IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 5, Ver. V (Sep. - Oct. 2016), PP 64-69
www.iosrjournals.org

Combining Arabic Nested Noun Compound and Collocation Extraction Using Linguistic and Statistical Approach

Maryam Yaseen Al-Mashhadani[1], Luma Adnan Al-Sagban[2]
[1, 2]*Department of Computer Science, College of Education for Women/ University of Baghdad, Iraq*

Abstract : *Arabic multi-word expressions are the combinations of two or more terms that associated with each other as one concept. The process of extracting such expressions is challenging especially when the length of the combination is getting longer. Recently, researchers attempted to extract nested noun compounds which consists of two or more combinations of nouns. Such extraction process requires comprehensive analysis using linguistic and statistical approaches. However, the process of extraction in the state of the art have extended to include 4-gra and 5-gram candidates. This paper aims to combine the extraction of nested noun compound and collocation in order to extend the process of extraction to include 6-gram and 7-gram. For this manner, a linguistic approach comprises of various kinds of pattern has been used, as well as, three statistical measures have been utilized including NC-value, LLR and PMI. Results shown that the proposed method has the ability to extend the extraction to include longer candidates.*

Keywords: *Multi-Word Expressions, Nested Noun Compound, Collocation, Linguistic Patterns, Statistical measures*

I. Introduction

Multi-word expressions are the combination of two or more words that organized with a blank space (e.g. prime minster), with no space (e.g. hyperlink) or with a punctuation (e.g. so-called) [1]. Arabic is one of the common languages that used widely around the world and contains tremendous kinds of multi-word expressions. One of the common types of multi-word expressions is the noun compounds which consists of two or more nouns such as "المدرسة الدولية / international school" [2]. Another type is the collocations which consists of a verb and other nouns or adjectives such as " صرح الرئيس الامريكي / US president announces" [3]. Such two types of multi-word expressions have been addressed widely by many researchers for the Arabic language [4-7].

Recently, a surge research has caught several attentions which is the automatic extraction of Arabic nested noun compounds [8, 9]. Nested noun compounds are the combination of multiple noun compounds which leads to long noun compound consisting of 4 or more noun compounds. For example, the noun compounds "نتنياهو رئيس الوزراء بنيامين / Israeli prime minister Benjamin Netanyahu" such compounds could be turned into two noun compounds of " رئيس الوزراء / prime minster" and " بنيامين نتنياهو / Benjamin Netanyahu". Apparently, extracting such nested noun compounds is a challenging task regarding to the comprehensive analysis that should be take a place in order to distinguish the sequence of multiple noun compounds. The problem is extended when treating nested collocation which consists of collocation that attached with nested noun compounds for example, " صرح رئيس الولايات المتحدة باراك أوباما / United states president Barak Obama announces". As stated in the example such combination of the collocation "صرح رئيس / president announces" with the nested noun compound " رئيس الولايات المتحدة باراك أوباما / president Barak Obama" contains 6-grams words. This requires an extension in the process of the extraction.

This paper aims to overcome this problem by proposing a hybrid method of linguistic approach and statistical approach. The linguistic contains a filtering approach that generates patterns using POS tagging, while the statistical approach consists of NC-value, Log-likelihood Ratio (LLR) and Point-wise Mutual Information (PMI). The paper is organized as Section 2 Related work, Section 3 Methodology, Section 4 Results and Section 5 Discussion.

II. Related Work

Several studies have been proposed for the automatic extraction of multi-word terms in Arabic for instance, Attia [10] proposed a linguistic approach based on the regular expression in which the multi-words are being processed lexically. Boulaknader et al. [11] proposed a linguistic and statistical approach for extracting the Arabic multi-word terms. The authors have utilized some linguistic patterns such as Noun + Noun and Noun + Adjective, then the statistical measures accommodated a ranking mechanism in order to attain the co-occurrence among the patterns. Bounhas & Slimani [12] developed a rule-based approach based on the linguistic patterns and statistical measure of LLR. Finally, Attia et al. [4] proposed a hybrid method of statistical and linguistic approach in order to identify multi-word terms in Arabic. Particularly, the authors have utilized a lexicon-based approach in order to identify the named entities which significantly contributes toward extracting the multi-word

terms. However, the above studies have focused on Arabic multi-word in general rather than specific types. Other researchers have concentrated on specific types such as Saif & Aziz [3] who have proposed a combination method for extracting Arabic collocations. Such method comprises of lemmatization approach and POS tagging for identifying the linguistic patterns. Consequentially, multiple statistical measures have been carried out in order to rank the candidates produced by the linguistic patterns. On other hand, El Mahdaouy et al. [6] have proposed a method for extracting Arabic noun compounds. The author utilized both linguistic approach (i.e. POS tagging) for identifying the linguistic patterns, as well as, statistical measures for the ranking process.

