IOSR Journal of Research & Method in Education (IOSR-JRME)
e-ISSN: 2320–1959.p- ISSN: 2320–1940 Volume 9, Issue 2 Ser. II. (Mar. - Apr .2019), PP 55-63
www.iosrjournals.org

A Non-Compositional Formulae List for EFL Undergraduates

Wenhua Hsu
Applied English Department, I-Shou University, Taiwan
Corresponding author: Wenhua Hsu

Abstract: *This paper aimed to establish a pedagogically useful list of the most frequent semantically non-compositional formulae for undergraduates in an EFL context, in which English is not an official language. The formulae list was derived from a corpus containing 20 million words of 200 textbooks across 40 academic subjects. In consideration of non-compositional formulae in widespread use, the researcher applied a set of criteria when using the program Collocate for selection in the candidate list. Based on frequency, meaningfulness, well-formedness and non-compositionality, 450 formulaic expressions of 2 to 5 words were chosen and they accounted for nearly 2% of the total words in the corpus. As with other individual word lists, this formulae list may serve as a reference for English for Academic Purposes courses.*

Keywords: *formulaic language, lexical coverage, non-compositionality*

Date of Submission: 24-03-2019 Date of acceptance: 08-04-2019

I. Introduction

A written or spoken discourse is not only made up of individual words but also a large number of multi-word sequences, in which some of the words frequently co-occur with others and form relatively fixed word combinations. This phenomenon is generally referred to as formulaic language and each individual case of formulaic language is called a formula or a formulaic sequence (Schmitt, 2010).

This research aimed to identify non-compositional formulaic expressions widely used in college specialist textbooks for EFL students to facilitate reading comprehension. Among a plethora of multi-word combinations, the researcher-teacher was more concerned with the multi-word sequences which may pose reading difficulty if they are not known. Not all formulae are equally semantically compositional. Semantic compositionality signifies how easily a multi-word sequence can be interpreted from its component words. Conversely, semantic non-compositionality denotes that the individual words of a phrase do not help each other to reveal the meaning as a whole. Martinez and Murphy (2011) pointed out that non-compositional formulae negatively affect reading comprehension, especially when they are composed of the most frequent general words and hidden in known words. Students may presume that they are familiar with these very common words (e.g. *as, in, of, that, well*) but actually they are not acquainted with the words in combination (e.g. *as well, as well as, in that, as of*) and deduce a wrong meaning. Non-compositional formulae may traverse various academic disciplines along with high-frequency general words. If no distinction is made between individual words and multi-word sequences, the latter may be misinterpreted.

As such, this research focused on a semantically non-compositional subset of formulaic language. It sought to answer the following two questions.

RQ1. What are the most frequent non-compositional formulaic expressions in specialist textbooks?

RQ2. What is the text coverage of the most frequent non-compositional formulae in the Specialist Textbook Corpus?

II. Literature Review

Formulaic language is ubiquitous. Using the London-Lund Corpus, Altenberg (1998) concluded that various multi-word combinations account for 80% of the total words in the corpus. Erman and Warren (2000) estimated that 55% of the words in an English discourse form parts of prefabricated multi-word units. Addressing native-like fluency, Pawley and Syder (1983, p. 214) proposed a possible explanation that native speakers have thousands of "lexicalized sentence stems" and other formulaic strings at their disposal. Individual words are merely the tips of phraseological icebergs and native speakers' mental lexicons may contain as many formulaic expressions as individual words (Martinez & Schmitt, 2012)..

In spite of the prevalence of formulaic language, there has hitherto been little agreement on what multi-word sequences are regarded as formulaic language, since linguistic researchers divide in what they consider formulaic. Idioms, phrasal verbs, set phrases, binomial expressions and proverbs are relatively fixed multi-word sequences and display one aspect of formulaic language. Similarly, lexical bundles (Biber, Conrad, & Cortes,

2003, 2004; Hyland, 2008), collocations (Altenberg, 1993; Howarth, 1998) and n-grams (Stubbs, 2007) are also subsets of formulaic language, since they are highly recurrent multi-word combinations.

Biber, Johansson, Leech, Conrad and Finegan (1999) first distinguished collocations from lexical bundles. Collocations are statistically associations between two words that are variable and not idiomatic. Namely, words can be associated with several other collocates and retain their own meanings. In contrast, lexical bundles are three or more contiguous words that occur repeatedly together and can be viewed as extended collocations. Most lexical bundles are semantically compositional (e.g. as can be seen) and are usually structurally ill-formed. They may straddle two adjacent phrases (e.g. an important role in the, is one of the).

