Establishment of a List of Non-Compositional Multi-Word Combinations of English for Journalism Learners

Wenhua Hsu
I-Shou University, Kaohsung, Taiwan

Abstract

This research describes an attempt to establish a pedagogically useful list of the most frequent semantically non-compositional multi-word combinations of English for Journalism learners in an EFL context, who need to read English news in their field of study. The list was compiled from the NOW (News on the Web) Corpus, the largest English news database by far. In consideration of opaque multi-word combinations in widespread use and pedagogical value, the researcher applied a set of selection criteria when using the corpus. Based on frequency, meaningfulness, and semantic non-compositionality, a total of 318 non-compositional multi-word combinations of 2 to 5 words with the exclusion of phrasal verbs were selected and they accounted for approximately 2% of the total words in the corpus. The list, not highly technical in nature, contains the most commonly-used multi-word units traversing various topic areas and newsreaders may encounter these phrasal expressions very often. As with other individual word lists, it is hoped that this opaque expressions list may serve as a reference for English for Journalism teaching.

Keywords: Semantic non-compositionality; Multi-word combinations; Formulaic language.

1. Introduction

A text or a 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 (Schmitt, 2010).

Formulaic language is ubiquitous and makes up a large proportion of any discourse (Nattinger and DeCarrico, 1992). Drawing upon the London-Lund Corpus, (Altenberg, 1993) estimated that various multi-word combinations accounted for as high as 80% of the total words in the corpus. In their study on the idiom principle, Erman and Warren (2000) reported that more than 55% of the words in an English text comprised prefabricated multi-word expressions.

Despite the prevalence of multi-word combinations, there has hitherto been little consensus on the definition, since researchers differ in what they consider formulaic. Wray (2002), identified over 50 terms to describe this phenomenon. Under divergent interpretations, recurrent multi-word combinations have been labeled in a range of ways: collocations (Altenberg, 1998; Howarth, 1998), lexical bundles (Biber et al., 2003; 2004; Hyland, 2008), clusters (Scott, 1996), formulaic sequences/formulae (Martinez and Schmitt, 2012; Wray, 2002), sentence stems (Pawley and Syder, 1983), prefabricated units/prefabs (Cowie, 1998) and n-grams (Stubbs, 2007) (a sequence of n words, where n can be any positive integer). Each of them reflects one aspect of formulaic language.

Due to its multiplicity, it is difficult to lend some consistency to every single instance of formulaic language. Therefore, this research used multi-word combinations as an umbrella term to refer to miscellaneous fixed combinations of words. Among a plethora of multi-word expressions, the researcher was more concerned with the word combinations that may pose reading comprehension problems if they are not known. Not all multi-word combinations are equally semantically compositional or transparent. Martinez and Murphy (2011), pointed out that opaque formulaic sequences may negatively affect reading comprehension or cause deceptive comprehension, especially when they are composed of the most frequent general words and concealed in the known words. English learners may presume that they are familiar with these very common words (e.g. as, of, in, well, that) but actually they are not acquainted with the words in combination (e.g. as of, as well, as in, that) and deduce a wrong meaning. If no distinction is made between individual general words and general word combinations, the latter may be overlooked or misinterpreted.

As such, this research focused on a semantically non-compositional subset of formulaic language. However, it excluded phrasal verbs, 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.

RQ1. Apart from phrasal verbs, what are the most frequent non-compositional multi-word combinations in English news articles?
RQ2. Apart from phrasal verbs, what is the lexical coverage of the most frequent non-compositional multi-word combinations in English news articles?
2. Literature Review

