On How to Identify Useful Collocations and the Multi-word Units They Occur In

James Rogers

Abstract

There is growing awareness in language education of the importance of collocation and formulaic language knowledge. Such knowledge is viewed as being essential to achieving fluency in a foreign language. However, various issues have led to a lack of research and resources despite the awareness of the value of such knowledge. Thus, educators and students alike are impeded when they approach the direct learning of this essential feature of language. The number of collocations can run into the hundreds of thousands, and thus it is difficult to narrow down the most useful collocations. This study will give a brief overview of previous collocation/formulaic language research to give insight into identifying the most useful items, and then present a step-by-step methodology to do so.

Keywords

collocation, formulaic language, multi-word units, high-frequency vocabulary, corpora

Introduction

Recently, more researchers are recognizing the value of collocations for second language learners. Lewis (2000) states, "Teaching collocation should be a top priority in every language course" (p. 8). This view stems from the realization that much of language consists of prefabricated chunks, and that collocation is one of the most important kinds of chunks. Thus collocation and the formulaic language, or *multi-word units* (MWUs), are both central to fluency.

Hoey (2005) and Hill (2000) agree that such knowledge plays a central role in language.

The central role that collocations/MWUs plays in helping second language learners attain fluency is multi-faceted. Multiple researchers cite how competent use of formulaic language helps the language learner to sound natural (Durrant & Schmitt; 2009, Wray, 2002; Cowie, 1998). In addition to aiding learners in making more native-like selections, the use of collocation has been shown to make language processing more efficient (de Glopper, 2002; Nation, 2001).

Despite teachers realizing the importance of collocations, their learners have still struggled to obtain collocational fluency (DeCock et al., 1998; Kallkvist, 1998; Waller, 1993). For example, Nesselhauf (2005) examined a 150,000 token learner corpus written by advanced German learners of English, and found that a quarter of the 2,000 verb-noun collocations were wrong, and a third deviant. These findings indicate that even advanced students still struggle with collocations, and this is a major barrier towards obtaining native-like fluency.

There are many factors that prevent students from attaining collocational fluency. Not only are collocations/MWUs a complex phenomenon (Hill, Lewis & Lewis, 2000), there is also a severe lack of emphasis on them (Gitsaki, 1996; Nesselhauf, 2005). One reason why practitioners do not emphasize collocations/MWUs despite being aware of their importance is that there are still very few studies that identify which are the most frequent (Durrant & Schmitt, 2009), and the studies that have been conducted all lack in comprehensiveness or are flawed in some way. Many limit their scope to a specific type of multi-word unit. For instance, Biber, Conrad, and Cortes (2004) only found 172 ‘lexical bundles’, limiting themselves by a very conservative cut-off of 40 occurrences per million and only considering four-word sequences. Simpson and Mendis’ (2003) search for fixed, institutionalized, semantically opaque, academic idioms only identified 238 such items. Aghbar (1990) and Bahns and Eldaw (1993) only

examined verb-noun collocations, while Channell (1981) only examined adjective-noun collocations. These studies produce results in stark contrast with claims that there are tens of thousands of collocations in the native lexicon (Bahns, 1993) or even hundreds of thousands (Hill, 2000). While there is an abundance of collocation dictionaries available, they tend to present users with too much information. For instance, Kjellmer's (1994) collocation dictionary contains over 85,000 entries, and pinpointing the most useful collocations from such a large dataset is clearly not an easy task. This lack of resources that specify useful collocations is thus clearly connected to the sheer number of items researchers must deal with. Shin's (2006) study was a good first step in alleviating these issues, but his study was limited by only examining the most frequent 1,000 types of English. Thus a more comprehensive list is still needed.

In addition to the lack of a comprehensive list, many questions remain as to the criteria and resources that should be used to create one. This study will thus provide a step-by-step process by which the most common collocations and the MWUs they occur in can be identified.

Background to the research problem

Defining collocations and MWUs

One weakness in many of the previous studies results from the lack of consensus on defining what a collocation is. For example, many researchers define collocations by their tendency to frequently co-occur (Hoey, 1991; Jones and Sinclair, 1974; Firth, 1957), while others use syntactic structures (Gitsaki, 1996; Zhang, 1993). Some researchers even resort to a combination of both frequency data and syntactic patterning to identify collocations (Lesniewska & Witalisz, 2007). The numerous terms used to describe MWUs, such as "'combinations of lexical items" (Korosadowicz-Struzynska, 1980), "conventionalized language forms" (Yorio, 1980),

"prefabricated language chunks and routinized formulas" (Nattinger&DeCarrico, 1992), "phrase patterns and sentence patterns" (Twaddell, 1973), and "fixed expressions" (Alexander 1984; Kennedy, 1990), are also problematic in that they often overlap in what they described. Shin (2006) and Cowan (1989) both stated that there is too much variability in researchers' definitions of 'collocation'. However, appropriate terminology alone will not solve this issue in that, as we will see below, there is a significant amount of variability in collocation/MWU type, and it is difficult if not impossible to create a single unifying definition.

