependent is a prepositional phrase, *na alarm* lit. *on alarm*
  - *cholera <u>wie</u>* lit. *cholera knows*'I have no idea'→ the single dependent is the nominal subject *cholera*
- (PT) *cometer <u>um crime</u>* lit. *to commit a crime* → one dependent
- (RO) *a face <u>faţă</u>*lit. *to make face*'to to deal with'→ the single dependent is a noun phrase, *faţă*
  - *a ieşi <u>înainte</u>* → the single dependent is an adverb, *înainte*
- (SV) *att ge **upp*** 'to give up' → the single dependent i s the particle *upp*
- (SL) *gre **za*** 'it is about' → the single dependent is a particle, *za*
  - *smejati **se*** 'to laugh' → the single dependent is a reflexive clitic, *se*
  - *imeti **mačka*** 'to have a hangover' → the single dependent is a noun, *maček*

↳ **NO** ⇒ annotate as ID
- (BG) *на стар <u>краставичар</u> <u>краставици</u> продавам* lit. *to an old cucumber seller cucumbers to sell* 'to try to cheat a more experienced person' → two dependents, *на стар краставичар* (PP) and *краставици* (NP)
  - *прочитам <u>от корица</u> <u>до корица</u>* 'to read from cover to cover' → two dependents, *от корица* (PP) and *до корица* (PP)
  - *правя (нечий) <u>живот</u> <u>черен</u>* lit. *make someone'l life black* 'to ruin someone's life' → two dependents, *(нечий) живот* (NP) and *черен* (small clause)
- (DE) *die <u>Katze</u> <u>aus dem Sack</u> lassen* 'to let the cat out of the bag' → two dependents <u>die Katze</u> and <u>aus dem Sack</u>

- (EN) *to make ends meet* → two dependents, *ends* and *meet*
  - *to let the cat out of the bag* → two dependents, *the cat* and *out of the bag*
- (IT) *non dire gatto se non ce l'hai nel sacco* lit. *don't say cat if you don't have it in the bag* 'don't count your chickens before they're hatched' → two dependents gatto and nel sacco
- (PL) *chować głowę w piasek* lit. *to hide head in sand* 'to pretend not to see a problem' → two dependents, *głowę* lit. *head* and *w piasek* lit. *in sand*
  - *bać się własnego cienia* lit. *to fear SELF one's own shadow* 'to be very timid' → two dependents, *się* lit. *SELF* and *własnego cienia* lit. *own shadow*
- (PT) *tapar o sol com a peneira* lit. *to hide the sun with a sieve* 'to sugar-coat' → two dependents
- (RO) *a da bir cu fugiții* lit. *to give tribute with fugitives the* 'to disappear' → two dependents, *bir* and *cu fugiții*
  - *a-i ieși ochii din cap* lit. *to his come out eyes the from head* 'to stare' → three dependents, *i*, which is a non-RCLI, *ochii*, and *din cap*
- (SV) *att sätta sig upp mot någon* lit. *to sit oneself up against someone* 'To defy someone' → two dependents, *sig* and *upp*
- (SL) *skrivati glavo v pesek* lit. *to hide head in sand* 'to pretend not to see a problem' → two dependents, *glavo* lit. *head* and *v pesek* lit. *in sand*
  - *vlečeš me za nos* lit. *you are pulling my nose* 'you're pulling my leg' → two dependents, *me* lit. *me* and *za nos* lit. *my nose*

The test covers only lexicalized dependents. There may be other, non-lexicalized dependents, which the test ignores. We explicitly call the non-verbal elements *dependents* 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.

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 clitic** - apply IReflV tests. If the outcome is negative, discard the VMWE candidate.
  - (BG) *страхувам се* lit. *fear myself.REFL* 'to be afraid'
    - *радвам се* lit. *feel joy myself.REFL* 'to feel joy'
  - (DE) *sich wundern* 'to wonder', *sich schämen* 'to be ashamed'
  - (EN) English does not have IReflV expressions
  - (FR) *se suicider* 'to suicide', *s'évanouir* 'to faint'
  - (IT) *suicidarsi* 'to suicide', *vergognarsi* 'to be ashamed'
  - (PL) *bać się* lit. *fear SELF* 'to be afraid'
  - (PT) *suicidar-se* 'to suicide', *queixar-se* 'to complain'
  - (RO) *a se sinucide* 'to commit suicide' with obligatory ACC reflexive clitic
    - *a se holba* 'to stare' with obligatory ACC reflexive clitic
  - (SL) *čuditi se* 'to wonder', *smejati se* 'to laugh', *onesvestiti se* 'to faint'
- **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 an* 'he begins', *er hat angefangen* 'he has begun' → in German, VPCs may occur separated or within one word, we annotate all occurrences!
    - *ich schlage vor* 'I propose'
  - (EL) *παίρνω μπρος, βάζω μπρος να, κάνω πίσω*
  - (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) *gre za* 'it is about'
    - *dati skozi* '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) *ритам камбаната* lit. *kick the bell* 'to die' → камбаната is a noun phrase composed of a single noun
    - *давам зелена светлина* lit. *give green light* 'to allow' → зелена светлина is a noun phrase composed of an adjective and a noun
    - *ставам на кайма* lit. *turn into mince* 'to be destroyed' → на кайма is a prepositional phrase composed of a preposition governing a noun
  - (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
    - *in Kraft treten* lit. *into*
  - (EN) *to make a wish* → *a wish* is a noun phrase composed of a determiner and a noun
    - *to take turns* → *turns* is a noun phrase composed of a single 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
    - *chodzić prostą drogą* lit. *to go (on) a straight road.INST* 'to avoid complications' → *prostą drogą* '(on) a straight road' is a noun phrase composed of an adjective and a noun in (instrumental)
    - *bujać w obłokach* lit. *to swing in the clouds* 'to fantasize' → *w obłokach* lit. *in the clouds* is a prepositinal phrase composed of a preposition and a noun
  - (PT) *tomar banho* 'to take a shower' → *banho* is a noun phrase composed of a single noun
  - (RO) *a rupe tăcerea* lit. *to break silence the* 'to start talking' → *tăcerea* is a noun phrase composed composed of a single noun
    - *a face baie* lit. *to do bath* 'to take a shower' → *baie* is a noun phrase composed of a single noun
  - (SL) *biti v dvomih* 'to doubt' → *v dvomih* lit. *in doubt* is a prepositional phrase composed of a preposition governing a noun
- **Other** - categorize as ID.
  - Adjective:
    - (BG) *излизам сух от водата* lit. *to come out dry from the water* 'to avoid taking responsibility'
      - *одирам жив* lit. *skin alive* 'to make somone suffer'
      - *гоня дивото* lit. *chase the wild.ADJ* 'to take risks' → *дивото* is a substantive
    - (DE) *rot sehen* 'to see red'
    - (EL) *τα βάφω μαύρα* 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'
    - (IT) *vedere nero* 'to see black'
    - (PL) *zrobić swoje* lit. *to do one's own* 'to do what one is supposed to do'
    - (PT) *pensar grande* 'to think big'

- (RO) *a vedea <u>roşu</u>* 'to see red'
  - *a o face <u>lată</u>* lit. *to CL.ACC make wide* 'to party'
- (SL) *narediti <u>svoje</u>* lit. *to do one's own* 'to do what one is supposed to do'

- **Verb**:
  - (BG) *не искам <u>и да чуя</u>* lit. *don't want to even hear* 'to oppose strongly' → *и да чуя* is a VP
    - *правя <u>сам да си говори</u>* lit. *make someone talk to himself* 'to drive someone crazy' → *сам да си говори* is a clause
  - (DE) *??*
  - (EN) *to make <u>do</u>*
  - (FR) *laisser <u>tomber</u>* lit. *let fall* 'to let down'
    - *vouloir <u>dire</u>* lit. *want say* 'to mean'
  - (IT) *lasciar <u>andare</u>* lit. *let go* 'to unhand'
    - *voler <u>dire</u>* lit. *want say* 'to mean'
  - (PL) *dać komuś <u>popalić</u>* lit. *to let someone smoke* 'to make someone's life hard'
  - (PT) *querer <u>dizer</u>* lit. *want say* 'to mean'
  - (SL) *<u>zagosti</u> jo komu* lit. *to play music to someone* 'play a joke on someone '

- **Adverb**:
  - (BG) *изваждам <u>наяве</u>* lit. *take out in the open* 'to uncover'
    - *хващам <u>натясно</u>* lit. *catch in a tight place* 'to coerce, to pressure'
  - (EL) *φέρω<u>βαρέως</u>*
  - (EN) *to get <u>well</u>*
  - (IT) *buttare <u>giù</u>* lit. *to throw down* ' to demoralize '
  - (PL) *chcieć <u>dobrze</u>* lit. *to want well* 'to have good intentions'
    - *robić komuś <u>dobrze</u>* lit. *to do someone.DAT well* 'to please someone'
    - *<u>źle</u>/<u>marnie</u> skończyć* lit. *badly finish* 'to come to a bad end'
  - (PT) *cair*
    - *bem*
    - lit. *fall well* 'to be appropriate'
  - (RO) *a se face <u>bine</u>* lit. *to himself make well* 'to get well'
    - *a face <u>bine</u>* lit. *to make well* 'to help'
  - (SL) *obrniti se <u>na bolje</u>* lit. *to turn for better* 'to be better'

- **Pronoun**:
  - (BG) *мързи <u>ме</u>* lit. *(it feels) lazy me.ACC* 'to be lazy'
  - (EL) *τα<u>καταφέρνω</u>* ' to make it '
    - *την<u>πατάω</u>* ' to fail '
  - (EN) *to make <u>it</u>*
  - (FR) *le faire* lit. *it make* 'to be enough/successful'
  - (IT) *far<u>cela</u>* lit. *to make it* ' to manage '
  - (PL) No example found in Polish
  - (PT) *dá-<u>lhe</u> João!* lit. *give to him/her, João!* 'show them what you got, João!'
  - (RO) *a o lua la <u>măsea</u>* lit. *CL.ACC.F.3SG take PREP tooth.ACC* 'to drink heavily' with the non-anaphoric feminine clitic 'o' functioning as an expletive

- Etc.

