AN ASSESSMENT OF LITERATURE ON THE EXTRACTION OF MULTIWORD EXPRESSIONS: FOCUS ON PEDAGOGICAL IMPLICATIONS

Ochu, Michael Chima
University of Botswana
ochumichaelc@gmail.com

Abstract

Multiword expressions are difficult to deal with. The difficulty in handling them ranges from definition to classification. Researchers in the areas of Natural Language Processing and Computational Linguistics have neither provided a universal definition of what multiword expressions are, nor have they clearly classified multiword expressions. The focus of this paper is not to provide a generally acceptable definition, nor a clear classification of multiword expressions. Rather, the paper provides a pathway of approaching multiword expressions in linguistic research, especially through their extraction. Through extensive literature review, we argue that multiword expressions can be dealt with statistically or linguistically. This paper concludes that whereas some researchers have argued for the use of statistics-based methods, others have adhered to linguistic approaches. However, this study does not argue that these two approaches to MWE extraction have not been used concurrently. There are some studies that have used a hybrid approach. Therefore, there is no simple and best method in the extraction of MWEs in a corpus. A study uses a specific approach according to its overall objective. Whatever approach in handling multiword expression is adopted, the overriding objective must be add pedagogical value to extracted multiword expressions.

Keywords: multiword expression, extraction, statistics, linguistics, pedagogy

Introduction

According to Muller, Ohenheiser, Olsen, and Raiser (2011), terminologies such as chunks, multiword units, fixed expressions, and phrasemes are used to refer to multiword expressions (henceforth MWEs). MWEs are delineated from various perspectives including semantics, syntax, and statistics. For example, Zinsmeister and Heid (2003) define MWEs, from a semantic point-of-view, as a group of words whose meaning cannot be easily predicted from the surface form. This simply means that the overall meaning of the expression cannot be obtained from the individual meanings of the words that make up the group. This definition is linked to the feature of non-compositionality of MWEs. The idea of non-compositionality suggests that the meaning of the MWE is not derived from the meaning of the constituting elements.

In defining MWEs from a syntactic approach, Fontenelle (1994) avers that MWEs are idiosyncratic syntagmatic combination of lexical items and are independent of word class or syntactic structure. On the one hand, the phrase ‘syntagmatic combination’ used in the definition above suggests that the words that make up MWEs adhere to the grammatical rules of the language, and the rules border largely on agreement between person and number and between subject and verb. Words that make up MWE group may or may not adhere to rules of 1st Person, 2nd Person, and 3rd Persons; rules of singular and plural; and rules of concord or agreement between a subject and its verb. On the other hand, the use of the word ‘idiosyncratic’ highlights that as much as certain language rules need to be adhered to, MWEs can exhibit defiance or anomaly.

From a statistics point of view, Baldwin, Bannard, Tanaka, and Widdows (2003) define MWEs as sequences of words that tend to co-occur more frequently than chance and either are decomposable into multiple simplex words or are idiosyncratic words. The high frequency of co-occurrence of simplex words, or constituent words that make up an MWE, is what Kim (2008) refers to as word combinations with surprising frequency. Logically, these frequently occurring expressions in a language are useful elements for language learners, especially for beginners. Language teachers need to focus on such elements because there is something spectacular about the frequently co-occurring patterns. Similarly, Xu, Lu, and Li (2006) define MWEs as frequent word combinations in natural language. These last two definitions suggest that frequency of occurrence is fundamental in the definition of MWEs from a statistical point-of-view.

Defining MWEs form different characterizing perspectives has made it difficult for scholars to attain a consensus on the precise definition of MWEs. To attain precision in the definition of MWEs falls outside the scope of this study. Larsen-Freeman and Celce-Murcia (2016, p: 44) refer to MWEs as ‘notoriously difficult’. However, after considering a number of definitions of MWEs in literature, this study concludes that MWEs are frequently co-occurring word patterns that have specific meaning. Our definition of MWEs here exhibits three defining features which include ‘frequency of co-occurrence’ (statistics); ‘specific syntactic patterns’ (syntax); and specific meaning (semantics). These features interact among themselves in defining MWEs. It is not exhaustive, but this is the definition adopted in this study. A specific definition of MWEs is important given the various ways in which linguists have approached it. This unit of language merits investigation in language research.

