Specification and Guidelines

IDIOM: Intelligent, Determined Investigators Of MWEs
Brynna Kilcline, Gabby Masini, Ruth Rosenblum, Annika Sparrell

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 two parts: finding multiword expressions, which we frame as spans, and labeling their type, which we define with attributes. We define multiword expressions as multiple words which act as a single semantic element, that are not decomposable into their individual parts to form a phrase of the same meaning.

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 Types of MWEs

4.1. 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.

(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.*

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:

(4) 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.

4.2. 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).

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

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

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

(8) ...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!)...

4.3. 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.

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

(10) ...you were heralded as a champion and an honorable man rather than getting shit on and *thrown in* jail.

4.4. Light verbs

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

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

(12) 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.

4.5. Other

This attribute covers any multiword expressions that fit our definition but not any of the attributes above – including expressions whose meaning is not derivable from the meaning of each word, but where the literal (compositional) reading is not possible.

(13) Well, I think the people of the 4th District, *by and large*, are ready for this war to come to end...

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

5 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.
- **Names.** We do not consider names to be MWEs because identifying them falls under the domain of named entity recognition.

6 Diagnostics

In case you are unsure if a particular expression is a MWE, consider whether the meaning as a whole is made up of the meanings of its parts. In order to do so, 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.
- For noun noun compounds, is the entire compound a subset of the last noun? For example, you can answer the question ”What kind of dough?” with the response ”Cookie dough,” so *cookie dough* is a subset of *dough* and we do not consider it to be a MWE. However, you cannot answer the question ”What kind of cream?” with the response ”ice cream,” so *ice cream* is a MWE.

7 References

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