Annotation guidelines for the PARSEME shared task on automatic detection of verbal multiword expressions

version 6.0

26 July 2016

Veronika Vincze, Agata Savary, Marie Candito, Carlos Ramisch, Fabienne Cap

These annotation guidelines are meant for the final annotation. You can see at a glance what is new in this version as compared to version 5 (used for the pilot annotation phase 1). See also Frequently Asked Questions for more detailed discussions of some notions and fuzzy cases.

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.

The purpose of this document is to define the annotation scope and to put forward a classification of VMWEs together with linguistic tests for VMWE identification and categorization. For the sake of simplicity, we cite mainly English examples here, with occasional references to other languages. However, language group leaders may adapt these guidelines to their language(s) of focus.

The notational convention used throughout the document is to display VMWEs in various colors depending on the language, e.g. for English in green, and to highlight in bold their lexicalized components (cf. Section 1.4). Counter-examples (i.e. expressions which resemble VMWEs but do not have the VMWE status) are highlighted in red, whatever the language. The language code appears in parentheses before each relevant example, except in English.

1. Definitions and scope

1.1 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 (e.g. *take, a, walk, astonishment*).
- Several tokens build up one word, e.g. abbreviations like *M|. (Mister), pp|.* (pages), possessives like *Pandora|'|s*, words with "accidental" separators like (FR) *aujourd|'|hui* 'today', inflected or derived forms of foreign names like (PL) *Chomsky|'|ego* 'of Chomsky', *SMS*|-|*ować* 'to write an SMS'. In this case we speak of a **multitoken word** (MTW).
- One token can contain several words, e.g. (EN) *don't = do not*, (DE) *Schulaufgabe = Schule+Aufgabe*, (IT) *della = de la*, (FR) *du = de le*. In this case we speak of a **multiword token** (MWT)[1]. Note that the precise word forms cannot always be

[1] See also the representation of MWTs in the Universal Dependencies.

straightforwardly deduced from the MWT containing them and vice versa, as in *don't, della, du*, etc.

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, as in *to **take** a **walk*** (4 words, 4 tokens), or not, as in *to **open a Pandora's box*** (5 words, possibly 7 tokens).
- A VMWE contains one (multiword) token, as in e.g. ***pretty-print***, (DE) ***vorbereiten*** lit. *pre-arrange* 'prepare', (FR) ***court-circuiter*** 'to short circuit'.

Note finally that multitoken words like (PL) *SMS-ować* 'to write an SMS' are not considered verbal MWEs since they contain one (multitoken) word only.

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.

1.2 Multiword expressions

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

- 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.
- They show some degree of orthographic, morphological, syntactic or semantic idiosyncrasy (see section 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*, etc.) are excluded from the scope of this study.
- At least two components of such a word sequence have to be lexicalized (see below). 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 below. Our underlying hypothesis is that (verbal) MWEs have some degree of semantic non-compositionality that implies limited flexibility.[2]

1.3 Verbal multiword expressions

Verbal MWEs (VMWEs) in this task include four syntactic types:

1. **Prototypical verbal phrases**, whose syntactic heads are verbs in finite forms and their other components are dependents of the verb. These phrases can be:
    - Partly saturated, i.e. only some of their arguments are lexicalized, e.g. ***made** a **decision**, **break** her **heart**, **took** this **to heart***, possibly their subjects, e.g. ***a little bird told** someone, the **problem lies** in something*.
    - Fully saturated, ***the early bird catches the worm***
2. **Meaning-preserving variants** of the following other syntactic categories
    - infinitives, e.g. *to **make** a **decision**, to **break** one's **heart***
    - nominal groups (headed by nominal complements from the prototypical VMWEs) with relative clauses, e.g., **decision** which he **made**, ***heart** which he **broke***
    - gerunds, e.g. ***decision making**, **heart breaking***
    - nominal and adjectival groups with participles, e.g. ***decisions** previously **made**, **heart-breaking**, **breaking** her **heart**, all **hearts broken** by him*

Note that expressions of the syntactic categories mentioned above are considered VMWEs only if they function as verb phrases (case 1) or nominal/participial phrases (case 2). Other kinds of variants are not considered VMWEs. This concerns e.g. nominalizations morphologically derived from verbs and describing an action or a state, e.g. *a take-off*, (FR) *une prise en compte* 'the fact of taking sth into account', (FR) *une mise à disposition* 'the fact of putting sth at one's disposal', as well as MWEs containing verbs but functioning as adverbials or nominals (other than in case 2), e.g. *forget-me-not,* (FR) *peut-être* lit. *may-be* 'maybe', *porte-feuille* lit. *carry-sheets* 'wallet'. Particular language teams may decide, however, to extend the annotation scope to this kind of 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 (cf. section 3) intact.