Jackendoff (1997) argues that MWEs are as much as single words in an individual’s lexicon. In other words, MWEs are ubiquitous in all manner of language use including informal conversations, formal speeches, academic writing, prose fictions, and emails among others.

The remaining part of this paper is divided into six sections. The first section focuses on the data collection approach used in the study. The second and third sections discuss literature on both statistics and linguistic MWE-extraction methods. The fourth section presents a summary of the discussions presented in the second and third sections. In the fifth section, the study highlights the pedagogical implications of the extraction of MWEs. Finally, the sixth section presents the concluding comments. The section that follows presents the systematic way in which data was collected in this study.

Methodology

In literature, MWE extraction approaches are discussed from two main perspectives. These are statistical and linguistic approaches (Kim 2008; Moiron and Tidermann 2006). There is also the hybrid method, which is the combination of both statistical and linguistic approaches. However, we focus on two main approaches. The reason is that after a thorough inquiry of the hybrid approach, we observe that the hybrid approach is either a statistical or linguistic method. Therefore, a review of literature on hybrid extraction approach is simply a repetition of what falls under statistical or linguistic approach. The literature searches were conducted using such online databases as IEEE, Science Direct, Ebscohost, Emerald, SAGE Journals and Springer Link among others. The objective of the search was to retrieve relevant materials in the extraction of MWEs. These online databases and other materials are all available at the University of Botswana Library and University of Botswana (online) Inter-library Loan.

Abstracts were accessed to identify relevant articles. Then, full articles were downloaded. Results of the search yielded different relevant materials like articles, books, and conference proceedings addressing MWE extraction approaches.

Statistical Approaches to MWEs Extraction

Often, the first step in statistical extraction approach is to compute frequency counts of words, tokens, or N-grams. Also, statistical approaches for extracting MWEs consist of applications of various statistical techniques such as Mutual Information (MI), n-gram statistics package, log-likelihood ratio, hypothesis testing, standard loglinear model, among others. These statistical techniques can perform various functions. For example, MI measures how much information a variable tells about another variable or measures the association ratio of just two variables.

Church and Hanks (1990) applied MI to extraction word associates automatically. The underlying idea of this approach is to divide the joint probability of a word pair (x, y) by the probabilities of observing *x* and *y* independently from one another. Put differently, MI measures the strength of association between words *x* and *y*. In a given corpus, MI is calculated on the basis of the number of times an item is observed the pair together versus the number of times you saw the pair separately. On the basis of the results reported in Church and Hanks (1990), MI is reported by the authors to an effective MWE extraction tool. In Pecina (2005), there is a comparison of 84 kinds of association measures for bi-gram MWE extraction in a Czech data. Since the MI statistical measure works well for word pairs, Pecina (2005)’s study concluded that in Czech data, MI has an excellent extraction capacity. However, one limitation of MI is that it does not extract MWEs consisting of more than two words. Schutze (1993) presented vector space distribution model, an approach which, in the view of Henderson and Popa (2016), creates vector-space representations that capture many forms of words that hang around one another. One good thing about this tool is that it captures MWEs consisting of more than two words, and at the same time statistically accounts for the associative consistency of the identified MWEs. In the same study, Schutze (1993) identified highly associated word pairs based on statistical measures. Then, two filters were applied to word pairs: one to extend them to MWEs of arbitrary length, while the other filter scans for syntactic consistency of the MWEs and automatic ally rejects invalid candidates.

In another instance of statistical MWE extraction, Dias, Lopes, and Gullimore (1999) introduced Mutual Expectation (ME) which has become a popular tool for such MWE extraction exercise. Their study has two objectives. The first objective was to extract MWEs by using ME as a statistical approach, while the second objective was to compare ME with LocalMax Algorithm. According to Dias et al. (1999), ME is a language-independent and statistically-based system that extracts multiword lexical units using candidates' frequency.

Silva and Lopes (1999) proposed LocalMaxs algorithm to extract both contiguous and non-contiguous multiword lexical units from corpora. Contiguous MWEs are uninterrupted sequences of words. On the other hand, non-contiguous MWEs are fixed sequences interrupted by one or two gaps filled in by interchangeable words that usually are synonyms. Local-Max identifies MWEs from a list of N-gram based on two assumptions. As proponents of LocalMax, Silva and Lopes (1999) stipulates two assumptions. The first assumption is that association measures show that the more cohesive a group of words is, the higher are the values of the association measures, thus allowing for multiword identification. The second assumption is that MWEs are groups of words that are highly associated with each other. From these assumptions, an N-gram *W*, for example, is a multiword if its association measure, *g (W),* is the local maxima. The basic idea behind LocalMaxs is that the association score of an *N-gram* should be a local maximum in three sequences as *N-1-gram, N-gram* and *N + 1 gram* which have same head word.

