Multi Word Expression Annotation Guidelines:

Your task is to **identify** each Multi Word Expression (MWE) in a sentence, and then to **categorize** each of the MWEs you identified into one of three categories: Fixed, Semi-Fixed, Flexible. Please read through this entire document, to give you a sense of the task before starting.

Step 1: Identifying MWEs

A MWE is a lexical item that can be decomposed into multiple words. In English, these expressions are made up of **tokens delimited by whitespace.**

Here are a few examples:

> *Mr. Rogers, hot tub, take over, pick up, pay attention (to), take pictures, put up with, at all, more or less, pick up where <we> left off, leave of absence, thank you, light as a feather, c'est la vie, good day*

In addition to being comprised of multiple words, these expressions must be idiomatic in some way. **An expression can demonstrate one or more of Lexical, Syntactic, Semantic, or Pragmatic idiomaticity.** Idiomatic means that some aspect of the meaning of the expression cannot be gathered simply by information about the individual tokens that make up the expression.

A brief explanation and examples for each category are given below. Further explanation and additional examples are given in the section *Questions to ask yourself when deciding if something is a MWE* below. Although you will not need to categorize the particular type of idiomaticity found in each expression, being familiar with the types will help you identify MWEs in the text.

Lexical idiomaticity occurs when a word in the expression is not in the English lexicon, and therefore the meaning of the expression cannot be derived from the meaning of the word.

Examples in this category: *ces't la vie*

Syntactic idiomaticity occurs when the syntactic categories of the words in the expression do not line up with the syntactic category of the expression itself.

Examples in this category: *at all, more or less*

Semantic idiomaticity occurs when the meaning of the expression cannot be derived from the meaning of the individual tokens.

Examples in this category: *take over, pick up where <we> left off, ...*

Pragmatic idiomaticity occurs when the expression is associated with some specific context, and might take on a different meaning if situated outside of this default context.

Examples in this category: *good day*

Questions to ask yourself when deciding if something is a MWE:

When trying to decide if an expression is a MWE, you should walk through each of these questions. They will help you consider each type of idiomaticity that the expression might be displaying. As soon your answer corresponds to the bullet under any of the questions, you know your expression is a MWE.

Are any of the words not part of the English lexicon? This one is really easy! It's as simple as checking whether any words in the expression are non-English words.

- If yes, it's lexically idiomatic, and should be marked as a MWE

Do the syntactic categories of each piece line up with the syntactic function of the expression?

For example: *by and large*, we have a Preposition *by*, a coordinating word *and*, and an Adjective *large*. These categories do not line up with the category of the phrase which is adverbial. This is opposed to a situation where we have *by the cat*, a P, D, N. These syntactic categories do line up with the syntactic category of the phrase which is Prepositional.

- If there is a mismatch, this is a MWE!

Can the meaning of the expression be derived from the meaning of its parts? This one is going to be EXTREMELY common, and also sometimes tricky to notice.

For example: *middle of the road*, this typically means non-extremism, which has nothing to do with being on a road or in the middle of it.

*** STOP AND THINK ABOUT THIS QUESTION***

It's very easy to read a sentence quickly and think 'I know what that means'. But this is an extremely important step where you need to consider whether the literal meaning of each token corresponds to the meaning of the expression. Read each sentence VERY slowly, and try to think about the meaning of each individual word.

- If you cannot compose the meaning from the parts, this is a MWE

Another very important Example:

He picked up his son from school.

Here the expression under consideration is *picked up*. A reasonable justification could be that *picking up (from school)*, involves meeting and taking someone somewhere else, not physically hefting them, so this is idiomatic. But ALSO

He picked up the book

The expression under consideration is again *picked up*. This time the meaning is to physically lift, but we have to break the expression down even further. *Pick* could mean to choose or to scratch at, among other things. The expression *pick up*, even when used to refer to physically lifting something, can't be derived from *pick* and *up*. You are not choosing in an upward direction when you pick something up. You are also not scratching at something in an upward direction. So this expression is also idiomatic and a MWE.

