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

This paper describes an attempt to establish a pedagogically useful list of the most frequent semantically non-compositional prepositional expressions for EFL undergraduates, who need to read English-medium specialist articles in their fields of study. Prepositional expressions are made up of a preposition plus a noun phrase and function as an adjective or adverb. The opaque prepositional expressions list was derived from the 112 million-word academic sub-corpus of the 560-million-token Corpus of Contemporary American English (COCA), which contains nearly 100 different peer-reviewed academic journals. In consideration of widespread use, the researcher applied a series of selection criteria when compiling the list. In accordance with frequency, meaningfulness, well-formedness and non-compositionality, 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, it is hoped that this prepositional expressions list may serve as a reference for English for General Purposes as well as general academic English.

Keywords: non-compositionality, formulaic language, lexical coverage

1. Introduction

Individual words are merely the tips of phraseological icebergs (Martinez & Schmitt, 2012). A discourse or a text is not only made up of individual words but also a large number of multi-word expressions, in which some of the words frequently co-occur with others and form relatively fixed multi-word combinations. This phenomenon is generally referred to as formulaic language (Schmitt, 2010). Using the London-Lund Corpus, Altenberg (1998) estimated that various multi-word combinations account for as high as 80% of the total words in the corpus. Erman and Warren (2000) reported that at least 55% of the words in an English text form parts of prefabricated multi-word units. To tackle the puzzle of native-like fluency, Pawley and Syder (1983, p. 214) came up with a possible explanation that adult native speakers have thousands of “lexicalized sentence stems” and other formulaic strings at their disposal.

Among a plethora of multi-word 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 interpreted from its component words. Conversely, semantic non-compositionality denotes that the meaning of a phrase as a whole contradicts the decoding of its constituent parts. Martinez and Murphy (2011) pointed out that non-compositional phrasal expressions negatively affect reading comprehension, especially when they are composed of the most frequent general words and hidden in the known words. Students may presume that they are familiar with these very common words (e.g. *that, well, of, as, in*) but actually they have no idea of the words in combination (e.g. *in that, as well, as well as, as of*) and deduce a wrong meaning. Multi-word expressions of general use may traverse various academic domains along with high-frequency words. If no distinction is made between individual general words and general multi-word combinations, the latter may be misinterpreted.

As such, this research focused on a semantically non-compositional subset of formulaic language, specifically opaque prepositional expressions. In this research, a prepositional expression is defined as phrases beginning with a preposition or functioning as a preposition. It excluded phrasal verbs and verbal expressions, since they form such a large subset of formulaic language that they merit separate research of their own. This research sought to answer the following two questions.

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

2. Literature review

Although formulaic language is ubiquitous, there has hitherto been little consensus on what multi-word combinations are counted as formulaic language. Language researchers differ in what they consider formulaic. Idioms, phrasal verbs, proverbs and binomial expressions are relatively fixed multi-word sequences and display one aspect of formulaic language respectively. In a similar vein, 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 lexical bundles from collocations. Collocations are statistically associations between two words that are variable and not idiomatic. For collocations, words can be associated with several other words (collocates) and retain their own meanings. Lexical bundles are three or more contiguous words that occur repeatedly together and can be regarded as extended collocations. Most lexical bundles are semantically compositional (e.g. *as can be seen, if you look at, in the form of the*) but are usually structurally incomplete. They may even straddle two adjacent

structural units (e.g. *an important role in the*, *to the fact that*, *is one of the*).

In contrast with collocations and lexical bundles, 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) found that some idioms whose individual semantic 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 exemplified above, formulaic language is multi-faceted. In some cases, formulaic expressions tend to abandon their semantically compositional meaning in favor of a holistic one (Nattinger & DeCarrico, 1992). If the meaning of a formulaic expression can be derived from the meanings of its components, it is compositional. A non-compositional phrasal expression is a semantically opaque multi-word unit where the meaning of the whole is not clear from the meanings of its parts. Lewis (1993) called the varying degrees of compositionality "a spectrum of idiomaticity" (p. 98).

Along the axis of idiomaticity, Howarth (1998) put forward a framework for categorization of multi-word units ranging from being least to most idiomatic: free combinations, restricted collocations, figurative idioms and pure idioms. At the extreme end of compositionality, free combinations deliver the literal meanings of their lexical components and allow substitution, having the highest degree of semantic transparency and flexibility (e.g. *free games*, *video games*, *indoor games*). Restricted collocations are word combinations in which some substitution is possible, but with some restrictions on substitution. Specifically, 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*). With little connection to the meanings of their constituent parts, pure idioms need to be explained and learned as whole units (e.g. *cut the mustard*, *red herring*).