In the second objective of their study, Dias, et al. (1999) compared ME with LocalMax algorithm introduced by Silva and Lopes (1999). The comparative results show that ME does not only give high precision and extraction rates, but also overcomes the problem of highly frequent words raised by LocalMax and tends to elect longer multiword units. Dias et al (1999), therefore, infer that ME is very suitable for extracting MWEs. One interesting thing about Dias et al. (1999) is the comparative approach their study took. This approach affords the study to weigh the merits and demerits that are associated with ME and LocalMax. As a result, the study was able come with an informed direction for further empirical activities. Furthermore, Zhang, Kordoni, Villavicencio, and Idiart (2006) argue for collocation optimization as an effective MWE extraction and propose that it determines the optimal length of an MWE based on association variation. Collocation optimization is an automatic search, detection, and extraction of semantically similar groups of words (Zhang et al 2006). The basic assumption for collocation optimization is that the association value will increase if a correct individual word is included in the MWE candidate, but the association value will decrease when an incorrect individual word is included in the MWE candidate. The extraction method, according to the authors, should be intensified when word extension from head word is moving within the span of an MWE. Otherwise, when word extension goes beyond the span of an MWE, its association decreases.

Smadja (1993) uses Xtract, which is able to retrieve a wide range of fixed patterns from corpora. Fixed patterns here refers to MWEs since fixedness is a fundamental characteristic of MWEs. In this method, word pair such as ***w1*** and ***w2*** could be considered as an MWE if only the two words are repeatedly used together within a single syntactic construct. This is a qualifying reason why they have a marked pattern of co-appearance. In other words, Xtract is an extraction technique that uses statistical information to extract fixed word pairs. Xtract in Smadja (1993) is comparable to MI used Church and Hanks (1989). Both extraction approaches deal with association between pairs of words. The word pairs are not just mere or random collocates, but are statistically determined to be pairs.

However, Xtract combines parsing and statistical techniques in labelling and filtering retrieved MWEs.
In a study of MWE extraction in a Chinese corpus, Zhang, Yoshida, and Tang (2008) used variance and a new statistical method called Augmented Mutual Information (AMI). AMI method was used in filtering criterion and variance as a secondary measure. This means that if a multiword candidate has a high AMI and low variance in its candidate set concurrently, the candidates are regarded as MWEs. Otherwise, it will not be regarded as an MWE. The conclusion of Zhang et al (2008) is that AMI has an approximate capability in the extraction of MWEs.

Choueka, Klein, and Neuwitz (1983) recommend the use of co-occurrence properties in extracting MWE. They assume that two or more words occur together with markedly high frequency if they form an MWE. In the current review, the above assumption is relevant because MWEs are about co-occurrence. The co-occurrence statistical approach forms the basis of a plethora of association measures and has been used by Pecina (2005) for the extraction of MWEs. Pecina (2005) opines that geometric co-occurrence properties have been found to be effective for extracting statistically-marked MWEs such as *shock and awe* as their co-occurrence tends to have abnormally high frequency relative to the alternative ordering.

A different dimension was adopted in Schone and Jurafsky (2001) wherein they evaluated a variety of MWE extraction approaches. The evaluated approaches include frequency, Pointwise Mutual Information (PMI), selectional association, symmetric conditional probability, Dice formula, log-likelihood ratios, Pearson’s chi-square, z-score and t-test. They showed that information-like approaches, particularly z-score, symmetric conditional probability, and chi-square perform better than the others but in general results. They also proposed two new approaches namely: joint probabilities and likelihood ratios.
Additionally, both Lin (1999) and Cruys, Paul, and Vitanyi (2007) used the principle of substitution to extract institutionalized MWEs. Both studies measure the differences between the distributional characteristics of MWEs and other similar fixed expressions obtained by substitution. For instance, *traffic signal* could have *traffic sign* and *traffic light* as similar collocations. If one of these collocations is highly preferred or occurs more frequently as compared to others, then it is likely to be an institutionalized MWE. The substitution tests measure this bias in preference for a collocation.