Different from the above two types of formulaic language, pure idioms are mostly defined in the sense of "opaque invariant word combinations" (Warren, 2005). However, not all idioms are invariable. Gibbs and Nayak (1989) discovered that some idioms of which individual components contribute to their overall figurative meanings are syntactically flexible, for example, 'Jack is sure to spill the beans before long' versus 'The beans that Jack spilled were far more confidential than he realized'.

As can be seen above, formulaic language is multi-faceted. In some cases, formulaic expressions tend to abandon their compositional meaning in favor of a holistic one (Nattinger & DeCarrico, 1992). If the meaning of a formula can be derived from the meanings of its components, it is compositional (semantically transparent). A non-compositional formula is a multi-word unit where the meaning of the whole is not clear from the meanings of its parts. Lewis (1993) dubbed the varying degrees of compositionality "a spectrum of idiomaticity" (p. 98).

In a similar vein, Howarth (1998) classified multi-word units into four categories according to idomaticity: pure idioms, figurative idioms, restricted collocations and free combinations. With little connection to their constituent parts, pure idioms need to be learned as whole units (e.g. cut the mustard, red herring). Contrary to pure idioms, free combinations deliver the literal meanings of their component words and allow substitution, having the highest degree of semantic transparency (e.g. free games, video games, indoor games). Between pure idioms and free combinations, restricted collocations are word combinations in which at least one word has a non-literal meaning and at least one word is used in its literal sense, and the whole combination is still transparent (Cowie, 1998) (e.g. keep an eye on, from door to door). Figurative idioms have metaphorical meanings in terms of the whole, which are separate from their literal meanings (e.g. in the doghouse, a house of cards).

This research leaned toward semantically non-compositional formulaic expressions because they form distinct meanings and can be learned like individual words. According to Nation (2006), lexical coverage is defined as ''the percentage of running words in the text known by the reader'' (p. 61) and generally regarded as a measure of whether a text is likely to be adequately understood. Running words here refer to individual words. When lexical text coverage with an emphasis on known words is calculated, multi-word expressions are not taken into account. As such, the lexical coverage of a text may be overestimated when non-compositional multi-word expressions are concealed in known words and their meanings as a whole happen to be unknown to learners. In this case, knowledge of non-compositional formulae may contribute to filling the chasm of text coverage that individual words fail to account for (Martinez & Murphy, 2011).

III. Research Method

3.1 The Corpus

Referring to required subjects in universities, the researcher compiled the College Textbook Corpus containing 240 textbooks in English across 60 subjects in 5 academic domains, totaling 24 million words. Each domain comprised 12 subjects with 4 principal textbooks being selected and having an approximately equal number of words after eliminating references, tables and figures.

3.2 The Procedure

Barlow's (2004) *Collocate* was used to retrieve multi-word units from the College Textbook Corpus. The span parameter for word length was set from 2 to 5, because frequencies drop drastically as word sequences are extended to five consecutive words or beyond (Hyland, 2008).

The frequency thresholds in past studies ranged from 10 to 40 times per million words. To prevent important multi-word sequences from being removed at the beginning, five times per million words were set to begin with. This decision was based on Nation's (2005) data figures. For a single word to enter the 5,000 most frequent word families, the word and its family member altogether need to occur at least 7.87 times per million words for inclusion in the fifth 1,000. Consequently, the cut-off was set at five times instead of 10 to 40 times. As far as 24 million words were concerned, appearing 120 times at the minimum was the selection threshold.

Since the goal of this research was to identify the formulaic expressions that are commonly used in college textbooks, formulae that occurred with a very high frequency but appeared in only one or two academic disciplines would not be taken into account. For the sake of widespread use, two decisions were made:

1. Even dispersion: Members of a formulaic expression in different inflectional forms taken together had to appear in each of the 60 disciplines across 5 academic domains.
2. Range: Members of a formulaic expression in different inflectional forms taken together had to appear in at least 120 out of the 240 textbooks across 60 disciplines.

