Chapter 10
Multiword Expressions in NLP: General Survey and a Special Case of Verb–Noun Constructions

Alexander Gelbukh
National Polytechnic Institute, Mexico

Olga Kolesnikova
National Polytechnic Institute, Mexico

ABSTRACT

This chapter presents a survey of contemporary NLP research on Multiword Expressions (MWEs). MWEs pose a huge problem to precise language processing due to their idiosyncratic nature and diversity of their semantic, lexical, and syntactical properties. The chapter begins by considering MWEs definitions, describes some MWEs classes, indicates problems MWEs generate in language applications and their possible solutions, presents methods of MWE encoding in dictionaries and their automatic detection in corpora. The chapter goes into more detail on a particular MWE class called Verb-Noun Constructions (VNCs). Due to their frequency in corpus and unique characteristics, VNCs present a research problem in their own right. Having outlined several approaches to VNC representation in lexicons, the chapter explains the formalism of Lexical Function as a possible VNC representation. Such representation may serve as a tool for VNCs automatic detection in a corpus. The latter is illustrated on Spanish material applying some supervised learning methods commonly used for NLP tasks.

1. INTRODUCTION

Many NLP applications deal with a natural language text as a "bag of words" where each string of letters between spaces, or a token, is viewed as an individual word associated with its proper semantics and syntactical functions. In *John suddenly kicked the ball*, the tokens *John*, *suddenly*, *kicked*, *the*, *ball*, are words contributing their own meaning to the overall meaning of the utterance. Thus, the semantics of the whole utterance is a composition or a "sum" of elementary meanings conveyed by each word in the utterance. However, the semantics of *John suddenly kicked the bucket* cannot be considered a "sum" of elementary meanings represented by individual words. Although

DOI: 10.4018/978-1-4666-6042-7.ch010

kicked the bucket consists of three words, this expression functions semantically as one word denoting the meaning *died*. Phrases like *kick the bucket*, *put on airs*, *bee in bonnet*, *up the creek* are called phraseological expressions, or idioms. Such linguistic phenomenon presents a challenge for natural language processing because they cannot be interpreted by a fully compositional analysis. The degree of cohesiveness between words in idioms is very high; the choice of such words is unmotivated and therefore cannot be predicted. For this reason, idioms are better viewed as a single word.

Sometimes, the degree of cohesiveness in expressions is not so strong as in idioms, and the semantics of a phrase is more transparent, though not completely clear, for example, for non-native speakers or language learners. *Mailing list*, *sign up*, *traffic light*, *kindle excitement* belong to such expressions. It has been a trend in NLP to group expressions with varying degree of cohesiveness under a single term "Multiword Expressions" (MWEs).

The main characteristic of MWEs is a lack of compositionality, though other features are put forward in alternative definitions of MWEs mentioned in Section 2. Section 3 presents MWEs classification, Section 4 discusses some problems posed by MWEs in NLP and their possible solutions. In Section 5, MWEs resources are described alongside with different encoding used to represent MWEs features in such resources. In Section 6, we consider statistic, rule-based and hybrid approaches to MWE automatic detection as well as evaluation methods used to estimate performance of MWE extraction techniques. Section 7 gives more detail on Verb-Noun Constructions (VNCs). Due to their frequency in corpus and unique characteristics, VNCs present a research problem in its own right. Having outlined several approaches to VNC representation in lexicons, Section 7 goes on explaining the formalism of Lexical Function as a possible VNC representation. Such representation may serve as a tool for VNC's automatic detection in a corpus. Section 8 illustrates the latter on Spanish material referring to results showed by some well-known supervised learning methods on the task of VNC automatic recognition while VNC was represented by means of the lexical function formalism. Section 9 concludes the chapter.

2. DEFINITIONS

Although MWEs are understood quite easily by intuition and their acquisition presents no difficulty to native speakers (though it is usually not the case for second language learners), it is hard to identify what features distinguish MWEs from free word combinations. Concerning this issue, such MWE properties are mentioned in literature: reduced syntactic and semantic transparency; reduced or lack of compositionality; more or less frozen or fixed status; possible violation of some otherwise general syntactic patterns or rules; a high degree of lexicalization (depending on pragmatic factors); a high degree of conventionality (Calzolari, Fillmore, Grishman, Ide, Lenci, MacLeod, & Zampolli, 2002).

No convention exists so far on the definition of MWEs but almost all formulations found in research papers emphasize the idiosyncratic nature of this linguistic phenomenon. Here are some definitions that are most frequently referred to in papers; we marked in boldface those concepts and properties that we think serve as the criteria for distinguishing MWE from compositional phrases:

- MWE are "recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages" (Smadja, 1993);
- MWE are "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag, Baldwin, Bond, Copestake, & Flickinger, 2002);