Whereas Lin (1999) uses PMI as the base association score, Cruys et.al. (2007) use a strength of association measure motivated by the idea of selectional preference of a constituent word for another.
McCarthy, Keller, and Carroll, (2003) use the semantic similarity measurement between phrasal verbs and their component words. They exploit contextual features and frequency information in order to assess meaning overlap. Their study established that human compositionality judgements correlate well with those measures that consider the semantics of the particle. The last statistical method highlighted in this discussion is the saturated loglinear model used in Blaheta and Johnson (2001). The model measures the strength between a word and other words that hand around it. Good results were reported in Blaheta and Johnson (2001) wherein the model was used to extract MWEs in English. One feature of the loglinear model is its control for the low frequency bias. And contrary to MI, this model is effective in extracting trigrams. It is important to highlight here that whereas some statistical MWE extraction measures are effective in handling words pairs, others are excellent in extracting strings beyond pairs. Next, we critically examine linguistic approaches to extracting MWEs.

Linguistic Approaches to MWEs

There are several linguistic approaches to the extraction of MWEs. Whereas some approaches have been repeatedly used in linguistic research, others have vanished into oblivion on the premise of inefficiency. Attia, Tounsi, Pecina, van Genabith, and Toral, 2010) propose two complementary approaches to extracting MWEs from available data resources in Arabic. The two approaches are Cross-lingual Correspondence Asymmetries (CCA) and Translation-based Approach. The first approach relies on the correspondence asymmetries between Arabic Wikipedia titles and titles in 21 different languages. The second approach collects English MWEs from Princeton WordNet

(henceforth PWN), translates the collection into Arabic using Google Translate mechanism, and utilizes different search engines to validate the output. PWN is a large lexical database of English nouns, verbs, adjectives, and adverbs grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. The database is provided by Princeton University.
The second approach used by Attia et al. (2010) is translation-based. The approach is bilingual and complements the cross lingual correspondence approach by focusing on compositional compound nouns. This method is partly like that of Vintar and Fisher (2008) who automatically extended the Slovene WordNet with nominal MWEs by translating the MWEs in PWN using a technique based on word alignment and lexico-syntactic patterns. However, Attia et al. (2010) departed from Vintar and Fisher (2008) in that instead of using a parallel corpus to find the translation, they used Google Translate. In simple terms, Attia et al. (2010) use cross lingual correspondence asymmetries and translation-based extraction, which are both linguistic approaches.

Similarly, Venkatapathy, Agrawal, and Joshi (2005) used Max Entropy classifier in extracting Hindi MWEs. They used the linguistic properties of verbs, semantic type of the object, case marker with the object, and similarity of the verb form of the object with the verb-object pair under consideration as features in a Max Entropy classifier. The Max Entropy classifier, developed by Suárez and Palomar (2002), is a discriminative classifier commonly used in Natural Language Processing, and Speech and Information Retrieval. One advantage of the maximum entropy framework is the ability to incorporate linguistic features or clues of MWEs. This is particularly effective when the corpus is tagged. In another approach, Moiron and Tiedemann (2006) used translation ambiguity to extract non-compositional MWEs. They posit that the non-compositional MWEs will have more translation candidates on account of more uncertainty in translation. This uncertainty is measured as translational entropy.

A preliminary extraction of MWEs in Dziob and Wendelburger (2016) was carried out using the set of MWeXtractor tools on the 70-million-word IPI-PAN and the plWordNet corpora of the Wrocław University of Technology, Poland. The result of the study yielded 607 MWEs. The authors, therefore, concluded that MWeXtractor is an effective tool in the extraction of Polish MWEs. Further inquiry reveals that in many respects, MWeXtractor and Sketch Engine do not differ greatly from each other. Sketch Engine is a tool which allows for the extraction of MWEs on the basis of their grammatical relations (Kilgarriff, Rychlý, Smrz, and Tugwell, 2004). The difference, however, is that MWeXtractor has elements of statistical measures such as word order exponential correlation and word frequency order. These statistical measures in MWeXtractor can be activated to statistical functions, otherwise the researcher limits him or herself to its linguistic tools. This is similar to Mike Scott's Wordsmith Tools which has both statistical and linguistic elements in extracting MWEs. Either element can be activated in a particular time for a specific purpose. According to Dziob and Wendelburger (2016), the modifications in MWeXtractor were imperative for good syntactic description and extraction of multiword candidates.