The decisions were admittedly arbitrary but in agreement with the goal of widespread use and stricter than previous research (e.g. Biber, Conrad & Cortes, 2004; Coxhead, 2000; Hyland, 2008), where the selection critera were established at having to occur in over a half of the total subjects, in over five different texts, and in at least 10% of texts to preclude idiosyncratic uses.

Another consideration was meaningfulness. The retrieved word strings must have meanings and can be learned as a whole. Therefore, Mutual Information (MI) was utilized to filter out free word combinations. A high MI indicates a stronger association between two words, while a lower one means that their co-occurrence is more likely a coincidence. Collocations with an MI higher than 3 are considered strong (Hunston, 2002). Accordingly, multi-word sequences with the MI lower than 3 (the default value) were removed at this stage. They were, for example, 'which is the' and 'to that of'.

Following that, the researcher referred to Shin and Nation (2008) as well as Martinez and Schmitt (2012) and formulated four post-hoc criteria to gauge meaningfulness, well-formedness, non-decomposability and non-compositionality in turn.

C1. Does the multi-word sequence convey a meaning?
C2. Does the multi-word sequence cross the boundary of a neighboring constituent?
C3. Does the construct of the candidate multi-word sequence behave like an individual word, which is unlikely to be further analyzed into the form-meaning link of its subparts?
C4. Is the candidate multi-word sequence semantically non-compositional?

For Q1 to Q4, the researcher and her colleague made an independent judgment on 6,000+ candidate multi-word sequences with a wide-range occurrence of at least 120 times and MI>3. The 3-point scale was used and the responses of yes, not sure and no were coded as 1, 0.5 and 0 respectively. When the answers of both raters were the same, which shows a clear-cut decision, the entry was either excluded from or included for further analysis. When there was no consensus between two raters or the answer was 'not sure', the entry was decided for tentative inclusion in the candidate list. A series of Cohen's Kappa statistics were undertaken as inter-rater reliability tests. The k values were all greater than 0.80, revealing a substantial level of agreement between the two raters.

3.3 Data processing

When compiling non-compositional formulaic sequences, a few modifications were made for pedagogical purposes. Multi-word sequences in different inflectional forms were combined together with an accumulative frequency, to form a single item with their lemma as the representative form. It was assumed that focusing on a single entry at a time may be simpler for EFL students to learn at the beginning.

Another major revision was made for two formulaic sequences being partially overlapping. A partial overlap occurs when a shorter formulaic sequence is subsumed in a longer one, each of which could occur as a meaningful unit (e.g. *as well* versus *as well as*). Both are different in meaning but either of them is non-compositional. In order to obtain an accurate frequency for '*as well*', subtractions were made from the frequency of '*as well as*'. The two phrases were compiled into the formulae list separately, since they can stand alone as an individual unit.

The selection of non-compositional formulaic expressions involved the following sequence: (1) frequency (five times per million words at the minimum for initial filtering), (2) even dispersion and wide range (across all of the 60 subjects in 5 academic domains and in over a half of the compulsory textbooks of the same discipline), (3) cohesiveness of words for meaningfulness (MI>3 and checked with C1), (4) well-formedness (C2), (5) non-decomposability (C3), and (6) non-compositionality (C4) (see the above-mentioned C1 to C4).

IV. Results and Discussion

4.1 The most frequent non-compositional formulae in English specialist textbooks

Totally, 450 non-compositional formulaic expressions of 2 to 5 words were selected from the Specialist Textbook Corpus and formed the formulae list. The list encompasses 251 two-word, 142 three-word, 55 four-word and 2 five-word phrases commonly used in college textbooks.

The RANGE program (Heatley, Nation & Coxhead, 2004) was administered to examine the vocabulary levels of the individual words of the non-compositional formulae. RANGE is installed with the ranked twenty-five 1,000 English word-family lists derived from the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) according to their occurring and dispersion in the corpora (Nation, 2012). Table 1 presents a picture of the vocabulary levels of the non-compositional formulae along the BNC/COCA word-frequency scale. The present formulae list consists of 1,197 running words and involves 357

word types as well as 311 word families. The BNC/COCA first 1,000 word families account for 89.13% of the total words in the formulae list and the second 1,000 make up 5.77%. The combined coverage percentage of the first 2,000 word families is 94.9%. The percentage of the third 1,000 word families is 2.18%, the third highest lexical coverage after the first 2,000 word families. After the first 4,000 word families, the coverage percentage of additional 1,000 word families decreases to less than 1%.