Formulaic language is multi-faceted. In some cases, formulaic expressions tend to abandon their semantically compositional meaning in favor of a holistic one (Nattlinger and DeCarriaco, 1992). Semantic transparency is related to semantic compositionality. Compositionality signifies how easily a multi-word combination can be interpreted from its component words. Conversely, non-compositionality denotes that the meaning of a multi-word unit as a whole contradicts the decoding of its constituent parts. Namely, the individual words of a multi-word unit do not help each other to reveal the meaning as a whole. 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 combinations 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 component words and allow substitution, having the highest degree of semantic transparency (e.g. video games, free 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 more or less transparent (Cowie, 1998) (e.g. keep an eye on, make a comeback). Figurative idioms have metaphorical meanings in terms of the whole, which are separate from their literal meanings (e.g. a house of cards, a smoking gun). With little connection to the meanings of their constituent parts, pure idioms need to be explained and learned as a whole (e.g. cut the mustard, red herring).

This research focused its attention on non-compositionality because non-compositional multi-word combinations form distinct meanings and can be learned like single words. According to Nation (2006), lexical text coverage is defined as “the percentage of running words in the text known by the reader” (p. 61) and regarded as an indicator of whether a text is likely to be adequately understood. When lexical text coverage with an emphasis on known words is calculated, multi-word combinations are not taken into account. As such, the lexical coverage of a text may be overestimated when non-compositional multi-word combinations are hidden 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 text coverage that individual words fail to account for (Martinez and 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.

The pre-determined cut-off points in the literature for frequency and dispersion have been arbitrary, subject to researchers’ goals. Biber et al. (1999), adopted a very flexible cut-off point at a minimum of ten times per million words across five or more texts. They found that 3-word bundles occurred over 60,000 times and 4-word bundles over 5,000 times per million tokens, accounting for approximately 21% of the 5.3 million words of the academic section of the Longman Corpus.

Biber et al. (2004), were more cautious in choosing lexical bundles from their corpora by setting a relatively high frequency cut-off at 40 times per million words. Following Biber et al., Hyland (2008) increased the cut-off value 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, when selecting lexical bundles in his 3.5-million-word corpus of research articles, Master’s dissertations and PhD thesis.

Present-day n-gram programs ensure the properties of frequency and multi-text occurrences but fail to adequately deal with meaningful retrievals. Purely based on statistical figures, a phrase extractor may generate a long list of multi-word sequences, part of which have little meanings (e.g. that do not, and there being) or part of which are grammatically ill-formed (e.g. was found in the, of the distribution of). Though frequent, such word combinations may not be “pedagogically compelling” (Simpson-Valch and Ellis, 2010) p. 493).

To identify the most frequent spoken collocations, Shin and Nation (2008), p. 341) proposed a set of selection criteria, one of which was “grammatical well-formedness” and involved a great deal of manual checking. From the British National Corpus spoken section, they targeted a sequence of words which do not span “immediate constituents” (two neighboring phrases/clauses) (Bloomfield, 1933), p. 161), because a grammatical well-formed word sequence is a comprehensible unit. For instance, ‘the fact that’ is more understandable than ‘fact that the’, since the retrieval of the former follows the dividing principle of immediate constituents.

By compiling a 25-million-token corpus of research articles across five academic domains, Durrant (2009) endeavored to make a listing of positionally-variable collocations for students from a wide range of departments. Relying on the log-likelihood and Mutual Information, he identified the most frequent 1,000 academic collocations (e.g. respect to, number of, effect on, effects on, was used). However, some collocations fail to contribute to the learning of grammatical patterns if they are not extended to longer word sequences (e.g. was used). Some collocations can be combined into one for learning together (e.g. effect on, effects on), while others are apparently incomplete so that they are not suitable for direct teaching (e.g. respect to, number of).

