A list of opaque prepositional expressions for EFL undergraduates

Wenhua Hsu
Department of Applied English
I-Shou University, Taiwan
Email: whh@isu.edu.tw

Abstract

The purpose of this paper was to establish a list of the most frequent semantically non-compositional prepositional expressions for EFL undergraduates, who need to read English academic texts in their fields of study. Made up of a preposition and a noun phrase, prepositional expressions usually function as an adjective or adverb. In this research, high-frequency prepositional expressions were retrieved from the 112 million-word academic section of the 560-million-token Corpus of Contemporary American English (COCA), which contains approximately 100 academic magazines and journals. In consideration of widespread use, the researcher applied a series of selection criteria (viz. frequency, meaningfulness, well-formedness and non-compositionality), while compiling the list. A total of 220 semantically non-compositional prepositional phrases of 2 to 5 words were chosen and they accounted for 1.02% of the total words in the COCA-academic. As with other individual word lists, this prepositional expressions list may serve as a reference for English for General Purposes as well as English for Academic Purposes syllabi.

Keywords: non-compositionality, formulaic language, lexical coverage

1. Introduction

Individual words may be deemed as the tips of phraseological icebergs (Martinez & Schmitt, 2012). A written or spoken discourse consists of not only individual words but also multiword sequences, in which some words frequently co-occur with others and form relatively fixed multiword combinations. Schmitt (2010) referred to this phenomenon as formulaic language (Schmitt, 2010). In earlier phraseological studies, Altenberg (1998) discovered that prefabricated multiword units make up as high as 80% of the total words in the London-Lund Corpus. In a similar vein, Erman and Warren (2000) estimated that multiword units account for at least 55% of the words in an English discourse. Concerning the issue of nativelike fluency, Pawley and Syder (1983, p. 214) pointed out that native speakers have thousands of “lexicalized sentence stems” at their disposal.

Among a wide range of multiword combinations, the researcher-teacher was more concerned with semantically non-compositional expressions, which may pose deceptive comprehension. Semantic compositionality signifies how easily a phrasal expression can be analyzed from its constituent parts. Conversely, semantic non-compositionality denotes that the interpretation of a phrase contradicts the meaning of its component words. Martinez and Murphy (2011) admonished that non-compositional phrasal expressions negatively affect reading comprehension, particularly when they are made up of the most frequent words and hidden in the familiar words. Non-beginners of English may assume that they have mastered most of the general words (e.g. *that, well, of, as*, *in*) but actually they have no idea of these words in combination (e.g. *in that*, *as well*, *as well as*, *as of*) and consequently infer a wrong meaning. Multiword expressions, in particular composed of general words, may traverse various topic areas along with their component words. If no distinction is made between individual words and their multiword combinations, the latter may be misinterpreted.

As such, this research aimed at an opaque subset of formulaic language, specifically semantically non-compositional prepositional expressions. In the present study, a prepositional expression is defined as a phrase beginning with a preposition or functioning as a preposition. It excluded phrasal verbs and verbal expressions, since they form another subset of formulaic language so as to merit separate research from prepositional phrases. This research addressed the two questions as follows.

1. What are the most frequent non-compositional prepositional expressions in academic English?
2. What is the coverage percentage of the most frequent non-compositional prepositional expressions in the Academic English Corpus

2. Literature Review

Although formulaic language is ubiquitous, little consensus has been hitherto reached concerning what multiword combinations are considered as formulaic language. Linguistic researchers diverge in what they deem formulaic. Phrasal verbs, idioms, slangs, proverbs and binomial expressions are relatively fixed multiword units and each shows one facet of formulaic language. Likewise, collocations (Altenberg, 1993; Howarth, 1998), lexical bundles (Biber, Conrad, & Cortes, 2003, 2004; Hyland, 2008) and n-grams (Stubbs, 2007) are different subsets of formulaic language as well, since they are quite fixed multiword combinations and highly recurrent.

