Vocabulary Learning and Instruction
Volume 4, Issue 2, December 2015
http://vli-journal.org

Article

On Using Corpus Frequency, Dispersion, and Chronological Data to Help Identify Useful Collocations

James Rogers[a], Chris Brizzard[a], Frank Daulton[b], Cosmin Florescu[c], Ian MacLean[a], Kayo Mimura[a], John O'Donoghue[d], Masaya Okamoto[e], Gordon Reid[a] and Yoshiaki Shimada[f]
[a]Kansai Gaidai University; [b]Ryukoku University; [c]University of New England; [d]Osaka Board of Education; [e]University of Manchester; [f]State University of New York at Albany

doi: http://dx.doi.org/10.7820/vli.v04.2.rogers.et.al

Abstract

This study analyzed corpus data to determine the extent to which frequency, dispersion, and chronological data can help identify useful collocations for second language learners who aim to master general English. The findings indicated that although various analysis levels of frequency and dispersion data are largely effective, the analyses could not identify useful collocations reliably. The findings also indicated that chronological data analysis is not as useful as dispersion analysis due to the amount of time it took versus the improvements that resulted from it. Ultimately, it was found that a manual analysis of data using native speaker intuition is unavoidable. This study highlighted the value and reliability of certain types of corpus data analysis, and also the necessity of labor-intensive, native speaker analysis for identifying useful collocations.

Keywords: corpus; frequency; dispersion; collocations; multi-word units; formulaic sequences.

1 Introduction

Comprehending and producing collocations is an essential skill for native-like fluency (Durrant & Schmitt, 2009; Wray, 2002). Knowledge of collocations helps the language learner sound more native-like and process language more efficiently (Nation, 2001a; Snellings, van Gelderen, & de Glopper, 2002). However, research has shown that many second language learners have significant trouble achieving collocational fluency due to a number of persistent hurdles (DeCock, Granger, Leech, & McEnery, 1998; Kallkvist, 1998). These include the complexity of how collocations function (Hill, Lewis, & Lewis, 2000), and also the lack of emphasis by teachers and material developers (Gitsaki, 1996; Nesselhauf, 2005). Adding to the problem is the fact that there are still very few studies that identify which are the most frequent (Durrant & Schmitt, 2009), and there is a lack of agreement on what criteria should be used to achieve this.

Thus, many questions persist regarding how students, teachers, and researchers should approach collocations. Corpora can help us tackle the multifaceted and

Please cite this article as: Rogers, J., Brizzard, C., Daulton, F., Florescu, C., MacLean, I., Mimura, K., . . . Shimada, Y. (2015). On using Corpus frequency, dispersion, and chronological data to help identify useful collocations. *Vocabulary Learning and Instruction*, *4* (2), 21–37. doi: 10.7820/vli.v04.2.rogers.et.al

complex issues that must be resolved to help students develop their collocational fluency. However, many questions still remain in regard to how to use corpus data to accomplish this. For instance, the ideal frequency cut-off in corpus data for identifying high-frequency collocational co-occurrence is still unknown, in that previous research has examined as low as two occurrences per million tokens (Liu, 2003) and as high as 40 occurrences per million tokens (Biber, Conrad, & Cortes, 2004). Furthermore, studies often disregard a collocation's distribution among a wide variety of genres, or *dispersion*. For instance, although Liu's (2003) study was quite comprehensive in regard to its frequency cut-off, it did not consider dispersion as Biber et al. (2004) did. In addition, considering whether a collocation occurs regularly over a number of years instead of sudden surges or a decline in frequency (*chronological data analysis*) has likewise been neglected. No previous collocation identification studies have used this criterion to our knowledge.

Thus, this paper examined the extent to which frequency, dispersion, and chronological data from corpora helped identify useful collocations to directly teach. Such items will have a good cost/benefit value for these learners; they will be worthwhile to study because they occur frequently in a wide variety of texts and will be chronologically stable.