In view of the multiplicity of formulaic language, this research leaned toward semantically non-compositional prepositional expressions because they form distinct meanings and can only be learned like single 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 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 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 multi-word expressions may contribute to filling the chasm of lexical coverage that individual words fail to account

for (Martinez & Murphy, 2011).

In the literature, there are two fundamental approaches used to retrieve recurrent multi-word combinations: a frequency-based approach and a phraseological approach (Nesselhauf, 2005). The former mainly relies on statistical measures as screening criteria, whereas the latter primarily resorts to linguistic analysis and hence manual examination is inevitable.

It is generally agreed that frequency is a good indicator in determining usefulness of a lexical item in terms of the return on learning effort. The pre-determined cut-off values for frequency and dispersion have been arbitrary, subject to researchers' goals. Biber et al. (1999) adopted a very flexible cut-off point at recurring at least ten times per million words and in five or more texts. Cortes (2004) was more conservative and opted for 20 times, when comparing the frequencies and functions of lexical bundles used in published and student disciplinary writing in history and biology. Biber, Conrad and Cortes (2004) were even more cautious in choosing lexical bundles from their corpora by setting a relatively high cut-off frequency at 40 times per million words. Hyland (2008) increased the frequency cut-off point from a minimum of 10 times to 20 times per million words and decided on the breadth of lexical bundles at occurring in at least 10% of the texts in the sample, when selecting lexical bundles in his 3.5-millon-word corpus of research articles, PhD theses and Master's dissertations.

Present-day phrase extractors ensure the properties of frequency and multi-text occurrences. Nevertheless, solely based on frequency, an n-gram program may generate long lists of word sequences, part of which have no meanings (e.g. *which is the, that do not*) and part of which span phrasal boundaries (e.g. *to change in the, found in the, of the two types of*). Though comprehensive, such lists may not, however, be "pedagogically compelling" (Simpson-Vlach & Ellis, 2010, p. 493).

To identify the most frequent collocations in spoken English, which need to be meaningful and comprehensible for deliberate learning, Shin and Nation (2008) drew upon a set of criteria and went through a great deal of manual checking. Among the six criteria they applied was "grammatical well-formedness" (p. 341). They targeted collocations which do not span "immediate constituents" (two neighboring structural units) (Bloomfield, 1933, p. 161), because a grammatical well-formed multi-word sequence is a comprehensible unit. For instance, '*to the extent that'* is more understandable than '*extent that the'*, since the retrieval of the former follows the dividing principle of "immediate constituents" (*Ibid*.).

By compiling a 25-million-token corpus of research articles across five university faculties (life sciences, arts & humanities, science & engineering, social-psychological and social-administrative), Durrant (2009) made an attempt on a listing of positionally-variable academic collocations for students from a wide range of departments. Purely relying on statistical measures to determine the strength of word co-occurrence, he identified the most

frequent 1,000 interdisciplinary two-word collocations. The great majority of word pairs in his listing are grammatical collocations (=763/1000), namely, the combination of one closed-class word and one open-class word (e.g. *this paper, number of, respect to, was used, effect on, effects on, note that, suggest that, suggests that, can be, show that,* and *no significan*t). As shown above, some collocations fail to contribute to the learning of grammatical patterns if they are not extended to longer sequences (e.g. *was used*). Some collocations can be combined into one item for learning together (e.g. *suggest/suggests that*), while others are apparently incomplete so that they are not suitable for direct teaching (e.g. *respect to* and *number of*). Still other collocations (e.g. *our results, our study*) are free combinations of two individual words, both of which have compositional properties associated with literal meanings and allow substitution. Though important in terms of frequency, free collocations with the highest semantic transparency may not be a top priority in deserving more attention than other collocations, if our students are already familiar with the component words of these decomposable collocations.

To tackle the teachability of multi-word sequences caused by purely frequency-based retrievals, Simpson-Vlach and Ellis (2010) took account of another quantitative measure and proposed the idea of Formula Teaching Worth (FTW). In their attempt to marshal an Academic Formulas List, mutual information (MI) and frequency were both factored in the multiple regression analysis. A FTW metric was thereby arrived at and then utilized to rank formulas according to relevance and usefulness, which were represented by MI and frequency in turn. MI is a statistical measure of the cohesiveness of words, signifying collocational strength and a degree of idiomaticity (Stubbs, 1995). Recurrent multi-word combinations with a high MI score are more likely to be meaningful and hence merit pedagogical attention. Simpson-Vlach and Ellis concluded that the FTW that combines frequency and MI may provide teachers with a basis of prioritization, when judging word sequences in terms of whether they are worthy of instruction.