According to Biber, Johansson, Leech, Conrad and Finegan (1999), two words that are statistically associated and remain their meanings are collocations. They are not idiomatic and can collocate with several other words. In contrast, lexical bundles are extended collocations, which are three or more word sequences that appear together repeatedly. Most collocations and lexical bundles are semantically compositional. However, lexical bundles are mostly ill-formed in terms of straddling two neighboring structural units (e.g. *is one of the, in the form of the, to the fact that, a role in the*)

As opposed to (extended) collocations, which are variable, Warren (2005) defined pure idioms as “opaque invariant word combinations” (Warren, 2005, p.11). Nevertheless, not all idioms are invariable. Gibbs and Nayak (1989) detected that some figurative idioms are flexible in syntax as shown in the cases of *John spilled the beans unintentionally* and *the beans John spilled are about to spread*.

Regardless of types of formulaic language, some multiword expressions are likely to favor a holistically different meaning and abandon the compositional meaning of their individual words (Nattinger & DeCarrico, 1992). As such, a phrase is compositional when its overall meaning can be parsed from the meanings of its component words. Conversely, a non-compositional phrase is an opaque multiword expression, in which the meaning of the whole cannot be deduced from the meanings of its constituent parts. As for the degree of compositionality, Lewis labelled it as “a spectrum of idiomaticity” (1993, p. 98).

Along the scale of compositionality, Howarth (1998) classified multiword combinations into four kinds, ranging from being least to most idiomatic. They are free combinations, restricted collocations, figurative idioms and pure idioms in turn. Free combinations, of which the component words deliver the literal meaning and allow substitution, have the highest degree of compositionality and hence semantic transparency (e.g. *free games, video games, indoor games*). Restricted collocations are more or less transparent multiword combinations with some restrictions on substitution. Concretely speaking, one or two words are used in the literal sense and one or two words have a non-literal meaning (Cowie, 1998) (e.g. *keep an eye on, from door to door*). Figurative idioms have both metaphorical and literal meanings, which may be used for pun or word play (e.g. *in the doghouse, a house of cards*). At the extreme end of a spectrum of idiomaticity, pure idioms have little connection to the meanings of their component words and therefore need to be learned anew (e.g. *cut the mustard, red herring*).

Among various types of multiword combinations, this research leaned toward semantically non-compositional prepositional expressions because they form a distinct subset of formulaic language and need to be learned like single words. If not known, they may cause comprehension problems. Relevant to adequate comprehension is lexical text coverage. Nation (2006) defined lexical text coverage as “the percentage of running words in the text known by the reader” (p. 61) and it is generally regarded as a measure of whether a text is likely to be adequately understood. Running words here refer to individual words. When adequate comprehension with an emphasis on how many individual words are known is measured by lexical coverage percentage, non-compositional multiword expressions are not calculated to avoid repeated counting. As a result, the lexical coverage of a text may be inflated when non-compositional expressions are hidden in familiar words and their meanings as a whole happen to be unfamiliar to learners. As such, knowledge of non-compositional multiword expressions may contribute to filling the chasm of text coverage that individual words fail to account for (Martinez & Murphy, 2011).

In the literature, frequency-based and phraseological approaches have often been used to analyze multiword units (Nesselhauf, 2005). The former resorts to statistical software to retrieve recurrent multiword units, while the latter relies on manual analyses of their discourse functions, patterns and structures.

Frequency has been generally agreed as an indicator of usefulness in terms of the return on learning efforts of a lexical item. The cutoff value for frequency is usually contingent on the researcher's decision and hence arbitrary. For instance, Biber et al.'s (1999) cutoff point was set at occurring 10+ times per million words and in 5+ texts. When studying the lexical bundles used in history and biology academic writing, Cortes (2004) opted for 20 times per million words. When compiling lexical bundles from the 3.5-millon-word corpus containing theses, dissertations and research articles (RAs), Hyland (2008) also increased the cutoff point from the occurrence of 10+ times to that of 20+ times per million words and decided on the dispersion cutoff at appearing in at a minimum of 10% of the sampled texts. More rigorously, Biber, Conrad and Cortes (2004) chose a cutoff frequency at 40 times per million words instead.