Section 6.2

Light verb constructions (LVC)

Light verb constructions (LVC) constitute a <u>universal</u> 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) **вземам решение** 'to make a decision'
     - **държа под контрол** 'to keep under control'
   - (DE) *zum **Einsatz** kommen* lit. *to the use come* 'to be called into action'
     - *eine **Rede** halten* lit. *a speech hold* 'to give a speech'
   - (EL) *παίρνω μία **απόφαση*** 'make a decision '
     - *δίνω στα **νεύρα*** ' give to the nerves'
   - (EN) *to **give** a lecture*
     - *to **come** into bloom*
   - (FR) ***faire** une **présentation*** lit. *make a presentation*
   - (IT) ***chiamare** in causa* lit. *to call in cause* 'to single out'
     - *fare una **passeggiata*** lit. *to make a walk* 'to have a walk'
   - (PL) **odnieść sukces** lit. *carry-away success* 'to be successful'
   - (PT) ***fazer** um **aborto*** 'to make an abortion'
     - ***estar** com **fome*** lit. *be with hunger* 'to be hungry'
   - (RO) *a duce <u>dorul</u>* lit. *to carry yearning.the* 'to miss somebody'
     - *a da <u>divorţ</u>* lit. *to give divorce* 'to divorce'
     - *a da <u>în clocot</u>* lit. *to give in boil* 'to come to the boil'
     - *a da <u>în fiert</u>* lit. *to give in boil* 'to come to the boil'
   - (SL) ***biti v dvomih*** 'to doubt'
     - ***imeti predavanje*** 'to give a lecture'

2. The noun *n* is <u>predicative</u>, ofter referring to an event (e.g. *decision, visit*) or a state (e.g. *fear, courage*).
   - (BG) **вземам решение** 'to make a decision' → noun refers to an act or event
     - **давам съгласие** 'to give permission' → noun refers to an act or event
     - **имам притеснения** 'to have concerns' → noun refers to a feeling or state
     - **имам готовност** 'to be ready' → noun refers to a feeling or state
   - (DE) *eine **Entscheidung** treffen* 'to make a decision' → noun refers to an event
     - ***Angst** haben* 'to have fear' → noun refers to a state

- (EL) *παίρνω* μία *απόφαση*, *κάνω* *βόλτα* → noun refers to an event
  έχω *αγωνία*, *κάνω* *κουράγιο* → noun refers to a state
- (EN) *to make* a *decision*, *to pay* a *visit* → noun refers to an event
  to have *fear*, *to have* *courage* → noun refers to a state
- (FR) *donner* un *conseil* 'give advice'→ noun refers to an event*avoir 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*'to lead negotiations'→ the noun refers to an event
  mieć rację lit. *to have right*'to be right'→ thre noun refers to a state
- (PT) *fazer* uma *prece* 'to make a prayer' → noun refers to an event
  ter *sintomas* 'to have symptoms' → noun refers to a state
- (RO) *a lua o decizie* 'to make a decision', *a face o vizită* 'to pay a visit'→ noun refers to an event
  a avea *curaj* → noun refers to a state
- (SL) *imeti* *pogum* '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. It may be "light" either per se, or when used in the specific context of the noun. This implies that *v*'s syntactic subject ❶ is *n*'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 that 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) *вземa решение* → *решението, което президентът* *взе* 'the decision that the president made'
- (DE) *eine Entscheidung treffen* → *die Entscheidung die der Direktor zu treffen hatte.*
- (EL) *παίρνωμία* *απόφαση* → *η* *απόφαση* που πρέπει κάποιος να *πάρει.*
- (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. *take shower* → *o banho* que eu *tomei* estava bom lit. *the shower which I took was good*
- (RO) *a lua o decizie*'to make a decision' → *decizia* pe care directorul trebuie să o *ia* 'the decison that the director has to make'.
- (SL) *sprejeti odločitev* 'to make a decision' → *odločitev, ki jo je moral sprejeti* direktor 'the decison 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) *вземам решение* 'to make a decision' vs *вземам отговорност* 'to take responsibility'
  имам право 'to be right' vs *притежавам право*
- (DE) *eine Entscheidung treffen* lit. *a decision meet* 'make a decision'vs.*eine Entscheidung machen* lit. *a decision make* vs. *einen Beschluss treffen* lit. *a resolution meet*
- (EL) *κάνω διάλειμμα* vs. *παίρνω διάλειμμα*
- (EN) *have a walk* vs *have a race*
  run a race vs *run a walk*
- (FR) *faire une marche* 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'vs.*fareuna decisione* lit. *make a decision* vs. *prendere una conclusione* lit. *take a conclusion*
- (PL) *wziąć udział* lit. *to take participation* vs. *wziąć uczestnictwo*
  mieć 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/dar uma caminhada* lit. *to make/give a walk*
- (RO) *a da divorț* lit. *to give 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 achievement). Therefore, we chose not to retain the selection of the verb as a criterion ❶ 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 12 - [**V-REDUC**: *Subj+v+n transformable to subj's n?*]
        ↳ **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 ⓘ argument?

↳ **YES** ⇒ continue to next test
- (BG) *поставям акцент* 'to emphasize' → event, with two arguments: the agent and the object being emphasized
  имам право → property, with one semantic argument: the possessor of the property
- (DE) *einen Besuch abstatten* 'to pay a visit' → event, with two arguments: the visitor and the visitee
  Angst haben 'to have fear' → property with one semantic argument: the entity having fear
  einen Blick auf etwas werfen lit. *a glance at sth. throw* 'to take a glance at sth' → an event with two arguments the entity glancing and the entity glanced at
- (EL) *κάνω μία επίσκεψη* lit. *to-make a visit* ' pay a visit, visit' → event, with two arguments: the visitor and the visitee
  έχω τη δυνατότητα lit. *to-have the ability* ' to be able ' → property, with one core semantic argument: the entity having the ability
  έχω μίσος → state, with two arguments: the entity being in state hate and the entity hated
- (EN) *pay a visit* → event, with two arguments: the visitor and the visitee
  have strength → property, with one semantic argument: the entity having strength
  take a glance at something → event, with two arguments: the entity glancing and the entity glanced at
- (FR) *avoir du courage* 'to have courage'→ state(property), with one argument: the entity having courage
- (IT) *fare una visita* → event, with two arguments: the visitor and the visitee
  avere forza → property, with one semantic argument: the entity having strength
  dare uno sguardo a qualcosa → event, with two arguments: the entity glancing and the entity glanced at
- (PL) *złożyć wizytę* lit. *to submit a visit* 'to pay a visit'→ event, with two arguments: the visitor and the visitee
  złożyć skargę lit. *to submit a complaint* 'to make a complaint' → event, with two arguments: the complaining person and the one he/she complains about
- (PT) *ter fome* lit. *to have hunger* 'to be hungry' → property, with one argument: the entity that is hungry
  ter idade para fazer algo lit. *to have age (to do something)* 'to be old enough (to do something)' → state, with one argument: the entity that is old enough
  We include as states and events (predicative nouns) the following classes: diseases (*gripe, trombose, infarto*), physical sensations (*fome, sede, sono*), emotions (*medo, paixão, nojo*), cognitive entities (*ideia, opinião, preocupação*), characteristics (*coragem, teimosia, fraqueza*), relations (*contato, conflito, amizade*) and communication (*conversa, discussão, briga*).
- (RO) *a face o vizită* lit. *to make a visit* 'to pay a visit' → event, with one argument: the entity that visits
  a avea curaj 'to have courage' → property, with one semantic argument: the entity having courage
- (SL) *imeti predavanje* 'to give a lecture' → event, with two arguments: a lecturer and the people who are attending the lecture

↳ **NO** ⇒ it is not an LVC
- (BG) *Иван хвърли боклука* 'Ivan 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 a semantic argument
- (EN) *Joe makes a cake* → physical entity (not event/state), even though *Joe* could be considered its semantic argument
  Joe experienced a tornado → event, but has no 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
  Joe ha vissuto un tornado → event, but has no semantic argument
- (PL) *złożyć kartkę* lit. *to fold a sheet*→ physical entity (not event/state)
  bić pianę lit. *to beat foam* 'to exaggerate about a problem'→ *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 prăjitură* '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) *поемам отговорност* 'to take responsibility' → the noun is literally understood
- (DE) *einen Besuch abstatten* → noun is literally understood
- (EN) *pay a visit* → noun is literally understood
- (FR) *rendre visite* → 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
  pobić rekord lit. *to beat a record* 'to break a record'→ the noun *rekord* is literally understood
- (PT) *tomar banho* → noun is literally understood
- (RO) *a face o vizită* lit. *to make a visit* 'to pay a visit' → noun is literally understood
- (SL) *imeti pogum* 'to have courage' → noun is literally understood

↳ **NO** ⇒ it is not an LVC
- (BG) *яхвам метлата* lit. *to get on the broom* 'to get very angry' → noun is not literally understood (it's an ID)
- (DE) *Herzklopfen haben* lit. *heartbeating have* 'to be in love' → noun is not used in one of its normal senses (it's an ID)
- (EN) *have kittens* '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. *have the heartbeating* 'to be in love' → noun is not used in one of its normal senses (it's an ID)
- (PL) *nadstawiać karku* lit. *to expose neck.GEN* 'to take personal risks'→ the noun *kark* lit. *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) *a face față* lit. *to make face* 'to 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 ⓘ 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) *взмам решение* 'make a decision' → *вземам* adds no meaning to *решение* 'decision' besides that of *performing an act*
  - *държа реч* 'to make a speech' → *държа* adds no meaning to *реч* besides that of *performing an act*
  - *поемам отговорност* 'to take responsibility' → *поемам* adds no meaning to *отговорност* besides that of *having a property*
- (DE) *eine Entscheidung treffen* lit. *a decision meet* 'to make a decision' → *treffen* adds no meaning to *Entscheidung* besides that of *performing an activity*
  - *Angst haben* 'to have fear' → *haben* adds no meaning to *Angst* besides that of *having a property*.
- (EN) *take a walk* → *take* adds no meaning to *walk* besides that of *performing an activity*
  - *make a decision* → *make* adds no meaning to *decision* besides that of *performing an activity*
  - *have fear* → *have* adds no meaning to *fear* besides that of *having a property*
  - *perform a check* → *perform* is a pure syntactic operator: in any context, it only bears tense and mood and never adds any sense to the noun
  - *commit a crime* → *commit* is a pure syntactic operator: in any context, it only bears tense and mood and never adds any sense to the noun
  - *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 place to 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*
- (FR) *ils ont du courage* lit. *they have some courage* → *have* adds no meaning to *courage* besides that of *having a property*
  - *ils reçoivent l'ordre de partir* lit. *they receive the order of leaving* 'they are ordered to leave' → *receive* adds no meaning to *order* besides indicating that the subject is the recepient of the *order*
  - *il a subi une intervention chirurgicale* lit. *he has undergone an intervention surgery* 'he underwent surgery' → *undergo* adds no meaning to *surgery* besides indicating that the subject is the patient of the *surgery*
- (IT) *fare una passeggiata* → *fare* adds no meaning to *passeggiata* besides that of *performing an activity*
  - *prendere una decisione* → *prendere* adds no meaning to *decisione* besides that of *performing an activity*
  - *avere paura* → *avere* adds no meaning to *paura* besides that of *having a property*
  - *eseguire un controllo* → *eseguire* is a pure syntactic operator: in any context, it only bears tense and mood and never adds any sense to the noun
  - *commettere un crimine* → *commettere* is a pure syntactic operator: in any context, it only bears tense and mood and never adds any sense to the noun
  - *fare una visita* → 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
  - *fare un discorso* → the verb in its usual sense means 'to make', but here it is not used in this sense and does not add any semantics to the "speech" event
- (PL) *oddać hołd* lit. *to give-back tribute* 'to pay tribute' → *oddać* lit. *give-back* adds no meaning to *hołd* 'tribute' besides that of *performing an activity*
  - *wystąpić z wnioskiem* lit. *to stand out with a proposal* 'to put forward a motion' → *wystąpić z* lit. *stand out with* adds no meaning to *wniosek* 'motion' besides that of *performing an activity*
- (PT) *mover uma ação judicial* lit. *to move a lawsuit* 'to sue' → *to move* adds no meaning to *lawsuit* besides that of *performing an activity*
  - *apresentar uma lesão* lit. *present a lesion* 'to have a lesion' → *to present* adds no meaning to *lesion* besides that of *having a property*
  - *estar com medo* lit. *be with fear* 'to be afraid' → *to be with* adds no meaning to *fear* besides that of *being in a state*
- (RO) *a avea curaj* lit. *to have courage* → *avea* adds no meaning to *curaj* besides that of *having a property*
  - *a lua o decizie* 'to make a decision' → *lua* adds no meaning to *decizie* besides that of *performing an activity*
- (SL) *imeti pogum* 'to have courage' → *imeti* 'have' adds no meaning to *pogum* 'courage' besides that of *having a property*
  - *sprejeti odločitev* 'to make a decision' → *sprejeti* in its usual sense means 'to receive', but here it is not used in this sense and does not add any semantics to event

