CHAPTER 2

REVIEW OF LITERATURE

In this chapter we have made literature survey of Multiword Expressions (MWEs) of both Indian and Foreign Languages along with Part of Speech tagging in Bengali and other Indian languages elaborately. Initially, we have endeavored to explain literature survey of Multiword Expressions and subsequently discussed Part of Speech tagging also.

The area of Automated Multiword Expression Detection using computational techniques has been upgraded from many years by contribution from several researchers. The study of Multiword Expressions is as old as linguistics but extraction of MWEs using Computer system is a recent phenomenon is an important research domain in the fields of computational system. The literature survey reveals the principled way to identify MWEs in different Indian languages as well as other foreign languages. Three types of MWEs namely, Noun-Noun (Compound noun), Noun-Verb (conjunct verb) and Verb(compound verb) sequences are examined Minia(2012). The focus is on the linguistic methods, like part of speech tagging, chunker and statistical methods like, point wise mutual information, log-likelihood for the extraction of MWEs.

The work on MWEs identification and extraction have been continuing in English for few years (Bannard *et al.,* 2003). Some of the MWEs extraction tasks in English have been cited in (N.Calzolari *et al.,* 2002). Indian languages (Hindi compound noun) MWE extraction has been studied in (Diab *et al.,* 2009.). Manipuri reduplicated MWE identification is discussed in (Kunchukuttan *et al.,* 2008) various statistical co-occurrence measurements like Pointwise Mutual Information (PMI), Log-Likelihood etc have been suggested for identification of MWEs. In Indian languages like Hindi,a considerable approach in compound noun MWE extraction (Nongmeikapam *et al.,* 2010) and a classification based approach for N-V collocation have been done, Sogou (2006). In Bengali, works on automated extraction of MWEs in are limited in number. There is no published work on Automated Multiword Expressions Detection in Bengali.

In literature Survey of MWEs, approaches for Multiword Expressions are:

1. **Statistical methods** are language independent and use collocation score measures like Pointwise Mutual Information or student's t test (Gelbukh *et al.,* 2005).

2. **Rules based approaches** (Dunning *et al.,* 1993): These approaches target some particular types of MWEs and extract them from the corpora taking clues from morpho syntactic properties. Such method depends heavily on POS taggers, chunkers, shallow or deep parser.

3. **Hybrid approaches** (Bannard *et al.,* 2003). These approaches recommended for MWEs extraction where pattern are identified at word level and POS tagging using statistical measures and joint properly to find out valid MWEs.

2.1 Different running projects on Multiword Expressions

To determine whether a given sequence of words are MWEs (e.g., ATM card vs the Big boss, ঝড়-বৃষ্টি, জীবন মরণ), it is required linguistics information for this work, while others required statistical methods of combining these words with some kinds of linguistic information such as syntactic and semantic properties or automatic word alignment techniques.

Statistical methods are frequently used for this task, because they can be independently applied to any language or different MWEs type. Though, there is no specific view about which measure is best appropriate for identifying and extracting MWEs in general. Since the theory of MWEs is recently developing and the significance of the problem is well accepted in the field of NLP, there is running work on MWEs in various projects that are increasing on a large scale in linguistically defined computational grammars Minia (2012), including

1. Par Gram Project at Xerox parc
(http://www.parc.xerox.com/istl/groups/nltt/pargram/)
2. The XTAG Project at the University of Pennsylvania
(http://www.cis.upenn.edu/˜xtag/)
3. The work on Combinatory Categorical Grammar at Edinburgh University
4. The LinGO Project (a multi-site collaboration including CSLI‟s English Resource Grammar Project — http://lingo.stanford.edu)

5. The FrameNet Project http://www.icsi.berkeley.edu/~framenet/), which is primarily developing large-scale lexical resources

These running projects are progressively doing research work in linguistically knowledgeable investigations of MWEs.

It is seen that the work of MWEs recognition is classified into two kinds: one for idiom types and other is idiom tokens. In idioms types, phrases that can be interpreted as idioms are found in text corpora, usually for lexicographers to compile idiom dictionaries. Previous studies have mostly focused on the idiom type identification (Baldwin *et al.,* 2003). However, there has been an increasing interest in idiom token identification recently Chakrabarti *et al*. (2006), Hoktoen and Eirik (1997). The idiom tokens identification is in an early stage for its development.

Previous approaches to MWEs identification were based on their collocational behavior (Church and Hanks, 1990). At first approaches were evaluated as Xtract (Smadja, 1993) in which, occurrences of word pairs with high frequency in a context of five words in a corpus are collected first, and these words are then ranked and filtered according to contextual considerations, including the part of speech in their neighbours.