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

[2] Light verb constructions like *to **make** a **presentation*** are notable exceptions of rather compositional MWEs, specially when very generic operator verbs are used. An extra hypothesis captures them, see Section 5.

1.4 Lexicalized components and open slots

Just like a regular verb, the head verb of a VMWE may have a varying number of compulsory arguments, i.e. arguments that have to 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** i.e. always realized by the same (possibly morphologically variable) lexemes (here: *by surprise* is lexicalized while *someone* is not).[3] Obviously, the head verb of a VMWE is itself also considered lexicalized. When it can be replaced by another verb, like in *to **make/take** a **decision***, we consider that these are two different, although possibly synonymous, VMWEs. Conversely, a component (of a compulsory argument) which can be realized by a free lexeme (i.e. 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:

- *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**.*

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*). Conversely, in the third case the preposition *in* introduces an open slot whose meaning compositionally combines with the meaning of the VMWE ***took part***. We say in this case that the preposition is selected by the VMWE but it is not considered part of it and should not be annotated. Namely, 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 (belonging to the phenomenon of valency) is considered a regular property of the grammatical system and is outside of our annotation scope.

Reflexive clitics also have a special status with respect to the lexicalization criterion. In some, e.g. Slavic, languages the same reflexive clitic is used regardless of the person and number (it inflects for case only): (PL) ***znajduję się*** lit. *find.1.SG self* 'I find myself', ***znajdujesz się*** lit. *find.2.SG self* 'you find yourself', ***znajdują się*** lit. *find.3.PL self* 'they find themselves'. In others, e.g. Romance or Germanic, reflexive clitics agree in person and number with the subject and the verb: (FR) *je **me trouve*** lit. *I self.1.SG find* 'I find myself', *tu **te trouves*** lit. *you self.2.SG find* 'you find yourself', (DE) *sie **wundert sich*** lit. *she wonders self.3.SG* 'she wonders', *ihr **wundert euch*** lit. *you.PL wonder.2.PL self.2.PL* 'you wonder'. It this case the clitic is realized by different lexemes, depending on the number and gender, i.e. it is, formally speaking, not lexicalized in expressions like (FR) ***se trouver*** 'to find oneself' and (DE) ***sich wundern*** 'to wonder'. However,

[3] This definition of a lexicalised component naturally extends to any syntactic type of MWE. Namely, the head of a (nominal, adjectival, prepositional etc.) MWE is lexicalized (always realized by the same lexeme) together with at least one component of at least one of its modifiers.

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.

1.5 Verbal multiword expressions versus collocations

As mentioned in the preceding section, collocations are not considered VMWEs in this task and should not be annotated. However, the boundary between both categories is not always easy to detect and should be handled with care. We understand **collocations** as combinations of words whose idiosyncratic behavior is mainly of a statistical nature. In other words, they 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. Combinations such as *drastically drop*, *the graphic shows, to take a bus* happen to be very frequent and are perceived as "frozen". However, applying lexical alternations to them (e.g. *significantly drop, drastically decrease, the diagram shows, the graphic illustrates*, *to take a couch* etc.) does not markedly impact their meaning.

The difficulty of distinguishing collocations from VMWEs lies in the fact that lexical variability is relevant to some VMWEs (e.g. *to* ***come in handy/useful****, to* ***stand firm/fast****, to* ***break*** *someone's* ***spirit/will****, to* ***take a cake/biscuit***). 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 whole semantic class for each of its required arguments. For instance, the verb *to take* meaning '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/a wafer,* **to stand hard/rigid/solid* etc.[4] See also test 2 in section 5.

1.6 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 issue (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 for success rather than giving oneself several different possibilities.

However, not all (verbal) metaphors are 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

[4] Henceforth, an asterisk (*) preceding a sentence will mean that the sentence is ungrammatical, while a dash (#) means an unexpected change in meaning with respect to the original expression.

to lift dark cloud of stress meaning 'to relax' and *my heart … dances with the daffodils*, meaning '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).

2. Textual annotation scope

In this annotation task, all occurrences of all syntactic types of VMWEs are to be annotated in the text.