Approaches to researching collocations/MWUs

The three main approaches to studying collocations are semantic, structural and lexical. In the *semantic approach*, collocation is viewed from a standpoint of being predictable by its semantic features (Robins, 1967). This approach aims to explain why particular lexical items occurred only with certain others. Gitsaki (1996, p. 35) points out that a weakness of this approach is that, "There is a large number of idiosyncratic co-occurrences or combinations that are arbitrarily restricted…they are left unexplained and marginal by semanticists." Gitsaki (1996) lists some examples, such as how *kick the bucket* and *blond hair* can only be used when referring to humans (p. 33). Lewis (2000) agrees, in that trying to use semantics to explain why certain words co-occur leads to, at best, "half-truths" (p. 13).

Meanwhile the *structural approach* utilizes grammatical patterns to explain collocation, and proponents believe that collocation is influenced by structure. Mitchell (1971) proposes that collocation be studied within these "grammatical matrices" (p. 48). Gitsaki (1996) agrees, in that his study of 275 Greek learners of English at three separate proficiency levels showed that the learners did not once use a number of particular collocation patterns, such as *adverb+adjective,*

and that these were avoided due to their structural and syntactic complexity and relative infrequency in English. However, Hill (2000) distances himself from "previously cherished structuralist ideas" and believes that instead of breaking down language into smaller and smaller categories, we should try to view language in the largest units possible (p. 48). Thus, this statement leads us to the lexical approach.

Regarding the *lexical approach,* Halliday and Sinclair (1966) begin to consider lexis as separate, but complementary to grammatical theory. They believe that it is necessary to consider collocation's influence on the organization of language because grammar alone was not enough to determine which lexical item would occur due to the idiosyncratic nature of collocations. Halliday (1966) cites how word choice can also be specified by collocational restrictions, in addition to structural and semantic limitations (p. 152). He gives the example of how *strong* is a member of a lexical set with *tea*, and *powerful* is a member of a lexical set with *car,* which cannot be explained by the structural or semantic approaches. Lewis (1993) states that language "consists of grammaticalised lexis, not lexicalized grammar" (p. vi). The *lexical approach* thus views lexis, and not grammar, as the overarching engine that organizes language.

Each of these approaches has its strengths and weaknesses, and their usage depends on the type of research being conducted. However the lexical approach does have advantages over the semantic and structural approaches, as is evident in figure 1. Figure 1 shows how the semantic approach would not identify *politics*, *character, himself,* or *herself* as collocates of the verb play since the items do not fall into the most common semantic groups. The weakness of the structural approach is also revealed in that it may not list less common grammatical categories, such as pronouns. Although each approach has its place in collocation research, the above examples highlight the significant advantages of the lexical approach.

Collocates of the verb *play*	Semantic Approach	Structural Approach	Lexical Approach
play {sports}/{instruments}/ {music}/{games}	O	O	O
play politics / play a character	X	O	O
play himself / play herself	X	X	O

Figure 1. The approaches' ability to identify common collocates of the verb 'play'

The value of collocation/MWUs

Collocational knowledge is of obvious value for the language learner. It has been referred to as a "decisive factor in developing fluency" (Almela & Sanchez, 2007, p. 37), awareness of it a matter of "first-rate importance" (McCarthy, 1984, p. 21), and essential even in early stages of language learning (Saville-Troike, 1984). Nation (2001a) states that a variety of knowledge is necessary to truly 'know' a word. This 'vocabulary depth' knowledge includes not only includes semantics, but also a word's pronunciation, orthography, word parts, concepts, associations, grammar, constraints on use, and possible collocates.

Learning collocation in comparison with isolated words has been found to actually be easier (Ellis, 2001; Lewis, 2000; Taylor, 1983). This may be because when learners utilize prefabricated language they are freeing up processing time (Almela & Sanchez, 2007; Lewis, 1993; Nation, 2001). For example, Bogaards (2001) found that multiword expressions

containing familiar words were retained 10% more than completely new single words immediately after a learning session and also 12.1% more in a delayed posttest 3 weeks later. Furthermore Furukawa et al. (1998) found that teaching students to utilize a chunking learning strategy improved 6th grade students' Stanford Achievement test scores by an average of 6.15 points. Sinclair (1991) refers to this in his 'idiom principle' as making "fewer and larger choices" (p. 113).

These studies highlight the value of collocational knowledge in language processing efficiency. This justifies greater emphasis on developing collocational fluency, such as with the direct teaching of the most useful collocations.

Regarding the direct teaching of collocations/MWUs

Although rote learning is dismissed by many as outdated, the direct teaching of certain collocations/MWUs may still be advantageous. Sokemen (1997) remarks that the anathema towards rote learning has actually led to a decrease in acquisition speed, and that now the pendulum is swinging back towards the middle for a more balanced approach. Shin (2006) agrees, stating that deliberate learning itself is not a problem, but rather a "lack of balance with other ways of learning" (p. 163).

Foremost teachers need to expose students to useful collocations, thus enabling students to fully acquire them. However, Nesselhauf's (2005) study reveals that exposure alone is insufficient. She argues that the direct teaching of collocations is essential for developing fluency. Likewise Doughty and Williams (1998), Ellis (1994) and Newman (1988) all argue that collocations should be taught directly. If encounters are left to chance, then as Wollard (2000) states, "Learning will be extremely haphazard and inefficient" (p. 26). Lewis (2000) remarks

that it may be weeks, months, or even years before students re-encounter a particular collocation. Gairns and Redman (1986) note that the most common way teachers deal with collocations is as they appear in the textbooks they use, and state that this is not ideal, if effective at all. Lewis (2000) and Wollard (2000) also agree, stating that directly focusing on collocations will bring students' attention to very high frequency words that they are already familiar with but do not realize are actually occurring formulaically.

