Multi-Word Expression

Parameswari Krishnamurthy

Materials taken from:

Rajeev Sangal. Processing of Multi Word Expressions (MWEs) 5 May 2012

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, Amalia Todirascu; Multiword Expression Processing: A Survey. *Computational Linguistics* 2017; 43 (4): 837–892. doi: https://doi.org/10.1162/COLI_a_00302

What are words?

- Traditional formal descriptions of individual human languages typically divide labor between a repository of words and their properties, called a **lexicon**, and a description of how such words combine to form larger units, called a **grammar**.[1]
- These two elements provide a systematic but finite basis for computing the properties of any syntactically legitimate sentence.
- Although grammatical theories differ about the nature of lexical versus grammatical information and their manner of interaction, a particular theory must establish what counts as a word in order to pin down what the lexicon should contain.

MWE

Multiword expressions (MWEs) are a class of linguistic forms spanning conventional word boundaries that are both idiosyncratic and pervasive across different languages.

The structure of linguistic processing that depends on the clear distinction between words and phrases has to be re-thought to accommodate MWEs.

The issue of MWE handling is crucial for NLP applications, where it raises a number of challenges.

Characteristics of MWE

- MWEs consist of several words (in the conventionally understood sense) but behave as single words to some extent.
- This is well illustrated by an expression like *by and large*, which any English speaker knows can have roughly equivalent meaning and syntactic function to *mostly*, an adverb.

Characteristics of MWE

Among the problematic characteristics of this expression are :

(1) syntactic anomaly of the part-of-speech (POS) sequence preposition + conjunction + adjective,

(2) non-compositionality: semantics of the whole that is unrelated to the individual pieces,

(3) non-substitutability of synonym words (e.g., *by and big*), and

(4) ambiguity between MWE and non-MWE readings of a substring *by and large* (e.g., *by and large we agree* versus *he walked by and large tractors passed him*).

- Although these characteristics by no means exhaust the list of peculiarities, the idiosyncratic nature of the expression is plain, leading us to ask where its pertinent characteristics should be stored.
- The traditional division of labor gives us two options—the lexicon or the grammar—but MWEs disrupt the tradition precisely because they are more than one word long (Sag et al., 2002).
- Their idiosyncrasy suggests that they belong in the lexicon, yet, being constructed out of more than one word, they would also fall within the traditional scope of grammar, even if constituted (cf. *by and large*) from non-standard sequences of syntactic categories.

As we shall soon see, the glue that can hold an MWE together often involves grammatical relations between the sub-parts, so that the structure of linguistic processing tasks such as parsing and machine translation (MT), which depends on a normally clear distinction between word tokens and phrases, has to be re-thought to accommodate MWEs.

The issue of MWE handling goes to the heart of natural language processing (NLP) where it raises a number of fundamental problems with a frequency that cannot be ignored.

Definitions and Categories

- Definitions of MWEs driven perhaps by their awkwardness—which causes trouble in many corners of linguistic study—their lack of homogeneity, and their surprising frequency.
- The awkwardness arises from the way in which they transcend boundaries imposed by the different subfields of morphology, lexicology, syntax, and semantics.
- Their lack of homogeneity has led to various categorization schemes that we discuss further subsequently.

1) “a multiword unit or a collocation of words that co-occur together statistically more than chance” (Carpuat and Diab, 2010)

2) “a sequence of words that acts as a single unit at some level of linguistic analysis” (Calzolari et al., 2002)

3) “idiosyncratic interpretations that cross word boundaries” (Sag et al., 2002)

4) “lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity” (Baldwin and Kim, 2010)

- The first two focus mainly on the essential *structural* aspects of MWEs evidenced by the unusual co-occurrence of two or more elements within a template of some kind. The complexity of the template can vary widely, from a simple sequence of two fixed words, to longer sequences of less tightly specified elements (e.g., lexemes) constrained by syntactic and/or semantic relationships, with the possibility of intervening gaps.
- The third definition emphasizes the essentially idiosyncratic semantic aspect of MWEs, evidenced by degrees of non-compositionality in arriving at the interpretation of the whole from the several parts.
- Baldwin and Kim (2010), which captures both of these aspects—that is, outstanding co-occurrence (i.e., collocation or statistical idiomaticity) and generalized non-compositionality (i.e., lexical, syntactic, semantic, and pragmatic idiomaticity). This definition also emphasizes that the anomalies of MWEs are manifest over different linguistic levels.

Types of MWE

- An **idiom** is a group of lexemes whose meaning is established by convention and cannot be deduced from the individual lexemes composing the expression (e.g., *to kick the bucket*).

