CHAPTER 3

MULTIWORD EXPRESSIONS

This chapter presents an overview of Multiword Expressions (MWEs), its characteristics, idiosyncrasies and different classes of MWEs. Its fundamental characteristics are formed into research directions. It also presents some applications of Multiword expressions in the field of computational linguistics.

3.1 Introduction

One of the important issues in Natural language processing is appropriate processing of Multiword Expressions as proposed by Baldwin and Kim (2010) as Multiword Expressions are lexical items that (a) can be decomposed into several lexemes and (b) show lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity. MWEs are made up of the combination of two or more than two words in which most of the time words lose their individual meaning and form a new resultant meaning. They are idiosyncratic in nature either by semantic, syntactic and lexical way. Sag *et al.* (2002) define MWEs as "peculiar interpretations that cross word boundaries (or space).

Concept of MWEs is closely associated with collocation concept. A collocation is a sequence of two or more consecutive words that has characteristics of a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components. For example, 'kick the bucket'', 'রাজ্য সরকার', 'উভয় সঙ্কট', 'গৃহপালিত', River bank and ATM Card etc.

The term MWE originally quoted by Firth as "We can know a word by the company it keeps". He said that "collocation of a given word are combination of the regular and customary places of that word" (Firth, 1957), (Ramisch, 2015). For example, New Delhi, San Francisco, পাপপুণ্য, খাওয়া দাওয়া , সুবর্ণ সুযোগ (golden opportunity), ঢল ঢল , ঝক মক[1] etc.

To detect MWEs, we need to test word to word translation into another language. If the translation is unnatural and ungrammatical, the original expression is

[1] **It is a Dhonnathak word in Bengali**

considered to be MWEs. Habitually a variety of linguistic expressions are used everyday both in spoken and written form of language which are MWEs. MWEs may be compound noun(e.g., movie display), compound verb (e.g., come up), Idioms (e.g., kick the bucket), institutionalized phrases (e.g., গৃহ মন্ত্রালয় (home ministry)) etc. Jackendoff (1997) noted that the occurrence of MWEs in a speaker's lexicon is of the equal order of size of single words.

Morphologically, some constituents of MWEs allow to occur inflection freely while preventing the variation of other constituents. MWEs may allow constituents to undergo atypical morphological inflections so that they would not undergo in separation. Syntactically, some MWEs act like phrases as a single concept, some occur in fixed order while others are semi fixed with various syntactic transformations. Semantically, the compositionality (features) of MWEs is changing, ranging from fully compositionality to non-compositionality (Bannard *et al*., 2003).

Idiomaticity of MWEs demands exhaustive enumeration of rule or grammar generation for machine understanding. For Machine Translation, database of a particular language is not enough. A database requires a single or multiword unit of such target language. Manual creation of database of MWEs is problematic due to the following reasons. First, the available resources of dictionaries and thesauri of different languages are not sufficient to create a database for MWEs processing. Secondly, it is time consuming and prone to errors. Third, it is difficult to choose whether a given expression satisfied the criteria of MWEs. Based on the collocation and anti-collocation frequency of measures MWEs are computed automatically from a corpus.

MWEs have a great issue in Natural Language Processing. Day by day some of the common words, that a native speaker uses in general and that have idiomatic nature, require to be machine translation for the purpose of language processing, but it is difficult in all languages due to recent growing research in MWEs both in India and other countries. Some of the English words are going to be converted to common usable Bengali words e.g. Big Bazar, Alert message (অ্যালার্ট মেসেজ), common people (আম আদমি), shopping Moll (শপিং মল) etc. Thus, these words are going to be idiomatic in nature which shows MWEs properties. Proper treatment of such words using computer system is very needful in present computational linguistics. MWEs vary from language to language based on lexical, syntactic, semantic properties of a particular

language. Isolation and identification of MWEs are based on the linguistics rules of that language.

MWEs are mostly used to improve ease of language, versatility and understandability of language use. Now-a-days, MWEs are used rapidly (either explicitly or implicitly) in machine translation to overcome the syntactic, semantic and realistic effect in the source and target languages. It has been shown that correct identification of MWEs influences the correctness of semantic tagging (Piao *et al*., 2003) and word arrangement in machine translation can be improved through a particular handling of the sentence structure and semantics of MWEs (Venkatapathy and Joshi, 2006).

3.2 Necessary and sufficient condition for Multiword Expressions

To identify a word to be MWEs, it needs to satisfy some necessary and sufficient conditions so that it fulfills MWEs criteria for any language. The necessary and sufficient conditions are:

Necessary condition

For a word sequence to be MWEs, it has to be separated by space/delimiter. This condition was taken in Kashmir Multiword workshop 2011. Since Multiword are collocations of individual words therefore space or delimiter will provide their collocations sequence.