Types of collocations/MWUs

Earlier research done without access to large corpora or computers defined collocations by their phraseological features. Some distinguish collocations from idioms (Liang, 1991), while others subdivide collocations in literals, figuratives, and core idioms (Grant & Nation, 2006). Biber, Johansson, Conrad, Leech, and Finegan (1999) deem collocations two-word phrases that co-occur, distinguishing them from idioms and lexical bundles, while others consider two *or more* frequently co-occurring words to be collocations (Conzett, 2000). Many researchers also delimit collocational searches by only considering content words (Woolard, 2000). Nesselhauf (2005) believes that "different types of collocations need different types of treatment," and thus a collocation's semantic transparency and congruency with the L1 must be considered (p. 271). Gitsaki (1996) and Shin (2006) agree.

There are issues involved with counting occurrences of collocations, such as considering 'constituent variation'. In other words, counting co-occurrences of *A* and *B* when they occur as *AB*, but also when they occur as *ACB*. Thus researchers such as Renouf and Sinclair (1991) use syntactic frameworks to grapple with such discontinuous sequences. Wilks (2005) used a more advanced approach by utilizing 'skipgram' searches, which can handle constituency variation.

For example, it could be argued that *close friends* and *close childhood friends* should be counted together since they are essentially the same collocation albeit with an adjective added. Cheng, Greaves, and Warren's (2006) 'concgramming' method was also a major advance, in that it counted co-occurrence not only by considering constituent variation, but also positional variation (*AB* and *BA*). This method thus counts instances of *childhood friends* and *friends from childhood* together. They believed that "searches which focus on contiguous collocations present an incomplete picture of the word associations that exist" (Cheng, Greaves, & Warren, 2006) in that the majority of their study's collocations were non-contiguous, showing both constituency and positional variation (p. 431).

Clearly there is a substantial amount of variability in the types of collocations/MWUs. However, despite difficulty in categorizing such types, all useful collocations exhibit a high-frequency of co-occurrence. The above studies elucidate how consideration for both constituent and positional variation, in addition to frequency of co-occurrence, helps to reveal patterns of co-occurrence.

Identifying collocations through corpora

One of the most useful resources for identifying common collocations/MWUs is corpora (Meijs, 1992; Noel, 1992; Francis, 1993). Shin (2006) states that a large corpus with a large variety of texts is essential for producing reliable data. Thus Kjellmer's (1994) use of the 1-million-token *Brown Corpus* (Nelson and Kucera, 1979) may not have produced the most reliable data, despite it being one of the largest, reliable corpora at the time. Through computer technology larger and larger corpora have been compiled. In recent years, many researchers have relied on the 100-million-token *British National Corpus*, or *BNC*, for collocational research (e.g., Durrant &

Schmitt, 2009; Shin, 2006). However, the *BNC* stopped being developed in 1993 and has been referred to as being past its sell-by date (Kilgarriff, Atkins & Rundell; 2007). Davies' (2008) *Corpus of Contemporary American English*, or *COCA*, can be considered a better choice as it is four times larger than the *BNC*, and it is still being added to today. Furthermore, it has a wide and balanced dispersion in regards to genres and spoken versus written content, and while not downloadable due to copyright, its data can be freely analyzed via an online interface.

Criteria for useful collocation/MWU identification

There are various criteria that can help identify common collocations/MWUs. Although previous researchers have not consistently used all of the following criteria, each criterion has been proven valuable in identifying useful collocations/MWUs.

Frequency

Because of their large number, determining a frequency cut-off is a necessary step in identifying the most useful collocations/MWUs to directly teach/study. Biber, Conrad, and Cortes (2004) set a self-admittedly conservative cut-off at 40 occurrences per million. Cortez (2002) limited examined items to 20 occurrences per million tokens, Biber, et al. (1999) considered up to ten occurrences, Shin (2006) examined as low as three occurrences, and Kjellmer (1987) collected data for items occurring two times per million. But questions still remain as to how low a frequency cut-off can go and still contain mostly useful collocations.

Dispersion

Both Kjellmer (1984) and Nation (2001a) state that a collocation's range, or balanced dispersion

in many different categories of text, is a necessary criterion for identifying useful collocations. Considering dispersion can thus help to identify collocations/MWUs with general usefulness. Dispersion data can also be utilized in the opposite sense to help identify items that have a limited range, and thus have value to those in a specialized area, such as teachers/students who are studying English for specific purposes.

L1

Researchers must also consider L1-L2 '*collocational congruency*', or how similar/dissimilar a collocation/MWU's translation is in the learner's native tongue. Gitsaki (1996) highlights how "In English people 'draw conclusions' while the Greeks 'bga;zounsumpera;smata' [take out conclusions]" (p. 3-4). Both Nesselhauf (2005) and Feyez-Hussein (1990) found that approximately 50% of collocational errors were due to L1 influence, and thus such items should receive more teaching time. Al-Zahrani (1998), Bahns (1993), and Biskup (1992) all call for increased emphasis on non-congruent collocations.