2 Research Topic Background

2.1 What is a Collocation?

This paper defines collocations in the traditional sense by frequency of co-occurrence (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Shin, 2006), while at the same time counting co-occurrence from a more modern perspective as lemmatized *concgrams*, as per the methodology set forth in Rogers, et al.'s (2014) study. Cheng, Greaves, and Warren (2006) define concgrams as "all the permutations of constituency and positional variation generated by the association of two or more words" (p. 411). *Constituency variation* (AB, ACB) involves a pair of words not only co-occurring adjacent to one another (*lose weight*) but also with a constituent (*lose some weight*). *Positional variation* (AB, BA) refers to counting total occurrences of two or more particular lexical items that include occurrences on either side of each other. Thus, *provide you support* and *support you provide* would both be included in the total counts for a multi-word units (MWU) concordance search for the lemma *provide* and *support*. This study also does not limit itself to including only collocations with high predictability such as *crux/matter*. This is due to the unreliability of statistical measures of association, discussed further below. In addition, this paper includes semantically transparent lemma pairs as collocations. Grant and Bauer (2004) refer to such word combinations as *literals*, or MWUs which are both compositional (the meaning of the whole can be deduced from its parts) and non-figurative. For example, with the MWU *eat breakfast* you literally *eat breakfast*. On the other end of the spectrum would be core idioms, non-compositional but also non-figurative MWUs, such as *kick the bucket* (you literally do not *kick* anything, there is no actual *bucket*, and it is impossible to deduce the meaning from the MWU's parts). There are clear rationales for considering such items, which are discussed further in this paper as well.

2.2 The Lack of Resources

Knowledge of collocations and formulaic language that have frequent co-occurrence is of obvious value to the language learner. Such knowledge has been referred to as a decisive factor in developing fluency (Almela & Sanchez, 2007). However, despite growing recognition of collocational fluency, few resources are available to guide collocation selection. Resources do exist, but they often only have hundreds of items and thus are far from being considered as comprehensive resources for helping learning to master the collocations of high-frequency vocabulary. In addition, very large dictionary-like resources with tens of thousands of items also exist, but clearly such massive contents are not practical in regard to direct instruction. One exception would be Shin (2006), but this study only aimed at created a collocation inventory for beginners and only involved L1-L2 congruency analysis with Korean.

In addition, many of the studies that have been conducted have been referred to as flawed in some aspect or lacking comprehensiveness (Durrant & Schmitt, 2009). Collocation can be viewed as intertwined with formulaic language, depending on one's definition of collocation. However, many formulaic language studies limit their scope to a specific type of multi-word unit. For instance, Biber et al. (2004) only found 172 "lexical bundles", limiting themselves to a cut-off of 40 occurrences per million and only considering four-word sequences. Such findings pale in comparison to the estimated collocations of native speakers, which have estimated to be in the hundreds of thousands (Hill, 2000).

2.3 Criteria for Identifying Useful Collocations

There are various ways to identify useful collocations. The simplest and most common involves frequency data from a corpus (Biber et al., 1999; Shin, 2006). While setting a frequency cut-off is unavoidably "arbitrary" (Nation, 2001a, p. 180), for teaching a cut-off must be set in regard to the practical limitation of how many items can be directly taught during limited classroom time. Nation (2001a, p. 96) suggested 2,000 word families as "practical and feasible" in regard to direct teaching, while Nation (2001b) suggested a limit of 3,000 word families.

Other researchers use statistical measures of association, such as how Lorenz (1999) utilized t-scores and mutual information data (MI) to identify high exclusive co-occurrence. MI measures the strength of co-occurrence between collocates. In other words, it measures "if the relative proportion of mutual occurrences of some words is large compared with their total frequencies" (Shin, 2006, p. 31). For instance, *tit/tat* has a very high MI of 15.01 (Davies, 2008). However, t-scores and MI can be problematic. MI emphasizes collocations whose components are not often found apart (Stubbs, 1995); thus, word pairs that clearly collocate but also have high frequencies might be excluded. Conversely, collocations with high t-scores will tend to be high-frequency words, and the measure may fail to identify collocations that have high frequencies of co-occurrence but low frequencies as individual words. Durrant and Schmitt (2009) give the following examples to highlight the issues with these statistical measures of association: "pairs like *good* *example, long way*, and *hard work* attain high t-scores but low MI scores, while pairs like *tectonic plates* attain the reverse" (p. 167).