Although a variety of multiword expression tools/software are very helpful in automatic extraction, yet merely based on frequency, the result may end up in long listings of multiword sequences, some of which may be meaningless (e.g. *that do not, which is the, of which is*) and some of which may straddle two adjacent phrases (e.g. *to change in the, found in the, of the two types of*). In spite of thoroughness, such listings may not be "pedagogically compelling" (Simpson-Vlach & Ellis, 2010, p. 493).

To compile a list of the most frequent spoken collocations for direct learning, Shin and Nation (2008) applied a series of criteria and underwent laborious manual examination. Among the six criteria they drew upon was "grammatical well-formedness" (p. 341). They made it a point to target the multiword sequences that do not span "immediate constituents" (namely, two adjoining phrases) (Bloomfield, 1933, p. 161), since the selected grammatical well-formed multiword sequences must be comprehensible. For example, '*extent that the*' is not easy to understand in comparison with '*to the extent that*', because the retrieval of the latter conforms to the rationale of "immediate constituents" (Bloomfield, 1933, p. 161).

By creating a 25-million-word corpus of RAs across five academic domains, Durrant (2009) endeavored to establish a list of positionally-variable academic collocations for college undergraduates from different departments. Only measuring the strength of word co-occurrence with the aid of statistical tools, Durrant (2009) filtered out the most frequent 1,000 two-word collocations across many academic disciplines. Amid the 1,000 collocations, a great majority of them (763/1000) are grammatical collocations, viz. the combination of one open-class word and one closed-class word (e.g. *number of, respect to, effect on, note that, suggest that, can be, show that, was used, no significant, effects on, suggests that* and *this paper*). As can be seen, if some of the 1,000 collocations are not extended to longer sequences (e.g. *can be*, *was used*), they may fail to contribute to the learning of grammatical patterns. Likewise, some of the 1,000 collocations need to be combined for better learning (e.g. *suggest(s) that*, *effect(s) on*), while other collocations are noticeably incomplete so as not to be helpful for immediate learning (e.g. *number of* and *respect to*). Still some collocations (e.g. *this paper, our results*) are free combinations of two single words, both of which remain their literal meanings and may be already known to our learners. Despite being significant in terms of occurring frequency, free collocations, having the highest semantic transparency and allowing substitution, may not be a priority in a learning sequence or deserve more attention than other collocations, especially when learners are already familiar with the component words of these decomposable, free collocations.

To tackle the issue of teachability, Simpson-Vlach and Ellis (2010) adopted an additional quantitative measure and put forward the notion of Formula Teaching Worth (FTW). In addition to frequency, mutual information (MI) was also factored in the multiple regression analyses of multiword sequences. A FTW formula was therefore proposed to rank formulaic sequences based on usefulness (denoted by frequency) and relevance (represented by MI). Mutual information (MI) is a statistical measure of cohesiveness of words, showing collocational strength and thereby a degree of idiomaticity (Stubbs, 1995). A recurrent multiword unit with a high MI value is more likely to be meaningful and is thus worth pedagogical attention. Accordingly, Simpson-Vlach and Ellis (2010) advocated that the FTW metric with the integration of MI into frequency measures may help teachers to prioritize multiword units, when judging them from the perspective of whether they merit instruction.

Despite the idea of the FTW, semantic opacity was not taken into account. Instead, Martinez and Schmitt (2012) attempted to identify the most frequent opaque phrasal expressions that are compatible with the existing single word lists along the word-frequency scale of the British National Corpus (BNC). Making reference to Wray and Namba's (2003) eleven criteria concerning whether a multiword combination is a formulaic sequence, Martinez and Schmitt (2012) created a set of post-hoc criteria to minimize intuitions. There were three core criteria and three auxiliary criteria being used subsequent to a frequency-oriented screening. The six criteria centered on the judgment of whether a multiword sequence is a Morpheme Equivalent Unit and then whether it is semantically transparent or potentially "deceptively transparent" (Laufer, 1989, p. 11).

