Annotation guidelines for the PARSEME shared task on automatic detection of verbal Multi-Word Expressions

version 5.0

4 March 2016

Veronika Vincze, Agata Savary, Marie Candito, Carlos Ramisch

This is a new version of the annotation guidelines meant for the pilot annotation phase 2. To see in a glance what is new in this version as compared to version 4 (used for the pilot annotation phase 1) see this document.

In this shared task, we aim at identifying verbal Multi-Word Expressions in running texts. They are of particular interest to the PARSEME COST action (www.parseme.eu) since they frequently introduce discontinuity and long-distance dependency issues, which are central to deep parsing.

The purpose of this document is to summarize the properties of several categories of verbal MWEs and to provide basic annotation guidelines for them. For the sake of simplicity, here we focus on English examples, with occasional comments on other languages. However, language group leaders may adapt these guidelines to their language(s) of focus.

1. Definitions and scope

1.1 Words and tokens

While the definition of a 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 *aujourd|'|hui* 'today' (FR), inflected or derived forms of foreign names like *Chomsky|'|ego* 'of Chomsky' (PL), *SMS|-|ować* 'to write an SMS' (PL). In this case we speak of a **multi-token word** (MTW).
- One token can contain several words, e.g. *don't=do not* (EN), *Schulaufgabe* (DE), *della = de la* (IT), *du=de le* (FR). In this case we speak of a **multi-word token** (MWT)[1]. Note that the precise word forms cannot always be 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 (multi-word) token, as in e.g. *pretty-print*, *court-circuiter* 'to short circuit' (FR).

Note finally that MTWs like *SMS-ować* 'to write an SMS' (PL) are not considered verbal MWEs since they contain one (multi-token) 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 Verbal multi-word expressions

Multi-Word Expressions (MWEs) are understood here as (continuous or discontinuous) sequences of words of words with the three compulsory properties:

- Their component words include a head word and at least one other syntactically related word. Most often the relation they maintain if 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 and semantic idiosyncrasy (see chapter 6) with respect to what is deemed 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 the open slots.

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

1. Prototypical verbal MWEs (VMWEs) - MWEs which function as (possibly unsaturated) verb phrases i.e. their syntactic heads are verbs in finite forms and their other lexicalized

[1] See also the representation of MWTs in the Universal Dependencies: http://universaldependencies.org/u/overview/tokenization.html

components are dependents of the verb (e.g. *made a decision, break her heart, took this to heart*).

2. Nominal, participial and other syntactic variants of prototypical VMWEs maintaining their idiomatic reading, e.g. *decisions which we made, decision making, heart-breaking*.
3. Partly lexicalized sentential MWEs with lexicalized subjects, e.g. *a little bird told someone, the problem lies in something*.
4. Fully lexicalized sentential MWEs (or sentential VMWEs for short) e.g. *the early bird catches the worm, better late than never*.

Note that the four syntactic categories mentioned above are considered VMWEs only if they function as verb phrases (case 1), nominal/participial phrases (case 2) or sentences (cases 3 and 4). MWEs containing verbs but functioning as adverbials or nominals (other than in case 2) are not considered VMWEs, e.g. (FR) *peut-être* (lit. *may-be* 'maybe'), *porte-feuille* (lit. *carry-sheets* 'wallet').

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* and *surprise*, are lexicalized while *someone* is not).[2] Obviously, the head verb of a VMWE is itself also considered lexicalized. When it can be replaced by another verb, like in t*o 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 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****.*

Prepositions are notoriously hard to classify as lexicalized components vs. open slots. In the first, second and fourth example above, the prepositions *by* and *in* are lexicalized since they introduce lexicalized complement (*the horns, surprise* and *pocket*). However in the third case the preposition *in* introduces an open slot whose meaning compositionally combines with the meaning of the VMWE *took part*. We say in this case that the preposition is selected by the VMWE but it is not considered part of it.

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 is considered a regular property of

[2] 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.

the grammatical system, as long as the preposition is not compulsory (unlike in *to rely on*) and it does not markedly modify the meaning of the verb. For instance the preposition *across* is selected by its governing verb in *to travel across a country* while it is lexicalized in *to come across an old photograph*. One of the tests to tell a selected from a lexicalized preposition is to check if it can be omitted without markedly changing the meaning of the verb (e.g. *to travel there* vs. *to come across this*). Other languages may have specific tests, for instance in the French verb *participer à* 'to take part in' the preposition *à* is selected by the verb but it is not lexicalized since pronominal variants omitting the preposition are allowed (*y participer* 'to take part therein'). Section 3.4 discusses non-compositional verb+preposition combinations, which are to be annotated in this task.