Fazly and Stevenson (2006) investigate the idiomaticity of verb phrases by combining the degree of syntactic fixedness with the variability that MWEs exhibit with respect to the selection of their lexical components. They use lexical and syntactic fixedness as partial indicators of non-compositionality. The method they used is the same as Lin's (1998) automatically generated thesaurus. Automatically generated thesaurus is used to compute a metric of lexical fixedness. Lexical fixedness is a measurement of the deviation between the PMI of a verb-object phrase and the average PMI of the expressions resulting from substituting the noun by its synonyms in the original phrase. This measure is like Lin's (1999) proposal for finding non-compositional phrases.

The assumption postulated in Fazly and Stevenson's (2006) is that non-compositional expressions score high in idiomaticity, that is, a score resulting from the combination of lexico-syntactic fixedness and semantic non-compositionality. Sometimes, syntactic flexibility forms part of the idiomaticity score. A syntactic flexibility score measures the probability of seeing a candidate in a set of pre-selected syntactic patterns. Their study report 80% accuracy in distinguishing literal from idiomatic expressions in a test set of 200 expressions. In their view, the performance of both metrics, idiomticity

score and syntactic flexibility, is stable across all frequency ranges. Although Fazly and Stevenson's (2006) study encroaches on statistical approach, but their study has a largely linguistic objective.
In addition, Hashimoto and Kawahara (2008) use token classification for the extraction of Japanese MWEs of all types. They apply a supervised learning framework using support vector machines based on TinySVM with a quadratic kernel. After annotating a web-based corpus for training data, they identify 101 MWE types with estimated corresponding 1000 examples. They use two types of features: word sense disambiguation (WSD) features and idiom features. On the one hand, the WSD features comprise some basic syntactic features such as Part of Speech (POS), lemma information, token n-gram features, in addition to hypernymy information on words as well as domain information. On the other hand, the idiom features were mostly characteristics that show voice, negativity, modality, in addition to adjacency and adnominal features. Hashimoto and Kawahara (2008) report results in terms of accuracy and rate of error reduction. Their overall accuracy is of 89.25% using all the features.

Another way to use knowledge on syntactic variation to extract MWEs is presented in Bannard (2009). He distinguishes three different types of syntactic variation, namely: the addition, dropping or determiner variation, and modification of the noun. Noun modification could take place by introducing an adjective, and by a passivation of the verb. The variation is determined by counting the presence or absence voice, negativity, modality, adjacency and adnominal features. If the phrase exhibits less flexibility than expected based on the flexibility of its parts, it is assumed to be a valid MWE.

Lexical choice process was used by Pearce (2001) to describe a method to extract collocations from corpora by measuring semantic compositionality. Semantic compositionality is the principle that the meaning of an expression is only a function of the meanings of its parts together with the method by which those parts are combined (Pelletier 1994). This definition of what semantic compositionality is, remains vague or underspecified because several points such as what counts as a part and as a meaning, and what kind of function is allowed are not clearly addressed. Nevertheless, Moon (1998) had subscribed to the use of semantic compositionality in handling MWEs. The underlying assumption is that a fully compositional expression allows synonym replacement of its component words, whereas a MWE does not. Using lexical choice process, Pearce (2001) measures the degree to which a collocation candidate allows synonym replacement. The measurement is also used to rank MWE candidates relative to their compositionality.

Furthermore, Xhai (1997) appealed to semantic compositionality in extracting lexicalized noun phrases as type of MWE. For example, Xhai (1997) opines that *white house* is a lexical atom because the phrase conceals its reference to the mansion where the President of the United States lives. Yet, semanticists would argue that *stock market* is lexicalized, too, because it is a noun-noun compound that refers to a persistent concept, not just in American culture, but in the field of finance. Thus, a semantic feature for recognizing lexicalized noun phrases that distinguish between *white house* and *stock market* performs an interesting task.