Related to this study is Ackermann and Chen's (2013) interdisciplinary Academic Collocation List (ACL). They created a corpus of over 25 million words from the curricular section of the Pearson International Corpus of Academic English (PICAE). After the initial statistical computation, a panel of experts was consulted to refine and systemize multiword expressions. Following that, Ackermann and Chen (2013) compiled 2,468 most frequent lexical collocations into the ACL to help EAP learners increase their collocational competence in academic English.

The present research supports Ackermann and Chen's (2013) idea that it is only through manual scrutiny that a collocation list can be pedagogically helpful. Despite being relevant to EAP learning, a quick glance at the ACL shows that some academic collocations (e.g. *further research, academic writing*) may have already been within our learners' grip and may not be their concern any longer. The inclusion of free collocations into the ACL makes the listing so unwieldy as to overburden learners. They may therefore distract their attention from the collocations they need imminently. As such, the researcher-teacher would like to suggest that since time is limited in class, there is no need for concentrating on transparent multiword expressions that are made up of known words,.

In view of the fact that not all formulaic sequences are equally important in the learning stages, this research adopted the criterion of semantic non-compositionality and targeted at prepositional expressions as a point of departure. The prepositional phrases list was intended for matriculating undergraduates as the next cohort of lexical items to learn after the most frequent 3,000 word families.

3. Research Method

3.1. The Corpus

The Corpus of Contemporary American English (COCA) is one of the largest and most balanced corpora of American English. The corpus contains 560+ million words of text (20 million words each year since 1990 to the present) and is divided into various sub-corpora equally, including fictions, popular magazines, newspapers, spoken and academic texts. The present prepositional expressions list was derived from the 112 million-word academic sub-corpus of the COCA, which comprise approximately 100 academic journals across nine academic domains (education, history, social sciences, political sciences, humanities, philosophies/religions, sciences and technologies, medicine as well as miscellaneous subject areas). The COCA-academic was utilized for its free access, contemporariness (involving the time since 1990 onwards) and large-size data including a variety of academic subjects, which caters to the need for EAP learners of different academic backgrounds.

3.2. The Procedure

The software *AntGram* (Anthony, 2018) was administered to retrieve recurrent multiword expressions (n-gram) from the corpus. The span parameter for word length was set at 2 to 5 words, since frequencies decrease sharply as word strings are extended to five words or more (Hyland, 2008). Although 5-word strings may be rare, they were included in the initial sifting for thoroughness sake.

As aforementioned in the literature review, the frequency cutoff thresholds ranged from 10 to 40 times per million words. To prevent important multiword bundles from being discarded at the very beginning, a loose criterion was determined to begin with, i.e. five times per million words. Five times per million words were based upon the frequency scale of the BNC ranked 1,000-word-family lists (Nation, 2005). For an individual word to enter the most frequent 5,000 word families, the word and its family members altogether need to occur at a minimum of 7.87 times per million words for inclusion in the BNC 5th 1,000-word-family list. Accordingly, the cutoff value was decided on five times as opposed to 10 to 40 times in past studies. As far as the 112 million words in the current corpus were concerned, 560 occurrences at the minimum was the threshold for inclusion.

Since the research purpose was to identify the most commonly-used prepositional expressions in academic English, multiword units that appeared with a very high frequency but occurred in only one or two academic domains would not be taken into account. In consideration of widespread use, two selection criteria were considered:

1. Uniform dispersion: Prepositional expressions with similar meanings but in different forms altogether had to appear in each of the nine academic domains.
2. Wide range: Prepositional expressions with similar meanings but in different forms altogether had to appear in 50+ out of the 100 academic journals across nine academic domains.

The two considerations were admittedly arbitrary but consistent with the current selection goal (common and widespread use). They were even more rigorous than previous studies (e.g. Biber, Conrad & Cortes, 2004; Coxhead, 2000; Hyland, 2008), in which the cutoff thresholds were set at having to appear in over a half of the total academic subjects, in 5+ different texts, and in over 10% of texts to preclude the possibility of an author's idiosyncrasy.