Table 1: Tokens and coverage at each of the BNC/COCA base word lists for the formulae list

BNC/COCA base word lists	Tokens	% coverage in tokens	Cumulative % coverage	Number of word families in the formulae list
1st 1,000	1,106	89.13%	89.13%	216
2nd 1,000	73	5.77%	94.90%	44
3rd 1,000	29	2.18%	97.08%	24
4th 1,000	16	1.28%	98.36%	15
5th 1,000	7	0.56%	98.92%	7
6th 1,000	2	0.16%	99.08%	1
7th 1,000	1	0.08%	99.16%	1
8th 1,000	1	0.06%	99.22%	1
9th 1,000	2	0.12%	99.34%	2
11th 1,000	1	0.08%	99.42%	1
12th 1,000	3	0.24%	99.66%	3
13th 1,000	4	0.22%	99.88%	3
15th 1,000	1	0.08%	99.96%	1
33rd compounds	1	0.03%	99.99%	1
34th abbreviations	1	0.01%	100%%	1
Total	1,197	100%		311
Note: None of the words in the formulae list belong to the BNC/COCA 10th, 14th, 16th ~25th 1000.				

As is evident in Table 1, a large number of non-compositional formulae are composed of very general words, most of which (94.9%) are from the first 2,000 most frequent word families in the BNC/COCA. The pairings or strings of content words and function words form a common pattern in the formulae list, for example, much as (=though), as well as, in order to, there + be, and to do with. Among the examples, the everyday words as, well, order, do, much and there do not have an independent meaning but are a component of a repertoire of multi-word combinations that make up a discourse, as Sinclair (1991) has claimed. Without specialist knowledge involved, these semantically non-compositional multi-word sequences occur across a wide range of subjects with their high-frequency component words.

Furthermore, it is worth noting that the formulae list includes 13 academic words from the Academic Word List (AWL) (Coxhead, 2000), which only covers 1.57% of the total words in the formulae list (see Table 2). This suggests that the meanings of multi-word sequences containing AWL are mostly literal and semantically compositional. The little overlap of the non-compositional formulae with the AWL argues for the essentiality of the non-compositional formulae because this formulae list is intended for another level of lexical learning beyond the most frequent 2,000 word families and the AWL.

Table 2; Non-compositional formulae containing the AWL words

Formulae	Total occurrences including inflections	Formulae	Total occurrences including inflections
consist of	2,173	**specific** to	297
range from … to …	2,002	take…for **granted**	281
prior to	1,707	the **odd(s)**	281
a **range** of/ a ~ **range** of	1,485	(be) **identical** [to/with]	260
impose (…) [on/upon]	1,068	the **bulk** of	246
assuming that	717	at **odds** with	207
consist with	613	**consist** in	156
on **behalf** of	444	on one's **behalf**	133
(be) **coupled** with	406	**granted** that	125
a **couple** of	337	in the first **instance**	110

Note: The AWL words are in bold.

The length of multi-word units has some impact on semantic transparency. When formulaic sequences become longer, their potential for ambiguity and polysemy will decrease. As for how long non-compositional formulae can be, the data shows that two 5-word formulaic sequences stretched from 3-word formulae can still be semantically non-compositional while retaining a cross-disciplinary attribute, as shown in the cased of and *as far as…be concerned* and *have ~ to do with*.

Concerning the structure, a great majority of 2-word formulae (233 out of 251) are grammatically-conditioned pairs, namely a content word combined with a function word, in contrast with merely 19 lexical

collocations, a content word combined with a content word (e.g. *no matter, simply put, so far, very few*). Amid grammatical collocations, phrasal verbs are in the majority (89/233=38%) (e.g. *account for, carry out, cater to, cope with,*) and phrasal prepositions come second (24/233=10.3%) (e.g. *as for, as per, according to, apart from*), followed by a preposition + a noun (18/233=7.7%) (*at times, at once, in question, in place*), being the third.