Semantic transparency

Semantic transparency, or how literal/figurative a collocation/MWU is, is yet another criterion which must be considered. Shin (2006) states, "Different categories of multi-word units need to be treated in a different way when they are taught and learned" (p. 141). Gitsaki (1996) cites semantically opaque examples, such as 'foot the bill' and 'high explosive', and their obvious potential to mislead (p. 49). Thus whether an item is semantically transparent, and whether students are aware of this, can affect a collocation's learning burden. Items with high learning burdens thus deserve more classroom time.

Colligation

Colligation refers to when it may be better to consider one or more components of a multi-word unit not simply as a single lexical item, but rather as an interchangeable grammatical category or set of lexical items, as Gitsaki (1996) does. There are varying types of collocation that may benefit from such an analysis. For instance, certain collocational pairs that include numbers can have each instance of a different number counted together, since replacing one number for another does not alter the meaning. If such a step is taken, co-occurrence can more accurately be recognized. Thus, when appropriate, colligation should be considered to help identify the MWU that best represents how a collocational pair co-occurs.

Research Question

What steps can be taken to identify the most useful collocations, and the MWUs they occur in for the purpose of direct study?

Methodology

Materials

In this study, the source for collocational lemma pairs was Davies' (2010) *word list plus collocates*, which consists of 739,255 lemma pairs. In addition, Davies' (2008) *COCA* was also used to collect concordance and dispersion data. The concordance software *Antconc* (Anthony, 2011) and the text editor software *Textcrawler* (2011) identified colligational issues.

Procedure

Davies' (2010) collocation list was utilized as a starting point and a frequency cut-off was set. Although frequency cut-offs are ultimately "arbitrary" (Nation, 2001a, p.180), a practical limit must be set. Nation (2001a) suggests 2,000 word families as "practical and feasible" (p.96) in regards to direct teaching, while Nation (2001b) suggests a limit of 3,000 word families. Thus assuming the collocations selected were deemed useful, this study aimed for 2-3,000 word families.

A number of frequency cut-offs were piloted to determine how many useful collocations there were at each level. The study began at the highest cut-off set by Biber, Conrad, and Cortes (2004) of 40 occurrences per million tokens, and progressed to Kjellmer's (1987) two occurrences per million. Then the 25,000 word family BNC and COCA list in the *vocabprofile* program (Cobb, 2013) was utilized to determine how many word families the collocations consisted of to ensure that those selected did not exceed 3,000 word families.

Only content words (i.e., nouns, verbs, adjectives, adverbs) were considered. Duplicate entries were also removed, since often the collocation that occurred was a node word itself within the most frequent 5,000 lemma of Davies' (2010). The 'usefulness' of a sample of the pairs were then judged by a native speaker to ensure that the list was not overly inclusive. A sample of the resulting list was then examined for dispersion, L1 congruency, semantic transparency, and colligational issues. Finally, a sample of pairs were selected and input into the *COCA*'s online interface to extract example sentences. 1,000 example sentences occurring from 1990-2009 were collected for each pair and were rendered into a text file, which was then processed with the concordance software to help identify the most common MWU the collocational pairs occurred in. It should be noted that sentences occurring between 2010 and

2012 were excluded to ensure study replicability; the *COCA* divides data into four-year sections, and 2010-2014 was yet to be completed at the time of this study.

Results

The cut-off of two occurrences per million tokens utilized in Davies' (2010) resulted in a list of lemma pairs consisting of only 1,671 families. It was thus determined that a more inclusive cut-off could be considered given the pedagogically feasible goal of teaching between 2,000 and 3,000 word families. Pairs occurring once per million tokens consisted of 2,540 families, and pairs occurring once per 500,000 tokens consisted of 4,122 families. Therefore, the cut-off of one occurrence per million tokens was determined to be ideal.

When the lemma pairs remaining at this cut-off point were processed with the *vocabprofile* program, it was found that these covered 75.6 percent of the top 3,000 word families. Also of note is the fact that 97.8 percent of the tokens in the lemma pair list occur within the top 3,000 word families. A more detailed breakdown of the data can be seen in figure 2 below.

Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	806 (32.59)	1095 (38.17)	17461 (69.15)	69.15
K-2 Words :	704 (28.47)	847 (29.52)	4945 (19.58)	88.73
K-3 Words :	595 (24.06)	660 (23.00)	2280 (9.03)	97.76
K-4 Words :	207 (8.37)	211 (7.35)	302 (1.20)	98.96
K-5 Words :	91 (3.68)	91 (3.17)	104 (0.41)	99.37

K-6 Words :	38 (1.54)	40 (1.39)	46 (0.18)	99.55
K-7 Words :	13 (0.53)	13 (0.45)	13 (0.05)	99.60
K-8 Words :	9 (0.36)	9 (0.31)	10 (0.04)	99.64
K-9 Words :	4 (0.16)	4 (0.14)	4 (0.02)	99.66
K-10 Words :				
K-11 Words :	2 (0.08)	2 (0.07)	2 (0.01)	99.67
K-12 Words :	2 (0.08)	2 (0.07)	2 (0.01)	99.68
K-13 Words :	1 (0.04)	1 (0.03)	1 (0.00)	
K-14 Words :	1 (0.04)	1 (0.03)	1 (0.00)	
K-15 Words :				
K-16 Words :				
K-17 Words :				
K-18 Words :				
K-19 Words :				
K-20 Words :				
K-21 Words :				
K-22 Words :				
K-23 Words :				
K-24 Words :				
K-25 Words :				
Off-List:	??	44 (1.53)	80 (0.32)	100.00
Total (unrounded)	2473+?	2869 (100)	25251 (100)	100.00