Previous research on identifying useful collocations has led to various other criteria and sub-categories. For example, some researchers subdivided collocations into literals, figuratives, and core idioms (Grant & Bauer, 2004). They explain that if each word in an MWU is replaced with its definition, and the meaning of the word does not change, then it is a "literal". If it is possible to understand the meaning of an MWU by recognizing an untruth and pragmatically reinterpreting it in a way that correctly explains the MWU, then it is a "figurative". If only one word in an MWU is either literal, then such an MWU would be a "ONCE". Finally, they explain that if the MWU did not fall into any of these above categories, then it should be considered a "core idiom". However, while these semantic sub-categories of collocations do exist, researchers such as Wray (2000) insist that we deal with semantically transparent items in addition to those that are opaque. Nesselhauf (2005) agrees, finding that students tend to assign literal meaning to collocations with a figurative meaning, and vice versa.

L1-L2 collocation congruency (i.e., how similar/dissimilar a collocation's translation is in the learner's native language) is another criteria considered by many researchers. Feyez-Hussein (1990) found that approximately 50% of collocation errors were due to L1 influence. Thus, whether or not a collocation is semantically transparent or is a free combination becomes moot when the collocation differs greatly in comparison to how it is said in the learner's L1.

Notably, Nation (2001a) states that a collocation's balanced dispersion in many different categories of text is a necessary criterion for identifying useful collocations. Such collocations can easily be identified when corpora provide *dispersion* data, or the distribution of frequency among genres within the corpus. Gries (2008) believes that dispersion data analysis is essential as well, stating that raw frequency data can be misleading in regard to a word's general importance when the dispersion of its frequency data is unbalanced. However, only a few small-scale studies on identifying useful collocations have utilized dispersion data from corpora to delimit their selections of useful collocations. One such study is Cortes (2002). However, its corpus consisted of only approximately 360,000 tokens. Biber et al. (2004) also employed dispersion criteria, but their corpus consisted of two million tokens. Furthermore, not only has dispersion not been adequately applied to identify useful collocations, neither has chronological stability. Furthermore, any cut-off set for dispersion or chronological data will also be unavoidably arbitrary. For instance, Nation and Hwang (1995) specifically state that their choice of vocabulary occurring in 10 out of 15 sections of the corpora in their study for balanced dispersion was arbitrary.

In regard to frequency, Cortes (2002) set a frequency cut-off of 20 occurrences per million, and Biber et al. (2004) set theirs at 40 occurrences per million. Other studies were more inclusive. Shin (2006) set a cut-off of three occurrences per million, and Liu (2003) at two occurrences per million. However, the massive amount of data to be examined remained an issue. For example, despite examining significantly more items than previous studies (this study examined one occurrence per million tokens), items occurring approximately half as often could still be considered to have value to language learners. For example, the lemma pairs *nice/ vacation*, *finish/workout*, and *tend/exaggerate*, all occur approximately only once per two million tokens (Davies, 2008). Practically speaking, a native speaker would not consider these as low-frequency language not worthy of direct learning despite their low frequencies of co-occurrence. The frequency cut-offs used in the above previous studies could thus be considered conservative.

Thus, a significant gap exists in the research as to the extent that frequency, dispersion, and chronological data from corpora can help identify the most useful collocations. This brings us to this study's research questions.

3 Research Questions

1. To what extent can utilizing corpus frequency data help identify useful collocations?
2. To what extent can utilizing corpus dispersion data help identify useful collocations?
3. To what extent can utilizing corpus chronological data help identify useful collocations?

4 Methodology

4.1 Materials

This study utilized data from the *Corpus of Contemporary American English* (*COCA*) (Davies, 2008). The COCA provides collocation lists that have been compiled with consideration for constituency and positional variation, and this was one of the reasons why it was chosen as a data source. This study thus utilized Davies' (2010) *Word List Plus Collocates*, a lemmatized concgram list. It consists of the most frequent 739,255 collocates that co-occur with the most frequent 5,000 lemmas in the corpus.