1.3 Verbal multi-word 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 (components of a syntagm) 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 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 in the case of collocations than of the 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/a 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. See also test (2) in section 6.

1.4 Verbal multi-word 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 *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 expression, 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

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 prepositions are pointed at but case suffixes are not. Thus, in t*o rely on sth*, the verb and the preposition are integral parts of the VMWE (see section 3.4), while in the Hungarian *döntést hoz valamiről* 'decision-ACC bring something-ELA'='make a decision', only *döntést hoz* is annotated, even if the delative case suffix is also lexically determined.

Reflexive pronouns and **prepositions** need to be handled with special care. Verb+pronoun and verb+preposition combinations are annotated only if the verb alone is a cranberry word, or if the pronoun or the preposition markedly changes the meaning of the verb - see sections 3.3 and 3.4 for details. Additionally, in some languages prepositions are homonymic with particles (*to do in*) and should be tested according to the specific guidelines sketched in section 3.5. Note that in this version of the guidelines we no longer give a special status to selected (non lexicalized) prepositions.

Both **continuous** and **discontinuous** lexicalized components of VMWEs are annotated.

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 one (or at most two, in case of hesitation) of the categories described in the following section.

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*
- 3 quasi-universal categories, valid for some language groups or languages but not all
    - verb-particle combinations (**VPC**), e.g. *to do in*;
    - inherently pronominal verbs (**IPronV**), e.g. (FR) *se suicider* 'to suicide'
    - inherently prepositional verbs (**IPrepV**), e.g. *to come across sth*, *to rely on sth*
- 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, fortune favors the bold, better late than never*

This section contains a general description of these categories except the language-specific ones. It is meant to provide intuitions as to the nature of these categories rather than a formal list of sufficient or necessary defining conditions. Then, in sections 4 and 5, more rigorous generic and category-specific tests are listed that can be used to practically identify and categorize verbal MWEs during manual annotation.

3.1 Light verb constructions

Light verb constructions (**LVCs**) constitute a universal category and have the following general characteristics:

1. They are formed by a verb and its argument containing a noun. The argument is usually a direct object (*to give a lecture*) but sometimes also a prepositional complement (*to come into bloom*) or a subject (*the problem lies in sth*).
2. Both the verb and the noun (included in the complement) are lexicalized.
3. The verb is “light”, i.e. it contributes to the meaning of the whole only to a small degree (e.g. aspectual information).
4. The noun has one of its regular meanings (which can be retrieved even in the absence of the verb).
5. The noun is predicative, i.e. takes at least one syntactic argument, and, when used with the light verb, one of its arguments becomes also a syntactic argument of the verb (e g. in *to pay a visit to a friend* the prepositional phrase *to a friend* is an argument both of *pay* and of *visit*). Also, the subject is usually an argument of the noun (here, the one who pays is also the one who visits). See section 7.2 for details.
6. The noun typically refers to an action or event.

As in most other VMWEs, the nominal and the verbal component of such constructions can be separated from each other in context (e.g. in passive sentences: *a decision was made by the committee*).

Many authors make a distinction between support verbs and light verbs, still others differentiate between true light verbs and vague action verbs. Here, however, we take a comprehensive approach and those verb + argument combinations that fulfill most of the above linguistic criteria are annotated (see section 7.2 for details).

3.2 Idioms

Idioms constitute another universal category. An **idiom** (**ID**) is a VMWE composed of a head verb (possibly phrasal) and at least one of its arguments. The complement can be of different types:

- subject, e.g. *a little bird told someone*
- direct object, e.g. *to kick the bucket*
- indirect object, e.g. *to throw someone to the lions*
- 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*.

The complement can be realized by syntactically different structures:

- nominal phrase, e.g. *to kick the bucket*
- prepositional phrase, e.g. *to throw someone to the lions, to take something with a pinch of salt*
- adjectival phrase e.g. *to be born under a lucky star*
- relative clause e.g. *to know on which side the bread is buttered*
- etc.

Several lexicalized complements of different functions and structures can co-occur (*to let the cat out of the bag, to cut a long story short, to call it a day*).