Figure 2. Word frequency breakdown of lemma pairs occurring once per million tokens according to *vocabprofile's* 25,000 word families of the BNC and COCA

It was found that 40,196 pairs occurred once per million tokens. Due to the large number of items, this list was simply scanned by an experienced, native-speaking teacher of English for usefulness, and the vast majority were found useful and worthy of direct teaching. However, some pairs had unbalanced dispersion, with the vast majority of their occurrences falling into one particular genre. These items have limited usefulness for learners of general English and were thus excluded. After removing such items, as well as duplicate entries, and only considering content words, 12,604 pairs remained. The concordance software used in this study also proved highly useful in identifying the most common MWU that these collocational pairs occurred in, often identifying MWUs that are difficult even for a native speaker to recall.

Due to time limitations, a small sample of these pairs was then scanned for possible ways in which the number of items could be reduced. It was found that dispersion, L1 congruency, semantic transparency, and colligation consideration were all viable methods to delimit target items. These criteria truly helped to identify the most common MWUs in which collocations occur, and helped pinpoint items which pose a higher learning burden and items which could be considered as more useful than others depending on the context of the learning environment. These steps are discussed in more detail below.

Discussion

One of the necessary steps to identifying high-frequency collocations/MWUs is to set a frequency cut-off. The frequency cut-off utilized in this study resulted in very good coverage of high-frequency vocabulary, in that 94.75 percent of the lemma pairs identified fell within the top 3,000 word families. The lemma pairs also exhibited good coverage of the top 3,000 word families, with 75.6 percent of the word families being represented in the lemma pair list.

However, the large number of items identified presents a challenge. The vast majority of items were deemed useful, even in the lower frequency range of one occurrence per million running words. In fact, this study found that useful collocations can still be found as low as one occurrence per hundred thousand tokens, such as *nice/vacation*, *finish/workout*, and *tend/exaggerate* (Davies, 2010). However setting a more inclusive frequency cut-off would then create a list consisting of more than 2-3,000 word families, which would not be practical in terms of direct instruction. This abundance of useful items poses a serious barrier both research and the study of collocation/MWUs. Therefore, further steps to focus on items with higher learning burdens, or items that have more usefulness for specific learning contexts, must be taken. Such steps include dispersion data analysis, L1 congruency analysis, and semantic transparency analysis.

While analyzing the full list of collocational pairs, dispersion data from the *COCA* (Davies, 2008) was collected for specific pairs that seemed potentially unbalanced. For instance, dispersion data for the pair *respondent*/*indicate* indicated that the two items frequently collocate, but were highly unbalanced in regards to distribution. The *COCA* divides its corpus into spoken, fiction, magazine, newspaper, and academic genres, and 97.2% of the occurrences for these two items occurred within the academic genre. Such a pair, while useful in the academic realm, would be excluded from collocates with well-balanced dispersion. However those involved with

academic writing or scientifically related English would find such a pair of high value. This highlights how dispersion data can help to further identify the most useful items high-frequency collocations.

Some high-frequency collocations exhibited L1 non-congruency. Not only was there non-congruency in idiomatic collocational pairs, but also in semi-figurative and even literal collocations. For instance, there is no direct translation for the collocational pair *double/standard*, which occurred within the frequency cut-off. The best translation would be 偽善者 [*gizensha*], of which the meaning is closer to the idiom *a wolf in sheep's clothing* rather than *double standard.* Non-congruent, semi-figurative collocational pairs included *take/medicine*, which when translated into 薬を飲む [*kusuriwonomu*] and literally means *drink medicine.* Furthermore, a number of literal collocations also did not translate directly, such as *setting/sun* as in *the setting sun* corresponding to 沈む夕日[*shizumuyuuhi*]. Not only is *shizumu* often translated as *sink* rather than *set,* but the L1 is also problematic in that *sun* is typically represented as *evening sun* [*yuuhi*] in Japanese. Yet another example is the literal collocational pair *married/couple*, which is typically 夫婦 [*fuufu*], literally *husband-lady*. All of these examples of non-congruency help to highlight how focusing especially on items that are non-congruent can efficiently direct students' focus.

Regarding semantic transparency, most items were either literal or semi-figurative. Very few fully idiomatic collocational pairs were found, which is in line with previous research. However as indicated above, even literal collocations can be problematic for learners. After a contrastive analysis with the L1, items can then be analyzed for semantic transparency to identify those that have an even higher learning burden. Indeed idiomatic collocational pairs that are non-congruent with the L1 will probably be the most difficult for students to master, followed by

non-congruent figuratives, non-congruent semi-figuratives, and finally non-congruent literals. Identifying such items clearly helps learners concentrate on those with higher learning burdens.