In addition, the COCA divides itself into five separate genres: spoken, fiction, magazine, newspaper, and academic, and these sections all have nearly as many tokens in total and per year. The division of data into these sections made the dispersion analysis in this study possible. Also, the COCA also divides itself into chronological sections of four years per section. This study utilized the completed chronological sections from 1990 to 2009.

4.2 Procedure

This study began by piloting various frequency cut-offs on Davies' (2010) collocation list. The aim of this study was to find a frequency cut-off which resulted in the vast majority of collocates identified being judged as useful and worthy of direct instruction, and also consisting of between 2,000 and 3,000 word families.

Frequency cut-offs were piloted to determine how many useful collocations were at each level. The study took a cue from previous research and started at Biber et al. (2004) cut-off of 40 occurrences per million tokens, and continued to

Kjellmer's (1987) two occurrences per million. Cobb's (2015) *Vocabprofile* programme, which consists of the most frequent 25,000 word families in the *BNC* and *COCA*, was utilized to determine the total word families the collocations consisted of to avoid exceeding 3,000 word families while not falling below 2,000 word families. After identifying a cut-off that resulted in between 2,000 and 3,000 word families, the list was then examined by a native speaker for general item usefulness to ensure that the list was not overly inclusive. For example, if a frequency cut was too inclusive, it could include very low-frequency collocates of high-frequency vocabulary that would be of little value to learners (see Table 1).

This study found that utilizing a frequency cut-off of one occurrence per million tokens was ideal. How this cut-off was decided upon is explained further in Section 4.1 of this paper.

Often, the collocation that occurred was a node word itself within the most frequent 5,000 lemma of Davies' (2010). Therefore, the list also includes many duplicate entries such as *take/walk* and also *walk/take*. Such duplicates were first removed.

Then, dispersion and chronological data for identified collocates were collected from the COCA. Its interface allows users to extract dispersion data for five genres: spoken, fiction, magazine, newspaper, and academic. The interface also allows for the extraction of chronological data in four-year increments: 1990–1994, 1995–1999, 2000–2004, 2005–2009, and 2010–2012. Since the four-year section 2010–2013 was yet to be completed, its data were not included in this study.

Various parameters were then piloted to determine the cut-off point for balanced dispersion and chronological data distribution; these ratios were trialed, and the items flagged at each ratio were examined using native speaker intuition to judge whether it was overly inclusive or exclusive.

As for dispersion and chronological data, a range of parameters were tested due to the gap in research with the corpus used in this study. As with frequency cut-offs, any cut-off set for dispersion or chronological data will also be unavoidably arbitrary. This study experimented with parameters that best approximate balanced distribution.

For dispersion data, the parameters required that a specific percentage of the total occurrences had to occur in a majority of the *COCA*'s genres: three or more out of the five genres. Native speaker intuition was used to determine the best percentage cut-off. The lemma list was examined for items specialized in nature, and a number of these items were found to have approximately 5% or less of their occurrences in three or more of the genres. Thus, dispersion data were analyzed at three separate percentages to determine the most useful parameter: less than 10%, 5%, and 2.5% of total occurrences in three or more genres. Then pairs flagged at these parameters were examined to determine if they truly were specialized by a native speaker, and thus not worthy of direct instruction for a general English course. Next, all remaining items in the list were also scanned by a native speaker to determine if the parameters were not able to identify items that were actually specialized. Finally, all items identified as being unbalanced by a native speaker were examined to determine if they fell into a common genre (e.g., academic language).

Table 1. Examples of High- and Low-Frequency Collocates of the Lemma *Play* in the COCA (Davies, 2008)

Rank	Collocate	Frequency
1	*role*	20,747
2	*game*	8,536
99	*gin*	82
100	*hoops*	80

Table 2. System for Rating the Value of Collocates for Learners of General English