Sufficient Condition

1. **Non-compositionality:** It implies that meaning of a Multiword Expressions cannot be derived from its constituents. For example, a gentle man (ভদ্রলোক). That is, meaning of the MWEs should be different from their individual words collocation because lexical properties (syntactic and semantic class) of MWEs need to be different from the word lexical properties of individual word.
2. **Fixity of expression:** It implies that constituents of MWEs cannot be replaced or modified by its synonyms or other words. For example,

    a) Attorney General

We do not say: General Attorney

b) Extremely sorry

We do not say: sorry extremely

c) গৃহ মন্ত্রালয় (home ministry)

We do not say: মন্ত্রালয় গৃহ (ministry home)

Thus, it is seen that if MWEs are replaced directly or indirectly, meaning will be completely unnatural and ungrammatical from the context. Thus, fixity of an expression is very important in case of MWEs (Minia, 2012).

3.3 Linguistic properties of Multiword Expressions

While studying linguistic properties of MWEs, we follow the definition of MWEs given by Sag *et al.* (2002) as Multiword Expressions are lexical items that (a) can be decomposed into many lexemes and (b) show lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity

The Linguistic properties of MWEs reveal the decomposition of lexemes. For a language like English, interpretation of decomposability into lexemes is that MWEs must in themselves be made up of two or more than two whitespace delimited words. For example, Advocate General is strongly an MWE as it is consist of two lexemes (advocate and general), while fused words such as green house are not classified as MWEs[1]. In the language like German, the frequency of occurrence of compound nouns, for example Kontaktlinse "contact lens" (the concatenation of Kontakt "contact" and Linse "lens"), without whitespace delimitation, means that we tend to eliminate this restriction and allowing MWEs as a single-word. Again, language like as Japanese and Chinese (Baldwin and Bond 2002; Xu *et al*., 2006), save this artificial consideration.

Decomposition of an expression into many lexemes is still appropriate and may lead to the conclusions. For example, in 'multiword expression' both multiword and expression are stand alone multiple lexemes but in 'school head' School is standalone lexeme, but head is not (Baldwin *et al.*, 2010).

[1] In practice, s significant subset of research on English noun compounds has considered both fused and whitespace-separated expressions.

One of the important features in MWEs is idiomaticity. Baldwin and Kim, (2010) explained details accounts of idiomaticity in its various manifestations. They are described as below.

3.3.1 Idiomaticity

In the perspective of MWEs, Idiomaticity is defined as lexico-syntactic, semantic, pragmatic and statistical markedness (Katz and Postal, 2004; Chafe, 1968; Cruse, 1986; Jackendoff, 1997). That is Idiomatic MWEs is derived from basic properties of the component lexemes. Lexico-syntactic idiomaticity implies that the MWEs have peculiar nature given the syntax of the individual simplex words. For example, 'by and large' is an idiom. Other example like 'state government' is an entirely unsurprising combination of the nouns state and government, whereas 'by and large' is a coordination of a preposition and an adjective to form an adverbial phrase, Thus it is not predicted by Standard English grammar rules. As such, 'state government' is not lexico-syntactically idiomatic while 'by and large' is. Semantic deviation commonly happens in idioms such as in one's cloths, where the semantics is not rapidly predictable from the simplex semantics of cloth. Pragmatic idiomaticity occurs in a rigid set of situated expressions such as 'good morning' and 'all board'. That is, these MWEs are associated with very particular situations and odd with other context.

Statistical idiomaticity occurs with MWEs such as 'buy and sell' where they occur with uncommonly high frequency compared to alternative forms of the same expression. It is perfectly acceptable to say buy and sell, but this idiom is considered as trading in English. Idiomaticity is closely related with compositionality, in which we determine the extent to which the features of the part of MWEs are combined to guess the features of the whole. Compositionality is regularly constructed to apply in semantic idiomatic expression that maintains same level of idiomaticity.

Types of idiomaticity explained by (Baldwin and Kim, 2010) are as follows:

3.3.1.1 Lexical Idiomaticity

Lexical idiomaticity is defined as when one or more component of MWEs are not as a part of the English lexicon. Example like, ad hoc is lexically considered as

MWEs in that neither of its components (ad and hoc) are standalone English words. Lexical idiomaticity ultimately gives outcome of both syntactic and semantic idiomaticity because there is no lexical idea related with individuals components to predict the behavior of the MWE. Thus, it is one of the transparent properties of MWEhood and is non-decomposable.

3.3.1.2 Syntactic Idiomaticity

Syntactic idiomaticity happens when the sentence structure of the MWE is not derived from its components directly (Katz and Postal, 2004; Chafe, 1968). For example, 'to and fro', is syntactically idiomatic in that it is adverbial in nature, but it consists of different composition of a preposition (to) and an adjective (fro). In addition, take a walk is not syntactically considered as it is a simple verb+object mixture which is the resultant of a transitive verb (take) and a countable noun (walk). Syntactic idiomaticity also occurs at the analysis level of MWEs having syntactic properties which are different from their component words, e.g., verb-particle constructions and determinerless prepositional phrases.