Recently, some researchers have concentrated on the Arabic nested noun compound to generate more than bi-gram candidates such as 3-gram, 4-gram and 5-gram. For instance, Al –Balushi et al. [8] have used linguistic and statistical approaches for identifying nested noun compounds. Specifically, the authors have extended the patterns to include Noun + Noun + Noun + Noun and Noun + Adjective + Noun + Noun + Noun in order to generate the 4-gram and 5-grams. Then, three statistical measures have been used for the ranking process. Similarly, [9] have attempted to enhance the effectiveness of extracting Arabic nested noun compounds by using optimized statistical approaches and lexicon-based approach. The lexicon aims to provide the named entities which facilitates the process of extracting 4-gram and 5-gram candidates.

III. Methodology

The methodology of this study consists of four main phases as depicted in Fig. 1. The first phase is associated with the corpus that will be used in the experiments. Second phase is associated with the preprocessing tasks that aim at turning the data into more suitable form for processing. Third phase is associated with generating and filtering the candidates using the linguistic patterns. Fourth phase is associated with ranking the candidates that have been produced by the linguistic patterns. However, these phases are being tackled in more detail in the following sub-sections.

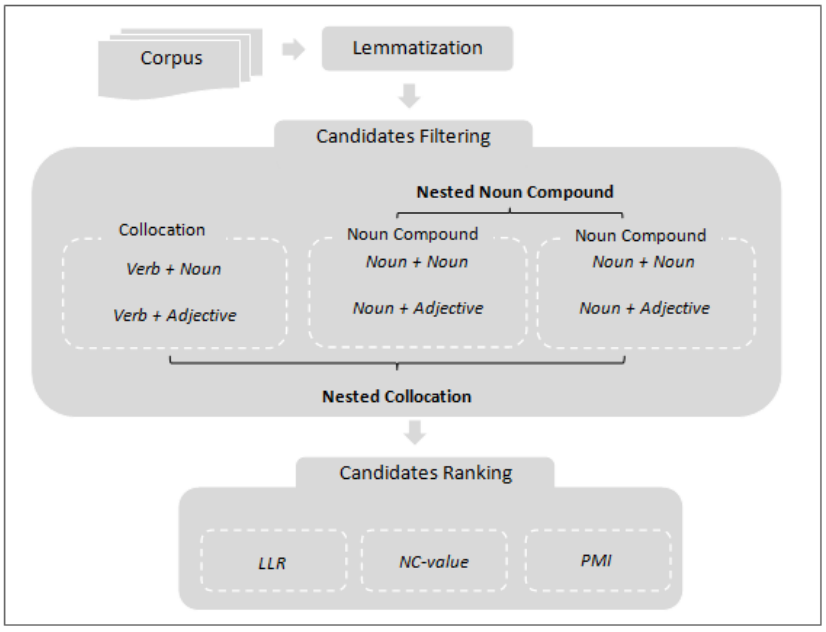

Figure 1. Methodology of the proposed method

3.1.Corpus

This phase aims to discuss the data that would be used which comprises of Arabic newspapers text that collected by Saif & Aziz [3]. Basically, two Arabic media agencies have been used for the collection which are Almotamar.net and Al-jazeera.net.

3.2.Lemmatization

In any natural language, words are being formed with different derivations such as study and studies. Even though, these words have the same meaning but it is difficult to process such word morphologically. Therefore, lemmatization task has been proposed in order to eliminate the inflectional derivations in which the word studies is being lemmatized into study by removing the 'ies' and replace it with the 'y' letter [13]. In this manner, the corpus will undergo a lemmatization task in order to retrieve the root of each words. A novel Arabic lemmatization approach has been used in this study which has been introduced by [14]. Table 1 shows an example of lemmatizing words.