To tackle the problem of teachability, Simpson-Valch and Ellis (2010) proposed the notion of Formula Teaching Worth (FTW) by incorporating the Mutual Information (MI) score into their weeding procedure in lieu of a merely frequency-based approach. MI is a statistical measure of the cohesiveness of words, which signifies the degree to which the words are bound together (Stubbts, 2007). In one of their cases, the word sequence ‘with which the’ occurred more frequently than expected (passing a certain threshold of both frequency and range). In contrast, the expression on the other hand cohered much more than would be expected by chance based on the high MI score. The
expression ‘with which the’ would come at the top if frequency is a top priority in ranking formulaic sequences, while on the other hand ranks high if the MI score is considered first. In the light of identifiable meanings, the latter seems to be more noteworthy for teaching than the former. After a series of reliability and validity checks, Simpson-Valch and Ellis (2010) concluded that the FTW that combines frequency and MI may provide teachers with a basis of prioritization, when judging multi-word sequences in terms of whether they are pedagogical compelling.

Also relevant to this study is cross-disciplinary Academic Collocation List (ACL). Ackermann and Chen (2013), compiled a corpus of over 25 million tokens from the Pearson International Corpus of Academic English (PICAE). Through MI and t-score as initial filtering and then a panel of experts for screening, they retrieved 2,468 most frequent lexical collocations, which were claimed to be immediately operationalizable for EAP teachers to help students increase collocational competence in academic English. Despite the relevance of the ACL for learners with academic goals, the ACL including free word combinations (e.g. further research, academic writing) seems to be so unwieldy as to possibly overburden students before they concentrate on the collocations they may need imminently.

The review of previous studies has helped to shape the present approach to selecting recurrent multi-word combinations for inclusion in the list for pedagogical purposes. 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 as a point of departure.

3. Research Method
3.1. The Corpus

The NOW (News on the Web) Corpus is the largest, well-balanced English news corpus to date. At the time of doing this research, it has already had 7.3 billion words of data retrieved from web-based newspapers from 2010 to the present time. Automated scripts run every day to add texts to the corpus, so the corpus is continually growing by 140 to 160 million words each month. Due to everyday update, the corpus reflects contemporaneity and modernity of English as time goes on. This has important implications for the learning of non-compositional multi-word expressions, since the very low frequencies in the NOW Corpus may indicate that these phrases may be of little pedagogical value.

3.2. The Procedure

The selection of recurrent multi-word combinations for inclusion in the list involved quantitative and qualitative approaches. The frequency measure resembled those of lexical bundles used in past studies in some ways. To lessen subjectivity, we referred to Shin and Nation (2008) as well as Ackermann and Chen (2013) and thereby formulated two questions to guide the judgment. They were used to gauge meaningfulness and well-formedness, after candidate multi-word sequences were initially identified.

The software Collocate Barlow (2004) was used to retrieve multi-word sequences from the downloaded NOW Corpus for offline use. The span parameter for multi-word length was set from 2 to 6 words. Frequencies drop drastically as word sequences are extended to five words or beyond (Hyland, 2008). Though recurrent 6-word combinations may be relatively rare, they were also included for thoroughness.

The next decision was what frequency level was to be used as a cut-off. Since there were other sifting measures, a less rigorous criterion was set to begin with, namely five times per million words. For a single word to enter the BNC first 5,000 most frequent word families, the word and its family members altogether need to occur at least 7.87 times per million words (Nation, 2012). Consequently, the cut-off was set at a minimum of five times rather than 10 to 40 times as in previous research (specifically, a minimum of 36,500 times as far as 7.3 billion words were concerned).

After the frequency-based measures, the strength of word co-occurrence was taken into account. There are several statistical measures to determine collocational strength. MI indicates the degree of mutual dependence of two or more words. The t score and log-likelihood ratio (LLR) are two measures of certainty of a word pairing. MI tends to give high scores to collocations having less frequent components but having strong associations between words, whereas the t score and LLR are sensitive to frequency in the sense that higher scores are associated with higher frequency of occurrence, and hence their scores are often high for functional/grammatical collocations. In consideration of possible multi-word combinations with less frequent constituent words, MI was adopted for ranking. MI complements the frequency measure. Frequency screening favors word sequences that may occur due to the high frequency of their components and may not have distinctive meanings (e.g. of which the). Since higher MI means greater association between words than is expected by chance, recurrent multi-word combinations with a high MI score are more likely to be meaningful. According to Hunston (2002), collocations with an MI score greater than 3 are considered strong. Therefore those candidate word sequences at the top of the ranked list by MI may be close to being integral in meaning. As a result, those multi-word combinations with both high frequency and high MI were first chosen while those appearing at the bottom of both frequency and MI rankings were removed. Multi-word combinations with the MI score lower than the default value (=3) were eliminated at this stage (e.g., with which can be).