We annotate, as integral parts of VMWEs, all lexicalized elements that can form a separate word. For instance, lexicalized particles are annotated, but case suffixes are not. Thus, in *to **put** something **up***, the verb and the particle are integral parts of the VMWE (see section 6.5), while in (HU) ***döntést hoz** valamiről* lit. *decision-ACC bring something-ELA* '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. Verb+pronoun and verb+particle combinations are annotated only if the verb alone is a cranberry word, or if the pronoun or the particle markedly changes the meaning or the syntactic behavior of the verb - see sections 6.4 and 6.5 for details. Additionally, in some languages particles are homonymic with prepositions (e.g. *to **get up** a petition vs. to get up a hill*) and should be tested according to language-specific guidelines (linked from section 6.1). Note that in this version of the guidelines verb+preposition combinations (*to rely on somebody, to come across something*) are no longer considered VMWEs. Prepositions are parts of VMWEs only if they introduce lexicalized complements (*to **take** somebody **by surprise***).

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 the internal syntactic structure of its components. We do annotate, however, VMWEs embedded in other VMWEs, e.g. the VMWE *to **let the cat out of the bag*** contains the embedded VMWE ***let out*** and both are to be annotated as two different VMWEs.

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

3. Categories of verbal MWEs

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

- 2 **universal** categories, i. e. valid for all languages participating in the task
    - light verb constructions (**LVC**), e.g. *to **give a lecture***
    - idioms (**ID**), e.g. *to **go bananas**, **fortune favors the bold***
- 2 **quasi-universal** categories, valid for some language groups or languages but not all
    - inherently reflexive verbs (**IReflV**), e.g. (FR) ***se suicider*** 'to suicide'
    - verb-particle combinations (**VPC**), e.g. *to **do in***
- 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 e.g. ***drink and drive**, to **voice act**, to **pretty-print**, to **short-circuit**, to **tumble dry***

In sections 5 and 6, rigorous generic and category-specific tests are listed that can be used to practically identify and categorize verbal MWEs during manual annotation.

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[5] 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 *lexicalised*, that is, if they are omitted, the VMWE does not occur anymore. 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, following the decision trees and detailed tests in sections 5 and 6.

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

[5] Or an infinitival/nominal/gerund/participial variant of a verb.

Decision tree 1: Identification

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

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

↳**YES** ⇒ Annotate as a VMWE and go to **test 6 - [HEAD]**

↳**NO** ⇒ Apply **test 2 - [LEX:** *Regular replacement of a component unexpected meaning shift?***]**

↳**YES** ⇒ Annotate as a VMWE and go to **test 6 - [HEAD]**

↳**NO** ⇒ Apply **test 3 - [MORPH:** *Regular morphological change unexpected meaning shift?***]**

↳**YES** ⇒ Annotate as a VMWE and go to **test 6 - [HEAD]**

↳**NO** ⇒ Apply **test 4 - [MORPHSYNT:** *Regular morphosyntactic change unexpected meaning shift?***]**

↳**YES** ⇒ Annotate as a VMWE and go to **test 6 - [HEAD]**

↳ **NO** ⇒ Apply **test 5 - [SYNT:** *Regular syntactic change unexpected meaning shift?***]**

↳ **YES** ⇒ Annotate as a VMWE and go to **test 6 - [HEAD]**

↳ **NO** ⇒ Apply the **LVC hypothesis - [***Candidate has operator verb + activity or state noun?***]**

↳ **YES** ⇒ Assume a VMWE and go to **test 6 - [HEAD]**

↳ **NO** ⇒ It is not a VMWE, **exit**

Decision tree 2: Categorisation

↳ 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**

↳ **NP or PP** ⇒ Apply **LVC-specific decision tree ⇒** *Answer positive?*

↳ **YES** ⇒ Annotate as a VMWE of category **LVC**

↳ **NO** ⇒ Annotate as a VMWE of category **ID**

↳ **Other category** ⇒ Annotate as a VMWE of category **ID**

5. Generic tests for identifying VMWEs

In order to decide if a candidate is a VMWE, we apply the following generic idiosyncrasy tests. If a candidate expression passes at least one test from 1 to 5, we consider it to be a VMWE, and it can further be categorised 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
  - *to* ***go astray*** - *astray* is not a stand-alone word (test passed)
- **NO** ⇒ further tests are required
  - *to* ***go on*** - both *go* and *on* are a stand-alone words (test not passed)
  - *to go away* - both *go* and *away* are stand-alone words (test not passed)

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. It only occurs in a particular expression (or a closed list of expressions) and can never be found in different contexts.