In the above studies, semantic opacity was not considered. In contrast, Martinez and Schmitt (2012) sought to identify the most frequent non-transparent phrasal expressions that are compatible with existing wordlist along the British National Corpus (BNC) word-frequency scale. Referring to Wray and Namba's (2003) eleven criteria regarding whether a word string is a formulaic sequence, Martinez and Schmitt established six post-hoc criteria to minimize intuitions. Three core criteria and three auxiliary criteria were used after a frequency-based search. They were mainly related to the judgment of whether an expression is a Morpheme Equivalent Unit, semantically transparent or potentially "deceptively transparent" (Laufer, 1989, p. 11).

More recent and relevant to this study is Ackermann and Chen's (2013) cross-disciplinary Academic Collocation List (ACL). They compiled a corpus of over 25 million tokens from the written curricular component of the Pearson International Corpus of

Academic English (PICAE). Through statistics as initial computation and subsequently relying on a panel of experts for refinement and systematization, Ackermann and Chen retrieved 2,468 most frequent lexical collocations to help students increase their collocational competence in academic English.

This research lends support to the view of Ackermann and Chen (2013) that only through human intervention can a collocation listing be of pedagogical use. Although they emphasized the relevance of the ACL for learners with academic goals, a look at the ACL shows that some of the ACL (e.g. *academic writing, further research*) are probably already within our students' grip and may no longer be their concern. The inclusion of free collocations leads to the ACL so large as to be unwieldy and possibly overburden students before they concentrate on the collocations they may need imminently. In this regard, the researcher is inclined to believe that there is no need to focus attention on transparent word sequences derived from known words, since class time is limited.

In view of the fact that not all multi-word units are of equal importance to learners with specific purposes, this research adopted 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 set of lexical items to learn after the most frequent general words.

3. Research method

3.1. The Corpus

The Corpus of Contemporary American English (COCA) is a large, balanced corpus of American English. The corpus contains more than 560 million words of text (20 million words each year since 1990 to the present) and it is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts. The present prepositional expressions list was derived from the 112 million-word academic sub-corpus of the COCA, which contains nearly 100 different peer-reviewed 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 chosen because of its free access, large size, contemporariness (involving the years since 1990 onwards) and comprehensive data covering a variety of academic disciplines, which caters to the need for academic English.

3.2. The Procedure

The software *AntGram* (Anthony, 2018) was used to retrieve recurrent multi-word expressions (n-gram) from the 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). Though 5-word sequences may be relatively rare, they were

included in the initial screening for completeness.

The frequency thresholds in past studies ranged between 10 and 40 times per million tokens. To preclude important word strings from being removed at the initial stage, a less rigorous criterion was set to begin with, namely five times per million tokens. This decision was based on the frequency index of the BNC ranked 1,000-word-family bands (Nation, 2005). 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 a minimum of five times rather than 10 to 40 times as in previous research. As far as the 112 million words in the present corpus were concerned, appearing 560 times at least was the selection threshold.

Since one of the present goals was to identify the prepositional expressions that are widely used in academic English, multi-word units that occurred with a very high frequency but appeared in only one or two academic domains would not be considered. In consideration of widespread use, two decisions were made:

1. Even dispersion: Prepositional expressions with similar meanings but in different forms taken together had to appear in each of the nine academic domains.
2. Range: Prepositional expressions with similar meanings but in different forms taken together had to appear in at least 50 out of the 100 academic journals across nine academic domains.

The decisions were admittedly arbitrary but relatively in agreement with the present goal (frequent and widespread occurrence) and more rigorous than the practice in the literature (e.g. Coxhead, 2000; Biber, Conrad & Cortes, 2004; Hyland, 2008), where the screening requirements were established at having to occur in at least a half of the total subject areas, in at least 5 different texts, and in at least 10% of texts to guard against idiosyncratic uses.

Another consideration given to recurrent multi-word sequences was meaningfulness. The multi-word sequences retrieved must have meanings and be learned as a whole. This principle would make them comparable to a list of individual words. Before manual checking, the measure *Mutual Information* (MI) was utilized to do the initial screening to filter out free word combinations.