Concordance data the collocational pairs was collected from the *COCA* (Davies, 2008) and processed with the concordance software to identify their most common MWUs. This revealed that, even for quite common collocations, it can be difficult to identify their most common MWUs, and the software helped produce more accurate results. For example, rather than a native teacher processing data for the collocational pair *by/large*, their intuition will be sufficient in identifying the MWU *by and large*. However there remain some cases where native intuition cannot be relied upon. For instance, *back/foot* is included within the frequency cut-off, but as can be seen in figure 3 below, it can be difficult to think of quite common MWUs that the analysis identified.

Multi-word units	Percentage of occurrence in 1,000 example sentences
back on [possessive pronoun] *feet*	38%
get back on [possessive pronoun] *feet*	14.6%
feet back	13.4%

Figure 3. MWUs for the verbs *back* and *foot*

At first glance, a native speaker may simply guess that these two words commonly occur in a MWU such as *move 10 feet back* or *the back of my foot,* but when possessive pronouns are

counted together we see that [*get*] *back on their feet* is the MWU most representative of how these two word collocate. In fact, even if a native speaker was aware that the two collocations were being used in this way, he/she may have chosen [*get*] *back on his feet* as an example. From a native speaker's perspective, this seems like an appropriate choice, while in reality the corpus data shows that *their* is actually the most common possessive pronoun in this MWU pattern, and it is actually twice as common as *his* (see figure 4 below).

Multi-word units	Percentage of occurrence in 1,000 example sentences
back on their feet	10.6%
back on its feet	9.8%
back on my feet	6.2%
back on his feet	5%

Figure 4. Most common possessive pronoun in the MWU *back on* [possessive pronoun] *feet*

Colligation analysis thus also helps extract accurate examples of collocational co-occurrence. Another specific example is an analysis of the collocational lemma *early* and *century*. When 1,000 example sentences in which *early* and *century* co-occur were analyzed with the concordance software without consideration for colligation, the results clearly became skewed, and the actual most common MWU was not identified as the most frequent. However, when text-editing software is used to replace all instances of numbers with a representative

marker (in this case [*year*]), the results were clearly improved (see figure 5 below).

Concordance software alone		Concordance and text-editing software	
% of occurrence in 1,000 example sentences	Multi-word unit with co-occurrence of *century* and *early*	% of occurrence in 1,000 example sentences	Multi-word unit with co-occurrence of *century* and *early*
10.7%	*century earlier*	19.2%	*early in the* [*year*] *century*
9.5%	*a century earlier*	10.7%	*century earlier*
8.5%	*early in this century*	9.7%	*early* [*year*] *century*
7.3%	*early in the century*	9.5%	*a century earlier*
6.4%	*centuries earlier*	8.5%	*early in this century*
5.0%	*early in the 20th century*	8.3%	*early as the* [*year*] *century*
		8.3%	*as early as the* [*year*] *century*
		7.3%	*early in the century*
		6.4%	*centuries earlier*

Figure 5. Concordance software alone versus concordance and text-editing software to deal with colligational issues

Conclusion

This study presented a step-by-step process in which the most useful collocations of English and the MWUs they occur in can be identified. It explored previous research and discussed some of the unavoidable barriers that prevent teachers and students from easily identifying useful collocations. Although the large number of items involved with mastery of high-frequency useful collocations/MWUs appeared most daunting, this study provided a methodology that helps to identify items having higher learning burdens. This methodology also highlighted the importance of colligation, and conducting MWU searches using concordance software even on seemingly transparent collocational pairs to more accurately identify the most common MWUs of any collocational pair.

Clearly the criteria identified in this study should be helpful, although this study was limited in that it only applied the criteria to a small sample of the items identified. Furthermore, frequency data indicates there being up to ten times as many useful items. Therefore, much larger-scaled research should be conducted, and collaboration is necessary to deal with time-consuming criteria such as L1 contrastive analysis, etc.

Many unanswered questions remain in regards to useful collocation/MWU identification. However, those answers have significant potential for improving the efficacy of language learning and thus should be explored.

References

Aghbar, A.A. (1990). *Fixed expressions in written texts: Implications for assessing writing sophistication.* Paper presented at a Meeting of the English Association of Pennsylvania State System Universities, October 1990.

Al-Zahrani, M. S. (1998). *Knowledge of English lexical collocations among male Saudi college students majoring in English at a Saudi university* (Unpublished doctoral dissertation). UMI, Ann Arbor, MI.

Alexander, R.J. (1984). Fixed expressions in English: Reference books and the teacher. *ELT Journal, 38*(2), 127-134.

Almela, M., & Sanchez, A. (2007).Words as "lexical units" in learning/teaching vocabulary. *IJES, 7*(2), 21-40.

Anthony, L. (2011). *Antconc*. Retrieved from http://www.antlab.sci.waseda.ac.jp/software.html

Bahns, J. (1993). Lexical collocations: A contrastive view. *English Language Teaching Journal, 47*(1), 56-63. doi: 10.1093/elt/47.1.56

Bahns, J. & Eldaw, M. (1993). Should we teach ESL students collocations? *System, 21*(1), 101-114. doi: 10.1016/0346-251X(93)90010-E

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405. doi: 10.1093/applin/25.3.371

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

Biskup, D. (1992). L1 influence on learners' renderings of English collocations: a Polish/German empirical study. In P.J.L. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 85-93). Houndmills: Macmillan.