Another factor concerning the multiword retrieval was meaningfulness. The retrieved word strings must convey a meaning. This selection requirement would make them equivalent to an individual word list, in which each item can be learned like a single word. Before the experts' examination, the statistical measure *Mutual Information* (MI) was used to do the initial filtering to remove multiword combinations at random.

A large MI score signifies stronger cohesiveness between two or more words, while a small value means that their co-occurrence is more likely a coincidence. Multiword combinations with an MI value greater than 3 are regarded as a strong association (Hunston, 2002). As a consequence, those multiword combinations with both a high MI value and a high frequency number were first filtered in while those appearing at the bottom of both rankings were discarded. Multiword combinations with the MI value lower than the default 3 were removed at this stage, as in the cases of '*to that of*' and '*with which the*'.

Following that, four questions were formulated based on Wray and Namba (2003), Shin and Nation (2008) as well as Martinez and Schmitt (2012) to guide the experts' judgement of potential prepositional phrases for selection in the listing. The four post-hoc questions were used to evaluate prepositional properties (Q1), meaningfulness (Q2), well-formedness (Q3) and semantic non-compositionality (Q4).

Q1. Does the candidate multiword combination begin with a preposition?

Q2. Does the candidate multiword combination have a meaning?

Q3. Does the candidate multiword combination span two neighboring phrases?

Q4. Does the meaning of a candidate multiword combination not remain or marginally remain when each component word is decoded with its core meaning?

For Q1 to Q4, the researcher-teacher and her colleagues made an independent judgment on nearly 4,000 candidate multiword combinations with a wide dispersion and high frequency of 560+ times and MI>3. The 3-point scale for the answers *yes*, *not sure* and *no* was measured as 1, 0.5 and 0 respectively. When the responses of raters were the same, which confirmed a clear-cut decision, the candidate multiword combination was either selected in for further examination or removed from the candidate list. When there was little consensus among the raters or the response was 'not sure', the candidate word string was decided for tentative inclusion in the list for later scrutiny. The experts' vetting for meaningful word strings at this phase led to a decrease of four-fifth of raw entries.

For Q1 to Q4, Cohen's Kappa statistical measures were conducted on the SPSS for an inter-rater reliability check. The k values were 0.98, 0.99, 0.96 and 0.92 in turn (all >0.80), demonstrating a substantial degree of agreement among the raters.

All in all, the inclusion of opaque prepositional expressions in the listing entailed the following steps: (1) frequency (f >5 times per million words for initial filtering), (2) wide and even dispersion (in over a half of the journals of the same subject and across each of the nine disciplinary domains), (3) co-occurrence of words for meaningfulness (MI>3) and (4) decision on prepositional properties, meaningfulness, well-formedness and non-compositionality with the aid of Q1 to Q4.

4. Results and Discussion

4.1. The most frequent non-compositional prepositional expressions in academic English

Two hundred and twenty non-compositional prepositional expressions of 2 to 5 words were ultimately selected from the COCA-academic and formed the phrase list. The list consists of the most commonly-used 77 two-word, 95 three-word, 43 four-word and 5 five-word opaque prepositional expressions in academic English.

The RANGE program (Heatley, Nation & Coxhead, 2004) was run to measure the vocabulary levels of the component words of the non-compositional prepositional phrases. This software is incorporated with the ranked BNC/COCA twenty-five 1,000-word-family lists derived from the British National Corpus and the Corpus of Contemporary American English based on their frequency and dispersion in the corpora (Nation, 2012). Table 1 presents a snapshot of the word levels of the opaque prepositional expressions along the BNC/COCA word-frequency scale. The present list involves 655 running words, 209 word types and 196 word families. The BNC/COCA 1st 1,000 word families covered 87.18% of the total words in the prepositional expressions list and the 2nd 1,000 word families account for 7.18%. Taking together, the coverage of the BNC/COCA 2,000 is 94.36%. The percentage of the 3rd, 4th and 5th 1,000 word families is 1.37% respectively, the third highest coverage percentage after the first 2,000 word families. After the first 6,000 word families, the coverage percentage of additional 1,000 word families decreases to lower than 1%.