Moreover, it is worth mentioning that several 2-word phrases containing Latin words used in Academic English were initially filtered in because of high frequency and high MI. They were ultimately selected in the formulae list due to the fact that they cannot be further decomposed into subparts and may impede reading comprehension if students do not know them, for instance, *ad hoc, bona fide, et al, vice versa, per se* and *post hoc.*

The most common pattern of the 3-word formulae is a passive verb followed by a preposition requiring a noun phrase or by an infinitive-to for completion. In the present specialist textbook corpus, past participle phrases come from a reduction of an adjective clause by omitting the relative pronoun and the copula-be form and are used as a post-nominal adjective phrase to modify the preceding noun. For completeness sake, they are presented as (be) + past participle + preposition or infinitive-to, as in the instances of (be) accustomed to, (be) bound to, (be) concerned with and (be) composed of. When the verb-be is added, they form the passive and can stand alone appearing in an independent clause. In addition, it should be noted that the frequent use of the passive voice without a by-phrase seems to be one of the grammatical features in academic prose. Therefore, this reveals a different picture of how we do the passive drills with a grammar book (the passive followed by a by-phrase) and how the passive is used in authentic discourse (the passive followed by a preposition other than *by* or followed by an infinitive-to).

The three patterns as ~ as, a ~ of, and by + noun phrase are also prolific among the 3-word sequences, as in the examples of *as far as, as soon as, as much as, a couple of, a host of, a range of, by means of, by way of* and *by virtue of*. These three patterns contribute to the coverage of a subject, the description of quantity or an approach.

For 4-word sequences, the prepositional phrase is the most common structure, comprising 56% of all forms in the category of 4-word formulae (=31/55). They are, for example, *in the event of/that, in the light of, in the wake of, on the grounds of/that, on one's own account* and *with a view to*.

As discussed, the structural types of the formulae are proliferous and it may not be easy to fold them into a compact categorization. The current formulae list is similar to Martinez's (2012) phrasal expression list in non-compositionality. However, only 208 of the 450 formulae overlap with Martinez's 505 phrasal expressions. Part of the difference between the two non-compositional formulae lists lies in the different sources of data. The current corpus was limited to both academic and written texts, whereas Martinez's was inclusive of speech.

4.2 The coverage of the most frequent non-compositional formulae in specialist textbooks

The formulae list contains a total of 450 phrases of 2 to 5 words with an accumulation of 136,008 individual instances and 405,034 running words, which makes up almost 2% of the total words in the Specialist Textbook Corpus.

A short excerpt from the corpus is presented below. This passage was selected from a textbook in relation to operation system. The non-compositional formulae are in bold and underlined, and may give us a picture of the most frequent non-compositional phrases used in college specialist textbooks.

Outsourcing is sending work outside the firm **rather than** having it handled by the firm's employees. **By virtue of** outsourcing, a firm's capacity needs may be reduced **a lot**. The decision **as to** where to locate is critical. Firms compete with **one another** by keeping labor, transportation **as well as** distribution costs low. **There have been** many impressive examples of savings and other benefits from outsourcing. Many firms have suffered from the costs of overcapacity as demand has fallen, continuing to pay heavy fixed costs even as plants are idle. Capacity can be a problem **as well in terms of** rising demand. As General Motors Corp. **appeared to** face the best of times, it added one-third of of work crew, recalling 1,000 workers who had previously been **laid off**. Mahadevan (2010, p. 312)

Among the 130 running words, eleven non-compositional phrases (26 words in total) belong to the formulae list. Their coverage in the passage is 20% in tokens (=26/130). Without recognition of the eleven phrases, an EFL undergraduate majoring in operation management may not be able to read this excerpt effectively.

At first glance, approximately 2% text coverage of the most frequent non-compositional formulae in the Specialist Textbook Corpus does not seem to be worth highlighting. However, not having a grip of them may hinder reading comprehension. English-native children regard a vocabulary load of two unknown words per hundred words as difficult reading (Carver, 1994). Some scholars (Hu & Nation, 2000; Schmitt, Jiang &

Grabe, 2011) view one unknown word in every fifty words (98% coverage) as the threshold necessary for adequate comprehension. If 2% unknown words are a critical benchmark for unassisted understanding of a text, then the present non-compositional formulae should not be ignored. For this reason, the researcher would like to propose the inclusion of them in English for Academic Purposes syllabi.