Rating value in regard to direct teaching
1. Provides no value whatsoever if directly taught. None of the examined items fell into this category so an example cannot be provided. However, this rating was included to possibly deal with any items in the list that represented corpus “noise”. In other words, these would include mistakes in the data or in how the data was compiled, which would result in the inclusion of items that are clearly not part of natural language but are the result of the fact that the compilation of a corpus is a mere attempt at emulating balanced natural language use.
2. Provides little value if directly taught. For example, *note/supra* was found to have extremely unbalanced dispersion data, occurring mostly in the academic section in the COCA. Because of its specialized usage in an exclusive genre, native speakers may not even be aware of its meaning. Thus, clearly such an item will be of little value to a learner of general English.
3. Provides questionable value if directly taught. For example, *lemon/zest*. This item occurs often but with unbalanced distribution data in the COCA because of inclusion of many magazines with recipes in the COCA’s “magazine” section, and is of clear questionable value for learners trying to master general English.
4. Provides value, but with limitations if directly taught. For example, *championship/year*, despite having unbalanced dispersion data distribution, would be of value to teach because of the generalness of the term and ubiquitous nature of sports in society. The term is general in that it can be used to describe all sports despite the fact that sports itself is somewhat specialized.
5. Provides clear value if directly taught. For example, how *email/address*, despite having unbalanced chronological data distribution, is clearly a valuable and stable item to learn.

A similar methodology was employed for chronological data analysis. Again, native speaker intuition was used to determine the best percentage cut-off. First, the lemma list was examined using native speaker intuition for pairs which were either dated, too modern or only occurred during a specific time period. Very few such items existed, but the items that were found had approximately 5% or less occurrences in one or more of the four chronological sections. Just as dispersion data were analyzed, chronological data were also analyzed to find items having less than 10%, 5%, and 2.5% of total occurrences in one or more sections. Then pairs flagged at these parameters were examined to determine if they truly were dated, too modern, or not useful because they only occurred during a specific time period by a native speaker, and thus not worthy of direct instruction for a general English course. Next, all remaining items in the list were also examined by a native speaker to determine if the parameters were unable to identify items that were dated, too modern, or had little value because they only occurred during a specific time period. Finally, all items identified as being unbalanced by a native speaker were examined to determine whether they were either dated, only occurred during a specific time period, or were too modern. How these parameters were decided upon is explained further in Sections 4.2 and 4.3 of this paper.

Last, to determine the extent to which the dispersion and chronological data distribution cut-offs truly identified items that were not worthy of direct instruction, the collocates were then judged by a native speaker in regard to their usefulness. Each item was given a rating (see Table 2) in regard to its value for learners of general English.

After being rated, any items flagged by each of the cut-off parameters that were rated 1 or 2 were tallied. Furthermore, any items not flagged by the cut-off parameters that received ratings of 1 or 2 were also tallied. These two steps would then be used to judge the cut-off parameter’s ability to identify collocations that truly are of little or no use for general learners of English in regard to balanced dispersion and chronological data.

5 Results

5.1 Frequency Cut-off Results

After trialing the various frequency cut-offs used by previous researchers, it was found that the cut-off of two occurrences per million tokens resulted in a list of lemma pairs consisting of only 1,671 families. Taking a cue from the recommendation of directly teaching between 2,000 (Nation, 2001a) and 3,000 (Nation, 2001b) word families, it was therefore determined that a more inclusive cut-off could be used. A cut-off of once per million tokens and once per 500,000 tokens was then piloted, which resulted in 2,540 families and 4,122 families, respectively. The cut-off of one occurrence per million tokens was thus determined to be ideal.

Cobb’s (2015) *Vocabprofile* programme showed that these pairs covered 75.6% of the top 3,000 word families. It should also be noted that 97.8% of the tokens in the lemma pair list occur within the top 3,000 word families. An analysis of the data is presented in Table 3.

After duplicate entries and proper nouns were removed, the cut-off resulted in 14,035 pairs being included. Due to the large number of items, this list was checked by an experienced, native-speaking teacher of English for usefulness, and the vast majority were found useful and worthy of direct teaching. Therefore, it was confirmed that the frequency cut-off was not too inclusive.