Idioms typically have both a literal and an idiomatic reading, thus they are closely connected to the phenomenon of a metaphor (see also section 1.4). This often makes them semantically totally non-compositional, i.e. none of their lexicalized components retains any of their original meanings.[3]

3.3 Inherently pronominal verbs

Inherently pronominal verbs (**IPronV**) is a quasi-universal category, i.e. it applies to some language groups or languages participating in this task but not all. It includes verbs combined with a reflexive clitic that:

- is compulsory, i.e. the verb alone is a cranberry word, like in (FR) *se suicide* 'to suicide'
- or markedly changes the meaning of the verb, like in (FR) *s'apercevoir ≠ apercevoir* 'realize' ≠ 'see'

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

Detailed guidelines on inherently pronominal verbs have been put forward on the basis on 5 Romance and 1 Slavic language. They contain precise tests to distinguish inherently pronominal verbs from others. They are meant to be extended to other languages and language families.

3.4 Inherently prepositional verbs

Inherently prepositional verbs (**IPrepV**) constitute another quasi-universal category. For its definition, we rely on the guidelines of a related DIMSUM initiative. IPrepVs are verb+preposition combinations in which

- the dependents of the preposition are not lexicalized (unlike in *to come of age*) and
- the preposition is "integral", i.e. "it cannot be omitted without markedly altering the meaning of the verb"; for instance the preposition *across* is integral in *to come across an old photograph*, while it is not in *to come across fields*.

Note that prepositional verbs, in which the preposition opens a slot for a complement, should not be mistaken for verb-particle constructions (VPCs). For instance *to carry on* (to continue) is a verb-particle construction, while *to rely on sth* is a prepositional verb. Tests allowing to distinguish particles from prepositions have been defined in a separate document.

Detailed guidelines on integral prepositions have been developed in DIMSUM shared task for English. Their adaptation to other languages is left to the individual language teams, as for all other universal categories.

3.5 Verb-particle constructions

Verb-particle constructions (VPCs), like *to put off*, *to blow up*, *to do in,* etc., constitute another quasi-universal category. They are pervasive in English, German, Hungarian and possibly some other languages but irrelevant to or very rare in Romance and Slavic languages or in Farsi and Greek for instance. Note that VMWEs from this category should not be mistaken for prepositional verbs like *to come across sth*, *to rely on sth*, etc. - see section 3.4. Namely, a particle, contrary to a preposition, cannot introduce a complement (*to do sb in, *to do in sb*). Guidelines for VPCs with linguistic tests in English and Hungarian have been described in a separate document. They should be adapted to each other language to which this category is relevant.

3.6 Language-specific categories

Language-specific categories can be proposed for annotation in this task provided that their are carefully defined and accompanied by linguistic tests that allow to distinguish them from other categories. A possible example of a category would be are **compound verbs**, i.e. non-compositional constructions composed of only (or at least two) verbs. They seem frequent in Romance languages and in Farsi, and exist in other languages, too. Examples include (FR) *va savoir* 'go know'='I have no idea', *go figure, make do*.

3.7 Other verbal MWEs

This category is meant to contain VMWEs which do not fit to the preceding categories, including notably:

- verbal expressions with **no lexicalized complements** such as *drink and drive*, *to tumble-dry* etc.
- **proverbs**, i.e. sentences expressing facts thought to be true by most people, e.g. *Fortune favors the bold*, possibly with omitted head verbs, e.g. *better late than never, loin des yeux, loin du coeur* 'far from the eyes, far from the heart' (FR).
- totally lexicalized and often morphologically and syntactically **frozen phrases**, e.g. *The pleasure is mine. I tu jest pies pogrzebany*. 'and here is the dog buried'='here is the essence of the problem' (PL)
- **exclamations**, e.g. *I beg you pardon! Co ja widzę!* 'what do I see?!' = 'what a surprise!' (PL)

Most VMWEs in this category are fully lexicalized (they have no open slots for required complements of the head verb). Additionally, sentential VMWEs from the 3 last tokens are also fully morphologically and syntactically frozen, e.g. *#the pleasures are mine*, *#the worm was caught by the early bird*. Some of them are semantically compositional, e.g. *to drink and drive*, *fortune favors the bold* but note that -- according to tests 2), 3 and/or 5) in section 6 -- they still qualify as MWEs. Some others, similarly to idioms, have a literal meaning inducing a metaphorical interpretation, e.g. *Early bird catches the worm. Rome was not built in a day.* See section 7.4 for more details.