A high MI means a stronger association between two or more words, while a lower one indicates that their co-occurrence is more likely due to chance. According to Hunston (2002), collocations with an MI no less than 3 are considered strong. Accordingly, those word sequences with both high frequency and high MI were first chosen while those appearing at the bottom of both rankings were removed. Multi-word sequences with the MI lower than the default value (=3) were eliminated at this phase. They were, for example, ‘*with which the’* and *‘to that of’*.

Subsequently, we referred to Martinez and Schmitt (2012), Shin and Nation (2008) as well as Wray and Namba (2003) and thereby formulated four questions to guide the decision

of potential prepositional phrases for inclusion in the list. They were used to gauge prepositional properties (Q1), meaningfulness (Q2), well-formedness (Q3) and semantic non-compositionality (Q4). The four post-hoc questions were:

Q1. Does the candidate multi-word sequence begin with a preposition?

Q2. Does the candidate multi-word sequence convey a meaning?

Q3. Does the candidate multi-word sequence cross the boundary of an immediate structural unit?

Q4. Is the candidate multi-word sequence semantically non-compositional? That is, the meaning as a whole does not remain or marginally remains when each component word is decoded with its core meaning.

For Q1 to Q4, the researcher-teacher and her colleague made an independent judgment on approximately 4,000 candidate multi-word sequences with a wide-range occurrence of at least 560 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 agreement between two raters or the answer was 'not sure', the entry was decided for tentative inclusion in the list. The manual vetting for meaningful and comprehensible units at this stage resulted in a reduction of three-fourths of raw items.

For Q1 to Q4, a series of Cohen's Kappa statistics were undertaken as inter-rater reliability tests. The k values were 0.98, 0.99, 0.96 and 0.92 respectively (all >0.80), revealing a substantial level of agreement between the two raters.

To sum up, the selection of opaque prepositional expressions involved the following sequence: (1) frequency (a minimum of five times per million words for initial screening), (2) even dispersion and wide range (across all of the nine academic domains and in at least half of the journals of the same discipline), (3) cohesiveness of words for meaningfulness (MI>=3 and (4) checked with Q1 to Q4 for prepositional nature, meaningfulness, well-formedness and semantic non-compositionality.

4. Results and discussion

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

A total of 220 non-compositional prepositional expressions of 2 to 5 words were ultimately chosen from the COCA-academic and formed the phrase list. The list encompasses 77 two-word, 95 three-word, 43 four-word and 5 five-word opaque prepositional phrases commonly used in academic English.

The RANGE program (Heatley, Nation & Coxhead, 2004) was used to examine the vocabulary levels of the individual words of the non-compositional prepositional phrases. This software 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 frequency and dispersion in the corpora (Nation, 2012). Table 1 presents a full picture of the vocabulary levels of the opaque prepositional expressions in the BNC/COCA word-frequency scale. The present list consists of 655 running words and involves 209 word types as well as 196 word families. The BNC/COCA first 1,000 word families account for 87.18% of the total words in the prepositional expressions list and the second 1,000 make up 7.18%. The combined coverage percentage of the first 2,000 word families is 94.36%. The percentage of the third, fourth and fifth 1,000 word families is 1.37% respectively, the third highest lexical coverage after the first 2,000 word families. After the first 6,000 word families, the coverage percentage of additional 1,000 word families rapidly reduces to less than 1%.

Table 1

Structure of the prepositional phrases list along the BNC/COCA word-frequency scale

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 is evident in Table 1, a large number of opaque prepositional expressions are composed of very general words, most of which (specifically, 94.36 %) are from the first 2,000 most frequent words in the BNC/COCA (e.g. *as of, as to, as per*). The everyday words *of, to* and *as* do not have an independent meaning but are a component of a repertoire of multi-word combinations that make up a text, as Sinclair (1991) has claimed. Without specialist knowledge involved, these semantically non-compositional word sequences occur across a wide range of subject areas with their high-frequency component words.

Concerning the structure of 2-word prepositional expressions, a vast majority of them (43out of 77) are the pattern *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 common structure, comprising 43% of all forms (=95/220). The most common 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 phrases contribute to the description of quantity, the coverage of a subject, an explanation or an approach.

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) light of, in the wake of, in the event of/that.*

As can be seen, the structural types of the prepositional expressions are proliferous and it may not be easy to fold them into a compact categorization.