Is the expression associated with a particular context?

For example: *all aboard* is a command used to instruct people to board some kind of a vessel. *Good morning* is a greeting typically used in the morning, and has a different connotation and default meaning than it would in a scenario like *I had a very good morning*.

- If it's associated with a particular context, this is a MWE.

Step 2: Categorizing the MWE

After you identify each MWE, you will label it as belonging to one of three categories: Fixed, Semi-Fixed, and Flexible.

The categories are described and exemplified below.

Fixed: These are expressions that do not undergo morphosyntactic variation or internal modification. They have completely fixed word order, and cannot be inflected or split up.
Examples: *ad hoc, by and large, *by and larger, *by and very large, large and by*

Take *by and large* as the example. We cannot inflect the adjective, for example, and result in a meaningful expression. We cannot add additional words to the expression. We also cannot change the word order.

Other listed examples in this class: *Mr. Rogers, at all, more or less, thank you, light as a feather, ces't la vie, good day.*

Semi-Fixed: These expressions can undergo some morphosyntactic variation like inflection. They might take differently inflected pronouns or determiners. They have hard restrictions on word order and the arguments that they take

Example: *kick the bucket, kicks the bucket, *the bucket was kicked*

Attorney general, attorney generals, part of speech, parts of speech

To stand in my/his/her/your/our shoes

Notice in the first example, we can inflect *kick*, but we cannot completely change the word order as with a passive construction. Similarly we can inflect nouns for number and pronouns for person.

Other listed examples in this class: *hot tub, take over, put up with*

Flexible: These expressions can undergo syntactic variation. Word order can change, they can take different arguments, adverbials can be added.

Example: *hand in the paper, hand the paper in.*

The plane took off, the plane took right off

A demo was given, how many demos did he give, give a clear demo

Other listed examples in this class: *pick up, pay attention to, take pictures*

Questions to ask yourself, in order, when deciding MWE type:

When you are categorizing a MWE, ask yourself these questions, in order. They will help you identify the type.

Can I inflect this? Can I make it plural? Change its tense? Add -er?

- If no, then it's a fixed expression!

Can I change the word order? Can I insert new words, like adverbs, into the phrase?

- If no, then it's a semi-fixed expression!

If you can change it enough to get to this point in the questions, then it's probably a flexible expression!

Specific Instructions for MAE and the task:

- The task is split up by sentence, but each file will contain 20 sentences.
    - At the beginning of each sentence you will see a text number and an id number in the form: 'text # reviews-######-####:'.
        - Please ignore this id number, and never mark it as part of a MWE.
    - Each sentence should be treated independently.
        - You should NEVER identify a MWE that spans more than one sentence.
        - Approach each sentence as if there were no surrounding context.
- To select a MWE, click and drag over the tokens that comprise the expression. You can then right click and select the type.
    - When selecting MWEs, include the full span of each token as it occurs in the expression. This includes inflection.
        - Example: *I'm going to pick up the kids*
        - Example: *Yesterday I picked up the book from the library*
            - Do NOT attempt to stem any of the tokens.
                - Ex: * *I picked up the book…*
    - Tokens may contain punctuation:
        - Example: *C'est la vie*
    - If a token ends with a punctuation mark because it is at the end of a sentence, the punctuation mark should not be included.
        - Example: *It's time to start winding down.*
    - If a token ends with a punctuation mark because it is part of the token, or the token is an abbreviation, the punctuation mark should be included
        - Example: *Mr. Rogers was …*
    - EXCEPTION: Possessive marking in English. If a possessive marker (*'s*) occurs at the end of a MWE it should NOT be included in the MWE.
        - Example: *Harry Houdini's final performance…*
- Certain Flexible MWEs may be discontiguous (more on this below). This is acceptable, gaps are permitted.
    - Each MWE can only have one gap.
        - You should never be marking a MWE of the form *part1 …. part 2 … part3*
    - To select a gappy MWE, click on Mode>Switch to discontiguous span selection mode. You can then select discontiguous tokens and select a type for them. If you are ever identifying a MWE with a gap, its type should be Flexible.
- There may be spelling and grammatical errors, as well as abbreviations and slang, as these texts are from informal online reviews.
    - Spelling errors should be ignored
    - Treat slang and abbreviations as you would any other token.