Test 2 - [LEX] - Lexical inflexibility

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

- **YES** ⇒ it is a VMWE
  - #*to allow the feline out of the container* (*to* ***let the cat out of the bag*** passes the test)
  - **to produce/build/create a decision* (*to* ***make*** *a* ***decision*** passes the test)
  - **to go upon* (*to* ***go on*** passes the test)
  - #*to take a cookie/a wafer* (*to* ***take a cake/a biscuit*** 'to do something worse than ever' passes the test)
  - **to stand hard/rigid/solid* (*to* ***stand firm/fast*** passes the test)
- **NO** ⇒ further tests are required
  - *to* ***commit*** *a* ***crime/suicide/theft/felony*** *etc.* (*to* ***commit*** *a* ***crime*** does not pass the test)
  - *to take a plane/bus/car, etc.* (*to take a plane* does not pass the test)

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 morphological 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
  - #*to kick the buckets* (*to* ***kick the bucket*** passes the test)
  - #*to prettier-print* (*to* ***pretty-print*** passes the test)
  - #*to take a turn* (*to* ***take turns*** passes the test)
- **NO** ⇒ further tests are required
  - ***make/makes/made*** *a/many/those/no* ***decision(s)*** (*to* ***make*** *a* ***decision*** does not pass the test)
  - ***make/makes/made*** *a/many/those/no* ***cake(s)*** (*to make a cake* does not pass the test)

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

- **YES** ⇒ it is a VMWE
  - #*I* ***give*** *you his* ***word*** *for that* (*I* ***give*** *you my* ***word*** *for that* passes the test)
  - #*I was pulling my leg* (*he was* ***pulling*** *my* ***leg*** passes the test)
- **NO** ⇒ further tests are required
  - *he* ***made*** *his/her/our/your* ***dreams come true*** (*he* ***made*** *his* ***dreams come true*** does not pass the test)
  - *he made his/her/our/your kids dream* (*he made his kids dream* does not pass the test)

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
  - #*he was speaking of the devil* (***speak of the devil*** passes the test)
  - #*bananas are gone* (*to* ***go bananas*** 'to get crazy' passes the test)
  - #*drive and drink* (***drink and drive*** passes the test)
  - #*the bucket was kicked* (*to* ***kick the bucket*** passes the test)
- **NO** ⇒ further tests are required
  - *her* ***heart*** *was* ***broken****, the heart that he broke,* ***heart breaking****,* etc. (*to* ***break*** *one's* ***heart*** does not pass the test)
  - *her car was washed*, *the car that she washed*, *car washing*, etc. (to wash one's war does not pass the test)

LVC Hypothesis

Does the candidate consist of a verb and a nominal complement, where the verb has a purely 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
  - In *to* ***make a decision***, *make* only expresses that an activity (*decision*) happened
  - In *to* ***have courage***, *have* only expresses that the subject has a property (*courage*)
  - In *to* ***commit suicide***, *make* only expresses that an activity (*suicide*) happened
- **NO** ⇒ the candidate is NOT a VMWE
  - In *to make a cake*, *make* has a concrete meaning and the thing being made (*cake*) is not an activity or state
  - In *to have neighbors*, *have* could be an operator verb, but *neighbors* are not activities or properties

    - In *to give hope*, *hope* is a state/property, but *give* adds inchoative (i.e. change-of-state) semantics to it

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.

6. Specific tests for categorizing verbal MWEs

Once a verbal MWE candidate has been pre-identified according to one of the criteria from the preceding section, the confirmation of its status as a VMWE, as well as its categorization can be based on category-specific tests proposed below.

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 categorisation 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
    - In *to **make a face**, make* is the head and the NP *a face* depends on it (test passed)
    - In *to **give up***, *give* is the head and *up* is a particle depending on it (test passed)
- **NO** ⇒ annotate as OTH
    - in ***pretty-print**,* there is an unusual case of an adjective modifying a verb (test not passed)
    - in ***drink and drive***, none of the verbs is clearly a headword, since there is no universally accepted syntactic representation of coordination (test not passed)

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:

- Since ***make decision*** passes the test, its variants like *the **decision** which was **made**, **decision-making**, the **making** of the **decision*** pass the test as well, even though there may be no verb (*the **making** of the **decision**)* or the verb may not be the syntactic head (*the **decision** which was **made***)

Test 7 - [1DEP] - Single dependent

Does the VMWE contain exactly one lexicalised syntactic dependent of the head verb?

- **YES** ⇒ continue to next test
  - in *to **make a face**,* the single dependent is a noun phrase, *a face* (test passed)
  - in *to **take into account**,* the single dependent is a prepositional phrase, *into account* (test passed)
  - in *to **take turns**,* the single dependent is a noun, *turns* (test passed)
  - in to **give up**, the single dependent is a particle, *up* (test passed)
- **NO** ⇒ annotate as ID
  - *to **make ends meet*** has two dependents, *ends* and *meet* (test not passed)
  - *to **let the cat out of the** bag* has two dependents, *the cat* and *out of the bag* (test not passed)

The test covers only lexicalised dependents. There may be other, non-lexicalised dependents, which the test ignores. We explicitly call the non-verbal elements *dependents* instead of *arguments* or *complements* because the traditional and polemic argument-adjunct distinction is irrelevant here. The outcome of the test is positive if the verb has a single lexicalised 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 that co-occurs with the head verb?