V. Conclusion

The major goal of this research was to create a semantically non-compositional subset of formulaic language for EFL undergraduates to learn after the most frequent 2,000 words and academic words. By means of a series of criteria, a total of 450 non-compositional formulaic expressions of 2 to 5 words were selected and they made up approximately 2% of the total words in the Specialist Textbook Corpus, although this can be as high as 20% in some cases. The formulae list contains the most commonly-used phrases across various academic fields. Nearly 95% of the non-compositional formulaic expressions are made of the BNC/COCA first 2,000 word families. Accordingly, the formulae list can bridge the gap between the coverage that the most general words can and cannot account for in a text. Regardless of academic majors, university students may encounter these phrases very often while reading English-medium textbooks in their fields of study. The current formulae list is short and may be a feasible option for all fields of students to learn in a short period of time.

Despite arbitrary decisions on cut-off values in the selection of the non-compositional formulae, there may be some advantages to overt instruction of these frequent expressions. The effectiveness of learning non-compositional phrases is worth investigation but beyond the present focus. It is hoped that the non-compositional formulae may provide some inspiration for teaching materials development for academic purposes as well as for future empirical studies.

References

[1]. Altenberg, B. (1993). Recurrent verb-complement constructions in the London Lund Corpus. In J. Aarts, P. de Haan, & N. Oostdijk, (Eds.), *English language corpora: Design, analysis and exploitation* (pp. 227-245). Amsterdam: Rodopi.

[2]. Altenberg, B. (1998). On the phraseology of spoken English: the evidence of recurrent word combinations. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 101-122). Oxford: Oxford University Press.

[3]. Barlow, M. (2004). Collocate [Computer software]. Houston: Athelstan. Available from http://athel.com/product_info.php?products_id=29

[4]. Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: An initial taxonomy. In A. Wilson, P. Rayson, & T. McEnery (Eds.), *Corpus linguistics by the Lune: A festschrift for Georffrey Leech* (pp. 71-93). Fankfurt: Peter Lang.

[5]. Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at …*: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371-405.

[6]. Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow, England: Pearson.

[7]. Carver, R. P. (1994). Percentage of unknown words in text as a function of the relative difficulty of the text: Implications for instruction. *Journal of Reading Behavior, 26*(4), 413-437.

[8]. Cowie, A. (1998). (Ed.). *Phraseologoy: Theory, analysis, and applications.* Oxford: Oxford University Press.

[9]. Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

[10]. Erman, B., & Warren, B. (2000). The idiom principle and the open-choice principle. *Text, 20*(1), 29-62.

[11]. Gibbs, R. W., & Nayak, N. (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology, 21*(1), 100-138.

[12]. Heatley, A., Nation, I. S. P., & Coxhead, A. (2004). RANGE [Computer software]. Retrieved from http://www.victoria.ac.nz/lal/about/staff/paul-nation

[13]. Howarth, P. (1998). Phraseology and Second Language Proficiency. *Applied Linguistics, 19*(1), 24-44.

[14]. Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

[15]. Hunston, S. (2002). Corpora in applied linguistics. Cambridge: Cambridge University Press.

[16]. Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4-21.

[17]. Lewis, M. (1993). *The lexical approach: The state of ELT and the way forward*. Hove, England: Language Teaching.

[18]. Mahadevan, B. (2010). *Operations management: Theory and practice*. Delhi, India: Pearson Education.

[19]. Martinez, R., & Murphy, V. A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly, 45*(2), 267-290.

[20]. Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299-320.

[21]. Nation, I. S. P. (2005). The BNC word family lists 14,000. Retrieved from http://www.victoria.ac.nz/lal/about/staff/paul-nation

[22]. Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? The Canadian Modern Language Review, 63(1), 59–82.

[23]. Nation, I. S. P. (2012). The BNC/COCA word family lists 25,000. Retrieved from http://www.victoria.ac.nz/lal/about/staff/paul-nation

[24]. Nattinger, J. R., & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

[25]. Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Hampshire, England: Palgrave Macmillan.

[26]. Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26-43.

[27]. Shin, D., & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal, 62*(4), 339-348.

[28]. Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

[29]. Stubbs, M. (2007). An example of frequent English phraseology: Distribution, structures and functions. In R. Facchinetti (Ed.), *Corpus Linguistics 25 years on* (pp. 89-105). Amsterdam: Radopi.

[30]. Warren, B. (2005). A model of idiomaticity. *NJES: Nordic Journal of English Studies*, *4*(1), 35-54.