Subsequently, meaningfulness, grammatical well-formedness and semantic non-compositionality guided manual checking. The multi-word combinations to be included in the list must have meaning(s) and can be learned as a whole. This criterion would help to make the present multi-word list comparable to an individual word list. To lessen subjectivity, four questions were used as selection criteria.

Q1. Does the candidate multi-word combination convey a meaning?
Q2. Does the candidate multi-word combination cross the boundary of an immediate constituent/phrase?
Q3. Does the construct of the candidate multi-word combination behave like an individual lexical item, which is unlikely to be further analyzed into the form-meaning link of its subparts?
Q4. Does the meaning of the candidate multi-word combination as a whole remain or marginally remain when each component word is decoded with its core meaning?

The researcher-teacher and her colleague made an independent judgment of each candidate word combination. 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 there was no agreement between the two raters or the answer was ‘not sure’, the entry was reserved for further examination.

For Q1 to Q4, a series of Cohen’s Kappa statistics were undertaken as inter-rater reliability tests. The k values were 0.91, 0.92, 0.87 and 0.89 respectively (all >0.80), revealing a substantial level of agreement between the two raters.

3.3. Data Processing

To make the list serve the pedagogical purpose, two major modifications were made. One revision was undertaken for partial overlap. It refers to a situation where a longer phrase was the combination of two or more shorter phrases, each of which could occur as an independent subset of the longer one. Take due to the/an absence of as an example again. One of its subset due to appeared 1,123,999 times, while the other three, the/an absence, absence of and the/an absence of appeared 68,255, 220,825 and 388,967 times respectively. The prepositional phrase due to may have been connected with other nouns or noun phrases other than the/an absence of. Similarly, the/an absence of was one of the combinations in connection with absence of, for example, a complete absence of, a total absence of and an absence of. The absence is a free word combination, so it was not included in the current list. Since the four phrases due to, the/an absence of, absence of and due to the/an absence of can stand alone as a meaningful unit, they were separately examined based on their respective occurring frequency for decision whether to be included in the list.

To make the list more compact, a word sequence in its usual form and its possible variants with the same meaning were combined. The examples include based on/upon, even if/though, and so on/forth, with the first word appearing more frequently than the second word.

To sum up, the selection of frequent non-compositional multi-word combinations involved the following sequence: (1) frequency (a minimum of 5 times per million words for initial screening), (2) cohesiveness of words for meaningfulness (MI>=3 and checked with Q1), (3) well-formedness (Q2), (4) non-decomposability (Q3) and (5) semantic non-compositionality (Q4). Step 1 resulted in an effective frequency threshold at having to occur over 36,500 times; Steps 2 and 3 led to effective MI greater than 6.

4. Results and Discussion

4.1. The Most Frequent Non-Compositional Multi-Word Combinations in English News Texts

A total of 318 non-compositional expressions of 2 to 5 words were ultimately chosen and formed the multi-word combinations list. The list consists of 153 two-word, 103 three-word, 56 four-word and 6 five-word phrasal expressions commonly used in English news articles.

The RANGE program (Healey et al., 2004) was used to examine the vocabulary levels of the individual word tokens of the frequent non-compositional multi-word combinations. 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). The multi-word combinations list consists of 869 running words and involves 335 word types as well as 298 word families. The BNC/COCA first 1,000 word families account for 87.72% of the total words in the present list and the second 1,000 make up 5.15%. The combined coverage percentage of the first 2,000 word families is 92.87%. The percentage of the third 1,000 word families is 2.42%, the third highest lexical coverage after the first 2,000 high-frequency word families. After the first 4,000 word families, the coverage percentage of additional 1,000 word families rapidly reduces to less than 1%.