- **Reflexive clitic -** apply IReflV tests. If the outcome is negative, discard the VMWE candidate.
  - Impossible in English
  - (FR) ***se suicider*** lit. *SELF suicide* 'commit suicide', ***s'évanouir*** 'lose consciousness'
- **Particle -** apply VPC tests. If the outcome is negative, discard VMWE candidate.
  - *to **give up**, to **look forward** to*
- **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.
  - in *to **make a wish**, a wish* is a noun phrase composed by a determiner and a noun
  - in *to **take turns**, turns* is a noun phrase composed by a single plural noun
  - in *to **take** it **into account**, into account* is a prepositional phrase composed by a preposition governing a noun
- **Other -** categorize as ID.
  - Adjective: *to **stand firm**, to **see red***
  - Verb: *to **make do***
  - Adverb: *to **get well***
  - Pronoun: *to **make it***
  - Etc.

Reflexive clitics are a special type of object pronoun that refers to the subject of the verb. 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:

- FR *me, te, se, nous, vous*
- PT *me, te, se, nos, vos*
- PL *się, sobie*

See the guidelines of IReflV category for more details.

Particles are hard to distinguish from homographic prepositions, e.g. *to **get up** a petition vs. to get up a hill*, (DE) *ich **schlage vor** allen zu verzeihen* 'I propose to forgive everyone'*, ich **schlage** vor allen Dingen **vor*** 'I propose prior to anything'. 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, e.g. (DE) *den See umfahren* 'to drive around the lake' vs. *das Schild **um|fahren*** 'to drive over the sign'. Most tests discriminating particles from prepositions and prefixes are language-specific and should be proposed by the individual language team. See a separate document for language-specific tests in English and German.

6.2 Light-verb constructions[6]

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

1. They are formed by a verb *v* and a noun n, which either directly depends on v (*to **give** a **lecture***), or is introduced by a preposition (*to **come into bloom***).
2. The noun *n* refers to an event (*decision, visit*) or a state (*fear*, *courage*).

[6] This chapter was significantly modified wrt. the previous working version, still available here.

3. The verb *v* is “light”, i.e. it contributes to the meaning of the whole only by bearing tense and mode (it is "light" either per se, or when used in the specific context of the noun).[7] This implies that *v*’s syntactic subject is n’s semantic argument.[8]

The noun *n* functions as a regular syntactic dependent, so LVCs exhibit regular syntactic variants listed in section 1.3 (e.g. *the **decision** that the director has to **make***).

In many cases of LVCs it can be said that there is some degree of selection of the verb by the noun. For instance ***have** a **walk**/*a race* and ***run** a **race**/*a walk*, (FR) ***faire** une **marche** lit. make a walk* 'take a walk', (FR) **procéder à une promenade lit. perform a walk*, but (FR) ***faire/procéder à** une **enquête*** 'make/perform an inquiry', and (FR) ***commettre** / *faire un **crime*** 'commit / *do a crime'. Yet some regularities exist, 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 for LVC categorization. Instead, the following decision tree should be applied, in order for a candidate to be annotated as an LVC:

LVC-specific decision tree:

Note: in this tree, one 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 no 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

---

[7] Many authors make a distinction between support verbs and light verbs, still others differentiate between true light verbs and vague action verbs. In this shared task, 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 take in causative support verbs (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.

[8] In a larger understanding of LVCs, any syntactic argument of *v* could play the role of *n*'s semantic argument, but we only retain the cases when *v*'s subject is concerned. This restriction facilitates the definition operational and easily applicable tests (notably test 12).

Test 9 - [N-EVENT] - Noun denoting an event/state

Does *n* refer to an event or state (including permanent or non-permanent properties, relations) with at least one semantic argument[9]

- **YES** ⇒ continue to next test
  - in ***pay a visit***, *visit* refers to an event, has two arguments: the visitor and the visitee (test passed)
  - in ***have strength***, *strength* refers to a property and has one semantic argument: the entity having strength (test passed)
  - in ***take a glance*** (at sg), *glance* refers to an event, with two arguments: the entity glancing, and the entity glanced at (test passed)
- **NO** ⇒ it is not an LVC
  - in *Joe made a cake*, *Joe* could be considered a semantic argument of the cake: the person who made the cake, but *cake* refers to a physical entity (test not passed, not an LVC)
  - In *Joe experienced a tornado*, *tornado* is an event but has no semantic argument (test not passed, not an LVC)

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

Is the noun *n* used in one of its original senses?