↳ **NO** ⇒ it is not an LVC
- (BG) *започвам играта* 'start the game, start playing' → *започвам* 'start' adds an aspectual meaning to the noun
- (DE) *eine Rede beginnen* 'to begin a speech' → *beginnen* adds an aspectual meaning to the noun *Rede*
- (EN) *to start a walk* → *start* adds an aspectual meaning to the noun
- (FR) *donner du courage* 'to give courage'
  - *donner son avis* 'to give one's opinion' → *donner* adds the information that the opinion is communicated
- (IT) *cominciare un ballo* 'to start a dance' → *cominciare* adds an aspectual meaning to the noun *ballo*
- (PL) *wymierzyć sprawiedliwość* lit. *to measure justice* 'to do justice' → *wymierzyć* lit. *measure* adds an aspectual meaning to *sprawiedliwość* 'justice'
  - *przejść na emeryturę* lit. *to cross to retirement* 'to take retirement' → *przejść* adds an inchoative (change-of-state) meaning to the noun
- (PT) *entrar com uma ação judicial* lit. *to enter with a lawsuit* 'to file a lawsuit' → *to enter* adds an aspectual meaning to the noun
  - *enfrentar uma ação judicial* lit. *to face a lawsuit* → *enfrentar* introduces the *patient* of a lawsuit, not the *agent*
  - *dar uma opinião* 'to give an opinion' → *to give* adds the meaning of *communication* which is not present in the name itself (one can **ter uma opinião** 'to have an opinion' without ccommunicating it).
- (RO) *a începe munca* lit. *to start work the* 'to start working' → *începe* adds an aspectual meaning to the noun

Note that this light semantics of the verb is either usual for that verb (i.e. the verb is a pure syntactic operator, like *commit*, *perform*), or happens in the context of the particular noun (e.g. for *pay* in *to* **pay** *a* **visit**).

Test 12 - [V-REDUC] - Verb reduction

Can an NP in which *v*'s subject becomes *n*'s dependent evoke the same event or state as the candidate construction does?

↳ **YES** ⇒ continue to next test
- (BG) *Иван* **пое** <u>отговорност</u> 'Ivan took responsibility' → *отговорността на Иван* — both refer to the same property/event
  - *Иван* **взе** <u>решение</u> 'Ivan made a decision' → *решението на Иван* — both refer to the same property/event
- (DE) *Paul hat eine* <u>Rede</u> *gehalten* 'Paul has given a speech' → *Paul's speech* both refer to the same speech event
  - *Ich habe ihm einen* <u>Besuch</u> *abgestattet* 'I have paid him a visit' → *mein Besuch* 'my visit' both refer to the same visiting event
- (EN) *Paul had a* <u>walk</u> → *Paul's walk* — both refer to the same walking event
  - *I* **paid** *him a* <u>visit</u> → *my visit* — both refer to the same visiting event
- (FR) *Paul a* **fait** *une* <u>enquête</u> 'Paul made an inquiry' → *L'enquête de Paul* 'Paul's inquiry'
  - *Paul* **procède à** *une* <u>perquisition</u> 'Paul makes a search' → *La perquisition de/par Paul* 'the search of/by Paul'
  - *Le général* **donne** *l'*<u>ordre</u> *de partir* lit. *The general gives the order to leave* 'The general orders to leave' → *l'ordre du général de partir* 'The general's order to leave'
  - *Les soldats* **reçoivent** *l'*<u>ordre</u> *de partir* lit. *The soldiers receive the order to leave* 'The soldiers are ordered to leave' → *l'ordre aux soldats de partir* 'The order to the soldiers to leave'
  - *Jean* **souffre de** <u>troubles</u> *psychiques* 'John suffers from psychic troubles' → *Les troubles psychiques de Jean* 'John's psychic troubles'
  - *Jean* **présente** *une* <u>hypersensibilité</u> lit. *John presents a hypersensibility* 'John has a hypersensibility' → *l'hypersensibilité de Jean* 'John's hypersensibility'
  - *Paul* **reçoit** *des* <u>menaces</u> *de (la part de) Pierre* 'Paul receives threats from (the part of) Peter' 'Paul is threatened by Peter' → *les menaces de Pierre à Paul* 'Peter's threats to Paul'
  - *Ce médicament* **présente** *un* <u>risque</u> lit. *This medicine presents a risk* 'This medicine poses a risk' → *le risque de ce médicament* 'this medicine's risk'

- (IT) *Paolo ha **fatto** una <u>conquista</u>* 'Paul made a conquer' → *la conquista di Paolo*
    - *Il generale **da** l' <u>ordine</u>di partire.* **lit. *The general gives the order to leave*** 'The general orders to leave ' → *L'ordine di/da parte del generale di partire*
    - *Paolo **riceve** delle <u>minacce</u> da (parte di) Piero* → *le minacce di Piero a Paolo*
- (PL) *Obecni **oddali** <u>hołd</u> poległym* lit. *The present gave-back tribute to the fallen* 'The audience payed tribute to the fallen' → *hołd obecnych* 'the tribute of the audience'
    - *Jan **miał na myśli** Marię* lit. *Jan had on thought Maria* 'Jan meant Maria' → *myśl Jana* 'Jan's thought'
    - *Jan **otrzymał wymówienie**elit. Jan received a dismissal* → *wymówienie dla Jana* 'dismissal for Jan'
    - *Inwestycja **przynosi zyski*** 'the investment brings profit' → *zyski z inwestycji* 'profit from the investment'
- (PT) *João **cometeu** um <u>deslize</u>* → *o deslize do João* — both refer to the same event
    - *O jogador **cobrou** um <u>pênalti</u>* lit. *the player charged a penalty kick* 'the player took a penalty kick' → *o pênalti do jogador* lit. *the player's penalty kick* — both refer to the same event
    - *João **tem** <u>consciência</u> do perigo* lit. *John has conscience of the danger* 'John is aware of the danger' → *a consciência do João sobre o perigo* 'John's awareness of the danger' — both refer to the same state
    - *João **recebeu** a <u>remuneração</u>* 'John received the remuneration' → *a remuneração do João* 'John's remuneration' — both refer to the same event
    - *O paciente **recebeu** a <u>visita</u> dos familiares* 'The patient received the visit of the relatives' → *a visita dos familiares ao paciente* 'the visit of the relatives to the patient' — both refer to the same event
    - *João **apresenta** <u>lesões</u>* lit. *John presents lesions* → *as lesões do João* 'John's lesions' — both refer to the same state
- (RO) *Paul **a făcut** o <u>plimbare</u>* 'Paul had a walk' → *plimbarea lui Paul* 'Paul's walk' — both refer to the same walking event
    - *i-am **făcut** o <u>vizită</u>* 'I paid him a visit' → *vizita mea* — both refer to the same visiting event
- (SL) *Zdravnik je **postavil** <u>diagnozo</u>* 'The doctor made a diagnosis' → *njegova diagnoza* 'His diagnosis' both refer to the same event
↳ **NO** ⇒ it is not an LVC
- (BG) *Иван **хвърли** <u>поглед</u> на вестника* 'Ivan threw a glance at the newspaper' → *#погледът на Иван върху вестника* — different semantics; and requires a different preposition
- (DE) *Paul hat einen guten <u>**Eindruck**</u> **gemacht**'Paul has made a good impression' → *#Paul's Eindruck auf seine Freunde* 'Paul's impression on his friends' has a different semantics
- (EN) *Paul **made** a good <u>impression</u>* → *#Paul's impression on his wife* — different semantics
- (FR) *Son comportement **porte** une <u>atteinte</u> grave à l'honneur des soldats* 'His behaviour seriously jeopardises the soldiers' honnour' → *#l'atteinte de son comportement* 'the jeopardy of his behaviour'
    - *Ce fait **attire** l'<u>attention</u> de la justice* 'This fact attracts the attention of the justice' → *?l'attention de la justice pour/sur ce fait* 'the attention of the justice on/about this fact'
- (PL) *Michael Phelps **pobił** <u>rekord</u> sprzed 2 tysięcy lat*'Michael Phelps broke the record from 2 thousand years ago' → *#Michael Phelps' record*
    - *Ulica **nosi** <u>imię</u> sławnego poety* lit. *The street carries the forename of a famous poet* 'The street carries the name of a famous poet.'→ *imię ulicy* 'the forename of the street'
- (PT) *O jogador **cobrou** uma <u>falta</u>* lit. *the player charged a foul* 'the player took a free kick' → *a falta do jogador* lit. *the player's foul* — the focus changes from taking a free kick to being one of the parts involved in a foul (it's an ID)
    - *O jogador **provocou** uma <u>lesão</u>* lit. *the player provoked a lesion* → *a lesão do jogador* lit. *the player's lesion* — In the reduced NP, the focus changes from hurting somebody else to getting hurt
    - *O músico **apresenta** suas <u>composições</u>* lit. *the musician presents his compositions* → *as composições do músico* lit. *the musician's compositions* — the reduced NP does not keep the sense of presenting, it is not refer to the same event as the verbal construction
- (RO) *Paul **a făcut** o <u>impresie bună</u>*'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

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 *r* be the semantic role that *s* plays with respect to the noun *n*. Is it prohibited for *r* 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 tests is that a semantic argument *n* cannot be realized as its syntactic dependent, since it is already realized as *v*'s 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 *"the visit of the Queen to the Prime Minister"*. When used in *to **pay** a **visit***, the visitor semantic argument is realized as the subject of *to pay* (*The Queen **paid** a **visit** to the Prime Minister*), and cannot be realized at the same time within the NP headed by *visit* (**The Queen paid a visit of the Lady to the Prime Minister*).

↳ **YES** ⇒ it is an LVC
- (BG) *Петър Стоянов **взе** <u>решението</u> да подпише договора* 'Ivan made the decision to sign the contract' + *<u>решението на президента</u> да подпише договора* → **Петър Стоянов <u>взе решението на президента</u> 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 <u>**Besuch**</u> **abgestattet*** 'the Queen has paid a visit to the Prime Minister' + *<u>ein Besuch der Dame</u> beim <u>Premierminister</u>* 'a visit of the Lady to the Prime Minister' **Die Königin hat <u>einen Besuch der Dame</u> beim <u>Premierminister</u> abgestattet*'*The Queen paid <u>a visit of the Lady</u> to the Prime Minister' — the visitor cannot be a modifier of *visit*
    - *Paul hat eine <u>**Entscheidung**</u> über das Budget **getroffen*** 'Paul made a decision on the budget' + *die <u>Entscheidung des Rates</u> über das Budget* 'the council's decision on the budget' → **Paul traf <u>die Entscheidung des Rates</u> über das Budget* '*Paul made the committee's decision on the budget' — the decision maker cannot modify *decision*
- (EN) *The Queen **paid** a <u>visit</u> to the Prime Minister* + *<u>a visit of the Lady</u> to the Prime Minister* → **The Queen paid <u>a visit of the Lady</u> to the Prime Minister* — the visitor cannot be a modifier of *visit*
    - *Paul **made** a <u>decision</u> on the budget* + *<u>the committee's decision</u> on the budget* → **Paul made <u>the committee's decision</u> on the budget* — the decision maker cannot modify *decision*
    - *Paul **leads** the <u>discussion</u>* + *<u>Peter's discussion</u>* → **Paul leads <u>Peter's discussion</u>* but Paul leads the <u>discussion of the committee</u> — the discussing entity can only modify *discussion* when the subject *Paul* is part of the *committee*
    - *Bjarnson **scored** a <u>goal</u>* + *<u>Arnason's goal</u>* → **Paul scored <u>Arnason's goal</u>* but Paul scored the <u>goal of Iceland</u> — the scoring entity can only modify *goal* in the last case, when they are part of the *Iceland* team
- (FR) *La ministre a **rendu** une <u>visite</u> aux victimes* + *<u>la visite de la ministre</u> aux victimes* → **La ministre a rendu une <u>visite du président</u> aux victimes* — the visitor cannot be a modifier of *visite*
    - *Bjarnson a **marqué** un <u>but</u>* + *<u>le but d'Arnason</u>* → **Paul a marqué <u>le but d'Arnason</u>* but Paul a marqué <u>le but de l'Islande</u> — the scoring entity can only modify *but (goal)* in the last case, when they are part of the *Iceland* team
- (IT) *Il primo ministro ha **preso** la <u>decisione</u> di dimettersi* ' the Prime Minister decided to resign '+ *<u>le dimissioni del governo</u>* 'the resignation of the government' → **Il primo ministro ha preso <u>la decisione del governo</u> di dimettersi* — the resigner cannot be a modifier of *resignation*
- (PL) *Paweł **złożył rezygnację** ze stanowiska dyrektora* lit. *Paweł submitted a resignation from the position of the director* 'Paweł tendered his resignation from the director position' + *rezygnacja Piotra*→ **Paweł **złożył rezygnację** Piotra ze stanowiska dyrektora* 'Paweł tendered Piotr's resignation from the director position' - the resignation cannot be modified by the resigning person