- A **light-verb construction** is formed by a head verb with light semantics that becomes fully specified when combined with a (directly or indirectly) dependent predicative noun (e.g., *to take a shower*).[2]

- A **verb-particle construction** comprises a verb and a particle, usually a preposition or adverb, which modifies the meaning of the verb and which needs not be immediately adjacent to it (e.g., *to give up*). Verb-particle constructions are also referred to as **phrasal verbs** in this article and elsewhere.

Types of MWE

- A **complex function word** is a function word formed by more than one lexeme, encompassing multiword conjunctions (e.g., *as soon as*), prepositions (e.g., *up until*), and adverbials (e.g., *by and large*).
- A **compound** is a lexeme formed by the juxtaposition of adjacent lexemes, occasionally with morphological adjustments (e.g., *snowman*).[3]
    - Compounds can be subdivided according to their syntactic function.
    - Thus, **nominal compounds** are headed by a noun (e.g., *dry run*)
    - whereas **noun compounds** and **verb compounds** are concatenations of nouns (e.g., *bank robbery*) or verbs (e.g., *stir fry*).
    - **closed compounds** when they are formed from a single token (e.g., *banknote*), and
    - **open compounds** when they are formed from lexemes separated by spaces or hyphens.
        - check-in, clean-cut, editor-in-chief
        - child care, work day, and time save

Types of MWE

- A **multiword named entity** is a multiword linguistic expression that rigidly designates an entity in the world, typically including persons, organizations, and locations (e.g., *International Business Machines*).

- A **multiword term** is a multiword designation of a general concept in a specific subject field[4] (e.g., *short-term scientific mission*).[5]

Types

  - *Kick the bucket* (meaning: to die)
  - *Break the ice* (meaning: to start a conversation in a relaxed way)
  - *Let the cat out of the bag* (meaning: to reveal a secret)
- **Phrasal Verbs:** Verb + particle combinations where the meaning may be literal or figurative.
  - *Look up* (literal: to search for; figurative: to improve)
  - *Break down* (literal: to fall apart; figurative: to lose control emotionally)

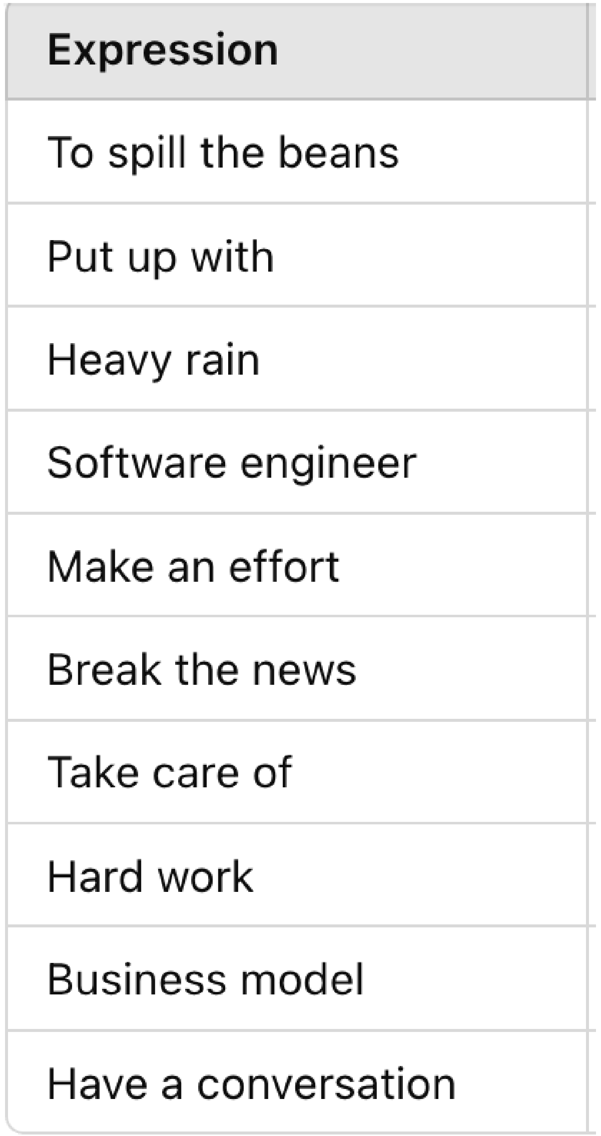
- **Idioms:** Fixed expressions where the meaning is not deducible from the individual words.