- **YES** ⇒ continue to next test
  - in ***pay a visit***, *visit* is literally understood (test passed)
- **NO** ⇒ it is not an LVC
  - in ***have kittens*** 'to be worried or angry', *kittens* is not used literally (test not passed, not an LVC)

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

Does *v* only bear tense and mood, and add no semantic to *n* other than the idea of an entity performing an activity / having a property / being in a certain state (depending on the noun's semantic type) ?

- **YES** ⇒ continue to next test[10]
  - in ***make a decision***, *make* adds no meaning to *decision* (test passed)
  - in ***have fear***, *have* adds no meaning to *fear* (test passed)
  - in ***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 (test passed)

---

[9] A semantic argument of a noun is semantically mandatory (a visit cannot hold if there is no visitor, or no visitee, *courage* applies to a human being ...), and its interpretation is dependent on the semantics of the noun. This rules out temporal or locative adjuncts: in "the walk of John in the forest in 1990", all three dependents (the entity walking, the time and place) are semantically necessary, but space-time localization is interpreted independently of the semantics of the noun.

[10] Note that in some of the examples below the LVC can be replaced by a single verb morphologically derived from *n*: *to* ***make a decision*** *= to decide, to* ***have fear*** *= to fear, to* ***pay a visit*** *= to visit*, etc. This test is sometimes used in the literature for LVC identification. Note however that it is neither sufficient nor compulsory, e.g. it does not apply to ***commit a crime***.

    - in ***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 (test passed)
    - in (FR) ***avoir du courage*** lit. *to have (some) courage*, *have* adds no meaning to *courage* (test passed)
    - in ***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 (test passed)
    - in ***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 (test passed)
- **NO** ⇒ it is not an LVC
    - in *to start a walk*, *start* adds the aspectual meaning to the noun (test not passed, it is not an LVC)

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
    - *Paul's walk* and *Paul **had a walk*** can refer to the same walking event (test passed)
- **NO** ⇒ it is not an LVC
    - *Paul made a good impression on his wife / *The Paul's impression on his wife* (test not passed, *not an LVC)*

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

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*?[11]

- **YES** ⇒ it is an LVC
    - *A visit of the Lady to the Prime Minister, The Queen **paid a visit** to the Prime Minister. - *The Queen paid a visit of the Lady to the Prime Minister* (the visitor cannot be a modifier of *visit*, test passed)

[11] An alternative formulation for this test is the following. Does *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). 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*).

    - *Paul **made** a **decision** on the budget. - *Paul made the committee's decision on the budget* (the decision maker cannot modify *decision*, test passed)
    - *Paul **leads** the **discussion**. - *Paul lead's Peter's discussion. - Paul **leads** the **discussion** of the committee* (the discussing entity can modify *discussion* only when *Paul* is part of the *committee*, test passed)
    - *Bjarnason **scored** a **goal**. - *Bjarnason scored Arnason's goal. - Bjarnason **scored** the **goal** of Iceland* (the scoring entity can modify *goal* only in the last case, when *Bjarnason* is part of the *Iceland* team, test passed)
- **NO** ⇒ it is not an LVC
    - *Paul transmitted the advice to his sister. - Paul transmitted Peter's advice to his sister* (the advice author can modify *decision,* test not passed, not an LVC)

6.3 Idioms

Idioms constitute another 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:

- subject, e.g. ***a little bird told** someone*
- direct object, e.g. *to **kick the bucket***
- circumstantial or adverbial complement e.g. *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***.
- etc.

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 (*to **pay** a **visit**, to **come into bloom***) and an ID (*to **kick the bucket**, to **come into play**, to **sleep like a log***), cf. section 6.2
- particle or reflexive pronoun - the VMWE is either a VPC (***set up***) or an IReflV (FR ***se suicider*** 'suicide'), 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, e.g. *to **take** something **with a pinch of salt***,
- adjectival phrase e.g. *to **come clean**, to **stand firm***
- verbal phrase e.g. *to **make do***, (FR) ***laisser tomber*** (lit. *to let fall*) 'to give up''
- relative clause e.g. *to **know on which side the bread is buttered***
- non-reflexive pronoun e.g. *to **make it***, (FR) ***l'emporter*** (lit. *to take it away*) 'to win', (DE) ***es gibt*** lit. *it gives* 'there is', (IT) ***prender-le*** lit. *to take it* 'to be beaten'
- etc.