Table 1

Vocabulary levels of the prepositional phrases list and the coverage percentage of each level

BNC/COCA base word lists	Tokens	% coverage in tokens	Cumulative % coverage	Number of word families
1st 1,000	571	87.18%	87.18%	130
2nd 1,000	47	7.18%	94.36%	36
3rd 1,000	9	1.37%	95.73%	8
4th 1,000	9	1.37%	97.10%	8
5th 1,000	9	1.37%	98.47%	7
6th 1,000	3	0.46%	98.93%	2
7th 1,000	0	0	98.93%	0
8th 1,000	0	0	98.93%	0
9th 1,000	1	0.15%	99.08%	1
10th 1,000	0	0	99.08%	0
11th 1,000	1	0.15%	99.23%	1
12th 1,000	0	0	99.23%	0
13th 1,000	1	0.15%	99.38%	1
14th~25th 1,000	0	0	99.38%	0
32nd compounds	1	0.15%	99.53%	1
33rd abbreviations	1	0.15%	99.68%	1
Not in the list	2	0.31%	100%	X
Total	655	100%		196

As shown in Table 1, most of the opaque prepositional expressions (specifically, 94.36 %) are made up of the BNC/COCA first 2,000 word families (e.g. *as of, as to, as per*). The everyday words *of, to* and *as* rarely appear alone. These general words are only a small part of a repertoire of multiword units that compose a text (Sinclair, 1991). Carrying little specialist knowledge, these semantically non-compositional prepositional phrases occur across a great variety of subject matter along with the most frequent general words.

Concerning the structure of 2-word prepositional expressions, a vast majority of them (43 out of 77) are the combination of *a preposition + a noun* (43/77=55.84%) (e.g. *above all, at times, in place, above board, in question, at once*), followed by phrasal prepositions (16/77=20.78%) (e.g. *as to, as per, as for, as with, apart from, according to*).

Among the prepositional phrases, 3-word sequences are the most commonly-used ones, accounting for 43% of all n-grams (=95/220). The commonest pattern of the 3-word prepositional expressions is *in + noun phrase*, as in the cases of *in a row, in case of, in a fashion, in line with*, followed by *on + noun phras*e, for instance, *on account of, on behalf of,* *on that note.* These prepositional phrases contribute to the explanation or reference of a subject and the description of an approach or quantity.

Four-word prepositional expressions are, for instance, *on one's own account, on one's own terms, on the grounds of/that, with a view to, in the event of/that, in (the) light of, in the wake of.*

Regardless of the number of words in the prepositional phrases, the combinational structures are proliferous because of many prepositions per se being the beginning word in the combination and it may therefore not be simple to classify them into only a few categories.

4.2. The coverage of the most frequent non-compositional prepositional expressions in the Academic English Corpus

The current opaque prepositional expressions list encompasses altogether 220 prepositional phrases of 2 to 5 words with a total of 1,142,406 running words and 376,223 instances, which accounts for 1.02% of the total words in the COCA-academic.

Below is a short excerpt from the present corpus. The following passage was extracted from a journal article in relation to operation management. The non-compositional prepositional phrases are in bold and underlined. They may give us a quick look at how the most frequent opaque prepositional expressions are used in academic English.

> 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 work crew, recalling 1,000 workers who had been laid off earlier. Mahadevan (2010, p. 312)

Amid the 128 words, six different non-compositional prepositional phrases (totally 16 words) are a part of the present list. The coverage of them in the passage is 12.5% in word-tokens (=16/128). Not recognizing these five non-compositional prepositional phrases, a management EFL undergraduate may not be able to read this passage fluently.

At first glance, a small percentage figure like 1.02% lexical coverage of the most frequent non-compositional prepositional expressions in the Academic English Corpus does not seem to be worth attention. Nevertheless, in some cases, they may cover as high as 12.5% of the total words. If they are not recognized, these prepositional phrases may deter adequate comprehension. On that note, the researcher-teacher would like to propose the incorporation of semantically non-compositional prepositional expressions into Academic English syllabi.