4. Syntactic variants of verbal MWEs

Verbal MWEs may occur in prototypical constructions (*make a decision, break one's heart, took off*), but also in meaning-preserving variants of other syntactic categories. We consider the following types of variants as to be annotated for all languages in this task:

- infinitives, e.g. *to make a decision, to break one's heart*
- nouns (which are complements in the prototypical VMWEs) with relative clauses, e.g., *heart which he broke, remark which he took to heart*
- gerunds, e.g. *decision making, heart breaking*
- participles, e.g. *heart-breaking, breaking her heart, decisions previously made, all hearts broken by him*

Particular language teams may decide to extend this list to other types of variants, for instance nominalisations 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'. It may prove useful in this case to introduce a new category for such variants (e.g. NVPC nominal verb-particle constructions) so as to keep the universal categories intact.

Like other VMWE occurrences, syntactic variants are only annotated if they contain more than one word (e.g. a *knockout* is not annotated as a VMWE if it is understood as one word - see also section 1.1).

5. Three-stage annotation process

We propose the following 3-step methodology for verbal MWE annotation:

- **step 1** – identify a candidate verbal phrase (or an infinitival/nominal/participial variant of a verbal phrase) or a sentence; we assume that the annotators have a sufficient linguistic knowledge to be able to perform this step
- **step 2** – check if it is a multi-word expression, i.e. if it has at least two lexicalized components and if it shows orthographic, morphological, syntactic or semantic idiosyncrasy with respect to what is deemed general grammar rules of the language; linguistic criteria for idiosyncrasy are given in section 6
- **step 3** – categorize it into one of the above categories according to the criteria in section 7

6. Generic criteria for identifying verbal MWEs (step 2)

In order to decide if a candidate verbal, nominal, participial or sentential expression is a MWE, we apply the following generic (i.e. independent of the specific VMWE category) idiosyncrasy tests.[4] If a candidate expression passes **at least one** test from 1) to 6), we consider that it is a VMWE. Note, however, that test 6) is category-specific and requires the application of further tests from section 7 (whose number and compulsory character depend on the precise category).

1) Cranberry word. Does the expression contain a component that does not have a status of a stand-alone word? If yes, then it is a MWE.

- *to go astray*

2) Lexical inflexibility. Does a replacement of one of the components by words taken from a relatively large semantic class lead to ungrammaticality or to a substantial change in meaning (i.e. a change which goes beyond the one expected by the substitution)? If yes, then it is a MWE.

- *# to allow the feline out of the container (to let the cat out of the bag)*
- **to produce/build/create a decision (to make a decision)*

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

3) Morphological inflexibility. Does a morphological change (that would normally be allowed by general grammar rules) lead to ungrammaticality or to a substantial change in meaning? If yes, it is a MWE.

- *# to kick the buckets (kick the bucket)*
- *# to prettier-print. (to pretty-print)*
- *to take turns, #to take a turn*

4) *Morpho-syntactic inflexibility*. Does a loss of agreement (that would normally be allowed by general grammar rules) between some components lead to ungrammaticality or a substantial change in meaning? If yes, then it is a MWE.

- *# I give you his word for that (I give you my word for that)*

5) Syntactic inflexibility. Do syntactic changes (that would normally be allowed by general grammar rules) lead to ungrammaticality or to a substantial change in meaning? If yes, it is a MWE.

- *#He was speaking of the devil. (Speak of the devil!)*
- *# Bananas are gone. (go bananas)*
- *# drive and drink (drink and drive)*

6) Semantic non-compositionality. Is the meaning of the whole unit impossible to deduce from the meanings of its parts and from its syntactic structure, i.e. does it have a non-compositional meaning? If yes, it is a MWE.

- *kick the bucket = die*
- *spill the beans = reveal a secret*
- *make it = succeed*
- *to do in = kill*

This test is largely based on intuition and can sometimes be hard to apply in practice. It might be simulated by syntactic tests which are category-dependent (see below).

7. Specific criteria for categorizing verbal MWEs

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

7.1 Light-verb constructions

The following tests apply provided that the pre-identified expression consists of a verb and its argument containing a noun. In order for a candidate to be annotated as a LVC, **the answer to the five questions below should be *yes*** (or *no* in the case of tests where it is explicitly marked).

7) Is the noun predicative, i.e. does it have at least one argument different from the possessor?

- in *pay a visit*, *visit* has two arguments (the visitor and the visitee) (LVC)

8) Is the noun used in its original sense?