    - For grammatical errors, mark MWEs according to your best guess at what the author was trying to convey.
    - Tokens that would typically be one word, but are represented as two in the text (Ex: *every one in attendance had some cake)*, should NOT be marked as MWEs.
- Do not expect to necessarily find a MWE in each text. Some texts will have none, others will have many.

Extremely Important things to consider for each sentence:

- MWEs must be made of multiple words. If you find yourself marking a MWE that is only made of one token, it is NOT a MWE.
- If you identify a discontiguous MWE (a MWE with a gap), then it should be categorized as Flexible.
- Classifying nominals:
    - Deciding whether or not certain nominals should count as MWEs will most likely be the most difficult part of this task. Please see the section "Noun noun compounds and other nominals" below for guidance on this.
- Verbs:
    - Check every single verb in the sentence for prepositions and particles.
        - What we do not want to include: productive verbal constructions in English
            - Passive constructions: *The cereal was eaten*
            - Perfective constructions: *I have eaten*
            - Progressive constructions: *I am eating*
        - What we do want to include:
            - Light verb constructions: *take a walk*
            - Verbs with particles: *look for*
            - Verbs with prepositions: *pick up*
    - Check each preposition that you find to see whether it is a part of the verb, or a part of a prepositional phrase adjunct.
        - Example: *I was working on my homework.*
            - Here *on*, is tied to the verb *work* forming the MWE *work on*
        - Example: *I was working on Tuesday*.
            - Here *on* is just part of a PP attached to the VP
    - If you identify a gappy MWE, it should be classified as Flexible, but gappy MWEs are not the only MWEs that should be classified as Flexible.
        - Example: *I dried out the sheets on the balcony.*
            - Here there is no gap in the MWE, but you should consider whether it is possible to insert something into the middle of the expression
            - *I dried the sheets out on the balcony*
                - Because of this, the expression should be classified as Flexible.
        - It is also good to try to insert an adverbial (*right* usually works well):
            - *I jumped in on the project.*
            - *I jumped right in on the project.*

Most common general categories:

- **Fixed:** Named Entities, prepositional phrase, conjunction phrase, proverbs and idioms
- **Semi Fixed:** Noun noun compounds, some verb phrases
- **Flexible:** Mostly verbs with particles

Additional Examples and Guidelines for Specific Cases:

Please read these before you begin, and refer back to them if you have questions or encounter cases you are unsure about while completing the annotation task. For each category, the typical type of MWE that it should be classified as is listed. This is to give you a place to start, but please consider each case carefully, as there may be exceptions that we have not listed here.

- If you are marking something as fixed, every instance of that MWE should have exactly the same form.
    - Example: If you encounter *John Smith*, and mark it correctly as a Fixed MWE, you should at no point be marking something like *John Smith's* as a different Fixed MWE. (more on this in the Named Entities section below)
    - In general, if you notice a similar expression with different inflection, this is a STRONG sign that the expression you are looking at is in fact NOT fixed, and should be categorized differently.

Named Entities - Fixed

- Named entities with more than one token in the name are MWEs and should be classified as Fixed.
- Notes: 's should not count as inflection. For example, we could say *The White House's decision…*, but we would still want to mark *The White House* as a Fixed expression, because we cannot say *the whiter house,* or *the white houses,* and have it keep the same meaning.
    - When we mark a MWE that is followed by 's, the 's should not be included in the MWE. For example: *John Doe's favorite restaurant…*
        - In a sentence like *How many John Smiths are there in attendance?*, we still want to count *John Smith* as a fixed MWE.