5.2 Dispersion Data Analysis Results

Out of all three parameters tested, the 5% or more cut-off in three or more genres was shown to be the most reliable in regard to both properly flagging items of little use for learners of general English, and not flagging items the native

Table 3. Word Frequency Breakdown of Lemma Pairs Occurring Once Per Million Tokens According to *Vocabprofile's* 25,000 Word Families of the BNC and COCA

Frequency level	Families (%)	Types (%)	Tokens (%)	Cumul. token%
K-1 Words:	806 (32.59)	1,095 (38.17)	17,461 (69.15)	69.15
K-2 Words:	704 (28.47)	847 (29.52)	4,945 (19.58)	88.73
K-3 Words:	595 (24.06)	660 (23.00)	2,280 (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)	2,473	2,869 (100)	25,251 (100)	100.00

speaker judged to be useful (see Table 4). At 5%, 845 items were considered to be either erroneously flagged or left unflagged by the parameters after native speaker analysis. The next most reliable parameter was at 2.5%, where a total of 1,283 items were considered either erroneously flagged or unflagged. At this parameter, the vast majority of the 1,283 items fell beyond the parameter (1,211) and thus it was not inclusive enough. The most unreliable parameter was 10%, where a total of 1,487

Table 4. Dispersion Data Analysis Results

Parameter	Accurately flagged	Items judged unbalanced by a native which were not flagged	Erroneously flagged	Total items parameters either did not flag or erroneously flagged
2.5%	616	1,211	72	1,283
5%	1,171	656	189	845
10%	1,618	209	1,278	1,487

Table 5. Most Common Types of Language from which Flagged Items are Derived

Parameter	Academic	Fiction	Food	Television
2.5%	238	26	249	56
5%	215	106	143	21
10%	300	17	54	4

items were either considered erroneously flagged or the parameters did not flag. Conversely, in comparison with the 2.5% parameter, 10% proved to be too inclusive in that the vast majority of the 1,487 items were erroneously flagged (1,278). In total, native speaker judgment identified 1,827 of the 14,035 pairs (13%) as having limited value for learners of general English.

When items were judged by a native speaker to determine their type of specialized language, four specific types accounted for the vast majority of items: academic language, descriptive language primarily used in fiction, language related to food, and language used primarily on television. Table 5 shows the number of items in each of these four types at all three parameters (non-combined).

A total of 1,539 flagged pairs at all three parameters were judged erroneously flagged by a native speaker. That is, the native speaker felt these items did have value for learners of general English. In addition, there were 209 pairs judged by a native speaker to be specialized and of little use to general learners that were not flagged at any of the three parameters.

5.3 Chronological Data Analysis Results

Out of all three parameters tested, the 2.5% and 5% or more cut-off in one or more chronological sections were shown to be the most reliable in regard to both flagging items of little use for learners of general English because of chronological issues, and not flagging items the native speaker judged to be useful for learners of general English (see Table 6). At both 2.5% and 5%, 100 items were either erroneously flagged or the parameters did not flag. At 10%, 145 items were either erroneously flagged or left unflagged. Only five items beyond the parameters tested were judged by a native speaker to be of little use for leaners because of chronological issues.

Despite the 2.5% and 5% parameters being the most reliable, the vast majority of the items flagged by all three parameters were found erroneously flagged by a

Table 6. Chronological Data Analysis Results

Parameter	Accurately flagged	Items judged unbalanced by a native which were not flagged	Erroneously flagged	Total items parameters either did not flag or erroneously flagged
2.5%	14	40	59	99
5%	13	22	77	99
10%	23	4	140	144

native speaker. These items were either useful to learners or did not actually exhibit chronological issues. Furthermore, the entire analysis only resulted in a total of 55 items being flagged for chronological issues, which amounts to only 0.39% of the total items examined.

6 Discussion

6.1 Frequency Data Analysis

Determining the extent to which frequency data can help inform useful collocation selection revealed both potential and limitations. First, it was shown that it is possible to set a frequency cut-off that results in a list of collocations that can be practically taught. What at first seemed an impractical amount of items to teach was in reality only 2,473 word families combining with each other in 14,035 different ways, which is within the 2,000–3,000 word family estimate of what can be taught directly. And while many useful collocations do occur beyond the frequency cut-off of this study, a list of collocations resulted that showed very good coverage of high-frequency vocabulary (75.6%) in addition to having 97.8% of the word families within the pairs being within the most frequent 3,000 word families. However, a number of other steps must still be taken to make the data practically usable, despite these positive results.