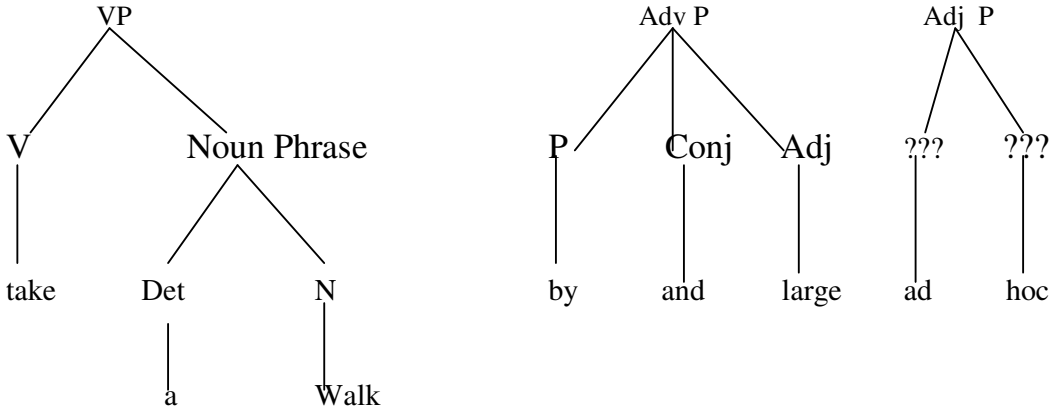

Figure 3.1: Examples of Syntactic non-markedness vs. markedness

3.3.1.3 Semantic Idiomaticity

Semantic idiomaticity is a reflection of the meaning of an MWE not being explicitly or implicitly derivable from its parts (Katz and Postal; Bauer, 1983). For

example, ‘dog in the manager’ usually indicates “self-seeker”, which we cannot expect from either dog or manager. Furthermore, at stake is not semantically noticeable as its semantics is fully identifiable from its parts. However, many cases are not transparent as these. The semantics of ‘live like cat and dog’ (“quarrel over a small matter”), viz., it is partly predictable from live like (“sustain” and hence “follow”), but not as immediately from cat and dog. These are the issues where a meaning of the components of MWEs works in this fashion. A part from these, additional semantic is there which is difficult to realize. One such example is car driver where car and driver both have their separate meanings, but in aggregate it implies a car driver is “one who drives a car” but not “one who drives like a car”.

Related with the issue of semantic idiomaticity there is a discussion on the notions of non-identifiability and figuration relatation. Figuration is the properties of components of MWEs having some metaphoric, hyperbolic or metonymic meaning in addition to their literal meaning (Liberman and Sproat, 1992). Example of semantic idiomaticity is given below Chakrabarty (2011):

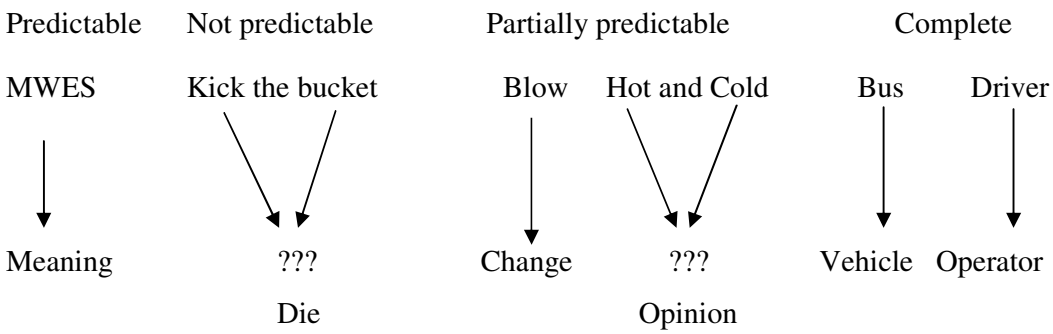

Figure 3.2 Examples of Semantics Idiomaticity

Non-identifiability (Nunberg *et al*., 1994) is the notion of the meaning of an MWE not being easily predictable from the surface form (components), similar to our definition of semantic idiomaticity. Example like, the meaning of ‘kick the bucket’ (means “die”) cannot be predicted from kick or bucket separately. Another example is hard up, where the parts (i.e., hard and up) do not semantically contribute to the meaning of the whole. This property relates closely to compositionality. That is, the meaning of MWEs can be envisaged from the parts when MWEs are compositional.

Hence, there is non-identifiability related with non-compositionality. Other examples of non-identifiable and non-compositional MWEs are ‘on roof’, control over and ‘by and large’.

Figuration (Fillmore *et al.,* 1988; Nunberg *et al.,* 1994) is an attribute of encoded expressions such as metaphors (viz. take the bull by the horns), metonymies (viz. give assistance) and hyperboles (e.g., of no value). It implies the properties of the components of an MWE having some metaphoric or hyperbolic meaning in addition to their literal meaning. Semantics of the MWE is achieved from the components through the sequence of metaphor, hyperbole or metonymy, although exact nature of the figuration may be more or less .Thus, figuration involves interactions between idiomatic and literal meaning.