If more than one argument of the head verb is lexicalized, then the candidate VMWE it is always classified as an ID, as in *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', (FR) *s'en aller* lit. *to go SELF from there* 'to leave', (FR) ***il y a*** lit. *it there is* 'there is'. Notably, sentential expressions with no open slots, such as proverbs and conventionalized sentences (***Rome was not built in a day, fortune favors the bold, the pleasure is mine, I beg you pardon!***), are included in the scope of IDs.
In case of several lexicalized arguments special care must be taken in identifying embedded VMWEs. For instance in (FR) ***se faire des idées*** lit. *to make SELF ideas* 'to imagine something false', neither *se faire*, nor *faire des idées* are autonomous VMWEs, so they should not be annotated as embedded. Conversely, *to **let out*** is a VPC and should be annotated as embedded in *to **let the cat out of the bag***.

Idioms whose head verb is the **copula** (*to be*) can pose special challenges because their complements may be (nominal, adjectival, etc.) MWEs themselves. For instance, in *it is double Dutch to me* the copula can be omitted as in *he seems to speak double Dutch*. 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. For instance, *to **be no chicken***, *to **be somebody***, *etc.* are idioms, while *to be double Dutch to someone* is not. 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 these, languages-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 section 1.6). This often makes them semantically totally non-compositional, i.e. none of their lexicalized components retains any of their original meanings.[12]

6.4 Inherently reflexive verbs

Inherently reflexive verbs (**IReflV**) is a quasi-universal category, i.e. it applies to some language groups or languages participating in this task but not to all. The list of relevant languages currently includes: Romance languages (French, Italian, Portuguese, Romanian, Spanish), some Slavic languages (Polish), some Germanic languages (German). This list is meant to evolve while the annotation guidelines become adapted to other languages.

IReflVs include verbs combined with a reflexive clitic that:

- is compulsory, i.e. the verb alone is a cranberry word, like in (FR) ***se suicider*** 'to suicide'
- or markedly changes the meaning or the subcategorisation frame of the verb, like in (FR) ***s'apercevoir*** ≠ *apercevoir* 'realize' ≠ 'see', (ES) *X **se olvidó** de Y* 'X SELF forgot of Y' vs. *X olvidó Y* 'X forgot Y'

[12] Some authors argue though that partial semantic compositionality can be obtained via decomposability, e.g. *to **spill** the **beans*** is compositional provided that *to spill* is paraphrased as *to reveal* and *the beans* as *a secret*.

See separate guidelines for inherently reflexive verbs for specific linguistic tests discriminating inherently pronominal verbs from others.

6.5 Verb-particle constructions

Verb-particle constructions (VPCs), sometimes called phrasal verbs or phrasal-prepositional verbs, like *to **put off***, *to **blow up***, *to **do in**,* (DE) ***um**|**fahren***, ***mit**|**kommen***, ***vor**|**bereiten***, etc., constitute another quasi-universal category. They have the following general characteristics:

- They are formed by a lexicalized head verb *v* and a lexicalized particle *p* dependent on *v*
- The meaning of the VPC is non-compositional. Notably, the change in the meaning of the *v* goes significantly beyond adding the meaning of *p* (e.g. *to **do in*** = to kill).

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, as in *Die Kinder sollen in der Schule **aufpassen*** 'The children must pay attention at school', *Herr Müller, **passen** Sie **auf**!* 'Mr. Müller, be careful'.

The first challenge in identifying a VPC is to properly distinguish the particle from a possibly homographic preposition, e.g. *to **get up** a petition* vs. *to get up a hill*, or a verbal prefix, e.g. (DE) *um-* in ***um**|**fahren*** vs. *umfahren*. Namely, a particle, contrary to a preposition, cannot introduce a complement (*to **do** sb **in**, *to do in sb*), and prefixes can never be spelled separately from the verb (**er fuhr den See um*), nor can the past tense of prefixed verbs be formed with the infix *-ge-* (**er hat den See **um**ge**fahren***, instead: *er hat den See umfahren* but: e*r hat das Schild **um**ge**fahren***). See the language-specific tests linked from section 6.1 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 (*lie down, come in*). Some combinations may have both compositional and non-compositional meanings depending on the context (e.g. *to put up a flag* vs. *to **put up** a friend for the night*), and only the latter should be annotated. 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 14 - [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 the meaning of the v+p construction **fail** to imply the meaning of a reformulation in which *v* appears without *p*?

- **YES** ⇒ it is a VPC
  - *to **do** somebody **in*** (= to kill) does not imply *to do somebody* (*to **do in*** passes the test, it is a VPC)
  - *to **check in** upon arrival* does not imply *to check upon arrival* (*to **check in*** passes the test, it is a VPC)

- **NO** ⇒ it is not a VPC
    - *to look up into the sky* implies *to look into the sky* (*look up* does not pass the test)
    - *to eat up the cookies* implies *to eat the cookies* (*to eat up* does not pass the test)

Language-specific tests for non-compositionality can be defined if needed - see the separate Hungarian and German-specific guidelines for VPCs.