TABLE 1. SAMPLE OF LEMMATIZED WORDS

Word	Meaning	Root	meaning
رؤساء	Presidents	رئيس	President
وزراء	Ministers	وزير	Minster
تصريح	Announcement	صرح	Announce
السياسي	The political	سياسي	political

3.3.Candidate Extraction

This phase aims to generate and filter the candidates of nested noun compounds and collocations based on specific patterns. Basically, this phase has been accommodated using Part-Of-Speech Tagging. POS tagging aims at identifying the grammatical class of the word such as verb, adjective or noun. Apparently, in order to generate the candidates of nested noun compounds and collocations, it is necessary to identify the grammatical tag of each word. In this manner, the Arabic POS tagger of AlGahtani et al. [15] has been used in this study. The following sub-sections are illustrating the linguistic patterns.

3.3.1 Collocations

Collocations are the combination of multi-word terms in which the first word is a verb such as 'اقرأ بعناية / read carefully' [16]. Such collocations are being formed using multiple linguistic patterns such as Verb + Noun (e.g. صرح أوباما / Obama announces) or using Verb + Adjective (e.g. تنديد قوي / strongly condemn). Obviously, these two patterns consist of two words or so-called bigram. Occasionally, collocations could be formed using three words or so-called 3-gram using the pattern of Verb + Noun + Adjective (e.g. ورد نبأ عاجل / receiving breaking news). Table 2 depicts some collocation candidate examples.

TABLE 2. SAMPLE OF LINGUISTIC PATTERNS FOR COLLOCATIONS

Pattern	Collocation	Meaning	Gram
Verb + Noun	أفاد مراسلنا	Reporter stated	Bi-gram
Verb + Adjective	تحريات مكثفة	Intensive investigations	Bi-gram
Verb + Noun + Adjective	صرح مصدر موثوق	Trusted source says	3-gram

3.3.2 Noun Compounds

Noun compounds are the combination of multi-word terms in which the first word is noun [17]. Basically, this type of multi-word terms is being formed using two main patterns which are Noun + Noun (e.g. المركز الأول / first class) and Noun + Adjective (e.g. المدير التنفيذي / executive manager). Apparently, both patterns produce bi-grams candidates. However, sometimes, the noun compounds could be formed as Noun + Adjective + Noun which is a 3-gram compound such as 'الرئيس الأمريكي أوباما / American president Obama' or using Noun + Noun + Adjective such as ' مدرب المنتخب الايطالي / Italian team coach'. Table 3 shows some examples of noun compounds.

TABLE 3. SAMPLE OF LINGUISTIC PATTERNS FOR NOUN COMPOUNDS

Pattern	Noun Compound	Meaning	Gram
Noun + Noun	الاسم الأول	First name	Bi-gram
Noun + Adjective	الجامعة الإسلامية	Islamic university	Bi-gram
Noun + Adjective + Noun	الرئيس السابق حسني	Former president Hosni	3-gram
Noun + Noun + Adjective	لجنة الألعاب الأولمبية	Olympic games committee	3-gram

3.3.3 Nested Noun Compounds

Nested Noun compounds are the combination of two or more noun compounds in which the sequence would contain 4, 5 or 6 words. Simply, attaching two patterns from Table 3 would directly lead to a nested noun compounds. For example, combining the two patterns Noun + Adjective (e.g. الرئيس الامريكي / American president) and Noun + Noun (e.g.باراك أوباما / Barak Obama) would lead to the nested noun compound of 'الرئيس الأمريكي باراك أوباما / American president Barak Obama' which is a 4-gram compound. Table 4 shows some examples of nested noun compounds.

TABLE 4. SAMPLE OF LINGUISTIC PATTERNS FOR NESTED NOUN COMPOUNDS

Pattern	Noun Compound	Meaning	Gram
Noun + Noun + Noun + Noun	رئيس الوزراء عبدالله غول	Prime minister Abdullah Gül	4-gram
Noun + Adjective + Noun + Noun	المنتخب الدولي لكرة السلة	International team of basket ball	4-gram
Noun + Noun + Adjective + Noun	رئيس اللجنة الدولية للسباحة	Head of swimming international committee	4-gram
Noun + Noun + Adjective + Noun + Noun	رئيس اللجنة الدولية لكرة القدم	Head of football international	5-gram

		committee	
Noun + Adjective + Adjective + Noun + Noun	الرئيس المصري الأسبق حسني مبارك	Former Egyptian president Hosni Mubarak	5-gram

3.3.4 Nested Collocations

Here is the contribution of this paper in which the extraction will be extended to include 6-gram and 7-gram candidates. This can be accommodated by combining both collocation (shown in Table 1) and nested noun compounds (shown in Table 4). Such combination is being occurred frequently specially in the news for example the sequence of ' صرح رئيس الوزراء الأسبق عبدالله النسور / former Jordanian prime minister Abdullah Ensour says' consists of a collocation of 'صرح رئيس الوزراء / prime minister says' and the nested noun compound of 'الأردني عبدالله النسور / Jordanian Abdullah Ensour'. Apparently, the sequence in the example is consisting of 7-gram compound. Therefore, the combination of collocation and nested noun compound would significantly extend the process of extracting multi-word terms to include more than 5-gram compounds. Table 5 shows some examples of the proposed nested collocation.