Pragmatic idiomaticity is concerned with an MWE being associated with a fixed set of situations or a particular context (Kastovsky, 1982; Saget *et al*., 2002). ‘good afternoon’ and ‘boarding time’ are examples of pragmatic MWEs: the first is a greeting associated specifically with afternoon and the second is a command associated with the specific situation of a train/flight to take passenger or departure of a train/flight from specific railway station platform or terminal of an airport. Pragmatic idiomaticity of MWEs are indistinct with (non situated) factual translations; e.g., good afternoon can mean “pleasing afternoon”.

3.3.1.4 Statistical idiomaticity

It occurs when a association of words occur with relative frequency of the component words or another phrasings of the same expression (Pawley and Syder, 1983). Cruse (1986) provides some good examples of statistical idiomaticity in the matrix of adjectives and nouns presented in Table 31.

Table 3.1: Examples of statistical idiomaticity ("+" = strong lexical affinity, "?" = marginal Lexical affinity, "−" = negative lexical affinity) (Cruse, 1986)

Word	unblemished	spotless	flawless	immaculate	impeccable
performance	-	-	+	+	+
argument	-	-	+	_	-
complexion	?	?	+	-	-
behaviour	-	-	-	-	-
kitchen	-	+	-	+	+
record	+	+	-	?	?
reputation	?	+	-	?	?
taste	-	-	?	?	?
order	-	-	?	+	+
credentials	-	-	-	-	_

This examples based on the bunch of near synonyms adjectives (unblemished, spotless, flawless, immaculate, impeccable) these examples are given and their affinity to pre-modify a range of nouns. Different noun have particular preferences for certain subject subsets of the adjectives as modifiers, as indicated by the cells in the matrix ("+" marks as a positive lexical affinity, "?" marks as a marginal lexical affinity, and "−" marks as a negative lexical affinity). For example, Impeccable has a strong lexical affinity with fulfillment (Impeccable shows a relatively common expression), whereas spotless has a negative affinity with credentials (spotless credentials is relatively infrequent). In other cases, there may be more or less frequent cases of linguistics, semantic or other basis for a particular adjective noun combination.

Statistical Idiomaticity is simply a perception of occurrence frequency in a combination. In addition, statistical idiomaticity is a regular graded occurrences and our observation is about lexical affinity in the table 3.1

In the table 3.1, grading of the inclination for each of the adjectives take place as a pre-modifier of record is shown in table 3.1,viz. 'impeccable' and 'spotless' are more probable selection than immaculate, which is in turn more probable than 'flawless'.

In some cases, statistical idiomaticity is the changed ordering of the components. For example, black and white is much more common in English than white and black, the overturn ordering does not hold the lexicalized semantics meaning of the word formation. While the overturn holds in the case of other languages such as Japanese and Spanish. In this work, we follow Sag *et al*. (2002) in referring to MWEs which are only statistically idiomatic (but not in lexico-statistically, semantically or pragmatically idiomatic) as collocations.

Statistical idiomaticity relates to the notion of institutionalization or conventionalization, it means a particular word formation coming to be used to refer to a given object (Fernando and Flavell, 1981; Nunberg *et al*., 1994). For example, traffic light is the conventionalized indicator for "a visual signal for road user to direct the flow of traffic at traffic points. There is no any appropriate reason why it wouldn't be called as a traffic controller or traffic director or crossroads regulator, but reason is that it is not referred to using all these expressions. Instead, traffic light was settled on as the recognized term for referring to the object. Likewise, it is random nature in the English language as we say 'many thanks' not as a several thanks, and 'salt and pepper' in first choice to 'pepper and salt[2].

Nunberg *et al.* (1994) considered collocation to be the most important property of MWEs. We consider collocation related with semantic, pragmatic and statistical idiomaticity, but we consider MWEs do not have any one of these three forms of markedness (e.g., MWEs which are strictly lexico-syntactically idiomatic are classified as MWEs in this research). Collocations are most transparent when observed in comparison with anti collocations.

Anti-collocations are lexicosyntactic form of collocations which have unpredictably low frequency (Pearce, 2001). For example, 'buy and sell' is an anti-collocation for 'sell and buy', and 'traffic director' is an anti-collocation for 'traffic light'.

[2] Which is not to say there wasn't grounds for the selection of the canonical form at its genesis, e.g., for historical, cross lingual or phonological reasons.

The use of the term collocation in our cases differs from the mainstream usage in the computational linguistics, where a collocation is often defined as an arbitrary and recurrent word combination that co-occurs more often than would be expected by chance (Choueka, 1988; Lin, 1998b; Evert, 2004).

In above, we have described four categories of Idiomaticity. These are brought together into a single table as shown below Chakrabarty (2010) as :

Table 3.2: Classification MWEs in term of their different Idiomaticity

	Lexico-sysntactic	Semantic	Pragmatic	Statistical
All abroad	-	-	+	+
Black and white	-	?	-	+
By and large	+	+	-	-
Kick the bucket	-	+	-	-
Social butterfly	-	+	-	+
Make out	-	+	-	-
Shock and awe	-	-	+	+
To and fro	+	-	-	+
Bus driver	-	+	-	+
Traffic light	-	-	-	+

In the Table 3.2, some examples like kick the bucket, make out and traffic light are marked with only one form of idiomaticity, which is sufficient for them to be classified as MWEs. Again, other MWEs such as shock and awe and to and fro are idiosyncratic in different ways, shock and awe show pragmatic idiomaticity because of its incident in the World War II in Japan, and to and fro as being lexico syntactically idiomatic by its nature.