  - *Strong tea* (not *powerful tea*)
  - • • *Make a decision* (not *take a decision* in American English)

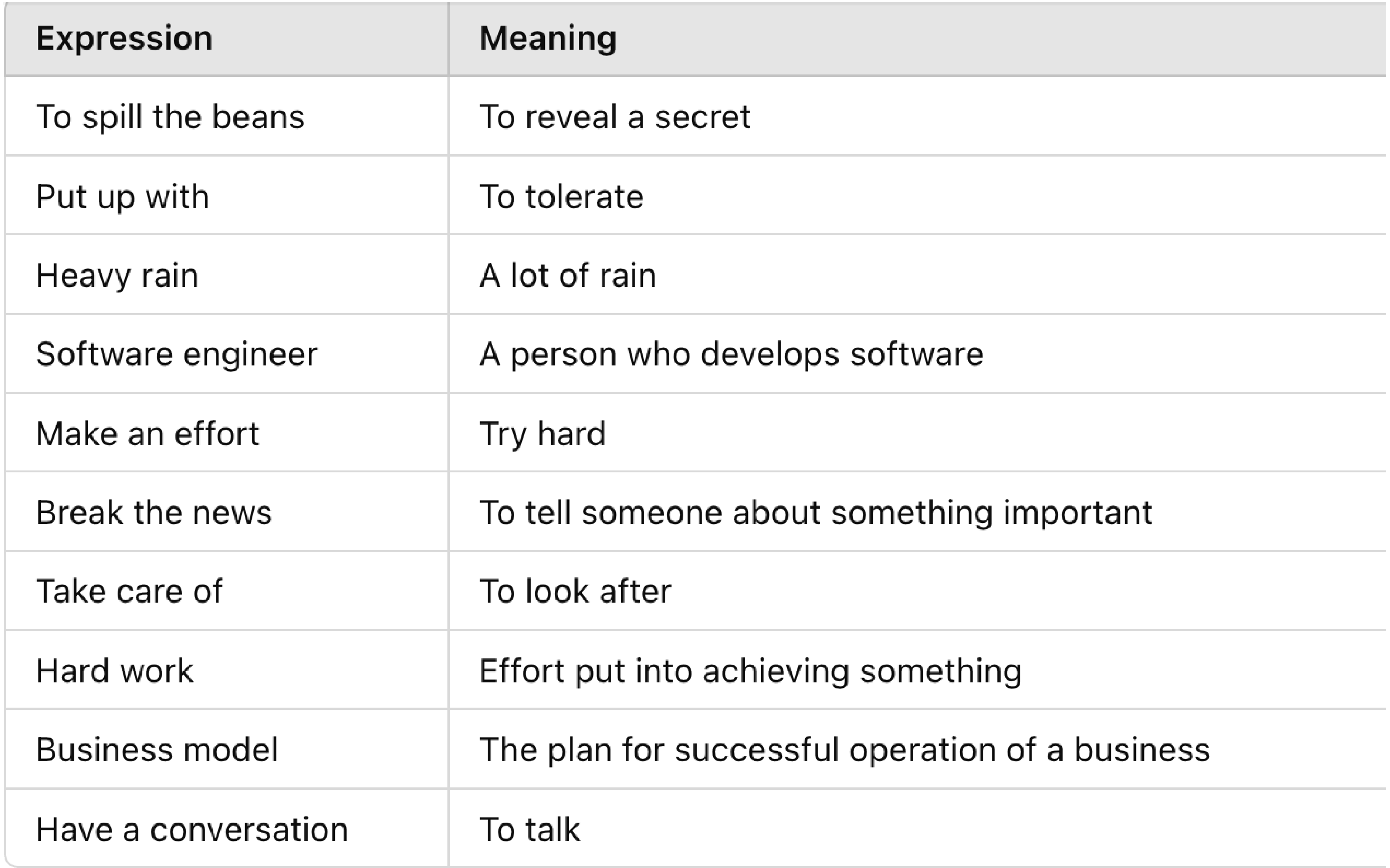
- **Collocations:** Words that frequently occur together.