As can be seen above, a large number of non-compositional multi-word combinations are composed of very general words, most of which (95.29%) are from the first 3,000 most frequent words in the BNC/COCA. The pairings or strings of content words (nouns, lexical verbs, adjectives or adverbs) and function words (determiners, conjunctions, prepositions, pronouns, auxiliary verbs, modals and quantifiers) form a common pattern in the present list, for example, much as (=though), as well as, in order to, there + be, and to do with. Among the instances, 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 text, as Sinclair (1991) has claimed. Without specialist knowledge involved, these semantically non-compositional multi-word combinations occur across a wide range of subject areas with their high-frequency component words.

Concerning the structure of 2-word combinations, a vast majority of them (132 out of 153) are grammatically-conditioned pairs, namely a content word combined with a function word, as opposed to only 21 lexical collocations, a content word tied with a content word (e.g., simply put, no matter, so far, very few). Phrasal prepositions come second (26/132=19.7%) (e.g. as for, apart from, as per, according to), followed by the pattern a preposition + a noun (20/132=15.2%) (e.g. at once, at times, in place, in question), being the third.
The three patterns as \textasciitilde as, a \textasciitilde of, and by + noun phrase are productive among the 3-word combinations, as in the cases of \textit{as far as}, \textit{as much as}, \textit{as soon as}, \textit{a host of}, \textit{a range of}, \textit{a couple of}, \textit{by means of}, \textit{by way of} and \textit{by virtue of}. These three patterns contribute to the description of quantity, the coverage of a subject or an approach.

For 4-word sequences, the prepositional phrase is the most common structure, comprising 57\% of all forms in the category of 4-word combinations (\textasciitilde=32/56). They are, for instance, \textit{on one’s own account}, \textit{in the event of/that}, \textit{in the light of}, \textit{in the wake of}, with a view to, \textit{on the grounds of/that}.

In the present list, two 5-word combinations extended from 3-word combinations can still be semantically opaque, as shown in the instances of \textit{as far as...be concerned} and \textit{have nothing/much/little/something to do with}.

4.2. The Lexical Coverage of the Most Frequent Non-Compositional Multi-Word Combinations in the English News Corpus

The present multi-word combinations list contains a total of 318 phrases of 2 to 5 words with an accumulation of 33,917,223 individual instances and 101,751,907 running words, which makes up almost 2\% of the tokens in the English News Corpus.

At first sight, 2\% lexical coverage in the English News Corpus does not appear to be worth noticing. However, if not recognized, the non-compositional multi-word combinations may impede comprehension. Native English-speaking children view a vocabulary load of two unknown words per hundred words as difficult reading (Carver, 1994). Some scholars (Hu and Nation, 2000; Schmitt \textit{et al.}, 2011) regard one unknown word in every fifty words (98\% lexical coverage) as the minimum threshold necessary for adequate comprehension. If 2\% unknown words are a critical benchmark for unassisted understanding of a text, then the present non-compositional multi-word combinations should not be neglected. As such, the researcher would like to propose the inclusion of the non-compositional multi-word combinations in English for Journalism syllabi.

5. Conclusion
5.1. Findings

The major concern of this research was to create a semantically non-compositional subset of formulaic language for English for Journalism learners for receptive use. By means of a principled set of criteria, a total of 318 multi-word combinations of 2 to 5 words were selected and they made up 2\% of the total words in the English News Corpus. The present list contains the most widely-used phrases across various everyday topics. As high as 95.29\% of the non-compositional multi-word combinations are made of the BNC/COCA first 3,000 word families. Accordingly, the present selected multi-word combinations can bridge the gap between the lexical coverage that the most general words can and cannot account for in a text. Irrespective of topic areas, English news readers may come across these phrases while reading everyday news. The present multi-word combinations list is short and may be a viable option for English for Journalism learners to learn in a short time.