Shin and Nation (2008) refer to collocations as having two parts: a pivot word and its collocate. In this study, pivot words were the top 5,000 most frequent lemma in the COCA and the collocates, the words that co-occur with these frequently. However, there was the issue of removing duplicates, or instances when a collocate of one pivot word is also a pivot word itself.

This is a time-consuming, manual process that is essential. Moreover, proper nouns also need to be removed. This step is also time-consuming because it must be done manually. It was also difficult to judge whether a lemmatized collocational pair is part of a larger proper noun without examining concordance data.

Thus, the answer to Research Question 1 is that frequency data can, to a large extent, help identify useful collocations. The limitation is that many of the items identified may have duplicate entries and proper nouns would also need to be removed. Finally, such a list may contain a significant amount of items that are of little value to learners of general English due to their specialized nature.

6.2 Dispersion Data Analysis

Considering a collocational pair's general value in regard to its usefulness across multiple genres proved to be an important criterion; it identified 13% of the 14,035 pairs as not being of significant value to general learners of English. However, dispersion data alone was not sufficient in identifying unbalanced items. Often the parameter set was either too inclusive or not inclusive enough, and thus items would be included that were of little value or items of little value were not identified for removal. The most reliable parameter was shown to be a cut-off of 5%

of occurrences across three or more genres. While the parameter was useful in helping to flag items to reconsider, native speaker judgments were unavoidable. The parameter could only flag 64.1% of the items that were truly of little value, while 10.3% of the items flagged were later judged to be valuable.

The largest group that had unbalanced dispersion data was pairs occurring mostly in the academic section. While these pairs would be highly useful for students who plan to do scientific research or read academic journals, such items may not be useful for more general language needs. Thus, identifying such genre-specific, unbalanced items can be extremely valuable, either to exclude them or even focus on them if appropriate.

The same can also be said for the large number of pairs that occurred mostly in the fiction section. They consisted of language employed by fiction writers to describe what the reader cannot see. Thus, these items do not occur often in any other genres. Again, their inclusion or exclusion depends on the course of study.

Biber, Conrad, and Reppen (1998) reminded us that large corpora can skew the type of data we are looking for. This was evident in the disproportionate amount of collocations related to cooking found in the magazine and newspaper sections. Since the magazines and newspapers sourced by the COCA regularly featured recipe articles, such items had disproportionate frequency totals. The pedagogical value of directly teaching such items to general learners is questionable except for those who plan to work in the food industry. Thus despite their high frequency, their pedagogical value is in doubt.

Items mostly occurring in the spoken section were also apparently influenced by the data source. The COCA sourced much of its spoken section data from television, and in particular, news or talk shows. Thus, the vast majority of the items with unbalanced dispersion in the spoken section consisted of the language newscasters or talk show hosts use, such as commercial break transitions, etc. The value of such items for learners of general English is also arguably low for second language learners, and their discovery shows the importance of dispersion data.

Also of note is how the COCA divides its genres, and the effects that it has on dispersion data. While much academic and fiction-related language was easily identified, the same cannot be said for other specialized genres, such as business-related collocations, despite it being a clearly specialized genre. Business-related terms were distributed throughout the spoken, magazine, and newspaper genres of the *COCA*, but not in particularly high-frequency counts in comparison with academic language, which had its own dedicated genre. Only a small portion of the spoken, magazine, and newspaper genres took its data from business-related sources, such as financial magazines. If the COCA were designed with this in mind, such language could have also been easily identified. Such data would be of clear value to the many learners of business English.

In summary, the data analysis showed that the most reliable parameter was able to identify 64% of the items deemed to be of little value for learners of general English by a native speaker. Thus, in regard to answering Research Question 2, the extent to which dispersion data can identify useful collocations is limited in that the parameter was only able to identify 64% of the items that needed to be excluded.

As with frequency data, a native speaker manual analysis of the items in regard to dispersion was deemed essential as well.