Bogaards, P. (2001). Lexical units and the learning of foreign language vocabulary. *Studies in Second Language Acquisition, 23,* 321-343. doi: 10.1017/S0272263101003011

Channell, J. (1981). Applying semantic theory to vocabulary teaching. *English Language*

Teaching Journal, 35, 115-122. doi: 10.1093/elt/XXXV.2.115

Cheng, W., Greaves, C. & Warren, M. (2006).From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics, 11*(4), 411-433. doi: 10.1075/ijcl.11.4.04che

Cobb, T. (2013). *Vocabprofile*. Retrieved from http://www.lextutor.ca/vp/

Conzett, J. (2000). Integrating collocation into a reading and writing course. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 70-86). Hove, England: Language Teaching Publications.

Cortez, V. (2002). Lexical bundles in freshmen composition. In R. Reppen, S.M. Fitzmaurice & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 131-145). Amsterdam: John Benjamins Publishing Company.

Cowan, L. (1989). *Towards a definition of collocation.* Unpublished MA Thesis, Concordia University, Montreal, Quebec, Canada.

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

Davies, M. (2010). *Word list plus collocates*. Retrieved from http://www.wordfrequency.info/purchase1.asp?i=c5a

Davies, M. (2008). *The corpus of contemporary American English: 425 million words, 1990-present*. Retrieved from http://corpus.byu.edu/coca/

DeCock, S., Granger, S., Leech, G. and McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp.67-79). London and New York: Longman.

Doughty, C. & Williams, J. (1998). Pedagogical choices in focus on form. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 197-

262). New York: CUP.

Durrant, P. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL, 47,* 157-177. doi: 10.1515/iral.2009.007

Ellis, N.C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33-68). Cambridge: Cambridge University Press. doi: 10.1017/CBO978 1139524780.004

Ellis, R. (1994). *The study of second language acquisition.* Oxford: Oxford University Press.

Feyez-Hussein, R. (1990). Collocations: the missing link in vocabulary acquisition amongst EFL learners. In J. Fisiak (Ed.), *Papers and studies in contrastive linguistics: The Polish English contrastive project, 26* (pp.123-136). Poznan: Adam Mickiewicz University.

Firth, J.R. (1957). A synopsis of linguistic theory. 1930-1955. In S*tudies in linguistic analysis* (pp. 1-32), reprinted in F. Palmer (Ed.), *Selected papers of J.R. Firth 1952-59* (pp. 168-205). London: Longman.

Francis, G. (1993). A corpus-driven approach to grammar: Principles, methods and examples. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 137-156). Amsterdam: John Benjamins.

Furukawa, J., Ford, B., Ayson, E., Cambra, K., Takahashi, L. & Yoshina, K. (1998). Effects of a cognitive processing strategy on spelling, definitions, and reading. *Paper presented at the annual meeting of the Hawaii Educational Research Association.* Honolulu, HI, January 17th.

Gairns, R. & Redman, S. (1986). *Working with words. A guide to teaching and learning vocabulary.* Cambridge: CUP.

Gitsaki, C. (1996). *The development of ESL collocational knowledge* (Unpublished doctoral

dissertation). University of Queensland, Brisbane, Australia.

Grant, L. and Nation, P. (2006) How many idioms are there in English? *ITL International Journal of Applied Linguistics, 151,* 1-14. doi: 10.2143/ITL.151.0.2015219

Halliday, M.A.K. (1966). Lexis as a linguistic level. In C.E. Bazell, J.C. Catford, M.A.K. Halliday & R.H. Robins (Eds.), *In memory of J.R. Firth* (pp. 148-162). London: Longman.

Halliday, M.A.K. & Sinclair, J.M. (1966). Beginning the study of lexis. In C.E. Bazell, J.C. Catford, M.A.K. Halliday & R.H. Robins (Eds.), *In memory of J.R. Firth* (pp. 410-430). London: Longman.

Hill, J. (2000). Revising priorities: from grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 47-67). Hove, England: Language Teaching Publications.

Hill, J., Lewis, M. & Lewis, M. (2000). Classroom strategies, activities and exercises. In M. Lewis (Ed.), *Teaching Collocation: Further developments in the lexical approach* (pp. 88-117). Hove, England: Language Teaching Publications.

Hoey, M. (2005). *Lexical priming. A new theory of words and language.* London: Routledge.

Hoey, M. (1991). *Patterns of lexis in text.* Oxford: Oxford University Press.

Jones, S. & Sinclair, J.M. (1974). English lexical collocations: a study in computational linguistics. *Catiers de Lexicologie, 23*(2), 15-61.

Kallkvist, M. (1998). Lexical infelicity in English: the case of nouns and verbs. In K. Haastrup and A. Viberg (Eds.), *Perspectives on lexical acquisition in a second language* (pp. 149-174). Lund: Lund University Press.

Kennedy, G.D. (1990). Collocations: Where grammar and vocabulary teaching meet. In S.

Anivan (Ed.), *Language teaching methodology for the nineties* (pp. 215-229). Singapore: RELC.

Kilgarriff, A., Atkins, S., & Rundell, M. (2007, July). *BNC design model past its sell-by*. Paper presented at 2007 Corpus Linguistics Conference, Birmingham, UK.

Kjellmer, G. (1994). *A dictionary of English collocations: Based on the Brown corpus*. Oxford: Claredon Press.