Proverbs and Idioms - Fixed

- Examples: *beggers can't be choosers, light as a feather*
- Examples: *forget it, no way, thank you, who knows, never mind, well done*

Prepositional Phrase - Fixed

- Examples: *just like, at all, under the weather, in charge of, at liberty to, at best*
- *Representing time: from time to time, on time, from now on, by far*
- *Representing location/transportation: on earth, on foot, in front of*
- *Representing payment: in cash*
- *Representing emotion: to my satisfaction(semi-fixed)*

Conjunction Phrase - Fixed

- Examples: *as well as, in addition to, in spite of, because of, so that,*

Noun Noun compounds and other nominals - Semi-Fixed

- Noun noun compounds are noun phrases that are made of two or more nouns.
- Examples: *mental health, golf club, hot dog*
- These are typically Semi-Fixed because most can be inflected for pluralization.
- Sometimes it will be difficult to decide whether or not a nominal expression is a MWE. The key is idiomaticity. Can the meaning be clearly derived from the component parts?
- A helpful technique can be trying to replace the modifier with a different token to see if it produces a similar and meaningful expression.
    - Example: *insurance companies*
        - This is a noun noun compound, but we can replace *insurance* with words like *electric, trading, paper,* for example, and the expression keeps the same kind of meaning, only different with respect to the modifier.
        - Because nouns can be attached to *companies* productively, we do not want to count this as a MWE.
    - Example: *family tree*
        - This is also a noun noun compound, but when we replace *family* with something like *oak,* or *apple*, the meaning completely changes.
        - We can again think of this in terms of idiomaticity, an *apple tree* is a tree that grows apples but a *family tree* is not a tree that grows families.
        - This should count as a MWE.
    - Example: *mental health*
        - *Mental health* has a more nuanced meaning than something like *women's health* does, for example.
        - This should be marked as a MWE.
- Deciding whether or not nominals should count as MWEs will probably be the most difficult distinction you have to make in this task. Think through idiomaticity, try replacing the modifying token, and make your best guess.

Adjective Noun compounds - Fixed or Semi-Fixed

- Examples: *last minute, red wine, every time*

Adjective Prepositions - Fixed

- Representing emotion: *satisfied with, happy with, serious about,*
- *Representing distance: close to, far from*
- *Representing comparison: less than, more than*

Verb Noun Phrases - Semi-Fixed or Flexible

- Verb noun phrases, which denote a state, event, or relation, indicates an abstract level of phenomenon and are usually idiomatic combinations of verb and noun
- Examples: *return the favor, pay attention to, kick the bucket*
- Verb noun phrases like *purchase purse, return book*, are not WMEs we are looking for, in that their objects can be substituted with different objects like *purchase stock/bonds*, *return the scooter*, and the meanings of the phrases are the same except for with different objects. To some extent, these verb noun phrases can be treated as word by word and it won't affect the meaning of the context.
- Similarly to nominal phrases, the key here is idiomaticity. Consider *return the book, return the key, return the paper*. All of these mean roughly *to bring back <object>*. Therefore these are likely not idiomatic, and not MWEs
    - These should be contrasted with *return the favor*, which has a much more specific meaning of *to do something nice for someone because they did something nice for you*. There is idiomaticity displayed in this expression that is not displayed by those previously listed. Therefore this expression is a MWE.

Modal/tense construction - Semi-Fixed

- Examples: *ought to, have to, want to, be supposed to*

Verbs with prepositions - Semi-Fixed

- Examples: *depend on, look for, work with*

Verbs with particles - Flexible

- Examples: *pick up, dry out, look up*

There is a very important distinction between verbs with prepositions and verbs with particles. Both are verbs followed by prepositions, but the ways/places in which they take objects are crucially different.