Appendix

there + be	97794	thanks to	1694	from time to time	480
such as	60842	above all	1688	for a while	474
et al	45724	the others	1656	lend oneself to	474
as well as	33902	there … to be	1641	get up	473
have to	31307	in question	1640	look forward to	472
(be) likely to/ (be) ... likely to	30725	(be) faced with	1637	with reference to	465
according to	27957	all but	1636	run into	464
rather than	26953	take over	1635	come to terms with	462
out of	21812	at once	1627	over and over	461
(be) used to	20551	only if	1608	in the event of	460
[quantifier] kind(s) of	18954	all over	1604	come across	455
in order to	18343	set out	1593	on board	449
a number of/ a ... number of	18163	(be) coupled with	1575	vantage point(s)	449
result in	16476	next to	1567	no idea	447
a few	16449	rest [on/upon]	1511	put forth	442
tend to	14606	or so	1476	[as/so] far as ~be concerned	434
due to	14460	give rise to	1467	stick to	432
appear to	14074	be subject to	1465	little [is/was] known about	429
in terms of	13611	no more	1463	get on	409
not only…but also	13333	as soon as	1462	as if it were	406
of course	12640	when it comes to	1458	(be) backed by	398
so that	12527	a bit	1452	set off	397
deal with	11978	on earth	1445	a little bit	397
serve as	11439	so as to	1429	sort out	395
engage in	11236	courtesy of	1419	by and large	394
so…that	11216	to the point	1417	third party	393
up to	10887	very few	1410	by the way	392
seek to	10305	in this regard	1390	get along	384
that is,	10238	(be) subjected to	1387	cutting edge/cutting-edge	383
in that	10118	work out	1322	a good deal of	380
consist of	10090	in the way	1307	(be) suited for	379
as well	9531	at the expense of	1302	no less than	375
no longer	9494	as though	1286	consist in	370
range from…to	9268	at present	1275	cater to	369
account for	9227	no more than	1243	would rather	365
as to	9223	a couple of	1242	hold out	361
at all	8730	and so forth	1242	yield to	361
each other	8617	free of	1235	in respect [of/to]	359
along with	8580	object to	1228	have~ bearing on	343
in addition to	8508	(be) charged with	1202	in accord with	339
take place	8234	resort to	1192	go off	337
(be) going to	7986	in the first place	1183	a wealth of	337
prior to	7844	in a manner	1180	on account	324
as…as possible	7639	insofar as	1165	step up	323
on the other hand	7638	now that	1160	all along	320
even though	7517	short of	1156	on account of	320
the latter	7235	have to do with	1153	look after	316
result from	7043	have to do with	1152	simply put	314
just as	7008	in the wake of	1136	hold true	313
call for	7007	and …alike	1127	run out of	311
even if	6999	in regard to	1127	(be) liable for	306
a range of/ a ... range of	6774	among others	1126	take account of	305
instead of	6620	take part in	1115	all right	302
apply to	6553	a host of	1090	a case in point	294
such…that	6360	per se	1086	in charge	291
the rest	6055	come up with	1085	in case of	290
in turn	5785	by no means	1084	bear in mind	289
the former	5709	figure out	1059	hang on	288
as a result of	5701	specific to	1050	put out	288
carry out	5578	build up	1037	put it another way	284
as if	5528	(be) bound to	1031	temper … with	280
[quantifier] sort(s) of	5362	for the sake of	1030	not … altogether	271
regardless of	5061	put on	1026	make up for	268
with respect to	4985	at stake	1021	a plethora of	261
the rest of	4882	catch up	1019	on one's behalf	260
not yet/ not … yet	4811	take care of	1018	high end/ high-end	259
impose [on/upon]	4767	the odd(s)	975	with a view to	254