Pecina (2008) made comparisons of 55 different association measures in ranking German Adj-N and PP Verb collocation candidates. He identified that combination of different collocation measures over a single collocation measure improves when standard statistical classification methods are used. In other survey (Chang *et al*., 2002; Villavicencio *et al.,* 2007) it is found that some collocation measures (including Pointwise mutual information and log likelihood) are better compared to others for identifying MWEs (Tsvetkov *et al.,* 2014).

Automatic MWE detection is a key factor in work on Collocational behavior. The idea behind this collocation approach is to compute the probability of the occurrence of a word pair in a combination to the probabilities of the individual words occurring independently (Church *et al.,* 1990).

To improve the quality of MWEs processing, existing linguistico-statistical approaches used part-of-speech taggers for handling certain categories of words, lemmatizers are used for recognizing all the inflected forms of a lexical item (Tsvetkov *et al.*, 2014).

Cook *et al.* (2007) focus on knowledge of general syntactic behavior of an idiomatic expression to decide whether the event of expression is used literally or idiomatically. They consider that in a good number of cases, idiomatic usages of an expression tend to occur in a few number of canonical forms for that idiom, in comparison with the literal usages of an expression which are less restricted syntactically, and can be expressed as a pattern diversity of inflected forms of the constituents.

Al-Haj and Wintner (2010) focused on morphological idiosyncrasies of Hebrew MWEs, and leverage such properties for automatic identification of a special construction, noun-noun compounds, for a given text. However, they did not considered for the semantics of the MWEs.

Lin (1999) performed his work on automatic identification of non-compositional phrases in an indirect way, via detecting non-productive phrases. The authors hypothesized that non-productive expressions are non-compositional. They also used distributional models, statistical measures and dependency triples. The statistical measures are based on the point wise mutual information of non-compositional phrases that differ significantly from the pointwise mutual information of phrases obtained by reducing each of their components with the 10 most similar words according to a corpus derived thesaurus (Lin 1998a); (Korkontzelos, 2010).

Ramisch (2015) on his MWEs studies gives the idea of Language phenomena into two classes which are Lexical level and syntactic level as found in formal languages. The lexical level considers words in a separate unit independent of their neighbour words. It deals with morphology, inflection (e.g., number, gender, verb tense), meaning of the words, word formation (prefixes, suffices). Lexical level checks validity of the word according to lexicon. Lexicon stands for dictionary. Syntactic level deals with formal relationships among words of sentences. Grammars are used to apply the rules that govern the position of words and phrases. Thus, Linguistic and computational approaches in grammar need to incorporate in MWEs representation in their approaches.

In corpus linguistics, MWEs play a vital role. According to Sinclair (1991), Language generation is supervised by two principles namely the open choice principle and the idiom principle. The open choice principle explains productivity, as speakers

can formalize many possible lexical units to express some information, while idiom principle constrains open choice by illustrating that these units are prefabricated. Thus, some constructions allow free variation while some cannot allow modification to some extent.

During 2000s, the Standford MWE project[1] has revived the importance of the NLP community in this area. One of the most well known publications of the MWE project is the famous "pain-in-neck" paper by Sag *et al.* (2002). It explained an general idea of MWEs characteristics and types and then represented some methods for dealing with them in the perspective of grammar engineering (Ramisch, 2015).

The MWE research community is organized to exchanges some common resources. Ideas on MWEs are shared in annual workshop on MWEs. It is a series of workshops that have been held since 2001 in combination with major computational linguistics conferences (Bond *et al*., 2003; Tanaka *et al.,* 2004; Rayson *et al.,* 2006; Gregoire *et al.,* 2007, 2008; Anastasiuo *et al.,* 2009; Laporte *et al.,* 2010; Kordoni *et al.,* 2011,2013,2014). The current research editions of the workshops reveal that there is a change of research on MWEs identification and extraction approaches that work towards applications oriented research.

The implementation of MWEs processing techniques and multilingual aspects are current issues in this field. In addition to the series of specialized workshops, main computational linguistics conferences such as COLING, ACL and LREC continuously feature papers on MWEs (Ramisch, 2015).

To enhance research on MWEs in computational linguistics various workshops and conferences were organized and special issues on MWEs have been published by renowned Journals in computational linguistics namely the Journal of Computer Speech and Language (Villavicencio *et al*., 2005), the Journal of Language Resources and Evaluation (Rayon *et al*., 2010), the Natural Language Engineering Journal (Szpakowicz *et al*., 2013) and the ACM transactions on speech and language processing (Ramisch *et al*., 2013).

[1]http://mwe.standford.edu/

2.2 Chapter Summary

In this chapter, we presented an elaborate literature survey on several issues related to multiword expressions in different languages along with Bengali MWEs. In particular, we started with reviewing methods and approaches for extraction and detection of MWEs and Multiword terms i.e. domain-specific MWEs. Methods for MWEs were classified as rule based method, statistical method and hybrid method respectively.