- *Traffic light*
  - *Data science*
  - *Coffee table*
- **Light Verb Constructions:** A verb (often *make, take, do, have*) combined with a noun where the verb contributes little semantic meaning.
  - *Make a mistake*
  - *Take a break*
  - • • *Do homework*

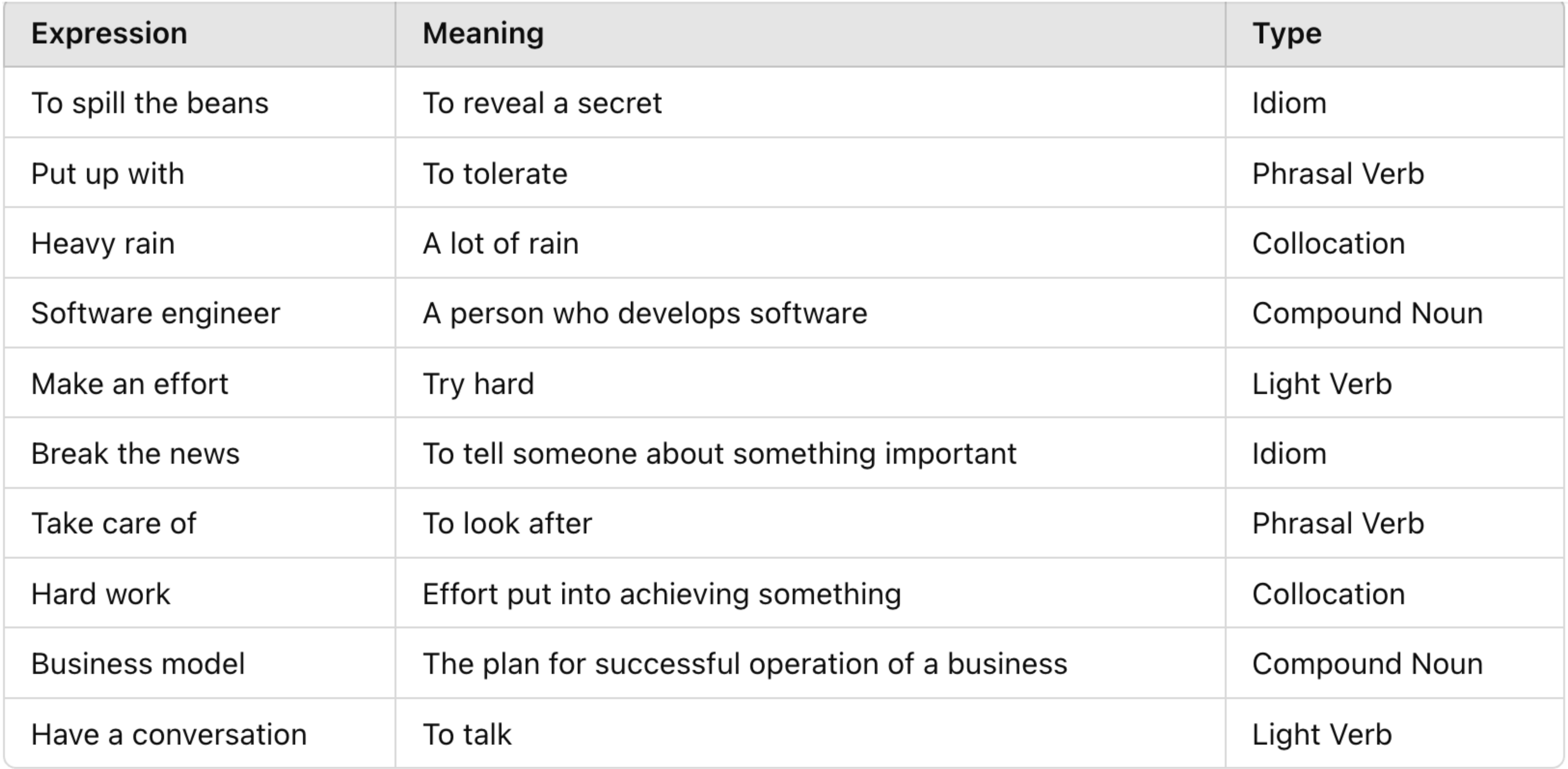
- **Compound Nouns:** Two or more words functioning as a single noun.



Expression
To spill the beans
Put up with
Heavy rain
Software engineer
Make an effort
Break the news
Take care of
Hard work
Business model
Have a conversation



Expression	Meaning
To spill the beans	To reveal a secret
Put up with	To tolerate
Heavy rain	A lot of rain
Software engineer	A person who develops software
Make an effort	Try hard
Break the news	To tell someone about something important
Take care of	To look after
Hard work	Effort put into achieving something
Business model	The plan for successful operation of a business
Have a conversation	To talk