3.4 Other Properties of Multiword Expressions

Baldwin and Kim (2010) explained some other general properties of MWEs which are: single word paraphrasability, provability and prosody. Compared with

idiomaticity, where some form of idiomaticity is a essential feature of MWEs, these other properties are optional. Prosody relates to semantic idiomaticity, while the other properties are not related with idiomaticity as explained above.

3.4.1 Cross Lingual Variation

Since there are noticeable changes in MWEs across languages (Villavicencio *et al*., 2004). In some cases, there is straight suiting for a cross lingual MWEs pair with related semantics. Some other MWEs are lexically similar but syntactically different.

There are many MWEs which have no straight translation equivalent to other languages. Similarly, there are terms which are realised as MWEs in one language but single word lexemes in other language, such as 'hard and fast' and its Bengali equivalent 'ধরাবাঁধা'।

3.4.2 Single-word Paraphrasability

Single word paraphrasability is the study that considerable numbers of MWEs can be paraphrased with a single word (Chafe, 1968; Nunberg et al., 1994). Though some other MWEs are single-word paraphrasable (e.g., run down means "cut"), Again, MWEs with point of view can sometimes be paraphrasable (e.g., bite the dust means "blamestorming"). Similarly, multiword non-MWEs can be single-word paraphrasable (e.g., not sufficient = "insufficient").

3.4.3 Proverbiality

Proverbiality is the capability of MWEs to "explain wholly a frequent circumstances of particular social importance in the good quality of its likeness or relation to a circumstances concerning unpleasant, actual things and relations" (Nunberg et al. 1994). For example, Verb Particles Constructions and idioms are often indicators of more casual situations. Nunberg *et al*. (1994) treat informality as a separate category, where we combine it with proverbiality.

3.4.5 Prosody

Prosody refer to the stress pattern of a language .MWEs exhibit different prosody i.e., stress patterns, from compositional language (Fillmore *et al*., 1988; Nunberg *et al*., 1994). For example, when the components are unequal distribution to the semantics of the whole, pattern of stress in MWEs can be marked, e.g., soft spot is prosodically marked (since stress on soft not in spot). Again, prosodically unmarked MWE is first aid and prosodically-marked non-MWE is surgical operation.

3.5 Collocations and Multiword Expressions

A collocation is a term in NLP which is directly related to our study of MWEs. A common definition for collocation is "an arbitrary and recurrent word combination" Benson (1990), it is a statistically idiomatic MWE. A significant difference is there among individual researchers, collocations are often distinguished from "idioms" or "non-compositional phrases" as basically they do not show syntactical idiomaticity, and if they are semantically idiomatic, it is a quite transparent process of figuration or metaphor (Choueka, 1988; Evert, 2004). While a lot of works have been done on collocations, based on preset constructional templates (e.g., adjective-noun or noun-verb collocations) finally, collocations form a proper subset of MWEs.

3.6 Types of Multiword Expressions

On literature survey of Multiword expressions it is seen that MWEs are categorized (Minia, 2012) as follows

1. Compound Noun

2. Conjunct Verb

3. Compound Verb

4. Reduplication

3.6.1 Compound Noun

Noun is considered as parts of the language which provide the vocabulary to describe things and concepts. New concept is the outcome of new nouns being added to the language. The new words are generated in language by combining existing nouns to form new compound nouns. A compound noun is a noun consisting of more than one free morpheme. For example,

Gloss: movie display

Translation: movie show

Compound nouns can be names (i.e., proper nouns) or common nouns that have become uses due to institutionalized usage or display semantic non-compositionality. Compound nouns consist of concatenated nouns, but they exhibit an internal hierarchical structure. Compound noun can generally be expressed recursively in terms of head-modifier relationships, giving rise to bracketed structures.

3.6.2 Conjunct Verb

Every language has a well defined syntax of lexicons. The conjugation of a verb shows a variety of forms, it can assume either by intonation or by grouping with parts of other verbs, to marks voice, mood, tense, number, and person and those are required to be added its infinitives and participles. In this case, meaning of the conjoined term is not strictly combinational. Conjunct Verb constructs different types of construction.

For examples,

1. He helped me with the car.
2. He gave me help with the car.

Normally, the noun is a true entity and it is not required to store it as a lexical unit all along with a co-occurring verb. It is necessary to divide true conjunct verbs from Noun-Verb sequences.

3.6.3 Compound Verb

This type of MWEs consist of Polar Verb and Vector Verb combination. Construction of Verb +Verb is complicated to express, as there are many serial sequences to be found in the languages. For example,

Ram is reading the book

রাম বই পড়িতেছে

Gloss: Up come

Translation: Come up

3.6.4 Reduplication

Reduplication is a linguistic phenomenon that occurs in language across all Indian languages. Reduplication is a morphological process in which the root or stem of a word, or part of it, is repeated. There is no standard classification exists, the following are the major classes of reduplications that commonly occur in Indian languages (Kmw, 2011).