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

5.2. Pedagogical Implications

Although the present multi-word combinations list provides a window to the Journalistic register, itemized phrasal expressions are still not enough for EFL undergraduates. As with the learning of individual words, the non-compositional multi-word combinations should be learned in context rather than in isolation. English for Journalism 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 \url{http://www.lextutor.ca/concordancers}; NOW at \url{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 of certain semantically non-compositional phrases. Classroom exercises using concordances may be undertaken, for instance, in gap-fill exercises. With more exposure to English news, EFL undergraduates will consolidate the lexical knowledge acquired from the present opaque multi-word combinations list.

References

Ackermann, K. and Chen, Y. (2013). Developing the academic collocation list (ACL)-A corpus-driven and expert-judged approach. \textit{Journal of English for Academic Purposes}, 12(4): 235-47.

Altenberg, B. (1993). Recurrent verb-complement constructions in the London Lund Corpus. J. Aarts, P. de Haan and N. Oostdijk, Eds., \textit{English language corpora: Design, analysis and exploitation}. Rodopi: Amsterdam. 227-45.

Altenberg, B. (1998). \textit{On the phraseology of spoken English: The evidence of recurrent word combinations}. A. P. Cowie Ed., \textit{Phraseology: Theory, analysis and applications}. Oxford University Press: Oxford. 101-22.

Barlow, M. (2004). \textit{Collocate, Computer software}. Athelstan: Houston.

\url{http://athel.com/product_info.php?products_id=29}
Biber, D., Conrad, S. and Cortes, V. (2003). *Lexical bundles in speech and writing: An initial taxonomy*. A. Wilson, P. Rayson and T. McEnery Eds., *Corpus linguistics by the Lane: A festschrift for Georffrey Leech*. Peter Lang: Frankfurt. 71-93.

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

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

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): 413-37.

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

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, 28(3): 157-69.

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

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

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, 19(1): 24-44.

Hu, M. and Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1): 403-30.

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

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

Martinez, R. and Murphy, V. A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly*, 45(2): 267-90.

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

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. Available: [http://www.victoria.ac.nz/lal/about/staff/paul-nation](http://www.victoria.ac.nz/lal/about/staff/paul-nation)

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

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

Pawley, A. and Syder, F. H. (1983). *Two puzzles for linguistic theory: nativelike selection and nativelike fluency*. In J. C. Richards and R. W. Schmidt (Eds.), *Language and communication*. Longman: London. 191-225.

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

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

Scott, M. (1996). *WordSmith tools; Computer software*. Oxford University Press: Oxford. Available: [http://lexically.net/wordsmith/downloads](http://lexically.net/wordsmith/downloads)

Shin, D. and Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal*, 62(4): 339-48.

Simpson-Valch, R. and 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 University Press: Oxford.

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

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press: Cambridge.
Appendix
The most frequent non-compositional multi-word combinations in the 7.3-billion-word English News Corpus in alphabetical order