TABLE 5. SAMPLE OF LINGUISTIC PATTERNS FOR COLLOCATIONS

Pattern	Noun Compound	Meaning	Gram
Verb + Noun + Noun + Adjective + Noun + Noun	ندد رئيس الوزراء الماليزي نجيب عبد الرزاق	Malaysian prime minister Najib Razak condemns	6-gram
Verb + Noun + Adjective + Noun + Noun + Noun	أصدرت اللجنة الدولية لحقوق الانسان بيانا	International committee for human rights declares a statement	6-gram
Verb + Noun + Adjective + Adjective + Noun + Noun	فاز الفريق الوطني العراقي لكرة القدم	Iraqi national football team wins	6-gram
Verb + Noun + Adjective + Noun + Adjective + Noun + Noun	أعلن الرئيس الأسبق للجنة المركزية لحركة فتح	Former head of the central committee of Fatah movement declares	7-gram
Verb + Noun + Noun + Noun + Noun + Adjective + Noun	صرح المتحدث باسم حركة المقاومة الإسلامية حماس	Spokesman of the Islamic resistance movement 'Hamas' says	7-gram

After generating all the patterns, multiple lists are being utilized. These lists aim at identifying the occurrence of each pattern. The occurrence is divided into two kinds; the single occurrence for each term (i.e. unigram list), and the co-occurrence between multiple terms. Table 6 shows these lists with their descriptions.

TABLE 6. GENERATED COUNT LISTS

Lists	Description
Unigram	Occurrence of each single term
Bigram	Co-occurrence between every two words
3-gram	Co-occurrence between every three words
4-gram	Co-occurrence between every four words
5-gram	Co-occurrence between every five words
6-gram	Co-occurrence between every six words
7-gram	Co-occurrence between every seven words

3.4.Candidate Ranking

This phase aims to rank the generated candidates produced by the linguistic filter. Basically, the process of ranking these candidates is mainly relying on the lists that describe each occurrence and co-occurrence of the patterns. In fact, the ranking process is being performed based on statistical measures that aim at identifying the strengthen correlation among the terms based on the occurrence and co-occurrence. Since, the occurrence and the co-occurrence are depending on the corpus in terms of the counting therefore, sometimes these measures may be defined as corpus-based measures. For this manner, three statistical measures have been used including NC-value, LLR and MI. These measures are being illustrated as follows:

3.4.1 NC-value

NC-value is an association measure that aims at determining the term-hood for the multi-word candidates. It has been introduced by Frantzi et al. [18]. The term-hood can be expressed as the C-value of the candidate which computed as:

$$C-value(a) = \begin{cases} \log_2(|a|).f(a) & if\ a\ is\ not\ nested \\ \log_2(|a|).(f(a)-g(a)) & otherwise \end{cases}$$

where $|a|$ indicates the number of words in term *a*, while *f*(*a*) is the number of occurrences of *a* and:

$$g(a) = \frac{1}{|T_a|}\sum_{b \in T} f(b)$$

where T(a) refers to the set of longer candidate terms into which a appears ($|T_a|$ is the cardinality of this set).