Within the linguistic approach of extracting MWEs, Pazienza, Pennacchiotti, and Zanzotto (2006) recommend the application of stop-list in a corpus. Stop-list is a process discarding unwanted words or terms that contain one of any undesirable linguistic elements. Scott (2015) WordSmith Tool has the stop-list feature which is designed to extract functional words, that is, words that are of very common usage in the language. Functional words are highly frequent elements in corpora. In the case of English, some functional words include: the, that, of, in, etc. They are largely drawn from prepositions, conjunctions, pronouns, articles, and demonstratives. By implication, the elimination of functional words simply suggests that frequent generic collocations are discarded. Content words, as opposed to functional words, are retained and validated by human experts. This approach is useful in extracting noun-noun MWEs such as in the case of Kim (2008). However, it will be difficult to extract MWEs in the form of Verb Particle Constructions (VPCs), which are made up of a verb and a preposition or an adverb. This is especially true in the case of a language like English that relies a great deal on VPCs.

Levin (1993) created semantic classification tool to extract MWEs based on their semantic properties. Semantic classification of verbs was used in Villavicencio (2005) to extract VPCs. The study identifies 3156 distinct VPCs across three electronic dictionaries and extends that total to 9745 through automatic extraction from British National Corpus. An extraction of VPCs is, in the view of this discussion, a part of MWE project. Finally, Taljard and de Schryver (2002) report that the initial stage of the investigation of what constitutes MWEs in Northern Sotho linguistics consisted of a manual excerption of terms from the linguistic texts. Manual excerption implies scrutiny of a text in order to identify terms which are relevant to a specific subject field; in this case, linguistics. This manual reading and marking were performed by professional terminologist, and the extracted MWEs were entered preliminary MWE list. To sum up, the extraction of MWEs within a linguistic approach should be able to parse the corpus or to identify at least PoS; to apply semantic features on the corpus; to implement other linguistic filters on the corpus, among others.

Summary of Review

Mutual Information (MI), Mutual Expectation (ME), and Local Maxs algorithm have been highlighted in literature to be effective statistical methods in the extraction of MWEs. In the first two, MWEs can only be extracted in pairs, whereas the last method can extract three multiword candidates. Like MI, Xtract used in Smadja (1993) is effective in extracting word pairs that make up MWEs. However, Xtract combines with parsing in labelling extracted MWEs. Augmented Mutual Information (AMI) is reported to have approximate capability in the extraction process. In addition, Schone and Jurafsky (2001) posit that z-score, symmetric conditional probability, and chi-square. Also, principle of substitution was recommended by both Lin (1999) and Cruys et al. (2007).

Some of the linguistic approaches articulated in this discussion include Cross-lingual Correspondence Asymmetries (CCA) and Translation-based Approach which are recommended by Attia et al. (2010). There also Max Enthropy classifier that has been used not only in Natural Language Processing, but also in Speech and Information Retrieval. Both syntactic and semantic, fixedness and compositionality, were used in Fazly and Sterenson (2006) who reported 80% of the two methods in the extraction of MWEs. Finally, stop-list was highlighted as useful extraction approach. In all of the foregoing, literature has not significantly directed attention to how teaching and learning can contribute to the growth and development of MWE. The section that follows highlights the pedagogical implications of extracting MWEs.

Pedagogical Implications to the Extraction of MWEs

As highlighted in the introductory section of this study, MWEs are as much as single words in an individual's mental lexicon. Siyanova-Chanturia (2017) maintains that MWEs in a individual's word base number in the hundreds of thousands. Therefore, given the magnitude of MWEs in the human lexicon, it is important that MWEs are taught and learnt just as other aspects of the language are taught and learnt. Concerning the use of MWEs in L2 scenario, Hinkel (2019) observes that there is low proficiency level which is occasioned by less frequent use of MWEs in both speech and writing. Put differently, L2 spoken and written productions contain few MWEs. In a related observation, Siyanova-Chanturia (2017) observes that L2 learners' use of MWEs is characterized by errors and the MWEs are typically incongruous. Peters (201) had argued that MWEs are inappropriately and infrequently employed. Basically, low frequencies and inaccuracy are well-established conclusions of the use of MWEs in L2 scenario. The implication therefore is that as more and more teachers are poised to teach MWEs, more and linguists will be bent on extracting MWEs for teaching and learning. Overall, learners' language proficiency will be enhanced.

In the light of the foregoing, it is imperative for teachers to collaborate with computational linguists in identifying MWEs for learning purposes. This can be done by brining learners' attention to the key lexical item in a MWE, and then highlighting other accompanying elements. In addition, frequent MWEs must be noted and taught explicitly. Teaching involves brining learners'' attention to their uses, forms, and functions. The assumption is that this will be productive as language users have MWEs in their mental lexicon.