- in *pay a visit*, *visit* is literally understood (LVC)
- in *have kittens* (to be worried or angry), *kittens* is not used literally (non-LVC)

9)When omitting the verb, e.g. in a possessive construction (genitive form of the noun or a noun introduced with a preposition like *of* (EN), *de* or *par* (FR)), can the resulting NP refer to the original event?

- *Paul's walk* and *Paul had a walk* refer to the same event - *Paul's walk* entails that *Paul had a walk* (LVC)
- *Paul's cake* and *Paul made a cake* do not necessarily refer to the same event - *Paul's cake* does not necessarily entail that *Paul made a cake* (non-LVC)

10) When used within the construction, can the noun have all the arguments that it could have if it were used without the verb? (NO)

- *Paul made a cake. - Paul made Joe's cake*. (non-LVC)
- *Paul had a walk. - *Paul had John's walk.* (LVC)
- *Paul made a decision - Paul made a decision on the budget. - *Paul made the committee's decision on the budget.* (LVC)
- *Paul leads the discussion. - Paul leads the discussion of the committee.* (non-LVC)

11) Is the verb semantically bleached (i.e. not used in its original sense(s))?

- *pay a visit != to spend some money on a visit* (LVC)
- *deliver a speech != to move a speech from one place to another* (LVC)

This criterion may not be applicable to verbs with a very general meaning such as *have*, *go* etc.

12) If the verb's complement is its direct object, can the construction be passivised (applicable only if the verb can be passivized)?

- *He made a decision. - A decision was made.* (LVC)
- *He kicked the bucket. - #A bucket was kicked*. (non-LVC)

Once the above compulsory properties have been verified for a MWE, the following <u>optional properties</u> may further confirm its LVC status.

13) Can a verb (derived from the same root as the nominal component) replace the construction?

- *to make a decision = to decide*
- *to have a walk = to walk*

14) Can the construction itself be nominalized?

- *decision making* (LVC)

- *#bucket kicking* (non-LVC)

15)Does the noun refer to an action/event?
- *have a walk* - *walk* is an event (LVC)
- *have a cat* - *cat* is not an event (non-LVC)

16) Can the construction be modified alternatively by an adjective or an adverb (without changing the meaning)?
- *He made a quick decision. = Quickly, he made a decision.* (LVC)
- *He had a nice cat. != Nicely, he had a cat.* (non-LVC)

17) Can the verb be ellipted? (NO)
- **Joe had a shower and Peter a walk.* (LVC)
- *Joe had a cat and Peter a dog.* (non-LVC)

Specific guidelines for common verbs with general meaning:

18) Is the noun predicative, i.e. does it have at least one argument different from the possessor?
- in *have a glance (at sg)*, *glance* has two arguments (the subject and the object) (LVC)

19) Is the noun used in its original sense?
- in *have a walk*, *walk* is literally understood (LVC)
- in *have kittens* (to be worried or angry), *kittens* is not used literally (non-LVC)

20) When omitting the verb (e.g. in a possessive construction), can the resulting NP refer to the original event?
- *Paul's walk* and *Paul had a walk* refer to the same event - *Paul's walk* entails that *Paul had a walk* (LVC)
- *Paul's cake* and *Paul made a cake* do not necessarily refer to the same event - *Paul's cake* does not necessírily entail that *Paul made a cake* (non-LVC)

21) Can a verb (derived from the same root as the nominal component) replace the construction?
- *to have a walk = to walk*

22) Does the noun refer to an action/event?
- *have a walk* - *walk* is an event (LVC)
- *have a cat* - *cat* is not an event (non-LVC)

Candidate	*make a cake*	*pay a visit*	*make a decision*
Predicative noun	no	yes	yes
Noun in the original sense	yes	yes	yes

Omission of verb (NP)	John's cake != John made a cake.	John's visit = John visited somebody.	John's decision = John made a decision.
All the arguments of the noun can occur	John made Paul's cake.	John paid a visit to Paul. but #John paid Mary's visit to Paul. (it can only mean that John paid Mary's costs).	John made a decision on the budget. But *John made the committee's decision on the budget.
Semantically bleached verb	no	yes	yes
Passivization	A cake was made.	A visit was paid.	A decision was made.
Synonymous verb	-	visit	decide
Nominalized construction	cake-making	?visit-paying	decision-making
Noun referring to an event	no	yes	yes
Modification	He made a nice cake. != Nicely, he made a cake.	He paid a short visit to Paul. = For a short time, he paid a visit to Paul.	He made an important decision. = Importantly, he made a decision.
Ellipted verb	cake != making a cake	visit = pay a visit	make a decision = decision
Status	not LVC, not ID	LVC	LVC

7.2 Idioms

The tests for categorizing a candidate MWE as an idiom consist often in distinguishing it from an LVCs of the same syntactic structure. The essential test is:

23) Is the noun used in its original sense?