Paweł prowadzi rozmowy → **Paweł prowadzi rozmowy Piotra* 'Paweł leads Piotr's talks' , *Paweł prowadzi rozmowy komisji* 'Paweł leads the talks of the commission' - the discussing entity *komisja*'commission' can only modify *rozmowy*'talks' if *Paweł* belongs to the commission.

*Jan otrzymał wymówienie*lit. *Jan received a dismissal* + *wymówienie dla Pawła* 'dismissal for Paweł' → **Jan otrzymał wymówienie dla Piotra*

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

    Pedro sofreu prejuízo com a compra lit. *Pedro suffered finantial loss with the purchase* + *o prejuízo do José* lit. *José's finantial loss* → **Pedro sofreu o prejuízo do José com a compra* — the *financial loss* cannot be modified by the affected entity

    A Maria fez um aborto lit. *Maria made an abortion* + *o aborto da Joana* lit. *Joana's abortion* → *#A Maria fez o aborto da Joana* — the noun cannot be modified by another patient

    O médico realizou o parto com sucesso lit. *The doctor performed the childbirth with success* + *o parto do Dr. Pedro* lit. *Dr. Smith's childbirth* → **O médico realizou o parto do Dr. Pedro com sucesso* — the *childbirth* could be modified by the mother (patient) but **not** by another doctor (agent).

- (RO) *Paul a dat sfaturi surorii sale*'Paul gave advice to his sister' + *sfatul lui Petre* 'Peter's advice' → *Paul a dat sfatul lui Petre surorii sale* 'Paul gave Peter's advice to his sister' — *sfatul* 'the advice' **cannot** be modified by its author

- (SL) *Učiteljica je sprejela odločitev v zvezi z nalogo* 'The teacher made a decision regarding the homework' + *dijakova odločitev v zvezi z nalogo* 'pupil's decision regarding the homework' → **učiteljica je sprejela dijakovo odločitev v zvezi z nalogo* — the decision maker cannot modify *decision*

↳ **NO** ⇒ it is not an LVC

- (BG) *Иван предаде решението на сестра су* 'Ivan transmitted the decision to his sister' + *решението на комисията* → *Иван предаде решението на комисията на сестра су* 'Ivan transmitted the decision of the comission to his sister' — the noun can be modified by the person performing the act/event (which can be different from the subject)

- (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

- (EN) *Paul transmitted the advice to his sister* + *Peter's advice* → *Paul transmitted Peter's advice to his sister* — the *advice* **can** be complemented by its author

- (FR) *Paul a transmis l'ordre aux soldats* 'Paul transmitted the order to the soldiers'+*l'ordre de Paul aux soldats* 'Paul's order to the soldiers' → *Paul a transmis l'ordre du général aux soldats* 'Paul transmitted the general's orders to the soldiers' — l' *ordre* **can** have as a complement the person who gave it

- (PL) *Paweł podważył niedawną decyzję*'Paweł questioned the recent decision' +*decyzja Piotra*'Piotr's decision'→ Paweł podważył niedawną decyzje PiotraPaweł questioned Piotr's recent decision — the decision **can** be modified by the decision maker

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

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

- (SL) *Janez mi je povedal mnenje o filmu* 'Janez told me his opinion on the movie'+ *Ninino mnenje o filmu* ' Nina's opinion on the movie' → *Janez mi je povedal 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

---

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) *броят му се ребрата* lit. *be counted someone's (possessive pronoun) ribs* '(someone) to be very thin and skinny'
    - (DE) *ein kleines Vöglein hat mir gezwitschert* 'a little bird told me'
    - (EN) *a little bird told* someone
    - (ES) *tu hora ha llegado* lit. *your time has arrived* 'your time has come'
    - (PL) *licho wie* lit. *devil knows*'I have no idea'
    - (PT) *a sua hora chegou* lit. *your time has arrived* 'your time has come'
    - (RO) *a șoptit o păsărică*lit. *whispered a bird little* 'a little bird told someone'
    - (SL) *srce mu je padlo v hlače* lit. *His heart fell into his pants* 'someone has lost courage'
- Direct object
    - (BG) *гушна букета* lit. *hug the bunch of flowers* 'to die'
    - (DE) *er hat den Schuss nicht gehört* lit. *he did the shoot not hear* 'it takes him a long(er) time to understand sth'
    - (EN) *to kick the bucket*
    - (ES) *estirar la pata* lit. *to strech the leg* 'kick the bucket'
    - (PL) *udać Greka* lit. *to pretend to be a Greek*'to pretend not to understand'
    - (PT) *bater as botas* lit. *to hit the boots* 'to die', *abrir mão de algo* lit. *to open hand (of something)* 'to give up (on something)'
    - (RO) *a arunca vina* lit. *to throw guilt-the* 'to blame'
    - (SL) *ustreliti kozla* lit. *to shoot the goat* 'to say or do something stupid'
- Circumstantial or adverbial complement
    - (BG) *удрям в гръб* lit. *hit in the back* 'to stab in the back'

        правя сам да си говори lit. *make (someone) to talk to himself* 'to drive (someone) crazy'
    - (DE) *etwas wie warme Semmeln verkaufen* lit. *sth. like warm bread rolls* 'to sell sth. fast and easy'
    - (EN) *to take* something *with a pinch of salt*, *to sell like hotcakes*, *to strike while the iron is hot*, *to come off with flying colors*
    - (ES) *coger* algo *con pinzas* lit. *to hang something with pegs* 'take something with a pinch of salt'
    - (PL) *wiercić* komuś *dziurę w brzuchu* lit. *to drill a hole in one's belly*'to intrusively solicit someone, to insist too much'
    - (PT) *levar em conta* lit. *to bring in account* 'to take into account'

        ir ao ar lit. *go to the air* 'to go on air'
    - (RO) *a lua în considerare* lit. *to bring in account* 'to take into account'
    - (SL) *spati kot ubit* lit. *to sleep like dead* 'to sleep soundly'

It is often challenging to distinguish IDs from other VMWE categories 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 IRefIV (reflexive pronoun), never an ID.

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

- Preposition governing a complex noun phrase
    - (BG) *удрям някого в гръб* lit. *hit someone in the back* 'to stab someone in the back'
    - (DE) *jmd 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* 'take something with a pinch of salt'
    - (PL) *dopiąć coś na ostatni guzik* 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 piciorul* 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) *schwarz 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'
        - *tykać cudze* lit. *to touch someone else's* 'to take something that does not belong to you'
        - *dopiąć swego* lit. *to button up one's own* 'to fulfill one's plans'
    - (PT) *to jogar sujo* lit. *to play dirty*
    - (RO) *a juca murdar* lit. *to play dirty*
    - (SL) *biti zelen od zavisti* 'to be green with envy'
- 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'
    - (PL) *dać komuś żyć* lit. *to let someone live* 'not to bother someone'
        - *można wytrzymać* lit. *one can stand* 'the situatiion is reasonably good'
    - (PT) *querer dizer* lit. *to want to say* 'to mean'
    - (RO) n.a.
    - (SL) n.a.
- 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. *know where to-step* 'to know the way to succeed in something'
        - *mostrar com quantos paus se faz uma canoa* lit. *show with how many sticks one makes a canoe* 'to punish or take revenge'
    - (RO) *a şti cu ce se mănâncă* lit. *to know with what CL.Refl. eats* 'to knwo what it is about'
    - (SL) *vedeti koliko je ura* lit. *to know what time it is* 'to realize the truth'
- Non-reflexive pronoun
    - (DE) *es gibt* lit. *it gives* 'there is'
    - (EL) *τα καταφέρνω, την πατάω*
    - (EN) *to make it*
    - (FR) *l'emporter* lit. *to take it away* 'to win'
    - (IT) *prender le* lit. *to take it* 'to be beaten'
    - (PL) Polish does not seem to have this type of VMWEs
    - (PT) *dá-lhe João!* lit. *give to him/her, João!* 'show them what you got, João!'
    - (RO) *a o şterge* lit. *to her delete* 'to fly the coop'
        - *a o întinde* lit. *to her extend* 'to fly the coop' synonymous expressions with the non-anaphoric feminine ACC personal clitic 'o' functioning as an expletive
    - (SL) *ucvreti jo* lit. *to escape her* 'to escape something/someone by running '

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*
    - *Fortune favors the bold*
    - *The pleasure is mine*
    - *I beg your pardon!*
- (PL) *trafiła kosa na kamień* lit. *met the scythe a stone* 'someone rude/dishonest came across someone else who used similar methods against him/her'
- (PT) *quem vê cara não vê coração* lit. *who sees face doesn't see heart* 'a person can lie/omit his/her feelings'
- (RO) *Urciorul nu merge de multe ori la apă* lit. *Pitcher-the not goes of many times at water* 'The pitcher goes so often to the well that it is broken at last'
- (SL) *Počasi se daleč pride* 'more haste less speed'
    - *Po toči zvoniti je prepozno* lit. *there is no use ringing the bells after hail* 'it is to late'

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 cut 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', *il y a* lit. *it has there* 'there is'
- (PL) *chować głowę w piasek* lit. *to hide head in sand* 'to pretend not to see a problem'
- (PT) *tapar o sol com a peneira* lit. *to hide the sun with a sieve* 'to sugar-coat'
- (RO) *a da bir cu fugiţii* lit. *to give tribute with fugitives.the* 'to back away'
- (SV) *att sätta sig upp mot någon* lit. *to sit oneself up against someone* 'to defy someone'
    - *att dra sitt strå till stacken* lit. *to draw one's straw to stack.the* 'to contribute (in a small way)'
- (SL) *beseda mi je ostala v grlu* lit. *word got stuck in my throat* 'I am speechless'

In case of several lexicalized 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 *aufstellen* lit. *to set up*
- (EN) *to let the cat out of the bag* → contains the VPC *to let out*