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

The present opaque prepositional expressions list contains a total of 220 phrases of 2 to 5 words with an accumulation of 376,223 individual instances and 1,142,406 running words, which makes up 1.02% of the tokens in the COCA-academic.

A short excerpt from the corpus is shown below. This passage was selected from an operation management journal article. The non-compositional prepositional phrases are underlined and in bold, and may give us a picture of the most frequent opaque prepositional expressions 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 previously been laid off. Mahadevan (2010, p. 312)

Among the 128 running words, six different opaque prepositional phrases (16 words in total) belong to the present list. Their coverage in the passage is 12.5% in tokens (=16/128). Without the recognition of these five opaque phrases, an EFL management major may not be able to read this excerpt effectively.

At first sight, an average of 1.02% lexical coverage of the most frequent non-compositional prepositional expressions in the Academic English Corpus does not appear

to be worth noticing. However, in some cases, they may account for as high as 12.5% of the total words, and if not recognized, these phrases may impede reading comprehension. As such, the researcher would like to propose the inclusion of semantically non-compositional prepositional expressions in Academic English syllabi.

5. Pedagogical implications

Although the present prepositional expressions list provides a window to the academic register, itemized phrases are still not enough for EFL undergraduates. As with the learning of individual words, the non-compositional prepositional expressions should be learned in context rather than in isolation. English teachers can raise their students' consciousness of how opaque phrases behave in context with the help of free online concordancers (e.g. Compleat Lexical Tutor at http://www.lextutor.ca/concordancers; GloWbe at http://corpus2.byu.edu/glowbe/; NOW at https://corpus.byu.edu/now/). By using corpora, students can gain direct access to abundant examples of authentic language, resulting in a better understanding of the use and patterns of certain opaque phrases. Classroom exercises using concordances may be undertaken, for instance, in a gap-fill exercise. With more exposure to academic texts in the years that follow, EFL undergraduates will consolidate the lexical knowledge acquired from this prepositional expressions list.

6. Conclusion

The principal concern of this study was to create a semantically non-compositional subset of formulaic language for EFL undergraduates for receptive use. By means of a principled set of criteria, a total of 220 opaque prepositional expressions of 2 to 5 words were selected and they made up 1.02% of the running words in the COCA-academic, although this can be as high as 12.5% in some cases. The present prepositional expressions list contains the most widely-used phrases across various academic fields. As high as 94.36% of the opaque prepositional expressions are made of the BNC/COCA first 2,000 word families. Accordingly, the present list can bridge the gap between the lexical coverage that the most general words can and cannot account for in a text. Irrespective of specialty, college students may come across these expressions while reading academic texts in their fields of study. The present phrasal expressions list is short and may be a viable option for all fields of students to learn in a short time.

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

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 opaque prepositional expressions in academic English

2-word	Freq.	**3-word**	Freq.	**4-word**	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 absence of	2814
as per	586	by and large	900	in the aftermath of	1193
as regards	710	by courtesy of	2091	in the event of	977
as such	4648	by means of	2460	in the event that	647
as to	11160	by no means	1702	in the face of	3294
as with	3656	by the way	898	in the first instance	617
as yet	1176	by virtue of	1564	in the first place	1817
aside from	1337	for a while	993	in the interest(s) of	1148
at all	10587	for the record	862	in the light of	1245
at issue	1854	for the sake of	1639	in the long run	1305
at once	2333	in a fashion	566	in the same breath	594
at par	1204	in a manner	1814	in the sense of	1085
at present	1924	in a nutshell	559	in the sense that	1353
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 accord with	836	in the way of	1158
below/under par	616	in accordance with	2596	on a par with	2173
by far	1299	in addition to	10329	on one's own account	868
close to	5032	in an instant	1490	on one's own terms	734
due to	17245	in any case	1721	on the brink of	2626
far from	3935	in case of	779	on the ground(s) of	679
for good	1369	in charge of	1413	on the ground(s) that	1250
for life	1276	in compliance with	699	on the one hand	3316
from scratch	1316	in due course	1915	on the other hand	9318
in arrears	1747	in favo(u)r of	4211	on the right track	2486
in case	1336	in good shape	605	on the same page	1803
in charge	781	in lieu of	735	on the verge of	2733
in order	3774	in line with	1575	out of the blue	571
in place	4694	in one's favor	646	out of the question	569
in point	1040	in order that	676	up in the air	588
in practice	3055	in order to	21757	with a view to	738
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		