Furthermore, NC-value aims at identifying the contextual information which can be expressed as N-value which indicates the measure of the terminological status of a context for given candidate term that can be computed as:

$$NC-value(w) = \alpha c-value(w) + (1-\alpha)\sum_{b \in C_w} f_w(b) weight(b)$$

3.4.2 Log-likelihood Ratio (LLR)

LLR is an association measure that has been characterized by a balanced indicator in terms of related multi-word in a nomeculture manner. It has been proposed by [19] in order to extract rare multi-word expressions statistically from a corpus. The log-likelihood mechanism is represented by utilizing the co-occurrence based on the contingency table in which the two words W1 and W2 are being addressed in terms of the occurrence and the co-occurrence. Let a is the number of contexts that the two words have been occurred together, b is the number of occurre3nce where W1 is being occurred independently, c is the number of contexts where W2 is being occured independently, and d is the number of contexts where both words are not being occurred. The log-likelihood ratio is being computed as follows:

$$LLR = 2((a\ln a + b\ln b + c\ln c + d\ln d + (a+b+c+d)\ln(a+b+c+d)) - ((a+b)\ln(a+b) + (a+c)\ln(a+c) + (b+d)\ln(b+d) + (c+d)\ln(c+d)))$$

3.4.3 Point-wise Mutual Information (PMI)

PMI is a statistical measure that aims at identifying the connectedness among the terms in accordance to the information theory proposed by [20]. It has been examined widely in terms of extracting multi-word terms regarding to its ability to utilize the information provided by the single occurrence of the two terms as well as the co-occurrence of the two terms together [21]. It can be computed as follows:

$$PMI(x,y) = log_2 \frac{p(x,y)}{p(x)p(y)}$$

where P(x) is the occurrence probability of word x and P(y) is the occurrence probability of word y in the corpus.

IV. Results

In order to evaluate the results of the statistical measures, the top-N approach proposed by [22] has been used to select the best ranked candidates for each measure. Hence, the best 100 candidates are being selected for every statistical measure. Consequentially, a manual tagging task is being conducted in order to annotate the correct candidates by 1 and the incorrect candidate by 0, then the precision is being calculated as follows:

$$Precision = \frac{Correct\ Candidates}{Total number\ of\ candidates}$$

Hence, the results of precision for NC-value, LLR and PMI are being depicted for multiple grams of candidates including 4-gram, 5-gram, 6-gram and 7-gram. Table 7 shows such results.

TABLE 7. RESULTS OF ASSOCIATION MEASURES

Association	4-gram	5-gram	6-gram	7-gram
NC-value	0.80	0.77	0.53	0.33
LLR	0.62	0.59	0.44	0.22
PMI	0.54	0.48	0.31	0.18

As shown in Table 7, the 4-gram candidates have obtained the greatest results compared to the other grams. This is due to the high probability of occurrence for such compounds compared to the longer candidates that may have a rare occurrence. However, NC-value has outperformed the other measures in terms of extracting the 4-gram by achieving 80% of precision, 5-gram by achieving 77% of precision, 6-gram by achieving 53% of precision, and 7-gram by achieving 33% of precision. On the other hand, LLR has outperformed the PMI for all gram candidates by obtaining 62% of precision for 4-gram, 59% of precision for 5-gram, 44% of precision for 6-gram and 22% of precision for 7-gram. Apparently, NC-value has obtained the superior results compared the other measures regarding to its ability to exploit the contextual information which facilitate the idenitificaiton of longer candidates.

V. Discussion

Although the results obtained for the longer candidates such as 6-gram and 7-gram are relatively low however, the proposed linguistic and statistical approach have shown a capability to identify complex and longer candidates compared to the related work. For example, Al –Balushi et al. [8] who have intended to extract nested noun compound including 4-gram and 5-gram, shown a precision of 29% for 4-gram candidates and 18% of precision for 5-gram candidates. As well as, Al-Mashhadani & Omar [9] who have attempted to

improve the extraction of nested noun compound using optimized statistical measures, have also identified compounds with maximum length of 4-gram and 5-gram by obtaining 72% and 65% of precision respectively. Obviously, the results of the proposed method of this study seems to be competitive and promising for future extraction of longer candidates.

VI. Conclusion

This study has proposed an extension for the extraction process of Arabic multi-word expressions. This can be represented by combining the extraction task of nested noun compounds with collocations. In fact, various linguistic patterns have been utilized based on POS tagging in order to generate longer candidates. As well as, three statistical measures have been used including NC-value, LLR and PMI. Results have shown that the proposed linguistic and statistical approach has the ability to generate longer candidates including 6-gram and 7-gram. However, for future research using more robust statistical measures would significantly improve the extraction of longer collocations.

References

[1] Dhekra Najar, Slim Mesfar, and Henda Ben Ghezela, "A Large Terminological Dictionary of Arabic Compound Words," in *International NooJ Conference*, 2015, pp. 16-28.doi.

[2] N Ababou and A Mazroui, "A hybrid Arabic POS tagging for simple and compound morphosyntactic tags," *International Journal of Speech Technology*, vol. 19, pp. 289-302, 2016.