- (FR) *se faire des idées* lit. *to make SELF ideas* 'to imagine something false' → contains the non-VMWEs *se faire* and *faire des idées*
- (PL) *bać się <u>własnego cienia</u>* lit. *to fear SELF one's own shadow* 'to be very timid' → contains the IReflV *bać się* lit. *to fear SELF* 'to be afraid'
- (PT) *virar-se nos trinta* lit. *turn-RCLI in-the thirty* 'to get by' contains the synonymous IReflV *virar-se* 'to get by' ≠ *virar* 'to turn/become'
- (RO) *a da cărțile pe față* lit. *to give cards.the on face* 'to reaveal one's true intentions' → contains the ID *a da pe față* 'to reveal'
    - *a-și da arama pe față* lit. *to give his/her copper.the on face* 'to reveal his/her true (evil) nature' → this is even more complicated since, besides the ID a **da pe față**, the IReflV has to be annotated as well - a three-level embedding
- (SL) *delati se norca iz koga* lit. *to make RCLI fool of someone* 'to make fun of someone' → contains the IReflV *delati se* lit. *to make oneself* 'to pretend'

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) *sei kein Frosch* lit. *be no frog* 'be no chicken' → idiom because #*kein Frosch* 'no frog' loses the meaning
- (EN) *to* **be no chicken** → idiom because #*no chicken* loses the meaning
    - *to* **be somebody** → idiom because #*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. *to be a ball* 'to suck/butter up'
- (PL) *być jedną nogą na tamtym świecie* lit. *to be with one leg in the other world* 'to be close to death' → idiom because #*jedna noga na tamtym świecie* 'one leg in the other world' loses the meaning
    - *być do rzeczy* lit. *to be to the thing* 'to be relevant' → non-VMWE because the copula can be omitted, as in *dał parę argumentów całkiem* **do rzeczy** 'he gave a couple of quite relevant arguments'
- (PT) *ser alguém na vida* lit. *to be somebody in life* '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* 'to be an untrustworthy person' → idiom because #*flor que se cheire* 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 biserică* lit. *to be door of church* ' to be honest' → idiom because #*ușă de biserică* loses the meaning
    - *a fi un papă-lapte* lit. *to be a eat-milk* ' to be a piker' → idiom because #*un papă-lapte* preserves the meaning
- (SL) *biti trn v peti komu* lit. *to be a thorn in somebody's heel* 'to be a big problem, obstacle' → 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.ⓘ

Section 6.4

Inherently reflexive verbs (IReflV)

Reflexive clitics (RCLI) are clitic pronouns that refer to the subject of the verb, like *oneself* 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 meaning. REFLV 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 <u>quasi-universal</u> 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', *divit se* 'to wonder'
        - (DE) *sich schämen* 'to be ashamed', *sich wundern* 'to wonder'
        - (ES) *suicidarse* 'to suicide', *abstenerse* 'to abstain'
        - (FR) *s'évanouir* 'to faint', *se suicider* 'to suicide'
        - (IT) *suicidarsi* '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 își însuși* 'to appropriate' with obligatory DAT reflexive clitic
        - (SV) *att försova sig* 'to sleep in'
            - *att gifta sig* 'to get married'
        - (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* ≠ *enthalten* 'to abstain ≠ to contain', *sich (um etw.) handeln* ≠ *handeln* 'to be ≠ to handle'
        - (ES) *recoger* ≠ *recogerse* 'to gather ≠ to go home', *empeñar* ≠ *empeñarse* 'to pawn ≠ to insist'
        - (FR) *s'apercevoir* ≠ *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 îndura* ≠ *a îndura* 'to have the heart ≠ to suffer'
            - *a se face* ≠ *a face* 'to become ≠ to make' even if it is inchoative (Dindelegan 2013: 79) a se face (=to become) is IReflV (it passes Test15)
        - (SV) *att känna sig ledsen/arg* 'to feel sad/angry ≠ to touch'
        - (SL) *tikati se česa* ≠ *tikati* 'to concern ≠ to refer', *pobrati se* ≠ *pobrati* 'to stand up ≠ to pick up'
- **Reciprocal ⇒ NOT ANNOTATED**
    - The RCLI has a sense of <u>mutually</u>:
        - (CS) *líbat se* 'to kiss each other', *potkávat se* 'to meet each other'
        - (DE) *sich küssen* 'to kiss each other', *sich treffen* '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'

- (IT) *baciarsi* 'to kiss each other'
- (PL) *całować się* 'to kiss each other', *spotykać się* 'to meet each other'
- (PT) *cumprimentar-se* 'to greet each other', *ver-se* 'to see each other'
- (RO) *a se saluta* 'to greet each other'
- (SL) *poljubljati se* 'to kiss each other', *srečati se* 'to meet each other'

- **Reflexive ⇒ NOT ANNOTATED**
  - The RCLI marks the <u>reflexive</u> or <u>reciprocal</u> construction, that is, the clitic plays the role of *self* in English
    - (CS) *mýt se* 'to wash oneself', *drbat se* 'to scratch oneself'
    - (DE) *sich waschen* 'to wash oneself', *sich kratzen* 'to scratch oneself'
    - (ES) *mirarse* 'to look at oneself', *vestirse* 'to dress oneself'
    - (FR) *se laver* 'to wash oneself', *se parler* 'to talk to oneself'
    - (IT) *lavarsi* 'to wash oneself', *vestirsi* 'to dress oneself'
    - (PL) *myć się* 'to wash oneself', *drapać się po głowie* 'to scratch oneself on the head'
    - (PT) *apressar-se* 'to hurry oneself', *vestir-se* 'to dress oneself'
    - (RO) *a se spăla* 'to wash oneself'
    - (SV) *att tvätta sig* 'to wash oneself'
    - (SL) *umivati se* 'to wash oneself', *praskati se* 'to scratch oneself'
- **Body part, also called possessive reflexive ⇒ NOT ANNOTATED**
  - Specific type of reflexive use in which the direct object ❶ is a <u>body part</u> or, more generally, an inalienable part of the subject
    - (CS) *mýt 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) *grattarsi 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) impossible, uses possessive instead
    - (RO) *a-și rupe mâna* lit. *RCLI.DAT 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 <u>syntactic alternation for transitive verbs</u>. Just like in regular passive alternation, the direct object of the transitive version appears as the subject of the REFLV version, and thus the verb agrees with the subject.
  - Differently from inchoative (see below), the subject of the transitive version is absent in the REFLV 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 synthetic passive ⇒ NOT ANNOTATED**
  - In some languages, middle alternation with preverbal subject sounds unnatural and middle alternation with postverbal subject ❶ 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 alquilan 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' ❶
    - (PT) *alugam-se casas* lit. *rent-RCLI houses* 'people rent houses'
    - (RO) *se vând bine apartamentele din blocurile noi* lit. *RCLI sell well apartments-the from blocks-the new* 'Apartments from new blocks sell well'
      se construiesc locuințe noi lit. *RCLI built houses new* 'new houses are built'
    - (SL) *nove hiše se gradijo* lit. *new houses RCLI built* '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'
      se trabaja mejor aquí lit. *RCLI works better here* 'people work better here'
    - (FR) *il 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'
      si affitta molte case lit. *RCLI rents many houses* 'people rent many houses'
    - (PL) *za dużo się pracuje* lit. *too much RCLI works* 'people work too much'
      bzdury się opowiada lit. *nonsense RCLI tells* 'people tell nonsense'
    - (PT) *dorme-se muito* lit. *sleeps-RCLI much* 'people sleep a lot'
      conta-se histórias lit. *tells-RCLI stories* 'people tell stories' ❶
    - (RO) *se lucrează până târziu* lit. *RCLI works until late* 'people work until late' transitive verbs can be impersonal in RO only when they are null-object verbs (se lucrează până târziu - *este lucrat până târziu) or when their subject is realized by a clause headed by a complementizer Dindelegan 2013: 174
      se suferă din cauza sărăciei lit. *RCLI suffer because of poverty* 'one suffers because of poverty' RO impersonal reflexive verbs are mostly intransitive Dindelegan 2013: 173
      se aleargă dimineața lit. *RCLI run in the morning* 'people run in the morning'
    - (SL) *govori se/govorijo se neumnosti* lit. *it says/they say RCLI silly things* 'people say silly things'
- **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 REFLV
    - the subject of the transitive version is not only absent, it is also semantically unclear or nonexistent
      - (CS) *dveře se otvírají* 'the door opens'
      - (DE) *die Tür öffnet sich* 'the door opens'

- (ES) *la puerta se abrió* 'the door opened'
- (FR) *la porte s'est subitement ouverte* 'the door suddenly opened'
- (IT) *la porta si apre* 'the door opens'
- (PL) *drzwi się otwierają* 'the door opens'
- (PT) *o vaso se quebrou* 'the vase broke'
- (RO) *mașina s-a stricat* 'the car broke down'
  - *ușa s-a deschis* 'the door opened'
- (SV) *dörren öppnar sig* ' the door opens'
- (SL) *vrata se odpirajo* 'the door opens'

IReflV-specific decision tree

↳ Apply test 14 - [**INHERENT**]
   ↳ **YES** ⇒ Annotate as **IReflV**
   ↳ **NO** ⇒ Apply test 15 - [**DIFF-SENSE**]
      ↳ **YES** ⇒ Annotate as **IReflV**
      ↳ **NO or UNSURE** ⇒ Apply test 16 - [**DIFF-SUBCAT**]
         ↳ **YES** ⇒ Annotate as **IReflV**
         ↳ **NO** ⇒
            ↳ verb has no subject ⇒ Apply test 17 - [**IMPERS**]
               ↳ **YES** ⇒ It is not a VMWE, **exit**
               ↳ **NO** ⇒ Annotate as **IReflV**
            ↳ verb has a subject ⇒ Apply test 18 - [**MIDDLE-INCHO**]
               ↳ **YES** ⇒ It is not a VMWE, **exit**
               ↳ **NO** ⇒ Apply test 19 - [**REFL**]
                  ↳ **YES** ⇒ It is not a VMWE, **exit**
                  ↳ **NO or UNSURE** ⇒
                     ↳ subject is SINGULAR ⇒ Apply test 20 - [**REFL-MUTUAL**]
                        ↳ **YES** ⇒ It is not a VMWE, **exit**
                        ↳ **NO** ⇒ Annotate as **IReflV**
                     ↳ subject is PLURAL ⇒ Apply test 21 - [**RECIPRO**]
                        ↳ **YES** ⇒ It is not a VMWE, **exit**
                        ↳ **NO** ⇒ Annotate as **IReflV**

Test 14 - [INHERENT] Inherent clitic

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

↳ **YES** ⇒ annotate as IReflV
- (DE) ***sich schämen*** ⇒ **schämen* 'to be ashamed'
  sich wundern ⇒ **wundern* 'to wonder'
- (ES) ***suicidarse*** ⇒ **suicidar* 'to suicide'
  abstenerse ⇒ **abstener* 'to abstain'
- (FR) ***s'évanouir*** ⇒ **évanouir* 'to faint'
  se suicider ⇒ **suicider* 'to suicide'
- (IT) ***suicidarsi*** ⇒ **suicidare* 'to suicide'
- (PL) ***dowiedzieć się*** ⇒ **dowiedzieć* 'to find out'
  bać się ⇒ **bać* 'to be afraid'
- (PT) ***queixar-se*** ⇒ **queixar* 'to complain'
  abster-se ⇒ **abster* 'to abstain'
- (RO) *a se teme* ⇒ **a teme* 'to be afraid'
  a îşi însuşi ⇒ **a însuşi* 'to appropriate'
- (SL) ***sramovati se*** ⇒ **sramovati* 'to be ashamed'
  čuditi se ⇒ **čuditi* 'to wonder'
↳ **NO** ⇒ next test

