Specification and Guidelines

1 Annotation Goal

Our goal is to identify and categorize all multiword expressions (MWEs) in our data. The annotated data will become training data for an ML model that will be able to identify different types of multiword expressions in text, which could be useful for downstream tasks such as machine translation, translation detection, synonym detection, and syntactic parsing.

2 Task Summary

The annotation task comprises finding multiword expressions and labeling their type, which we frame as spans. We define a multiword expression as a string of multiple words which act as a single semantic or syntactic element.

3 Corpus

The Corpus of Contemporary American English (COCA) contains 950 million words of data in American English from 1990 to 2019 in various genres, including spoken conversations, news and magazines, and literature (fiction). The data is tokenized and tagged for lemmas and parts of speech. Access to the full corpus requires paid registration, but we were able to download a free sample of about 1/100th of the corpus (8.9 million words).

4 Preprocessing

We took the plain text data of the sample corpus and ran the SpaCy sentence segmentizer on the data to obtain 10-sentence chunks of the text which we could present to an annotator. We then ran the SpaCy part of speech tagger on the data in order to gain some information on potential MWEs that are based on syntactic category, such as noun-noun compounds and verb-particle constructions. Another step taken in preprocessing was replacing sensitive information that was originally present in the data as a span of @-symbols with spaces in order for the tagger to function properly.

5 Annotation Process

You will be presented with a maximum of 10 sentences at a time. If you come across nested MWEs and the software allows, mark them both. If the words of a MWE are noncontinuous, mark the entire span, including the intermediary word that is not technically part of the expression (see example 13).

6 Types of MWEs

6.1. Idioms

These are phrases that could have a literal reading, but used in a context where their meaning is non-compositional – i.e. where the meaning of the entire expression cannot be derived from the composition of its parts. They are distinct from metaphors in that idioms are set expressions known to the speech community, whereas metaphors are more creative and often novel. If you suspect a phrase may be an idiom but haven't heard it before, look it up to see if it is a standard expression.

(1) We met at a bar mitzvah in the Valley—San Fernando, not Central—and *hit it off*.

(2) *When you get lemons, you make lemonade.*

(3) My response is: Deal! Maybe Mr. Schnatter, who's been both praised and pilloried for his position, was *on to something here*.

Expressions such as *a lot* and *a little* count as idioms for our purposes because when you replace one of the words with a synonym, the meaning changes. (*A small cake*, e.g., is not the same as *A little cake*.)

In some instances, only part of an idiom, or a variation on an idiom, will come up. In this case, you should highlight the maximum number of lemmas that are part of the known idiom. For example, you would want to mark the following phrase, which is based on the idiom ”a gift that keeps on giving,” like so:

(4) The story *that keeps on giving*

Note: In an example like the following, where the phrase *break a leg* appears, we would *not* want to tag it as a MWE (even though the same phrase can be idiomatic in other contexts), since here it is actually literal:

(5) It's like skiing lessons: If you take a class, you're probably less likely to break a leg. And even if you do learn first, you might still fall and break a leg.

6.2. Noun-noun compounds

These are expressions that are formed of a sequence of two or more nouns that come together to form a single noun (or what syntactically behaves as a single noun). Even in situations in which one of the nouns behaves adjectivally, we have decided to label all consecutive nouns that form a syntactic unit as a noun-noun compound, in order to have a simple but maximal definition of the category. Do not label proper nouns as being part of a noun-noun compound.

(6) ... back home watching the Dodgers by the *air conditioner* is where it's at.

(7) Natural lighting: is never natural, looks like a *soap opera* or a porno.

(8) Those movies are plain old straightforward *comic book character adaptations.*

(9) ...most 8 year olds would rightly mock me and throw rotten fruit – the statement is clearly wrong and no philosophical wrangling can make it less so – arguments in philosophy which appear to make it so are clearly spurious deviations into the realms of the *word salad* (the geography of which I am clearly well versed in!)...

Although most noun-noun sequences highlighted by our preprocessing will be true noun-noun compounds, not all sequences of nouns form a syntactic unit. The counterexample below is a good demonstration of two consecutive nouns which do not form a syntactic unit.

(10) forever changed the *way crimes* can be solved

6.3. Verb-particle constructions

These are expressions that contain a verb followed by a particle. They have some level of compositional meaning, and a constituent can appear between them. Besides examples with an extended phrase that encompasses an object like the third below, verb-particle constructions should function syntactically as a single verb, rather than a verb followed by a prepositional phrase.

(11) When she had *taken down* her clothes, her sewing, and her Bible from the shelf...

(12) Two weeks a year would be reserved for something special: working at a small hospital in rural Brazil *set up* by the grandfather of a good friend from medical school.

(13) What can we find out about this guy that 's going to *set him off.*

Contrast the above with the syntactic breakdown of the example below, in which *to the story* forms a syntactic constituent whereas *down her clothes* is not in the first example above.

(14) Jim *refers to* the story.

6.4. Light verbs

These expressions start with a verb that has little semantic content and is followed by another word or phrase that carries the entire expression's meaning.

(15) While there, she *gives birth* to her child.

(16) I just hate nubbly towels. Is that a crime? No, of course not. I'm, uh – I'm gonna *take a shower.* Oh, this is perfect.

(17) *Have fun.*

(18) I *threw a ball.*

Above, if *a ball* refers to a type of party then it is a MWE, but if *a ball* refers to a ball that is used in sports, then it would not be counted as a MWE as can be seen below.

(19) I *threw a football.*

Other examples of light verb constructions include *draw a breath, take a walk,* and *make fun.*

7 What is not a MWE

For this annotation task, there are some types of expressions that we do not count as MWEs.

- **Latin phrases and other borrowings.** If you come across a borrowed expression such as *quid pro quo*, do not tag it as a MWE. We consider borrowings to be a separate phenomenon.
- **Proper nouns.** We do not consider proper nouns or names to be MWEs because identifying them falls under the domain of named entity recognition.
- **Single words.** It may appear obvious at first that single words should not be marked as multiword expressions. However, some of our categories seem to encompass single words as well such as idioms. For example, we would mark *a lot* as an idiom multiword expression, but *lots* would not be marked.
- **Hyphenated text.** If a document refers to a concept (such as *multi-word* expressions) with a hyphen, it should not be marked as a MWE, while if it refers to the same concept elsewhere with a space, it should be marked as such (i.e. *multi word* expressions).

8 Diagnostics

In case you are unsure if a particular expression is a MWE, you can ask yourself the following questions.

- Can you replace one of the words with a synonym and have it retain the same meaning? If not, it is likely a MWE.
- Can the expression be translated literally into another language you know? If not, it is likely to be a MWE.

9 References

Davies, Mark. (2008-) The Corpus of Contemporary American English (COCA): 560 million words, 1990-present. Available online at `https://corpus.byu.edu/coca/`.