Kjellmer, G. (1987). Aspects of English collocations. In W. Meijs (Ed.), *Proceedings of the International Conference on English Language Research on Computerised Corpora* (pp. 133-140). Amsterdam: Rodopi.

Kjellmer, G. (1984). Some thoughts on collocational distinctiveness. In J. Aarts & W. Meijs (Eds.), *Computer corpora in English language research* (pp. 163-171). Bergen: Norwegian Computing Centre for the Humanities.

Korosadowicz-Struzynska, M. (1980).Word collocations in FL vocabulary instruction. *Studia Anglica Posnaniensia, 12,* 109-120.

Lesniewska, J. & Witalisz, E. (2007). Cross-linguistic influences on L2 and L1 collocations. *EUROSLA Yearbook, 7,* 27-48. doi: 10.1075/eurosla.7.04les

Lewis, M. (2000). Language in the lexical approach. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 8-10). Hove, England: Language Teaching Publications.

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

Lewis, M. (2000). There is nothing as practical as a good theory. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 10-27). Hove, England:

Language Teaching Publications.

Liang, S.Q. (1991). A propos du dictionnaire francais-chinois des collocations francaises. *Cahiers de Lexicologie, 59*(2), 151-167.

McCarthy, M.J. (1984). A new look at vocabulary in EFL. *Applied Linguistics, 5*(1), 12-22. doi: 10.1093/applin/5.1.12

Meijs, W. (1992). Inferences and lexical relations. In G. Leitner (Ed.), *New directions in English language corpora: Methodology, results, software developments* (pp. 123-152). Berlin: Mouton de Gruyter. doi: 10.1515/9783110878202.123

Mitchell, T.F. (1971). Linguistic 'goings on': Collocations and the other lexical matters arising on the syntagmatic record. *Archivum Linguisticum, 2,* 35-69.

Nation, P. (2001a). *Learning vocabulary in another language*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139524759

Nation, I.S.P. (2001b). How many high frequency words are there in English? In M. Gill, A.W. Johnson, L.M. Koski, R.D. Sell and B. Wårvik (Eds.) *Language, Learning and Literature: Studies Presented to Håkan Ringbom* English Department Publications 4, Åbo Akademi University, Åbo, 167-181.

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

Nelson, F. & Kucera, H. (1979). *The Brown corpus: A standard corpus of present-day edited American English.* Providence, RI: Department of Linguistics, Brown University.

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

Newman, A. (1988). The contrastive analysis of Hebrew and English dress and cooking collocations: Some linguistic and pedagogic parameters. *Applied Linguistics 9*(3), 293-

305. doi: 10.1093/applin/9.3.293

Noel, J. (1992). Collocation and bilingual text. In G. Leitner (Ed.), *New directions in English language corpora: Methodology, results, software developments* (pp. 345-357). Berlin: Mouton de Gruyter. doi: 10.1515/9783110878202.345

Renouf, A., & Sinclair, J.M. (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 128-143). Harlow: Longman.

Robins, R.H. (1967). *A short history of linguistics.* London: Longman.

Saville-Troike, M. (1984). What really matters in second language learning for academic achievement? *TESOL Quarterly, 18*(2), 199-217. doi: 10.2307/3586690

Shin, D. (2006). *A collocation inventory for beginners* (Unpublished doctoral dissertation). Victoria University of Wellington, Wellington, New Zealand.

Simpson, R. & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly, 37*(3), 419-441. doi: 10.2307/3588398

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

Sokmen, A. (1997). Current trends in teaching second language vocabulary. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: description, acquisition and pedagogy.* Cambridge: Cambridge University Press.

Taylor, C. (1983). Vocabulary for education in English. *World Language English, 2*(2), 100-104. doi: 10.1111/j.1467-971X.1982.tb00531.x

Textcrawler. (2011). Retrieved from http://www.digitalvolcano.co.uk/content/textcrawler

Twaddell, F. (1973). Vocabulary expansion in the TESOL classroom. *TESOL Quarterly, 10,* 19-32.

Waller, T. (1993). Characteristics of near-native proficiency in writing. In H. Ringbom (Ed.),

Near-native proficiency in English (pp. 183-293). Abo: Abo Akedemi University.

Wilks, Y. (2005). REVEAL: the notion of anomalous texts in a very large corpus. Tuscan Word Centre International Workshop. Certosa di Pontignano, Tuscany, Italy, 31 June–3 July 2005.

Woolard, G. (2000). Collocation – encouraging learner independence. In M. Lewis, (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 28-46). Hove, England: Language Teaching Publications.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511519772

Yorio, C.A. (1980). Conventionalized language forms and the development of communicative competence. *TESOL Quarterly, 14*(4), 433-442. doi: 10.2307/3586232

Zhang, X. (1993). *English collocations and their effect on the writing of native and non-native college freshmen* (Unpublished doctoral dissertation). Indiana University of Pennsylvania, Indiana, PA.

Author Bio

James Rogers is an assistant professor at Kansai Gaidai University with 10 years of experience teaching English in Japan. He is currently pursuing a PhD in education examining the high frequency collocations of English. In addition to collocations, his other research interests include corpus linguistics, C.A.L.L., vocabulary acquisition, and the use of psychology in the classroom.