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
- (DE) ***sich verstehen*** ≠ *verstehen* 'to get along well ≠ to understand'
- (ES) ***recogerse*** ≠ *recoger* 'to go home ≠ to pick up, to gather'
- (FR) ***s'apercevoir*** ≠ *apercevoir* 'to realize ≠ to see'
  s'agir ≠ *agir* 'to be ≠ to act'
- (IT) ***riferirsi*** ≠ *riferire* 'to refer ≠ to report, to tell'
- (PL) ***znajdować się*** ≠ *znajdować* 'to find oneself ≠ to be'
- (PT) ***encontrar-se*** ≠ *encontrar* 'to be ≠ to meet'
  referir-se ≠ *referir* 'to concern ≠ to refer'
- (RO) *a se îndura* ≠ *a îndura* 'to have the heart to ≠ to suffer'
- (SL) ***razumeti se*** ≠ *razumeti* 'to get along well ≠ to understand'
↳ **NO or UNSURE** ⇒ next test

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

Is the subcategorization frame of the simple verb <u>without the RCLI</u> 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

- (DE) *X **verliert sich** in Y* ⇔ *X verliert Y* lit. *X looses RCLI in Y* ⇔ *X looses Y*
- (ES) *X **se olvidó** de Y* ⇔ *X olvidó Y* lit. *X RCLI forgot of Y* ⇔ *X forgot Y*
- (FR) *X **se confesse** de Y* ⇔ *X confesse Y* (but *X confesse de Y) lit. *X RCLI confesses of Y* ⇔ *X confesses Y (but not *X confesses of Y)*
  - *X **se plaint** de Z* ⇒ **Y plaint (à) X de Z* lit. *X RCLI complains of Z* ⇒ **Y complains (to) X of Z* → the verb without RCLI, plus direct or indirect object. does not subcategorize for the PP with preposition *de*
    - *X **se refuse** à Vinf* ⇒ **Y refuse (à) X à Vinf* lit. *X RCLI refuses to Vinf* ⇒ **Y refuses (to) X to Vinf*
- (IT) *X **si è dimenticato** di Y* ⇔ *X ha dimenticato Y* lit. *X RCLI forgot of Y* ⇔ *X forgot Y*
- (PL) *X **tłumaczy się** z Y* ⇔ *X tłumaczy Y* lit. *X explains SELF of Y* ⇔ *X explains Y*
- (PT) *X **se esqueceu** de Y* ⇔ *X esqueceu Y* lit. *X RCLI forgot of Y* ⇔ *X forgot Y*
- (RO) *X **se gândeşte** la Y* ⇔ *X gândeşte că Y* lit. *X RCLI thinks of Y* ⇔ *X thinks that Y*

↳ **NO** ⇒ next test

Test 17 - [IMPERS ❶] - Impersonal

When you replace the RCLI by an underspecified subject such as one or people, does the sentence keep its meaning?

↳ **YES** ⇒ do not annotate as verbal MWE
- (DE) *hier tanzt **es** sich gut* ⇔ *hier tanzen die Leute gut* 'people dance well here'
- (ES) *se duerme mucho* ⇔ *las personas duermen mucho* 'people sleep a lot'
  - *se busca a actores* ⇔ *la gente busca a actores* 'people look for actors'
- (FR) *il se dit des bêtises* ⇔ *les personnes disent des bêtises* 'people say silly things'
- (IT) *si dorme molto* ⇔ *le persone dormono molto* 'people sleep a lot'
  - *si affitta molte case* ⇔ *le persone affittano molte case* 'people rent many houses'
- (PL) *pracuje się za dużo* ⇔ *ludzie pracują za dużo* 'people work too much'
  - *opowiada się bzdury* ⇔ *ludzie opowiadają bzdury* 'people tell nonsense'
- (PT) *dorme-se muito* ⇔ *as pessoas dormem muito* 'people sleep a lot'
  - *conta-se histórias* ⇔ *as pessoas contam histórias* 'people tell stories'
- (RO) *se lucrează până târziu* ⇔ *lumea lucrează până târziu* 'people work until late'
  - *se aleargă dimineaţa* ⇔ *lumea aleargă dimineaţa* 'people run in the morning'
- (SL) *govorijo se neumnosti* ⇔ *ljudje govorijo neumnosti* 'people tell nonsense'

↳ **NO** ⇒ annotate as IReflV

Test 18 - [MIDDLE-INCHO ❶] - Middle or Inchoative

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 ❶ the REFLV version? In other words, *people/somebody ❶ V [to] X ⇒ X REFLV*?

↳ **YES** ⇒ do not annotate as verbal MWE
- (DE) *man kann die Häuser gut verkaufen ⇒ die Häuser verkaufen sich gut* 'people can sell the houses well ⇒ the houses sell well'
  - *jemand öffnet die Tür ⇒ die Tür öffnet sich* 'somebody opens the door ⇒ the door opens'
- (ES) *la gente cuenta historias ⇒ se cuentan historias* 'people tell stories ⇒ stories are told'
  - *alguien abrió la puerta ⇒ la puerta se abrió* 'somebody opened the door ⇒ the door opened'
- (FR) *on vend bien Ce produit ⇒ ce produit se vend bien* 'people sell this product well ⇒ this product sells well'
  - *quelqu'un ouvre la porte ⇒ la porte s'ouvre,* 'somebody opens the door ⇒ the door opens'
- (IT) *qualcuno vende bene questo prodotto ⇒ questo prodotto si vende bene* 'someone people sells this product well ⇒ this product sells well'
  - *qualcuno apre la porta ⇒ la porta si apre* 'somebody opens the door ⇒ the door opens'
- (PL) *ktoś sprzedaje te domy ⇒ te domy się sprzedają* 'somebody sells these houses ⇒ these houses sell well'
  - *ktoś otwiera drzwi ⇒ drzwi się otwierają* 'somebody opens the door ⇒ the door opens'
- (PT) *alguém conta histórias ⇒ contam-se histórias* 'somebody tells stories ⇒ stories are told'
  - *alguém acalmou o menino ⇒ o menino se acalmou* 'somebody calmed the boy down ⇒ the boy calmed down'
- (RO) *cineva spune glume ⇒ se spun glume* 'somebody tells jokes ⇒ jokes are told'
  - *cineva a deschis uşa ⇒ uşa s-a deschis* 'somebody opened the door ⇒ the door opened'
- (SL) *nekdo pripoveduje šale ⇒ šale se pripovedujejo* 'somebody tells jokes ⇒ jokes are told'
  - *nekdo je odprl vrata ⇒ vrata so se odprla* 'somebody opened the door ⇒ the door opened'

↳ **NO** ⇒ next test

Test 19 - [REFL ❶] - Reflexive

When you replace the RCLI by *oneself only* or *to oneself only*, does it imply ❶ the REFLV version? In other words, *X V [to] himself only ⇒ X REFLV*?

↳ **YES** ⇒ do not annotate as verbal MWE
- (DE) *Paul kratzt nur sich selbst ⇒ Paul kratzt sich* 'Paul scratches himself'
- (ES) *Pablo se lava a sí mismo ⇒ Pablo se lava* 'Paul washes himself'
- (FR) *Paul ne soigne que lui-même ⇒ Paul se soigne* 'Paul heals himself'
  - *Paul ne parle qu'à lui-même ⇒ Paul se parle* 'Paul talks to himself'
- (IT) *Paolo cura solo se stesso ⇒ Paolo si cura* 'Paul heals himself'
  - *Paolo parla solo a se stesso ⇒ Paolo si parla* 'Paul talks to himself'
- (PL) *Paweł leczy tylko siebie ⇒ Paweł leczy się* 'Paul heals himself'
- (PT) *Paulo só lava a si mesmo ⇒ Paulo se lava* 'Paul washes himself'
- (RO) *Paul se spală doar pe sine ⇒ Paul se spală.* 'Paul washes himself'
- (SL) *Pavel praska samo sebe ⇒ Pavel se praska* 'Paul scratches himself'

↳ **NO or UNSURE** ⇒ next test ❶
  ↳ The subject is singular: test REFL-MUTUAL
  ↳ The subject is plural or coordinated (Bob and Alice): test RECIPRO

Test 20 - [REFL-MUTUAL ❶] - Reflexive-mutual

Is a reciprocal version possible? Namely: Is it acceptable to replace the singular subject by a plural and add *each other* to the REFLV form without changing the REFLV's meaning ❶?

↳ **YES** ⇒ do not annotate as verbal MWE ⓘ
- (DE) *Paul wäscht sich ⇔ Sie waschen sich gegenseitig / einander* 'they wash each other'
- (ES) *Pablo se lava ⇔ ellos se lavan mutuamente / los unos a los otros* 'they wash each other'
- (FR) *Paul se lave ⇔ ils se lavent mutuellement / les uns les autres* 'they wash each other'
- (IT) *Paolo si lava ⇔ essi si lavano reciprocamente / l'un l'altro* 'they wash each other'
- (PL) *Paweł się myje ⇔ oni myją się nawzajem* 'they wash each other'
- (PT) *Paulo se lava ⇔ eles se lavam mutuamente / uns aos outros* 'they wash each other'
- (RO) *el se spală ⇔ ei se spală unul pe altul* 'they wash each other'
- (SL) *Pavel se umiva ⇔ umivajo drug drugega* 'they wash each other'

↳ **NO** ⇒ annotate as IReflV

Test 21 - [RECIPRO ⓘ] - Reciprocal

Is it possible to remove the RCLI and replace ⓘ the coordinated subject (A and B) or plural subject (A.PL) by a singular subject (A or A.PL) and a singular object, often introduced by to/with (B or A.PL), without changing the REFLV's meaning ⓘ? That is:

- Coordinated subject: *A and B PronV ⇔ A V [to/with] B and B V [to/with] A*?
- Plural subject: *A.PL PronV ⇔ A.PL V [to/with] A.PL*?

↳ **YES** ⇒ do not annotate as verbal MWE
- (DE) *Paul und Anna umarmen sich ⇔ Paul umarmt Anna and Anna umarmt Paul* 'Paul and Anna hug each other'
  die Affen kratzen sich ⇔ die Affen kratzen die Affen 'the monkeys scratch each other'
- (ES) *Pablo y Ana se abrazan ⇔ Pablo abraza a Ana and Ana abraza a Pablo* 'Paul and Ann hug each other'
  los niños se abrazan ⇔ los niños abrazan a los niños 'the children hug each other'
- (FR) *Paul et Anne s'embrassent ⇔ Paul embrasse Anne and Anne embrasse Paul* 'Paul and Ann kiss'
  les jours se suivent ⇔ les jours suivent les jours 'the days follow each other'
- (IT) *Giovanni e Anna si baciano ⇔ Giovanni bacia Anna and Anna bacia Giovanni* 'John and Ann kiss'
  i giorni si seguono ⇔ i giorni seguono i giorni 'i giorni seguono l'un l'altro'
- (PL) *Paweł i Elena się całują ⇔ Paweł całuje Elenę and Elenę całuje Paweł* 'Paweł and Elena kiss'
- (PT) *João e Ana se beijam ⇔ João beija Ana and Ana beija João* 'John and Ann kiss'
  os presos se agridem ⇔ os presos agridem os presos 'the prisoners aggress each other'
- (RO) *Ion şi George se salută ⇔ Ion îl salută pe George and George îl salută pe Ion* 'Ion and George greet each other'
  participanţii se salută ⇔ participanţii îi salută pe participanţi 'the participants greet each other'
- (SL) *Pavel in Ana se objemata ⇔ Pavel objema Ano in Ana objema Pavla* 'Paul and Anna hug each other'

↳ **NO** ⇒ annotate as IReflV

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 (mesoclisis), 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
  - (ES) *darse cuenta de* 'to realize' ⇒ **darse N de* for any noun N
    meterse en líos 'to get in trouble' ⇒ REFLV not annotated in literal equivalents like *meterse en una tienda* 'to get in a store'