Verbs with prepositions:

*depend on the employee, depend on my brother, *depend the employee on, *depend my brother on, *depend really on, *depend right on*

A verb with a preposition can take different objects, as seen above with *my brother* and *the employee*, but the actual expression *depend on*, cannot be broken up. We cannot insert the object between the verb and preposition, or stick an adverbial in the middle of the expression. Therefore they are Semi-Fixed

Verbs with particles:

Pick up my son, pick up the cat, pick the book up, picked the snake right up

We can insert the object between the verb and the particle. We can also insert additional adverbials into the phrase. Therefore they are Flexible.

Light Verbs - Flexible

- Light verb constructions are MWEs, with verbs like do, give, have, make, keep and take… It might take forms of verb-noun, verb-noun-preposition, or verb-adjective. It can take different arguments, adjectives can be added.
    - Example: *Yesterday I took a walk.* *I made a severe mistake.*
    - Example: *do damage to, have a problem with, take care of, give speeches*
    - Example: *get ready, get done*

Quantity phrase - Flexible

- adjectives can be added
- Examples: *a number of, a large number of, an amount of, half an hour*

Notes:

Determiners should not be included in MWEs unless they are part of a named entity:

- Example: *The New York Times published…,* we want to include 'The' in the MWE
  A Dunkin Donuts offered…, we do not want to include "A".

With discontiguous MWEs do not include arguments if they can change freely with the expression BUT do include arguments if they can only be inflected within an expression:

- Example: *hand the paper in*
    - This expression can be thought of as *hand <NP> in.*
    - The correct annotation for this would be *hand the paper in,* because we do not want to include
- Example: *shot myself in the foot*
    - This expression can be thought of abstractly as *shot <oneself> in the foot.*
    - The correct annotation for this would be *shot myself in the foot,* because the <oneself> can only be inflected for person and number, it cannot be interchanged with any NP.
    - You will probably encounter this kind of case only very rarely, if at all.
    - If you find yourself marking something like this, make absolutely sure that you cannot replace the reflexive pronoun with some other NP.

Passive constructions do not count as MWEs:

- We do not count these as MWEs because they are the same for all verbs.
- Example: *the dog had been fixed*

    - We do not mark *been fixed*, as a MWE, even though technically both of those words make up the expression of the verb in English. This is a productive pattern that can be seen with all or most verbs.
- Similarly, any type of verb construction that takes more than one word should NOT be considered a MWE.
    - Examples: *have eaten, would eat, had eaten, will eat, do eat, etc.*
- Please note that this does not exclude things like *have a good time*
    - In this case *have* is not being used to form a perfective construction, it's being used in the expression *have a(n) <adj> time.*
    - This should be marked as a MWE.

Multi Word Expressions should not overlap:

- You should never mark MWEs with overlapping spans.
- Example: *surprise birthday party*
    - The argument could be made that there are two different MWEs in these three words, *birthday party* and *surprise birthday party*. This is true, but for the sake of simplicity we cannot have overlapping MWEs, so we should categorize this all together as one large MWE.

Take expressions in context:

- Example: *pick up where we left off*
    - The observant reader will notice that this is the same example as above, and we have said that this should be annotated as *pick up where we left off*.
    - It is important to take the context provided by the full sentence into account when identifying MWEs, because *pick up*, is also an expression that we would normally annotate as a MWE when it refers to retrieving an item from the floor, or going out to acquire an item. In this specific context though, it is neither of those things and actually belongs to a larger expression.

Mixed Metaphors

- There may be mistakes in use of MWEs, these should still be annotated as MWEs
- Examples: *hit the ground moving*
    - This should be *hit the ground running,* but it should still be annotated as a MWE.

Ditransitive verb phrases are not MWE: recommend to/tell to

- A verb like *I recommended the movie to my friend*, should not be considered a MWE because you can also say *I recommended the movie* and it keeps the same meaning, just less specific.

- Similarly, *I told the secret to my friend*, has a similar but more specific meaning than *I told the secret*.
- Any ditransitive verb phrases are not MWEs. Here is a link of ditransitive verbs: https://www.cse.unsw.edu.au/~billw/ditransitive.html