Concluding Comments

Thus far, we have succinctly described existing literature on the statistical and linguistic dimensions of the extraction of MWEs. In statistical methods, there are novel developments for multiword extraction exercises. Empirical evidences show that statistical methods include evolving association measures that rank association strengths of multiword candidates and new strategies to align the best linguistic elements as MWEs. There is a problem in extracting MWEs using statistical measures. This often involves deciding the cut-off boundary threshold. For example, one of the main problems facing statistical approaches is that it is difficult to deal with low-frequency of words that form MWEs in a corpus. Nevertheless, one beauty of the use of statistical measures in the extraction of MWEs is that it gives linguistic research an inter-disciplinary value. In essence, it expands the frontiers of linguistic research. On the contrary, the linguistic approach to the extraction of MWEs involves corpus annotation and PoS tagging.

Linguistic properties are important in the extraction of MWEs. The discussion on the use of linguistic methods in extracting MWEs shows that it is a promising approach to the extraction task. Therefore, we recommend future research will proceed in this direction.

This paper concludes that whereas some researchers have argued for the use of statistics-based methods, others have adhered to linguistic approaches. However, this study does not argue that these two approaches to MWE extraction have not been used concurrently. There are some studies that have used what Kim and Baldwin (2010) refer to as a hybrid approach. There is no simple and best method in the extraction of MWEs in a corpus. A study uses a specific approach according to its overall objective. Whatever extraction approach is adopted, it is imperative that extracted MWEs are used in language teaching and learning.

References

Attia, M., Tounsi, L., Pecina, P., van Genabith, J., Toral, A., (2010). Automatic extraction of Arabic multiword expressions. *COLING 2010 Workshop on Multiword Expressions: from Theory to Applications. Beijing,* China.

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: Analysis, acquisition and treatment,* 89–96, Sapporo, Japan.

Bannard, C. (2009). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions,* pp. 1–8, Prague.

Blaheta, D., and Johnson, M. (2001). Unsupervised learning of multiword verbs. In *39th Annual Meeting and 10th Conference of the European chapter of the Association for Computational Linguistics (ACL 39),* pp. 54-60, Toulousse, France.

Choueka, Y., Klein, S., and Neuwitz, E. (1983). Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal for Literary and Linguistic,* 4.34–38.

Church, K. and Hanks, (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Cruys, R., Paul, M., Vitanyi, B. (2007). The Google word similarity distance. *IEEE Transactional Knowledge and Data Engineering,* 19(3):370–383.

Dias, G., Lopes, G., and Guillore, S. (1999). Normalising the IJS-ELAN Slovene-English parallel corpus for the extraction of multilingual terminology. *In Proceedings of the CLIN '99 Computational Linguistics,* pp. 84 91, Netherlands.

Dziob, A., and Wendelburger, M. (2016). Extraction and Description of multiword lexical units in plWordNet 3.0. *Proceedings of the 8th Global WordNet Conference,* pp 87-92, Bucharest Romania.

Fazly, A., Cook. P., and Stevenson, S. (2006). Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the ACL 06 Workshop on a Broader Perspective on Multiword Expressions*, pages 41–48.

Fontenelle, T. (1994). What on earth are collocations: an assessment of the ways in which certain words co-occur and others do not. *English Today*, 10(4), 42–48.

Hashimoto, C., and Kawahara, D. (2008). Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proceedings EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing,* Pp. 992-1001. Honolulu, Hawaii.

Hinkel, A. (2019). Pedagogical approaches to teaching and learning multiword expressions. TESOL (1), pp 1-18.

Henderson, J., and Popa, D. (2016). A Vector Space for Distributional Semantics for Entailment. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp 2052–2062, Berlin, Germany.

Kilgarriff, A., Rychlý, P., Smrz, P., and Tugwell, D. (2004). The Sketch Engine. *Proceedings of EURALEX*, Lorient, France.

Kim, S. N. (2008). Statistical modelling of multiword expressions. Unpublished PhD Thesis. Department of Computer Science and Software Engineering, University of Melbourne.

Levin, B. (1993). *English Verb Class and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Lin, D. (1998). Automatic retrieval and clustering of similar words. *Proceedings COLING '98 Proceedings of the 17th international conference on Computational linguistics,* Vol. 2 pp. 768-774. Montreal, Quebec, Canada.