3.7 Onomatopoeic expressions

The constituent words imitate a sound, and the unit as a whole refers to that sound.

মিট মিট (Bengali).

Transliteration: mit mit

Gloss: twinkle twinkle

Translation: twinkle twinkle

3.8 Complete Reduplication

The individual words are meaningful, and they are repeated. For examples, খাওয়া দাওয়া, বড় বড়(big big), ধীরে ধীরে (slowly).

Transliteration: khaoa daoa

Translation: eat

3.9 Partial Reduplication

In this, only one of the words is meaningful, while the other word is constructed by partial reduplicating the first word. There are various ways of constructing such reduplications, but the most common type in Hindi is one where the first syllable alone is changed. For Example, বোকা সোঁকা (foolish)

Gloss: water voice

Translation: water

3.10 Semantic Reduplication

The semantic reduplication is related with semantic relations between paired words which are synonyms (ঘর-দুয়ার, house), antonym (দিন রাত, day and night). The pair of words in reduplication work as a single word syntactically and generally indicates a single concept. Thus, reduplicate expressions are truly MWEs.

3.11 Idiosyncrasies Observed in Multiword Expressions

Due to the peculiar nature of MWEs, idiosyncrasies are observed in multiword expressions in semantic, lexical, syntactic and statistical level which are as follows:

1. **Semantic**: It implies meaning and semantics which are not understandable from the composition of the meaning of the constituent words. For example, “show him the door”, the individual constituent words have no connotation to the actual meanings of the phrase.

2. **Lexical:** Collocations not generally observed in the language probably borrowed from other languages and institutionalization due to usage.

   For example, ad hoc

3. **Syntactic:** It means certain collocation do not follow the rules of the conventional grammar. Thus it gives a meaningful interpretation.

For example, By and Large

4. **Statistical:** A few collocations are compositional both semantically and syntactically. But, these collocations has a tendency to co-occur together more than what can be attributed to change. For example, it will be more likely co-occur *"*traffic signal*"* than "traffic lamp" or *"*traffic lights*"*, though they all mean the same thing. Such an idiosyncrasy arises because of the occurrence of collocation with the concept and or institutionalization of the collocation. Thus, statistical occurrence of collocation of word is a criterion of MWEs. Some instances of such collocations are *"*good morning*"*, *"*নববার্তা".

3.12 Characteristics of Multiword Expressions

MWEs have many characteristics due their idiosyncratic nature. These Characteristic vary based on the context. Following are the characteristics based on the non-compositionality of MWEs.

1. **Institutionalization:** MWEs occur in non-conventional usage and thus collocations show statistical significance in their occurrences. For example, traffic light, prime minister (প্রধান মন্ত্রী).

2. **Paraphrasability:** In many cases MWEs stand for single concept, it might be possible to paraphrase the MWEs with a single word. For examples, Red letter means marks.

3. **Substitutability:** Most of the time**,** MWEs opposed substitution of a component of word by a similar word. For example. 'Many thanks' cannot be replaced by "several thanks" or "much gratitude".

4. **Compositionality:** The level to which the features of the parts of MWEs combine to guess the features of the whole.

5. **Non-compositionality**: Non-Compositionality of the MWE into its constituents is one of the important characteristics of MWEs. For example, "kick the bucket", "spill the beans".

6. **Syntactic Fixedness:** It resists any further insertion into the MWEs. For example, Traffic signal, New York, New Delhi etc.

7. **Idiomaticity:** Idiomaticity is the fixation of the component lexemes, at a lexical, syntactic, semantic, pragmatic, and statistical level. For example, by and large.

3.13 Types of Multiword Expressions

Multiword Expressions have different types, depending on the context in which Multiword Expressions are used in the language. Various types of MWEs explained by (Baldwin *et al*., 2010) which are corresponding to the English MWEs and are as follows:

a. Nominal MWEs

Nominal MWEs are one of the most common MWE types, in terms of occurrence of token, types and their occurrence in the world's languages (Tanaka and Baldwin 2003).

b. Verb-Particle Constructions

Verb-particle constructions (i.e. VPCs) are consisting of verb and obligatory particle (s) such as hand over and take up (Bolinger, 1976b; Jackendoff 1997; Huddleston and Pullum 2002; Sag *et al.* 2002). The obligatory particles are usually intransitive prepositions, adjectives or verbs, as shown below:

verb + intransitive prepositions: *break into*, *take to*

verb + adjectives: *cut off*, *band together*

verb + verbs: *let eat*, *let run*

All these define construction of VPCs. Generally, VPCs are both idiosyncratic and semi-idiosyncratic combinations although some are adverbial and/or non-lexical particle cases (Dehe *et al.,* 2002). VPCs often involve subtle interactions between the verb and particle (Bolinger 1976b; Jackendoff, 1973; Fraser, 1976;Dehe, 2002). For example, the particle can impact on various properties of the verb, including, aspect (e.g., come vs. come out), reciprocity (e.g., ring vs. ring back) and repetition (e.g., work vs. work out).