5. Pedagogical implications

The current non-compositional/opaque prepositional expressions list provides an access to the academic register, yet the list itself does not suffice for EFL undergraduates. Like the learning of single words, the non-compositional prepositional expressions should be learned as an individual word. They should be taught and learned with examples in context rather than sentences in isolation. EAP instructors can raise their students' awareness of how opaque prepositional phrases function in context with the help of concordance lines, which can be freely accessed from these websites such as Compleat Lexical Tutor (http://www.lextutor.ca/concordancers), GloWbe (http://corpus2.byu.edu/glowbe/) and NOW (https://www.english-corpora.org/now/). By using concordance lines from free online corpora, students can observe abundant examples of authentic language, making an inference on their own of how these opaque prepositional phrases are used. Teachers can also use concordance lines derived from the said corpora and design classroom exercises from within, for example, cloze tests or gap-fill drills. With more exposure to academic texts in pursuing years, EFL learners will consolidate the lexical knowledge acquired from this prepositional expressions list.

6. Conclusion

The purpose of this study was to create a semantically non-compositional subset of formulaic language for EFL learners to facilitate reading comprehension of academic texts. Applying a principled set of criteria, the researcher selected a total of 220 non-compositional prepositional expressions of 2 to 5 words for inclusion in the list, which covered 1.02% of the total words in the COCA-academic. In some cases, the text coverage can be as high as 12.5%. The present list contains the most commonly-used opaque prepositional phrases across a great diversity of academic disciplines. As high as 94.36% of the opaque prepositional expressions are composed of the BNC/COCA first 2,000 word families. Accordingly, the present list can fill the rift between the coverage that the most frequent words can and cannot account for in a text. No matter what subject they major in, EFL undergraduates may encounter these non-compositional prepositional expressions very often while reading English-medium academic materials in their fields of study. Though the present prepositional expressions list is short, it may be a feasible option for college students to learn within a short period of time.

The effectiveness of learning opaque prepositional expressions is worthy of continuous attention but beyond the present concern. Last but not least, the researcher wishes to suggest that the present prepositional expressions list may serve as a reference for English for General Purposes as well as English for Academic Purposes teaching materials development.

References

Ackermann, K., & Chen, Y. (2013). Developing the academic collocation list (ACL)-A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*(4), 235-247.

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.

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.

Anthony, L. (2018). AntGram (Version 1.0.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

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.

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

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

Bloomfield, L. (1933). *Language.* New York: Henry Holt.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23,* 397-423.

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

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

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes, 28*(3), 157-169.

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

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

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

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

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

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for*

Specific Purposes, *27*(1), 4-21.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From human thinking to thinking machines* (pp. 316-323). Clevedon, England: Multilingual Matters.

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

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

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

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

Nation, I. S. P. (2005). The BNC word family lists 14,000. Retrieved from <https://www.victoria.ac.nz/lals/about/staff/paul-nation>.

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

Nation, I. S. P. (2012). The BNC/COCA word family lists 25,000. Retrieved from <https://www.victoria.ac.nz/lals/about/staff/paul-nation>.

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

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

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

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

Simpson-Valch, R., & Ellis, N.C. (2010). An academic formulas list: New methods in phraseology Research. *Applied Linguistics, 31*(4), 487-512.

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

Stubbs, M. (1995). Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language, 2*(1), 23-55.

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.

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

Wray, A., & Namba, K. (2003). Formulaic language in a Japanese-English bilingual child: A practical approach to data analysis. *Japan Journal for Multilingualism and Multiculturalism, 9*(1), 24-51.