- (FR) *se rendre compte de* 'to realize' ⇒ **se rendre N de* for any noun N
  - *s'arracher les cheveux* lit. *RCLI tear the hair* 'worry' ⇒ REFLV not annotated in literal equivalents like *s'arracher un ongle* 'to tear oneself's nail'
- (IT) *rendersi conto di* 'to realize' ⇒ **si rende N di* for any noun N
  - *si strappa i capelli* lit. *RCLI tear the hair* 'to worry' ⇒ REFLV not annotated in literal equivalents like *strapparsi un unghia* 'to tear oneself's nail'
- (PL) *zdawać sobie sprawę z* 'to realize' ⇒ **zdawać sobie N z* for any noun N
- (PT) *dar-se mal* 'to fail' ⇒ dar-se ADV intransitive is acceptable only for antonym *bem* 'well'
  - *meter-se numa fria* lit. *to get-RCLI in a cold* 'to get in trouble' ⇒ REFLV not annotated in literal equivalent like *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 IRefIV in syntactically comparable literal constructions, annotate both the IRefIV and the ID as embedded MWEs (rare):
  - (PL) *rozlatywać się w proch* lit. *scatter itself into dust* 'disappear'
  - (PT) *virar-se nos trinta* lit. *turn-RCLI in-the thirty* contains *virar-se* 'to get by' ≠ *virar* 'to turn/become'
  - (RO) *a i se face rău* lit. *to CL.DAT RCLI.ACC make ill* 'to feel sick' this is a case when both a non-reflexive, dative clitic and a RCLI.ACC appear in the structure; the REFLV is annotated as IRefIV; both the IRefIV and the ID are annotated as embedded MWEs; note that the non-reflexive clitic is also considered as part of an ID (6.4_R)
    - *a se duce pe apa sâmbetei* lit. *RCLI go on water-the Saturday-of* 'to get lost' the REFLV is annotated in literal equivalent *a se duce pe apa Bistriței* 'he goes on the river Bistrița' there is a notable difference in meaning betwee the non-REFLV a duce 'to take' and the REFLV a se duce 'to go'
  - (SL) *režati se kot pečen maček* lit. *to laugh RCLI like a baked tomcat* 'to laugh loudly' *režati se is IRefIV*

Overlap LVC - IRefIV

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

- (DE) *Fragen stellen* 'to ask questions' ⇒ *sich Fragen stellen* 'to doubt/hesitate'
- (FR) *poser des questions* 'to ask questions' ⇒ *se poser des questions* 'to doubt/hesitate'
- (RO) no examples found for RO

In this case, the whole construction, including the verb, the noun and the reflexive clitic, must be annotated as ID, since there are two syntactic arguments:

- (DE) *sich Fragen stellen* 'to doubt/hesitate'
- (FR) *se poser des questions*
- (RO) no examples found for RO

Notice that annotating only the verb and the RCLI as IRefIV would be wrong, since it will have a completely different meaning without the noun, sometimes even coinciding with another IRefIV:

- (DE) *sich stellen* 'to surrender'
- (FR) *se poser* 'to sit/lay down'

Dative clitics and double clitics

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

- (CS) *bát se* 'to be afraid'
- (PL) *bać się* 'to be afraid'
- (RO) *a se sinchisi* lit. *to RCLI.ACC care* 'to care'
  - *a se sfii* lit. *to RCLI.ACC be.shy* 'to be shy'
  - *a se căi* lit. *to RCLI.ACC repent* 'to repent'
- (SL) *bati se* 'to be afraid'

However, other cases can appear in IRefIV:

- (CS) *poradit si* lit. *to advise oneself.DAT* 'to manage'
- (PL) *radzić sobie* lit. *to advise oneself.DAT* 'to manage'
- (RO) *a-și însuși* lit. *to-RCLI.DAT appropriate* 'to appropriate' - with a Dative clitic
  - *a-și apropria* lit. *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 IRefIV:

- (PL) *przyglądać się sobie* lit. *to observe RCLI.acc RCLI.DAT* 'to observe each other'
  - *radzić sobie z sobą* lit. *to advise RCLI.DAT with RCLI.INST* 'to manage with oneself'
- (RO) n.a.
- (SL) *nasmehniti se sebi* '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 IRefIV 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'

- (IT) *prender-ci* lit. *to take to-it* 'to make the right choice'
  - *prender-le* lit. *to take it* 'to be beaten'
- (PT) *dá-lhe João!* lit. *give to-him/her, João!* 'show them what you got, João!'
- (RO) *a-i arde* lit. *to CL.DAT burn* 'to have a desire'
  - *a o lua pe jos* lit. *to take CL.ACC on foot* 'to walk' according to the current guidelines, such examples pass the ID tests (see also 6.3_B5); both have literal correspondents that are not characterized by an obligatory non-reflexive clitic: a arde 'to burn' and a lua 'to take'
    - *a-i repugna* lit. *to CL.DAT loathe* 'to loathe'
    - *a-i prii* lit. *to CL.DAT* 'to be favourable to sb.'
- (SL) *ucvreti jo* lit. *to escape her* 'to escape something/someone by running'

Section 6.5

Verb-particle constructions (VPC)

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

- (DE) *um|fahren* lit. *over|drive* 'to run over', *mit|kommen* lit. *with|come* 'to join', *vor|bereiten* 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) *gre za* lit. *it goes for* 'it is about', *biti ob* lit. *to be at* 'to lose'

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 *um|fahren* 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 umgefahren*, instead: *er hat den See umfahren* 'he drove around the lake' but: *er hat das Schild umgefahren* 'he run over the sign'
- (ES) n.a.
- (PL) n.a.
- (RO) n.a.

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:

- (DE) *er **legt** das Buch **ab*** ' he puts down the book', *er **kommt** ins Haus **rein*** lit. *he comes into the house* 'he enters the house'
- (EN) *to **lie down**, to **come in***
- (ES) n.a.
- (PL) n.a.
- (RO) n.a.
- (SL) ***prišel** je **do** hriba* 'He came to the hill'

Some combinations may have both compositional and non-compositional meanings depending on the context and only the latter should be annotated:

- (DE) *ein Schild **aufstellen*** 'to put up a sign' vs. *einen **Plan aufstellen*** 'to draw up a plan'
- (EN) *to **put up** a flag* vs. *to **put up** a friend for the night*
- (ES) n.a.
- (PL) n.a.
- (RO) n.a.
- (SL) ***gre za** vodičem* 'he follows the guide' vs. ***gre za** naše temeljno načelo* 'it is about our fundamental principle'

The essential compositionality test is to see if a sentence without the particle can refer to the same event/state as the sentence with the particle.

Test 22 - [V+PART-DIFF-SENSE] - Sense shift due to the particle

Does the particle provoke an unexpected change in meaning of the verb? I.e., does a sentence without the particle <u>fail</u> to refer to the same event/state as the sentence with the particle (special care must be taken when the same construction might or might not be a valid VPC depending on its context)?

↳ **YES** ⇒ it is a VPC
- (DE) *Der Lehrling **fängt** ein Praktikum **an*** lit. *the apprentice catches an internship on* 'the apprentice begins an internship' does not imply #*Der Lehrling fängt ein Praktikum* 'the apprentice catches an internship'

  *Die Bäuerin hat sich wieder **eingefangen*** lit. *the farmer's wife has herself again catched* 'the farmer's wife has calmed down again' does not imply #*Die Bäuerin hat sich wieder gefangen* 'the farmer's wife has catched herself again'

  *Der Schüler **legt** die Prüfung **ab*** lit. *the pupil lays the exam off* 'the pupil takes the exam' does not imply #*der Schüler legt die Prüfung* 'the pupil lays the exam'

  *Das Schiff **legt** vom Hafen **ab*** lit. *the boat lays from the harbor off* 'the ship leaves the harbor' does not imply #*das Schiff legt vom Hafen* 'the boat lays from the harbor'
- (EN) *to **do** somebody **in*** 'to kill sb' does not imply #*to do somebody*

  *to **check in** upon arrival* does not imply #*to check upon arrival*
- (ES) n.a.
- (HU) *A meccs után csak az edző nem **rúgott be*** 'Only the coach did not get drunk after the match' → *A meccs után az edző **berúgott*** 'The coach got drunk after the match' does not imply #*Az edző rúgott* 'the coach kicked'

  *Nem **jött be** ez a koktél nekem* 'I didn't like this cocktail' → ***Bejött** ez a koktél nekem* 'I liked this cocktail' does not imply #*Jött ez a koktél nekem* 'this cocktail bumped into me'
- (PL) n.a.
- (RO) n.a.
- (SL) *Ona **je za** enakopravnost* 'she is for equality' does not imply #*Ona je enakopravnost* 'she is equality'

  ***postavili** so ga **za** sodnika* lit. *they set him for a judge* ' they appointed him a judge' does not imply **Postavili so ga sodnika* lit. *they set him judge – no sense*

↳ **NO** ⇒ it is not a VPC
- (DE) *Der Bauer **fängt** die Hühner **ein*** lit. *the farmer catches the chickens in* 'the farmer catches the chickens' implies *der Bauer fängt die Hühner* 'the farmer catches the chickens'

  *Der Lehrer **legt** das Buch auf dem Tisch **ab*** lit. *the teacher lays the book on the table apart* 'the teacher puts the book away on the table' implies *Der Lehrer legt das Buch auf den Tisch* 'the teacher puts the book on the table'

  *Der Lehrer **legt** den Mantel **ab*** lit. *the teacher lays the coat off* 'the teacher takes off his coat' implies *Der Lehrer legt den Mantel* 'the teacher puts the coat'
- (EN) *to **look up** into the sky* implies *to look into the sky*

  *to **eat up** the cookies* implies *to eat the cookies*
- (ES) n.a.
- (HU) *A csatár nem **rúgta be** a helyzetét* 'The forward missed its chance to score a goal' → *A csatár **berúgta** a helyzetét* implies *A csatár rúgott* 'The forward kicked'

  *Nem **jött be** a szobába* 'He did not come into the room' → *(**Bejött** a szobába* 'he entered the room' implies *Jött a szobába* 'he came into the room'
- (PL) n.a.
- (RO) n.a.

Section 6.6

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 recommended 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*** lit. *to_sew and to_sing* 'easy as pie, a piece of cake'
- (PL) ***pluć i łapać*** lit. *spit and catch* 'to be lazy, to do nothing useful'
  - *coś kogoś **ani ziębi, ani grzeje*** lit. *something neither cools nor warms someone* 'someone is indifferent to something'
  - ***bądź tak dobry i** zrób coś* lit. *be so good and do somenthing* 'be so good as to do something'
- (PT) ***pintar e bordar*** lit. *paint and knit* 'to abuse'
- (RO) *a **tunat și** i-a **adunat*** lit. *it.has thundered and CL.ACC-it.has gathered* 'birds of a feather flock together'
  - ***seamănă, dar nu răsare*** lit. *sow.3SG (homonym of resemble), but not sprout.3SG* 'not to resemble'
- (SL) ***živi in pusti živeti*** 'live and let live'
- **compound verbs**, resulting usually from conversion of nominal compounds:
  - (DE)
  - (EN) *to **voice act***
    - *to **pretty-print***
    - *to **short-circuit***
    - *to **tumble dry***
  - (FR) ***court-circuiter*** 'to short-circuit'
  - (RO) *n.a.* there are no cases of compound hyphenated verbs in RO
  - (SL) *n.a.* there are no cases of compound hyphenated verbs in SL

No specific tests apply to this category. In other words an expression should be annotated as OTH if:

- it is of one of the syntactic/functional types of verbal multiword expression
- it is a VMWE, i.e. it fulfills one of the 5 generic tests for VMWE identification
- it <u>cannot</u> be classified into any universal (LVC, ID), quasi-universal (IReflV, VPC) or language-specific category

---

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 langue 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 <u>any</u> 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** <u>mit</u> ihm* 'I can no longer put up with him'
  - (EN) *I can no longer **put up** <u>with</u> 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 <u>prohibited</u> 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 <u>took</u> <u>at once</u> <u>off</u> my clothes.*
  *She always tries to **take in** her clients. *She tries to <u>take</u> <u>always</u> <u>in</u> 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** <u>recently</u> **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 <u>mit</u>?* lit. *come you with?* 'are you coming?'
  *Ich **schlage vor** allen zu verzeihen.* 'I propose to forgive everyone' *Ich **schlage** es <u>vor</u>* 'I propose it'
  *Der Mülleimer wurde **umgefahren**.* 'The trash bin was knocked down' *Er **fuhr** den Mülleimer **um**.* 'He knocked down the trash bin'
- (PL) n.a.
- (RO) n.a.