Different researchers define VPCs in different ways. VPCs are termed phrasal verbs by some researchers (Bolinger, 1976b; Side, 1990; McCarthy *et al.,* 2003) and

verb-particle constructions by others (Dehe et al., 2002; Bannard et al., 2003; Kim and Baldwin, 2007a).

In our thesis, we will refer to them exclusively as VPCs. One MWE type which relates closely to VPCs is prepositional verbs (Jackendoff, 1973; Baldwin, 2005b), which are similarly made up of a verb and selected preposition, but the preposition is transitive and selected by the verb (e.g., refer to, look into).

It is possible to distinguish transitive VPCs from prepositional verbs via their respective linguistic properties which are explained by (Bolinger, 1976b;Baldwin, 2005b):

i. When the object NP is not functioning as pronoun, transitive VPCs can occur in either the joined or tear word order, while prepositional verbs must always occur in the joined form.
ii. When the object NP is pronoun, transitive VPCs must occur in the tear word order while prepositional verbs must occur in the joined form.
iii. Mode of adverbs cannot occur between the verb and particle in VPCs, while they can occur with prepositional verbs. In this thesis, we will focus exclusively on VPCs where the particle is prepositional.

VPCs undergo morphological, syntactic and semantic variation. Morphologically, VPCs inflect for tense and number (e.g., come/comes /come/has come/is come/... off). Syntactically, VPCs undergo word order variation, and are internally modifiable by a small set of adverbs (e.g., without, ultimately, with open arms).

c. Light Verb constructions

Light-verb constructions (i.e., LVCs) are consist of a verb and a noun complement, usually in the indefinite singular form (Jespersen, 1965; Abeill's, 1988; Stevenson *et al.,* 2004).

The name of the construction based on the verb which is semantically whiten or 'light' in the sense that their part to the meaning of the LVC is comparatively small in relative to that of the noun complement. Our definition of *light-verb constructions* is in line with that of Huddleston and Pullum (2002). The principal light verbs are *do, give, have, make, put* and *take*, for each of which we provide a selection

of LVCs in (2.10)–(2.15). English LVCs generally take the form verb+*a/an*+object, although there is some variation here.

(2.10) do: *do a demo*

(2.11) give: *give a lecture*

(2.12) have: *have a rest*

(2.13) make: *make an offer, make a ring*

(2.14) put: *put the blame (on), put an end (to), put stop (to)*

(2.15) take: *take a walk, take a photograph (of)*

Morphologically, LVCs inflect but the noun complement tends to have fixed number and a inclination for determiner type (Stevenson et al. 2004). For example, put an end (to) undergoes full verbal inflection (put/puts/putting an end (to)), but the noun complement cannot be pluralized or modified derivationally.

d. Verb–Noun Idiom (VN Idioms)

A Verb–Noun idiom is an MWE whose meaning is fully or partially unpredictable from the meanings of its components (e.g., 'kick the bucket', 'broken fortune') (Nunberg *et al.* 1994; Potter *et al.* 2000). Huddleston and Pullum (2002) identified subtypes of idioms such as verbal idioms (e.g., pass away, get over, run away) and prepositional idioms (e.g. in person, under the weather) which we classify as VPCs/prepositional verbs and determinerless PPs, respectively. In our terms, therefore, idioms are those non compositional MWEs not included in the named construction types of VPCs, prepositional verbs, noun compounds and determinerless PPs.

All idioms are non-compositional (to varying degrees), we further categorize them into two groups: decomposable and non-decomposable (Nunberg *et al.* 1994). In decomposable idioms, given the explanation of the idiom, it is possible to correlate components of the idiom with separate elements of the idiom explanation based on semantics which is not easy to get to from the components in isolation.

e. Determinerless-Prepositional Phrases

Determinerless prepositional phrases (i.e.,D-PPs) are MWEs, that are consist of a preposition and a singular noun without a determiner (Quirk *et al.,* 1985; Baldwin *et al.,* 2006).Combinations of prepositions and singular nouns shows many problematic characteristics of Multiword Expressions namely the syntax and semantics of the construction is often, but not always idiosyncratic, and at the same time the constructions are to some degree productive or allow modification (Baldwin et al., 2003).

In the work of Baldwin et al., (2003) it is seen that D-PPs do not form a homogeneous group. They proclaimed that in principle, each formation of a preposition and a singular noun without a determiner is a PP-D, but these P-N combinations vary with respect to their syntactic and semantic markedness. In addition, it may be either the noun or the preposition which selects for the lack of a determiner. We can say that at least one more different has to be made, namely whether or not the PP as a whole is dependent on a verbal or nominal head.

