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

version 4.0

19 January 2016

Veronika Vincze, Agata Savary, Marie Candito, Carlos Ramisch

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[1] 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 characteristics of several classes 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

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

- They constitute syntagms (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).

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

- Prototypical verbal MWEs (VMWEs) - MWEs which function as verb phrases i.e. their syntactic heads are verbs in finite forms (e.g. *made a decision, break her heart, took this to heart*).

1 Www.parseme.eu

- Syntactic nominal, participial and other variants of prototypical VMWEs maintaining their idiomatic reading, e.g. *decisions which we made, decision making, heart-breaking.*
- Partly lexicalized sentential MWEs with lexicalized subjects, e.g. *a little bird told someone, the problem lies in sth*.
- Fully lexicalized sentential MWEs (or sentential VMWEs for short) e.g. the *early bird catches the worm.*

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 *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 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. 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').

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.

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 token. For instance lexicalized prepositions are pointed at but case suffixes are not. Thus, in *to come across something*, the verb and the preposition are integral parts of the VMWE, 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.

We give a special status to selected (non-lexicalized) prepositions but only if they are selected by VMWEs, i.e. such verbal expressions which contain more than one lexicalized components disregarding the selected preposition itself. For instance, in *to take part in sth* we do mark the selected preposition *in* (since both *take* and *part* are lexicalized) but in *to come in* we do not mark this preposition since this verb-particle combination is semantically compositional and *come* alone is not a VMWE.

Both continuous and discontinuous lexicalized components of VMWEs are also 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 5 categories of verbal MWEs:

- verb particle constructions (**VPC**), e.g. *to set sth up*
- light verb constructions *(**LVC**), e.g. to give a lecture*
- idioms (**ID**), e.g. *to go bananas*
- sentential MWEs (**SENT**), e.g. *fortune favors the bold, better late than never*
- other verbal MWEs (**OTH**), e.g. *drink and drive*

The following sub-section contains a general description of these categories. 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 Verb-particle constructions

Verb-particle constructions (also called phrasal verbs or phrasal-prepositional verbs) have the following general characteristics:

- They are formed by a head verb and a particle (cf. open question 4 in section 7).
- Both the verb and the particle are lexicalized.
- The meaning of the VPC is non-compositional. Notably, the change in the meaning of the verb goes significantly beyond adding the meaning of the particle (e.g. *do in = to kill*).
- The verb and the particle can sometimes be separated with a noun or pronoun without any change in meaning (e.g. spit *(it) out*).
- There is often an English synonym or a translational equivalent in another language which is a one-word unit or is a verb with a verbal prefix (e.g. *get away* and *escape*).
- A noun can often be derived from it (e.g. *breakthrough*).

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 the verb (*lie down, come in, rely on sth*). We do, however, mark (with a special label) prepositions selected by VMWEs (cf. section 2).

In Germanic languages and also in Hungarian, verb-particle constructions are usually spelled as one word but due to syntactic changes in the sentence, they can be separated. Here we focus only on cases when the two are separated, hence examples like *Die Kinder sollen in der Schule aufpassen*. 'The children must pay attention at school.' should not be annotated but examples like *Herr Müller, passen Sie auf! '* Mr. Müller, be careful' should be.

3.2 Light verb constructions

Light verb constructions 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*). 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).

3.3 Idioms

An **idiom** is a VMWE composed of a head verb (possibly phrasal) and at least one of its complements. 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. This usually makes them semantically totally non-compositional, i.e. none of their lexicalized components retains any of their original meanings.[3]

3.4 Sentential VMWEs

Sentential VMWEs are fully lexicalized sentences and include:

- 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 sentential VMWEs, additionally to being fully lexicalized, 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. *fortune favors the bold* 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.*

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.*

3.5 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.*

4. Syntactic variants of verbal MWEs

Verbal MWEs may occur in prototypical constructions (*to make a decision, to break one's heart, to take off)*, but also in meaning-preserving variants of other syntactic categories, including:

- infinitives, e.g. *to make a decision, to break one's heart*
- nouns 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*, *braking her heart, decisions previously made, all hearts broken by him*

Like other VMWE occurrences, they are only annotated if they are spelled as separate tokens/words.