[3] Abdulgabbar M Saif and Mohd JA Aziz, "An automatic collocation extraction from Arabic corpus," *Journal of Computer Science*, vol. 7, p. 6, 2010.doi:10.3844/jcssp.2011.6.11.

[4] Mohammed Attia, Lamia Tounsi, Pavel Pecina, Josef van Genabith, and Antonio Toral, "Automatic extraction of Arabic multiword expressions," in *In Proceedings of the 7th Conference on Language Resources and Evaluation, LREC-2010*, 2010 http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.187.575&rep=rep1&type=pdf.

[5] Khalid Al Khatib and Amer Badarneh, "Automatic extraction of arabic multi-word terms," in *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on*, 2010, pp. 411-418.doi.

[6] Abdelkader El Mahdaouy, Saïd El Alaoui Ouatik, and Eric Gaussier, "A Study of Association Measures and their Combination for Arabic MWT Extraction," in *Proceedings 10th International Conference on Terminology and Artificial Intelligence*, 2013, pp. 45-52.doi: http://hal.archives-ouvertes.fr/docs/00/88/11/75/PDF/MWT_Latex_7_pages.pdf.

[7] Abdelati Hawwari, Mohammed Attia, and Mona Diab, "A framework for the classification and annotation of multiword expressions in dialectal arabic," *ANLP 2014*, p. 48, 2014.

[8] Hamed Al-Balushi and Mohd Juziaddin Ab Aziz, "A Hybrid Method of Linguistic Approach and Statistical Method for Nested Noun Compound Extraction," *Journal of Theoretical and Applied Information Technology*, vol. 67, 2014.

[9] Maryam Al-Mashhadani and Nazlia Omar, "Extraction of arabic nested noun compounds based on a hybrid method of linguistic approach and statistical methods," *Journal of Theoretical and Applied Information Technology*, vol. 76, pp. 408-416, 2015.

[10] Mohammed Attia, "An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks," in *Challenges of Arabic for NLP/MT Conference, The British Computer Society, London, UK*, 200610.1.1.72.1482.

[11] Siham Boulaknadel, Beatrice Daille, and Driss Aboutajdine, "A Multi-Word Term Extraction Program for Arabic Language," in *International Conference on Language Resources and Evaluation*, 2008 http://dblp.uni-trier.de/db/conf/lrec/lrec2008.html#BoulaknadelDA08.

[12] Ibrahim Bounhas and Yahya Slimani, "A hybrid approach for Arabic multi-word term extraction," in *Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on*, 2009, pp. 1-8.doi.

[13] Tuomo Korenius, Jorma Laurikkala, Kalervo Järvelin, and Martti Juhola, "Stemming and lemmatization in the clustering of finnish text documents," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 2004, pp. 625-633.doi.

[14] Tarek El-Shishtawy and Fatma El-Ghannam, "An accurate arabic root-based lemmatizer for information retrieval purposes," *arXiv preprint arXiv:1203.3584*, 2012.

[15] Shabib AlGahtani, William Black, and John McNaught, "Arabic part-of-speech tagging using transformation-based learning," in *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, Cairo*, 2009, pp. 66-70.doi: http://www.elda.org/medar-conference/pdf/43.pdf.

[16] Stefan Evert, "Corpora and collocations," in *Corpus Linguistics. An International Handbook*, 200810.1.1.159.6220.

[17] Johan Bos and Malvina Nissim, "Uncovering noun-noun compound relations by gamification," in *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*, 2015, pp. 251-255.doi.

[18] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima, "Automatic recognition of multi-word terms:. the C-value/NC-value method," *International Journal on Digital Libraries*, vol. 3, pp. 115-130, 2000/08/01 2000.doi:10.1007/s007999900023 http://dx.doi.org/10.1007/s007999900023.

[19] Ted Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational linguistics*, vol. 19, pp. 61-74, 1993.

[20] Kenneth Ward Church and Patrick Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, pp. 22-29, 1990.

[21] Wen Zhang, Taketoshi Yoshida, Xijin Tang, and Tu-Bao Ho, "Improving effectiveness of mutual information for substantival multiword expression extraction," *Expert Systems with Applications*, vol. 36, pp. 10919-10930, 2009.

[22] Stefan Evert, "The statistics of word cooccurrences," Dissertation, Stuttgart University, 2005Retrieved from.