| 2-word combinations | Freq.  | 3-word combinations | Freq.  |
|---------------------|--------|---------------------|--------|
| a bit               | 138806 | a bit of            | 179945 |
| a few               | 1274641| a couple of         | 122901 |
| a little            | 352916 | a good deal         | 72157  |
| a lot               | 493480 | a great deal        | 237189 |
| a priori            | 80337  | a handful of        | 91697  |
| above all           | 156680 | a host of           | 111389 |
| according to        | 2146228| a little bit        | 58903  |
| across from         | 38908  | a lot of            | 341555 |
| ad hoc              | 79806  | a number of/ a ... number of | 1404455 |
| after all           | 295204 | a plethora of       | 48602  |
| all along           | 53071  | a range of/ a ... range of | 541881 |
| all but             | 152742 | a wealth of         | 54339  |
| all over            | 150318 | all manner of       | 39211  |
| all right           | 51708  | among other things  | 96317  |
| all too             | 99271  | and so forth        | 122901 |
| along with          | 678663 | and so on           | 194852 |
| among others        | 114115 | as a means          | 207878 |
| and ... alike       | 114191 | as a rule           | 39590  |
| and ... respectively| 72611  | as it is            | 41483  |
| any longer          | 43907  | as opposed to       | 232872 |
| apart from          | 188187 | as soon as          | 139563 |
| as for              | 229691 | as well as          | 2596488|
| as if               | 447512 | as/so far as        | 220678 |
| as of               | 162057 | as/so long as       | 318985 |
| as per              | 38226  | as...as possible    | 607394 |
| as regards          | 46255  | at odds with        | 75565  |
| as such             | 302929 | be about to         | 92985  |
| as though           | 126233 | be subject to       | 139790 |
| as to               | 727362 | be to blame         | 65335  |
| as well             | 750689 | by all accounts     | 36939  |
| as with             | 238325 | by and large        | 58676  |
| as yet              | 76625  | by means of         | 160315 |
| aside from          | 87153  | by no means         | 110934 |
| assuming (that)     | 261728 | by the way          | 58524  |
| at all              | 690023 | by virtue of        | 101922 |
| at once             | 152060 | course of events    | 36787  |
| at present          | 125400 | course(s) of action | 67840  |
| at stake            | 106163 | every bit as        | 36636  |
| at times            | 213937 | for a while         | 64735  |
| before long         | 38378  | for the sake of     | 106845 |
| bonne fide(s)       | 40574  | have got to         | 37621  |
| by far              | 84654  | have to do with     | 116085 |
| close to            | 327998 | have~ bearing on    | 54813  |
| courtesy of         | 136307 | in a fashion        | 36863  |
| cutting edge/cutting-edge | 57842 | in a manner        | 118205 |
| due to              | 1123999| in a nutshell       | 36409  |
| each other          | 681465 | in a row            | 41710  |
| et al               | 3491856| in a sense          | 89198  |
| even if             | 558922 | in accord with      | 54510  |
| even though         | 598154 | in accordance with   | 169177 |
| every other         | 79655  | in addition to      | 673210 |
| far from            | 256502 | in any case         | 112146 |
| follow suit         | 45649  | in case of          | 50799  |
| for good            | 89198  | in charge of        | 92076  |
| for life            | 83139  | in compliance with  | 45573  |
| free from           | 81927  | in favo(u)r of      | 274452 |
| free of             | 122371 | in lieu of          | 47921  |
| Expression | Frequency | Description |
|------------|-----------|-------------|
| given that | 210832    | in line with |
| granted that | 45573    | in one's favor |
| had better | 45119     | in order that |
| have to | 2399949   | in order to |
| high end/high-end | 48451 | in place of |
| if only | 100861    | in regard to |
| in case | 87077     | in respect [of/to] |
| in charge | 50875     | in return for |
| in place | 305959    | in spite of |
| in point | 67764     | in terms of |
| in practice | 199093   | in the way |
| in question | 153045   | in this regard |
| in return | 103295    | in this respect |
| in short | 245671    | in view of |
| in that | 795147    | kind of |
| in time | 231130    | little [is/was] known about |
| in turn | 466976    | little more than |
| in view | 37393     | may as well |
| insofar as | 117069   | no less than |
| instead of | 530217   | no more than |
| irrespective of | 75641  | not ... the least |
| just as | 559603    | not... at all |
| let alone | 94802     | of a kind |
| lots of | 98741     | on account of |
| may well | 227722    | on behalf of |
| much less | 69355     | on one's behalf |
| next to | 147516    | on one's own |
| no idea | 62690     | on top of |
| no longer | 747887    | over and over |
| no matter | 195760    | point(s) of view |
| no more | 139639    | pros and cons |
| no point | 40802     | quite a few |
| not ... altogether | 49360 | range from…to |
| not yet/ not … yet | 393208 | rule(s) of thumb |
| nothing but | 85941 | so as to |
| now that | 116691    | sort of |
| of course | 986157    | the bulk of |
| of late | 46785     | the rest of |
| of sorts | 45649     | the/an absence of |
| on account | 53374     | there … to be |
| on board | 62841     | to a degree |
| on earth | 138276    | to do with |
| once more | 70566     | to the point |
| one another | 369805 | with reference to |
| only if | 150621    | with regard to |
| or otherwise | 98514     | with respect to |
| or so | 140624    | |
| other than | 358748    | 4-word combinations |
| out of | 1680821   | a case in point |
| owing to | 80109     | a good deal of |
| per capita | 161148   | a great deal of |
| per se | 111086    | as a result of |
| prior to | 622920    | as if it were |
| provided/ providing that | 215755 | at the expense of |
| rather than | 2070188 | at the mercy of |
| regardless of | 412142 | be to blame for |
| relative to | 293462 | by the same token |
| right away | 47088   | can not help but |
| short of | 116388    | come to terms with |
| simply put | 52617     | come/get to grips with |
| so far | 288690    | from time to time |
| so that | 977599    | give a ... account of |