↳ **NO** ⇒ other tests are needed
- (DE) ***Kommst** Du **mit** jemandem?* 'Are you coming with someone?' **Kommst Du jemandem <u>mit</u>?*
  *Er **umfuhr** den ganzen See mit dem Fahrrad.* 'He drove around the whole lake with a bike' **Er fuhr ihn <u>um</u>.*
- (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 <u>auf</u>!* 'Be careful with the cars!' *Sie müssen auf die Autos **aufpassen**!* 'You must be careful with the cars!'
  *Er **fuhr** das Schild <u>um</u>.* 'He drove over the sign' *Er sollte das Schild nicht **<u>um</u>fahren*** 'He should not drive over the sign'
- (PL) n.a.
- (RO) n.a.

↳ **NO** ⇒ other tests are needed
- (DE) *Er **<u>um</u>fuhr** den ganzen See mit dem Fahrrad.* 'He rode around the whole lake with a bike' **Er fuhr den ganzen See mit dem Fahrrad <u>um</u>.*
  ***Sprechen** Sie <u>mit</u> ihm!* 'Speak with him!' **Sie sollen ihm <u>mit</u>sprechen.*
- (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 IReflV 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 underlined undexpected 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 multiword 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 multitoken word (since compound splitting might not be a standard feature of a tokenizer):

- (DE) *unter Drogeneinfluss stehen* 'to be under the influence of drugs'
    Heisshunger 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 ❶ about the mixed status of this MWT:

- (DE) *Drogeneinfluss* → MWT containing a lexicalized VMWE component *Einfluss* and an external word *Drogen*
    Heisshunger → 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) *odprawić mszę i pokutę* lit. *celebrate a mass and a penance*→ *odprawić* should be annotated both as part of *odprawić mszę* lit. *to celebrate a mass* and of *odprawić 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 IReflV, 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--

specific decision tree indicate that it belongs to the LVC category.

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 gibt Beschwerden* 'there are complaints'
- (EN) *there are complaints*
- (FR) *il existe des plaintes* lit. *it there has complaints* 'there are complaints'
- (PL) n.a.
- (PT) *há queixas* lit. *has complaints* 'there are complaints'
- (SL) *imeti pripombe* 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-REDUC]. 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. *it there has complaints* 'there are complaints', two dependents of the verb *a* lit. *has* are lexicalized: *il* 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-REDUC].
- 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--REDUC].
- 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?

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* 'maybe', *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 nouns 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 corresponging 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 (#*spill three/twenty beans*), it is possible to find some variation (*spill these/many/all/my/his/more/no 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 jgħid** ilbieraħ* lit. *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) *παίζω __το χαρτί__ του ευρωσκεπτικισμού* lit. *to-play the paper the.SG.GEN euroscepticism.SG.GEN* 'to use the asset of euroscepticism, to use euroscepticism 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__** *Coronairien 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** *uma* **__sessão__** *de fotos/autógrafos* 'to make a photo/autograph session'
  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 indiate 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.*

- (CS) *MWEs with **their lexicalized components** in Czech are indicated like this.*
- (DE) *MWEs with **their lexicalized components** in German are indicated like this.*
- (EL) *MWEs with **their lexicalized components** in Greek are indicated like this.*
- (EN) *MWEs with **their lexicalized components** in English are indicated like this.*
- (ES) *MWEs with **their lexicalized components** in Spanish are indicated like this.*
- (FA) *MWEs with **their lexicalized components** in Farsi are indicated like this.*
- (FR) *MWEs with **their lexicalized components** in French are indicated like this.*
- (HE) *MWEs with **their lexicalized components** in Hebrew are indicated like this.*
- (HR) *MWEs with **their lexicalized components** in Croatian are indicated like this.*
- (HU) *MWEs with **their lexicalized components** in Hungarian are indicated like this.*
- (IT) *MWEs with **their lexicalized components** in Italian are indicated like this.*
- (LT) *MWEs with **their lexicalized components** in Lithuanian are indicated like this.*
- (MT) *MWEs with **their lexicalized components** in Maltese are indicated like this.*
- (PL) *MWEs with **their lexicalized components** in Polish are indicated like this.*
- (PT) *MWEs with **their lexicalized components** in Portuguese are indicated like this.*
- (RO) *MWEs with **their lexicalized components** in Romanian are indicated like this.*
- (SV) *MWEs with **their lexicalized components** in Swedish are indicated like this.*
- (SL) *MWEs with **their lexicalized components** in Slovene are indicated like this.*
- (TR) *MWEs with **their lexicalized components** in Turkish are indicated like this.*
- (YI) *MWEs with **their lexicalized components** in Yiddish are indicated like this.*

Examples are preceded by the 2-letter language code in parentheses (e.g. EN for English). You can control what languages are shown and hidden by toggling the header buttons. See the section on notation for more information.

In order to see the ID of all examples, make sure the ID button is toggled on the header of the current page. Now look at the template above. You should see this ID: 8.2_A_template-mwe. The 8.2 represents the current section number (in bold in the TOC on the left). The letter A (or B, C, D...) indicates the position of the example inside this page. The name template-mwe is a more human-readable identifier for this example.

The spreadsheet

The spreadsheet can be accessed through this link to Google Docs. From time to time, the guidelines will be updated based on the contents of the spreadsheet.

The spreadsheet is divided into the following columns: *ID-section*, *ID-order*, *ID-name*, *lang*, *HTML-example* and *Status*. In order to edit an example, you need to look at its ID, and then find the appropriate place in the spreadsheet. For example, for the ID 8.2_A_template-mwe, you should look for the lines with ID-order 8.2 (towards the bottom of the spreadsheet). Then look for ID-order A on the second column. Check that the third column contains the ID-name template-mwe.

You will then see a sequence of examples, one for each language. The examples in the template above were collected from this spreadsheet. The rest of this page will teach you how to add you own examples to this spreadsheet.

When adding examples for your own language, we advise you to always start by copying an example that has already been filled in for another language, and then adapting it to your language. Remember that you should not translate an example, but rather find an example of the target phenomenon in your language, regardless if it is a direct translation or not. Therefore, before entering an example in the spreadsheet, you should always check its context using its ID. A quick way to do this is to search (Ctrl+F) the ID of an example in the full-text version of the guidelines (where the ID button is on).

If we notice something wrong or suspicious with your example, we may correct it (e.g. you forgot a closing <lex> tag). If we cannot correct the example, we will ask you to check it by using the last column of the spreadsheet, *Status*.

If you think that a phenomenon is not relevant for your language or that examples are not needed for a given phenomenon, just leave the corresponding cell empty.

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 `<nmwe>This is not an MWE</nmwe>` 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 `<nmwe>This is <u>not</u> an MWE</nmwe>` yields the following:

- (EN) *This is <u>not</u> 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 the tags `<trans>` and `</trans>`. English examples can also use the tag `<trans>` to indicate the meaning of an idiomatic expression. For example, the code `<lex>défendre</lex> son <lex>bifteck</lex> ⓘ <gl>defend one's beefsteak</gl> ⓘ <trans>to defend one's interests</trans>` generates the example below. Notice that the code for gloss and translation is only shown when the user hovers the gloss icon. For consistency, you should always follow this order: `original text ⓘ <gl>the gloss</gl> ⓘ <trans>the translation</trans>`.

- (FR) *défendre* *son* *bifteck* lit. *defend one's beefsteak* 'to defend one's interests'

Normal: Some examples are presented followed by an explanation, in *normal* font (black color). This is done by using the tags `<n>` and `</n>`. For example, the code `some words <n>→ further details</n>` generates this:

- (EN) *some words* → further details

Newline: Sometimes, one may want to add several examples for a single phenomenon in the same language. If they are rather long, they should be presented on separate lines using the tag `<br/>`. This tag is special as it does not come in pairs: you only write one tag with the slash at the end (technically, it is an empty XML element). For example, the code `example 1 <br/> example 2 <br/> example 3` will be rendered as follows:

- (EN) *example 1*
  example 2
  example 3

Inside normal text, you may also use tags such as `<i>` (italics), `<strong>` (bold), as well as other HTML tags. If another language is using a given tag for an example, you can use it too. Otherwise, try to stick to the established conventions.

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** <u>*einjagen*</u> lit. *to-someone chase-in fear* 'to frighten someone'
  jemanden einen **Besuch** <u>*abstatten*</u>
- (EN) *to* **go** <u>*astray*</u>
- (FR) *se* **mettre** <u>*martel*</u> **en tête** lit. *SELF put a hammer in head* 'to worry a lot'
- (PL) <u>*odsądzić*</u> *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* <u>*kvit*</u> '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 that the traditional notion of a light verb. Notably, aspectual light verbs (which adds aspectual semantics to the complement), as in *to* <u>*start*</u> *a* **walk**, *to* <u>*give*</u> **courage**, are not considered operators. Operators are typical head verbs of light-verb constructions:

- (DE) *eine* **Entscheidung** <u>*treffen*</u> 'to make a decision'
  Angst <u>*haben*</u> 'to have fear'
  ein **Verbrechen** <u>*begehen*</u> 'to commit a crime'
- (EN) *to* <u>*make*</u> *a* **decision**
  to <u>*have*</u> **fear**
  to <u>*commit*</u> *a* **crime**
- (PL) *oddać* **hołd** lit. *to give-back tribute* 'to pay tribute'
- (SL) <u>*priti*</u> **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* **Diagramm zeigt** 'the diagram shows', *mit einem* **Bus fahren** 'to take a bus'
- (EN) *the* **graphic shows**
  drastically drop
- (PL) *zalać* **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 heard 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 hold' 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 *decyzje, które podjął* 'decisions which he took' is *podjął decyzję* 'he took a decision'
- (PT) the canonical form of *decisão nunca antes tomada* 'decision never before taken' is *tomar uma decisão* 'take a decision'
- (SL) the canonical form of *odločitev, ki jo je sprejel* 'decisions which he took' is *sprejeti 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'm 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 Schild um|fahren* '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-specifc 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:

- (DE) *Ich gebe dir mein Buch* ' I give you my book' → *Ich gebe Dir mein(e) Publikation/Doktorarbeit/Kapitel/Novelle/Ausgabe* 'I give you my publication/thesis/chapter/novel/edition'
- (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) *dam 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* 'I give you my word' → *daję Ci wyraz/sylabę/czasownik* 'I give you a word/syllable/verb'
- (SL) *dam ti besedo* lit. *I give you a word* 'I give you my word' → *#dam ti izraz/zlog/glagol* '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) *jmd. 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 dem 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'