3.14 Classification of Multiword Expressions

To develop the lexicon of MWEs, it is customarily required to build up a classification MWEs which retain the properties of MWEs classes, but in that time allows for the training of information for a particular MWEs instances. Here, we have given the high level classification of MWEs as explained by (Bauer 1983; Sag et al., 2002) which is based on the syntactic and semantic properties of MWEs as shown in Fig.3.3

Classification of MWEs

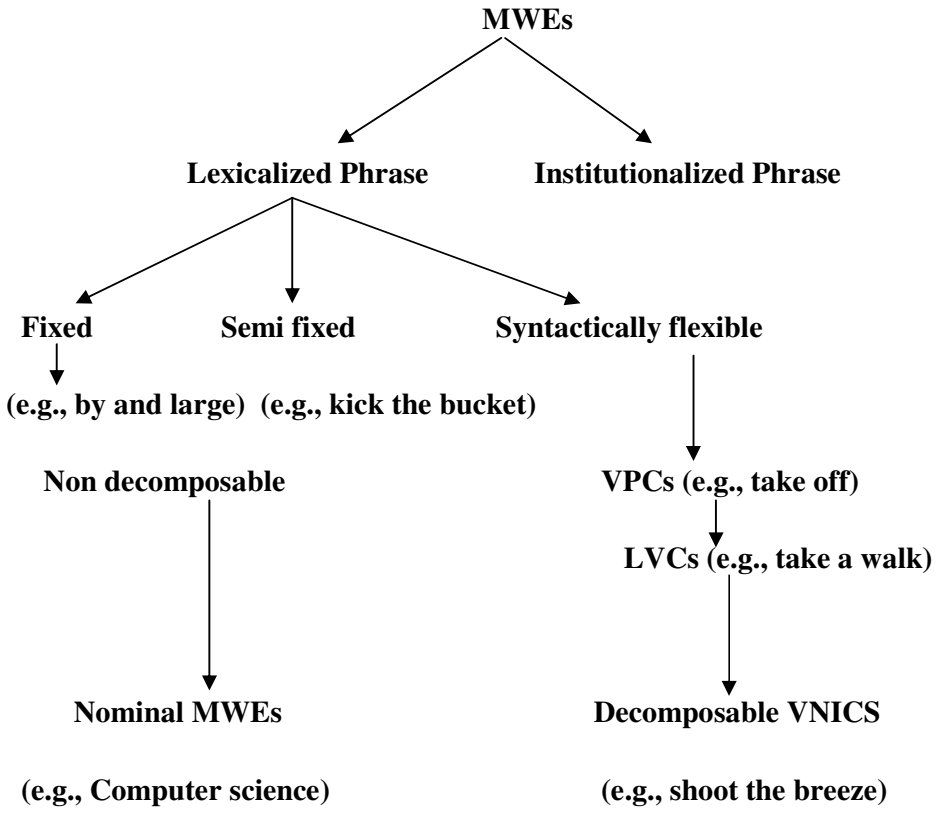

Figure 3.3 Classifications of MWEs (Sag *et al*., 2002)

The classification of MWEs into lexicalized phrases and institutionalized phrases based on either MWEs is lexicalized on the grounds of lexico-syntactic or semantic idiomaticity, or a uncomplicated collocation (i.e., only statistically idiosyncratic). In lexicalized phrases of MWEs, components have idiosyncratic syntax or semantics in part or in combination. Lexicalized phrases again divided into: fixed expressions (e.g., by train, at first, প্রতিশব্দ), semi fixed expressions (e.g., spill the beans, মোটর চালক) and syntactically-flexible expressions (e.g., add up, যুক্ত করা, persist in).

Fixed expressions consist of fixed strings that cannot be changed or modified morph syntactically. Example like, by and large cannot be modify morpho syntactically (e.g., by and larger) or cannot internally modifiable in reverse way(e.g., by and very larger). Non-modifiable determinerless prepositional phrases such as in order, of danger are also fixed expressions.

Semi-fixed expressions are lexically-variable MWEs that have strict restrictions on word order and composition, but undergo some degree of lexical variation such as inflection (e.g., kick/kicks/kicked/kicking the bucket vs. *the bucket was kicked), variation in reflexive pronouns (e.g.,in her/his/their cloths) and determiner selection (e.g., The Beatles vs. a Beatles album). Non-decomposable VNICs (e.g., kick the bucket, shoot the breeze) and nominal MWEs (e.g., Advocate general, part of speech) are also classified as semi-fixed expressions.

Syntactically flexible expressions are those MWEs, which undergo syntactic deviation, such as verb-particle constructions (VPCs), light-verb constructions (LVCs and decomposable idioms. The nature of the flexibility changes basing on construction types. Verb-particle constructions are syntactically flexible with respect to the word order of the particle and NP in transitive usages: hardness of hearing vs. hearing of hardness.

As described in Section 3.5 collocations (or institutionalized phrases) are MWEs that occur in a randomly, comparative to the component words or other phrasings of the same expression (i.e., they are strictly statistically idiosyncratic in nature), but some are unmarked. Examples include rank and file, salt and pepper, watch and word, 'many thanks' (ধন্যবাদ) and traffic light etc.

3.15 Chapter Summary

In this chapter, we present MWEs in detail viz. necessary and sufficient conditions for MWEs, linguistics properties of MWEs, types of MWEs, idiosyncratic properties of MWEs, characteristics of MWEs, types of MWEs and finally classification of MWEs explain with different examples. Based on these MWEs extraction and detection methods are applied in our experiments.