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. It is recommended not to redefine the universal and quasi-universal categories described above but to introduce new names and abbreviations in order to answer such needs.

6.7 Other verbal MWEs

This category is meant to contain VMWEs which do not fit to the preceding categories, i.e. whose lexicalized components do not include a head verb and at least one of its arguments. This includes:

- **coordinations** of verbs, e.g. *to **drink and drive***
- **compound** verbs (resulting usually from conversion of nominal compounds), e.g. *to **short-circuit**, to **pretty-print**, to **voice act***

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 from section 1.3
- it is a VMWE, i.e. it fulfills one of the 5 idiosyncrasy tests from section 5
- it cannot be classified into any universal (LVC or ID), quasi-universal (IReflV, VPC) or language-specific category

Glossary

Notion	Definition	Examples[13]	Comments
collocation	a combination of words (components of a syntagm) whose idiosyncratic behavior is mainly of a statistical nature; they are notably semantically compositional	*to take a bus* *drastically drop* *the graphic shows*	Verbs in a collocation select arguments taken from large semantic classes (although they prove statistically idiosyncratic only with few representatives of these classes).
lexicalized component (of a MWE)	a component of a VMWE which is always realized by the same (possibly morphologically variable)	*He **took** me **by surprise***	When the head word of a MWE can be replaced by another verb (*to make/take a decision*), we consider

[13] As in the whole document, the annotated components of VMWEs are highlighted in violet

	lexeme; the head verb of a VMWE is always lexicalized		that these are two different, although possibly synonymous, MWEs.
lexicalized preposition	a preposition which introduces a lexicalized complement	*He **took** me **by surprise***	Lexicalized prepositions are to be distinguished from the selected prepositions. Only the former are annotated.
metaphor	an expression in which one concept is seen in terms of the properties of the other	*to **take the bull by the horns*** *to lift the dark cloud of stress over one's head*	Some VMWEs are metaphors but some others are not and, importantly for this task, some metaphors are no MWEs; the distinction between the two categories is an open problem
multitoken word (MTW)	a word split by the tokenizer into several tokens (e.g. due to a tokenizer's imprecision)	*Pandora|'|s* (PL) *SMS|-|ować* (PL) *Skype|'|ować* (FR) *aujourd|'|hui*	MTWs are annotated in this task only if they are parts of VMWEs. Annotating them is optional and can be decided by each language group.
multiword expression (MWE)	a continuous or discontinuous sequence of words which: (i) is a syntagm (with possible open slots), (ii) shows some degree of orthographic, morphological, syntactic and semantic idiosyncrasy, (iii) has at least two lexicalized components, one of which is the head word		
multiword token (MWT)	a token containing several words (due to a contraction or a tokenizer's imprecision)	*don't* (FR) ***court-circuiter*** (IT) *della*	One MWT alone can sometimes be a VMWE (e.g. *court-circuiter* 'to short circuit'). Splitting MWTs into individual words is optional and has to be done prior to the annotation.
open slot	a required but non-lexicalized component of a VMWE	***Money burns a hole in Bob's pocket**.* *He **took** his sister **by surprise***	Open slots are not annotated in this task. See also selected prepositions, which are boundary cases between lexicalized components and open slots.
selected preposition	A preposition that is selected by a particular sense of a verb, i.e. belonging to its valency frame. It is usually lexically fixed (cannot be replaced by another preposition) but is not considered part of a VMWE unless it introduces a lexicalized complement.	*I don't want to participate in this.* *I don't want to **take part** in it.* *He **took** me **by surprise**.* *He grasped me by the wrist.*	Only selected prepositions introducing lexicalized complements (*to take sb by surprise*) are to be annotated. Those introducing open slots (*to grasp someone by the wrist*) are not annotated.
token	technical and pragmatic notion, defined according to more or less linguistically motivated clues and depending on the particular tokenization tool at hand	*do* *s* *don't* *della* *aujourd*	Tokenization errors should not be corrected by the annotators (and the evaluated tools), so as to allow easy comparisons of parallel annotations.
verbal multiword expression (VMWEs)	A MWE which functions as a (possibly unsaturated) verb phrase; syntactic variants of prototypical VMWEs, e.g. gerunds, relative clauses and participles are included in the VMWE category	*to **break** one's **heart*** *The **heart** that he **broke*** *He is an expert of **heart breaking*** *A **heart-breaking** news*	

word	linguistically (notably semantically) motivated unit; this notion, thus, language-dependent and annotation experts should have a clear idea of how to define it for their own language	*do* *not* *astonishment* (IT) *de* (IT) *la* (FR) *aujourd'hui*	