Lin, D. (1999). Automatic identification of non-compositional phrases. *In Proceedings of the 37th Association of Computational Linguistics* (ACL-1999), 317–324, College Park, Maryland, USA.

McCarthy, D., Keller, B., and Carroll, B. (2003). Detecting a continuum of compositionality in phrasal verbs. *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment.*

Moiron, B. and Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word alignment. *In Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context,* Trento, Italy.

Moon, R. (1998). *Fixed expressions and idioms in English*. A Corpus-based Approach. Oxford: Clarendon Press.

Müller, P., Ohnheiser, I., Olsen, S., and Rainer, F. (2011). Word formation. *An International Handbook of the Languages of Europe* [HSK series]. Berlin: De Gruyter

Pazienza, M., Pennacchiotti, M., and Zanzotto, F. (2006). Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. *Studies in Fuzziness and Soft Computing* 10.1007/3-540-32394-5_20.

Pearce, D. (2001). Synonymy in collocation extraction. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations,* Pittsburgh, USA, pp. 41–46.

Pecina, P. (2005). An extensive empirical study of collocation extraction methods. *In Proceedings of the ACL Student Research Workshop,* pp. 13–18, Ann Arbour, MI.

Pelletier, F. (1994). The principle of semantic compositionality. *Topoi*, 13:11–24.

Peters, E. (2016). The learning burden of collocations: The role of inter-lexical and intra-lexical factors. Language Teaching Research, 20, 113-138.

Schone, P. and Jurafsky, D. (2001). Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem? *Proceedings of Empirical Methods in Natural Language Processing*, pp. 100-108.

Schutze, H. (1993). Automatic word sense discrimination. *Computational Linguistics*, Vol. 24, No.1, pp 97-123.

Scott, M. (2015). *Updated Lexical Analysis Software*. Oxford: Oxford University Press.

Silva, J. and Lopes, G. (1999). A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. *In Proceedings of the Sixth Meeting on Mathematics of Language* (MOL6), pp. 369–381, Orlando, FL, USA.

Siyanova-Chanturia, A. (2017). Researching the teaching and learning of multiword expressions. Language Teaching Research Vol. 21(3) pp 289-297.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19 (1), 143–77.

Suarez, A. and Palomar, M. (2002). A maximum entropy-based word sense disambiguation system. In Chen, H.-H., & Lin, C.-Y. (Eds.), *Proceedings of the 19th International Conference on Computational Linguistics* (COLING´2002), pp. 960–966.

Taljard, E. and De Schryver, G-M. (2002). Semi-automatic term extraction for the African languages, with special reference to Northern Sotho. *Lexikos*, 12 (12): 44–74.

Venkatapathy, S., Agrawal, P. and Joshi, A.K. (2005). Relative compositionality of noun + verb multi-word expressions in Hindi. *In Proceedings of the International Conference on Natural Language* (ICON 2005).

Villavicencio, A. (2005). The availability of verb–particle constructions in lexical resources: How much is enough? *Journal of Computer Speech and Language Processing,* 19, 415–432.

Vintar, S. and Fiser, D. (2008). Harvesting multi-word expressions from parallel corpora. *In Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008), pp.1091–1096, Marrakech, Morocco.

Xhai, C. (1997). Exploiting context to identify lexical atoms: a statistical view of linguistic context. In *Proceedings of the International and Interdisciplinary Conference on Modelling and Using Context (CONTEXT-97),* pp. 119–129.

Xu, R., Q. Lu, and S. Li (2006). The design and construction of a Chinese collocation bank. *In Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006), Genoa, Italy.

Zhang, Wen & Yoshida, Taketoshi & Ho, Tu & Tang, Xijin. (2008). Augmented Mutual Information for Multi-word Extraction. *International Journal of Innovative Computing, Information and Control* Vol. 5, No 2, pp. 543-554.

Zhang, Y., Kordoni, V., Villavicencio, A., and Idiart, M. (2009). Automated Multiword Expression Prediction for Grammar Engineering. *In Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties,* pp. 36–44, Sydney, Australia, July. Association for Computational Linguistics.

Zinsmeister H. and Heid. U. (2004). Collocations of complex nouns: Evidence for lexicalization. In *Proceedings of KONVENS*-2004, Heidelberg. Springer.