- in *have a walk*, *walk* is literally understood (non-ID)

- in *have kittens* (to be worried or angry), *kittens* is not used literally (ID)

An additional frequent, but not compulsory, property of idioms, is described by the passivization test:

24) If the verb alone allows passivization, can the construction be passivized?

- *He made a decision. - A decision was made.* (non-ID)
- *He kicked the bucket. - #A bucket was kicked.* (ID)

Some of the LVC and ID-oriented tests are illustrated in the table below:

Candidate	*make a cake*	*make a meal*	*make a decision*
Omission of verb	John's cake != John made a cake.	John's meal != John made a meal of it.	John's decision = John made a decision.
Passivization	A cake was made.	#A meal was made. (acceptable only in the literal sense)	A decision was made.
Synonymous verb	-	-	decide
Status	not LVC, not ID	ID	LVC

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 *to be born under a lucky star* the copula can be omitted as in *this man born under a lucky star*. In this task we consider constructions with a copula VMWEs only if the complement does not retain the idiomatic meaning when used without the verb. For instance, *to be double Dutch to someone, to be no chicken, to be somebody, etc.* are idioms, while *to be born under a lucky star* is not.

Note, finally, 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.

7.3 Other verbal MWEs

There are other types of verbal MWEs that do not fit into the above categories such as *to make it, to make do, to voice act, to wait and see, to drink and drive, to tumble-dry, to pretty-print, the pleasure is mine, I beg you pardon!, better late than never.*

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

- it is of one of the 4 syntactic/functional types from section 1.2

- it is a VMWE, i.e. it fulfills one of the 6 idiosyncrasy tests from section 6
- it cannot be classified into any universal (LVC or ID), quasi-universal (IPronV, IPrepV) or language-specific category

8. Open questions

These annotation guidelines are meant to evolve during pilot annotation. The currently open questions include the following:

1. Some verbal MWEs are difficult to classify according to the above criteria, notably those where the verb keeps its original sense but the complement doesn't e.g. *to take sth easy, to take sth with a pinch of salt*, or where a metaphorical use is explicitly signaled, e.g. *to sell like hotcakes*. Maybe, by analogy to LVC (where the noun keeps its original sense but the verb doesn't), these MWEs should yield a separate class?

2. Concerning the lexical flexibility in chapter 5, could we define more precisely the "substantial change in meaning" on the basis of analogy? For instance while *feline* is more generic than *cat*, it is not true that *to let the feline out of the bag* is more generic than *to let the cat out of the bag*. Could this extend to other semantic relations like synonymy, antonymy etc.?

3. How should we annotate VMWEs with anaphora? E.g. *This decision was hard. But he took it.* Should it be annotated and if so, should the anaphora be resolved (by selecting also the noun "decision").

Glossary

Notion	Definition	Examples[5]	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) lexeme; if a VMWE contains a head verb, it is always lexicalized	*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 that these are two different, although possibly synonymous, MWEs.
lexicalized preposition	a preposition which governs a lexicalized complement or which is part of an inherently prepositional	*He **took** me **by surprise*** *You can always **rely***	Lexicalized prepositions are to be distinguished from the selected

[5] Annotated components of VMWEs are highlighted in bold

	verb (i.e. the verb alone is a cranberry word or the preposition markedly changes the meaning of the verb)	***on** him.* *I **came across** an old photograph.*	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 VMWEs are no metaphors and, importantly for this task, some metaphors are no MWEs; the distinction between the two categories is an open problem
multi-token word (MTW)	a word split by the tokenizer into several tokens (due to a tokenizer's imprecision)	*Pandora|'|s* (PL) *SMS|-|ować* (PL) *Skype|'|ować* (FR) *ajourd|'|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.
multi-word 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		
multi-word 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**.*	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. It is usually lexically fixed (cannot be replaced by another preposition) but it does not markedly change the meaning of the verb.	*I don't want to participate in this.* *I don't want to **take part** in it.*	Selected prepositions are to be distinguished from the lexicalized ones. Only the latter are 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 multi-word expression (VMWEs)	a MWE which functions as a verbal syntagm; usually it's head word is a verb (but there may be exceptions)		
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*	