Expression	Meaning	Type
To spill the beans	To reveal a secret	Idiom
Put up with	To tolerate	Phrasal Verb
Heavy rain	A lot of rain	Collocation
Software engineer	A person who develops software	Compound Noun
Make an effort	Try hard	Light Verb
Break the news	To tell someone about something important	Idiom
Take care of	To look after	Phrasal Verb
Hard work	Effort put into achieving something	Collocation
Business model	The plan for successful operation of a business	Compound Noun
Have a conversation	To talk	Light Verb

- MWEs can be characterized by a number of properties that on the one hand present challenges for MWE processing and the two use cases, namely, **parsing and MT**

Properties of MWEs

1) Collocation

- Arbitrarily prominent co-occurrence, that is, **collocation**, is one of the outstanding properties of MWEs.
  - For example, although the words *strong*, *powerful*, *intense*, and *vigorous* are (near) synonyms, only *strong* is usually used to magnify the noun *coffee* (Pearce, 2001).
  - This property has been heavily used by MWE discovery methods partly because it is easy to capture using statistical **association measures.**

- Conversely, prominent co-occurrence is problematic for MT, because word-for-word translation might lead to translations of words that are suitable individually, but that yield non-fluent or ambiguous translations of MWEs.
- For instance, the Italian expression *compilare un modulo* has to be translated into English as *to fill in a form* rather than the word-for-word translation *to compile a module*.
- **Non-substitutability**. It is not possible to replace part of an expression by a synonym or similar word. This property is generally modeled by variability or fixedness measures

2) Discontiguity,

- **Discontiguity**, whereby alien elements can intervene between core MWE components, is a challenge for MWE processing.
- For instance, the Portuguese expression *levou em conta* (*to take into account*) licenses a direct object that can either appear after the idiom, like in *ele levou em conta minha opinio* (*he took into account my opinion*) or between the verb and the fixed prepositional complement, like in *ele levou minho opinio em conta* (*he took my opinion into account*).
- Discriminating the intervening words from the core can be non-trivial but if they form a single syntactic constituent, as in the example, the task can be facilitated by syntactic analysis, thus creating an opportunity for parsing.

3) Non-compositionality

- **Non-compositionality** is prototypical in idioms such as the French nominal compound *fleur bleue* (lit. *blue flower*).
- This expression is used to characterize a sentimental and often naive person, so its meaning is completely opaque to speakers who only know the meanings of the individual words.
- This property is a challenge for MT because translating non-compositional MWEs through the individual words or structures will very often yield an inappropriate translation.
- The problem of non-compositionality of MWEs requires a strategy aiming to correctly identify the borders of MWEs and to find the associated sense of the expression.

4) Ambiguity

- **Ambiguity** is a challenge for many NLP tasks.
- The type of ambiguity that impacts MWE processing the most is the choice between a compositional and an MWE reading of a sequence of words,
- As illustrated by the sentence *I am struck by the way the rest of the world is confident of a better future.*
- In most cases the sequence of words *by the way* is an MWE with the approximate meaning of *incidentally*.
- However, in the example it is a regular prepositional complement of the verb *struck*.

4) Ambiguity

- In some cases, syntactic analysis can aid in determining whether the sequence of words should be recognized as an MWE.
- An analysis that takes *by the way* to be an MWE and thus an adverb in this case, will yield an ungrammatical sentence (which becomes clear when we replace *by the way* with *incidentally*: *I am struck incidentally the rest of the world …*).
- Parsing can help reveal the relevant subcategorization frame that includes the preposition selected by the verb.

5) **Variability**

- **Variability**, that is, the fact that MWEs allow for varying degrees of flexibility in their formation, poses great challenges for their identification.
- Searching for fixed forms only will lead to low recall, because the fixed form will fail to match all possible variations.
- For example, *een graantje meepikken* (lit. *to pick a grain with the others*) is a Dutch MWE meaning to benefit from something as a side effect.

5) **Variability**

- Just searching for the fixed string *graantje meepikken* will not identify *Zij pikken er hun graantje van mee* (lit. *they pick their grain of something with the others*), meaning that they are benefiting from something as a side effect.
- However, syntactic and semantic analysis can help us identify the parts of this MWE that allow for variation, here the determiner that can be changed into a possessive pronoun.

MWE Identification Methods

- **Rule-based methods** apply rules of various levels of sophistication to project MWE lexicons onto corpora
- **Classifiers** typically used for word sense disambiguation can be adapted to token-based MWE classification using contextual features .
- **Sequence tagging models**, inspired by POS-tagging, chunking, and named entity recognition, can be learned from manually annotated corpora using supervised techniques.
- Identification can also be performed as a by-product of *parsing*,

Hindi Examples

- Hindi_MWE_utf.txt