subject to	4618	in line with	975	in lieu of	252
one another	4502	take … for granted	974	by the same token	251
other than	4356	by virtue of	965	on one's own terms	251
belong to	4345	stand for	957	not ... the least	248
a little	4279	if only	951	the more…the less	248
look for	4226	do with	931	strike out/ strike ... out	247
welling-being/ well being	4156	all too	930	so as not to	245
a lot of	4129	lots of	923	in the short run	244
close to	3950	or otherwise	920	right away	241
too…to	3918	give way	909	make out	239
make up	3878	in view of	894	of late	237
[as/so] long as	3831	among other things	891	no choice but to	235
[auxiliary verb] + hardly	3708	(be) entitled to	882	get off	233
in place	3659	in this respect	882	go ahead	233
as such	3619	to death	877	as regards	230
after all	3517	(be) identical [to/with]	874	put down	227
relative to	3494	let alone	871	suit(s) against	224
to do with	3447	hold on	868	(be) attuned to	224
so far	3431	little more than	857	follow suit	222
on one's own	3406	owe…to	854	of sorts	222
with regard to	3394	be about to	847	can not help but	222
point(s) of view	3366	shed ~ light on	846	granted that	221
put it	3328	make sense of	840	lie with	221
(be) concerned with	3308	in charge of	835	in compliance with	221
take one's place	3301	a handful of	830	turn down	220
take on	3299	on top of	825	had better	215
the extent to which	3268	set forth	820	(be) liable to	213
in favo(u)r of	3243	(be) set to	809	tie up	205
would like	3149	(be) suited to	806	on the ground(s) of	204
appeal to	3114	the bulk of	804	in order that	201
assuming (that)	3075	for good	797	any longer	199
far from	3006	vice versa	797	map out	189
in short	2863	in a sense	797	catch up with	189
take... into account	2839	lay out	787	once and for all	186
get to	2785	rule out	787	the other way [around/round]	186
as with	2766	in the sense that	784	take the place of	182
a great deal	2751	(be) accustomed to	780	may as well	181
draw [on/upon]	2719	embark [on/upon]	776	know better	179
as opposed to	2694	aside from	770	make room for	179
in time	2671	in case	769	lay off	178
as for	2652	nothing but	754	tie in	178
may well	2626	in the long run	742	unless otherwise	177
not... at all	2598	by far	737	in the event that	176
happen to	2587	make one's way	728	in one's favor	175
consist with	2572	for life	717	of a kind	172
cope with	2572	free from	701	in a row	170
hold that	2544	on the ground(s) that	695	pros and cons	170
[as/so] far as	2533	in the light of	691	rule(s) of thumb	166
on the one hand	2473	a priori	680	(be) incumbent [on/upon]	165
to date	2472	owing to	677	give a ... account of	165
[provided/ providing] that	2468	ad hoc	673	[come/get] to grips with	161
in the face of	2454	every other	671	take charge of	159
at times	2444	[suppose/supposing] that	668	no point	158
[in/over] the course of	2409	hinge [on/upon]	652	make a point	158
given that	2403	in the aftermath of	646	make a point of	158
in spite of	2396	hold up	645	bona fide(s)	155
arise from	2383	(be) obliged to	632	call forth	154
as a means	2364	as yet	631	get used to	154
bring about	2304	irrespective of	618	pull up	152
make sense	2258	at odds with	617	at the mercy of	151
in practice	2248	in the way of	616	in the first instance	150
no matter	2204	bring up	608	put off	149
and so on	2192	in the interest(s) of	607	in so far as	144
manage to	2189	top-down/ top down	603	(be) versed in	140
(be) supposed to	2187	and …respectively	578	make the most of	140
stem from	2128	that is to say	576	all manner of	137
apart from	2104	a good deal	572	subject … to	136
(be) composed of	2101	long for	553	quite a few	136
give up	2096	in the sense of	553	boil down to	135
in the absence of	2041	once more	551	across from	133

take up	2038	used to	550	in a nutshell	133
a lot	2008	make one's point	545	come true	130
take advantage of	1995	a bit of	544	before long	126
pick up	1991	in a position to	543	as per	124
subject matter	1975	in one's own right	543	(be) to blame for	124
call [on/upon]	1915	much less	535	have got to	124
amount to	1898	take off	525	once in a while	124
pertain to	1874	run out	521	out of the question	124
yet to	1861	course(s) of action	515	in a fashion	123
in accordance with	1853	in point	514	in view	122
take in	1793	put forward	512	level off	121
such that	1790	(be) contingent [on/upon]	506	by all accounts	121
a great deal of	1776	in return for	501	put up with	121
as of	1759	back up	495	as a matter of course	121
on behalf of	1759	in place of	487	course of events	120
per capita	1747	come about	486	every bit as	120
the few	1741	turn in	484	be that as it may	120
by means of	1736	be to blame	483	in the last couple of	120
have [quantifier] to do with	1730	in return	481	in a manner of speaking	120

Wenhua Hsu. “A Non-Compositional Formulae List for EFL Undergraduates". IOSR Journal of Research & Method in Education (IOSR-JRME) , vol. 9, no. 2, 2019, pp. 55-63.