Appendix

The most frequent non-compositional/opaque prepositional expressions in academic English

2-word prep. phr.	Freq.	**3-word prep. phr.**	Freq.	**4-word prep. phr.**	Freq.
above all	2404	above and beyond	627	[in/over] the course of	3242
above board	599	against the grain	566	along the lines of	2520
according to	32928	ahead of time	1109	as a result of	7067
across from	597	among other things	1478	at the expense of	1955
after all	4529	around the corner	1434	at the mercy of	618
all along	814	as a means	3189	by leaps and bounds	2184
all but	2343	as a rule	607	by the same token	734
all over	2306	as opposed to	3573	for the time being	700
along with	10412	as well as	39837	from time to time	1000
among others	1751	at face value	2604	in a position to	1073
apart from	2887	at odds with	1159	in one's own right	1073
as for	3524	between the lines	924	in so far as	610
as of	2486	by all accounts	567	in the light of	1245
as per	586	by and large	900	in the event that	647
as regards	710	by courtesy of	2091	in the event of	977
as such	4648	by means of	2460	in the first instance	617
as to	11160	by no means	1702	in the face of	3294
as with	3656	by the way	898	in the first place	1817
as yet	1176	by virtue of	1564	in the interest(s) of	1148
aside from	1337	for a while	993	in the aftermath of	1193
at all	10587	for the record	862	in the same breath	594
at issue	1854	for the sake of	1639	in the sense of	1085
at once	2333	in a fashion	566	in the sense that	1353
at par	1204	in a manner	1814	in the absence of	2814
at present	1924	in a nutshell	559	in the long run	1305
at stake	1629	in a row	640	in the short run	726
at times	3282	in a sense	1369	in the wake of	1762
before long	589	in accordance with	2596	in the way of	1158
below/under par	616	in accord with	836	on a par with	2173
by far	1299	in addition to	10329	on one's own account	868
close to	5032	in any case	1721	on one's own terms	734
due to	17245	in an instant	1490	on the brink of	2626
far from	3935	in charge of	1413	on the ground(s) that	1250
for good	1369	in case of	779	on the one hand	3316
for life	1276	in due course	1915	on the other hand	9318
from scratch	1316	in compliance with	699	on the right track	2486
in arrears	1747	in favo(u)r of	4211	on the same page	1803
in case	1336	in good shape	605	on the verge of	2733
in charge	781	in lieu of	735	out of the blue	571
in order	3774	in line with	1575	out of the question	569
in place	4694	in one's favor	646	up in the air	588
in point	1040	in order that	676	with a view to	738
in practice	3055	in order to	21757	on the ground(s) of	679
in question	2348	in place of	1008	-----------------------------------	----------
in return	1585	in regard to	1752	**5-word**	Freq.
in short	3769	in respect [of/to]	860	[as/so] far as ~be concerned	947
in that	12200	in return for	1025	as a matter of course	560
in time	3546	in spite of	3227	in a manner of speaking	672
in turn	7165	in store for	1915	in the last couple of	739
in view	574	in terms of	16258	with a grain/pinch of salt	1232
insofar as	1796	in the bag	700		
instead of	8135	in the balance	2554		
irrespective of	1161	in the black	1982		

next to	2263	in the flesh	594		
nothing but	1319	in the loop	812		
of course	15130	in the picture	3774		
of late	718	in the pipeline	672		
of sorts	700	in the red	2548		
on account	819	in the way	1961		
on board	964	in this regard	2058		
on demand	1025	in this respect	1467		
on earth	2121	in view of	1481		
on end	1501	of a kind	642		
other than	5504	on account of	814		
out of	25788	on behalf of	2486		
owing to	1229	on one's behalf	745		
prior to	9557	on one's own	4400		
regardless of	6323	on one's plate	571		
as well	4502	on that note	1305		
subject to	5809	on the air	2307		
such as	71141	on the horizon	1926		
thanks to	2411	on the map	3052		
to date	3315	on the ropes	1490		
to death	1461	on the shelf	1932		
under fire	594	on the spot	2554		
under wraps	700	on top of	1401		
up to	13093	out of action	2486		
		out of hand	1753		
		out of order	4413		
		out of place	3052		
		out of pocket	711		
		outside the box	1316		
		over and over	978		
		over one's head	588		
		over the counter	2492		
		over the top	2352		
		to a degree	698		
		to the letter	1445		
		to the point	2089		
		up to speed	1310		
		with each other	2005		
		with one another	2610		
		with reference to	983		
		with regard to	4386		
		with respect to	6235		