Freq.
| **5-word combinations** | **Freq.** |
|-------------------------|----------|
| as a matter of course   | 34515    |
| as so far as ~be        | 61705    |
| be that as it may       | 31865    |
| have [quantifier] to do with | 159861 |
| in a manner of speaking | 30501    |
| in the last couple of   | 30880    |

| **Frequency** | **Meaning** |
|---------------|-------------|
| 878307        | so...that   |
| 108359        | specific to |
| 39135         | subject ... to |
| 178417        | subject matter to |
| 378591        | subject to |
| 4636856       | such as |
| 164405        | such that |
| 510526        | such...that |
| 45800         | suit(s) against |
| 79428         | suppose/supposing that |
| 157134        | thanks to |
| 804236        | that is, |
| 160694        | the few |
| 461220        | the former |
| 576796        | the latter |
| 102679        | the odd(s) |
| 154256        | the others |
| 487426        | the rest |
| 7435508       | there + be |
| 58600         | third party |
| 216058        | to date |
| 95257         | to death |
| 32557         | too...to |
| 74505         | top-down/ top down |
| 42241         | unless otherwise |
| 853389        | up to |
| 70491         | used to |
| 62841         | vantage point(s) |
| 135625        | very few |
| 89198         | vice versa |
| 343600        | welling-being/ well being |
| 267332        | would like |
| 56479         | would rather |
| 169782        | yet to |
| 159861        | have [quantifier] to do with |
| 30501         | in a manner of speaking |
| 30880         | in the last couple of |
| 88213         | in the event of |
| 77761         | in the aftermath of |
| 69960         | in one's own right |
| 63674         | in the absence of |
| 183415        | in the absence of such as |
| 214695        | in the face of |
| 40196         | in the first instance |
| 74808         | in the interest(s) of |
| 75489         | in the way of |
| 211287        | in/over the course of |
| 40802         | make a point of |
| 39438         | make the most of |
| 46633         | no choice but to |
| 1038643       | not only...but also |
| 47845         | on one's own terms |
| 44285         | on the ground(s) |
| 81473         | on the one hand |
| 607318        | on the other hand |
| 42922         | once and for all |
| 37318         | once in a while |
| 37090         | out of the question |
| 50344         | put it another way |
| 47391         | so as not to |
| 42619         | take the place of |
| 72460         | that is to say |
| 276345        | the extent to which |
| 47618         | the more...the less |
| 88213         | the more...the more |
| 42922         | the other way |
| 48072         | with a view to |