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
- **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
- **step 3** – categorize it into one of the 5 categories (LVC, VPC, ID, SENT, or OTH)

We assume that the annotators have a sufficient linguistic knowledge to be able to perform step 1. The following sections provide proposals of linguistic criteria for steps 2 and 3.

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]

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)*

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)*

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 sentence.

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 Verb-particle constructions

The following tests allow to properly identify prepositional particles in the case when they might be ambiguous with prepositions including prepositional phrases (PPs):

7) Are the verb and the particle separated? If yes, then it is a VPC.

   - *He **turned** the lights **on**.*
   - ***Call** me **up**!*

8) Does question formation leave the particle intact? If yes, it is a VPC.

   - *I went to the shop. - Where did you go? (non-VPC)*
   - *He **turned** the TV **off**. - What did he **turn off**? (VPC)*

9) Can a circumstantial PP be inserted between the verb and the (supposedly) preposition or can the preposition and complement move around the sentence? If NO, it is a VPC.

   - *I could rely on him at once. - I could rely at once on him. - On him I always rely.(non-VPC)*
   - *I **took off** my clothes at once. - *I **took** at once **off** my clothes. - ***Off** my clothes I always **take**. (VPC)*

Once we have checked that the pre-identified expression contains a verb and a particle, the following tests confirm its VPC status. These additional tests are neither sufficient not necessary criteria for VPC identification.

10) *Can a noun be derived from the construction? If yes, it is a VPC.*

- *to turn over -> turnover*
- *to work out -> workout*

11) *Is there a verbal synonym that can express the same meaning? If yes, it is a VPC.*

- *to run away = to escape*
- *to try out = to test*

7.2 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).

11) 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)

12) When omitting the verb (e.g. in a possessive construction), can the original action be semantically reconstructed?

- *Paul's walk* entails that *Paul had a walk. (LVC)*
- *Paul's cake* does not necessarily entail that *Paul made a cake. (non-LVC)*

13) 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)*

14) 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)*

15) If the verb's complement is its direct object, can the construction be passivised?

- *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 optional properties may further confirm its LVC status.

16) *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*

17) *Can the construction itself be nominalized?*

- *decision making (LVC)*
- *#bucket kicking (non-LVC)*

18) *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)*

19) 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)*

20) 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)*

7.3 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:

21) *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:

22) *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, non-ID	ID	LVC

7.4 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 main difference between idioms and other verbal MWEs is that idioms are more complex, i.e. they usually contain verbs with (syntactic) arguments which are at least partly lexicalized. Other verbal MWEs mostly contain verbs without any lexicalized arguments.

This is also the category for language-specific verbal MWEs – in case there are any.

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 complements 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. Is the VPC class general enough for this multi-lingual task? Shouldn't it rather be language-specific (e.g. English-specific) or language-group-specific (e.g. Germanic-specific). Slavic and Romance languages seems to have very few VPCs (if any).

3. Conversely, should we add other VMWE categories that are generic enough for many languages? One proposal would be “compound verbs” which are 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 *va savoir* 'go to know'='I have no idea'(FR), *go figure*, *make do*, *coś kogoś ani ziębi, ani grzeje* 'sth neither cools nor warms someone' = 'someone is indifferent to sth'

4. How to define a particle in section 3.1? Some authors define them as prepositions or adverbs. But it seems that, even if syncratic with prepositions, they have no prepositional behavior (they do not govern prepositional groups).

5. 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.?

6. Should we annotate MWEs whose heads are (finite) verbs but which function as adverbs or nominals, e.g. *peut-être* 'may-be'='maybe' (FR), *Bóg wie co* 'God knows what'='something unrealistic' (PL).