6.3 Chronological Data Analysis

Considering a collocational pair's balanced chronological data distribution, when determining its value for learners, proved to be much less effective than the dispersion data analysis, since only 0.39% of the 14,035 pairs were found to be either dated, too modern, or only occurred in a limited time span in the past. Furthermore, each parameter was shown to be quite unreliable in that the vast majority of the items it flagged as having unbalanced distribution was deemed valuable for learners of general English.

Often items erroneously flagged by the parameters were new collocations deemed by a native speaker to have high potential to be used regularly in the future, such as *internet/access*. The types of items that were accurately flagged or deemed by a native speaker to have chronological issues were mostly related to temporal events, such as with *new/millennium*. Items with sudden surges in frequency counts were mostly connected to political events, wars, or other time-sensitive events.

Some items were also deemed too modern, so their future value was unclear. For instance, *cell/embryonic* was flagged by one of the parameters and considered by a native speaker to be of questionable value. It may have high-frequency counts simply because it is a new technology and being discussed often, and it is unclear how whether the collocation will continue to be used. The science may become commonplace or outdated, and thus the term may not be discussed as often in the future.

Only a few items were considered as dated, such as *word/processor*. Notably, the corpus only provides data back to 1990. If older data were available, then there would be more dated collocations identified. However, within the data's 19-year span, very few dated collocations were found. In addition, if a more detailed chronological breakdown of data were available (i.e., a breakdown by year instead of four-year sections), a more in-depth analysis would have been possible.

As for Research Question 3, the data clearly demonstrated the limited efficacy of chronological data analysis. Not only was there a very small number of items that actually had chronological issues, all of the parameters tested were highly unreliable, thus again requiring native speaker judgment. Thus, this criterion was shown to be of limited value for useful collocation identification.

7 Conclusion

7.1 Summary of Results

This paper has described the extent to which frequency, dispersion, and chronological data can help identify useful collocations. The frequency cut-offs that were tested identified a particular cut-off that resulted in a practical number of useful items to be taught. It resulted in a list of collocations with very high coverage of high-frequency vocabulary. However, many useful collocations were also shown to remain beyond the frequency cut-off. It furthermore showed that some highly time-consuming steps were still required to make the results usable, such as removing duplicate entries and proper nouns.

Frequency data analysis alone proved insufficient in producing a list of collocations that all have value to learners of general English. The dispersion data analysis conducted in this study identified 13% of the items as not being of value for learners. Although one particular dispersion data cut-off was more reliable in identifying items deemed by a native speaker to be unbalanced in their usage across genres of English, this parameter could not identify all of the items a native speaker deemed as being of little value due to unbalanced dispersion. This parameter failed to identify 35.9% of the total items deemed to be of little value due to unbalanced dispersion. Furthermore, 13.9% of the items this parameter did identify as having unbalanced dispersion data were actually judged to be of value, and thus had been erroneously flagged. So despite being useful in flagging many of the items that truly had little value for learners, such data cannot be considered reliable. Native speaker analysis, a time-consuming manual process, was thus shown to be necessary.

This study also revealed that dispersion data analysis can not only identify useful items for learners of general English, but can also identify specialized vocabulary, which is prominent in academic language and fiction writing, etc. Identifying such specialized vocabulary is not only useful for teachers who need to exclude it from a more general language course, but it is conversely useful for specialized classes needing to focus on highly salient collocations. The analysis also revealed that even data in large corpora can exhibit skewed frequencies. Any corpus is only as good as its source, and dispersion data analysis can identify deficiencies in corpora, such as the COCA's heavy inclusion of food/recipe-related language.

This study's chronological data dispersion analysis was shown to be of far lesser value in comparison with the genre dispersion data analysis. In total, only 0.39% of the items examined were found to have chronological issues. The parameters tested were also shown to be quite unreliable, in that the vast majority of items they flagged as having unbalanced chronological data distribution were either deemed of value to learners or to not have chronological issues in the first place. Therefore, it is less clear whether examining collocations for chronological balance is warranted or productive.

As mentioned earlier, the number of items to be examined when determining useful collocations is staggering. The goal of this study was to identify collocations that could be practically taught. So while this study examined significantly more items in comparison with previous research, the resulting list cannot be considered completely comprehensive. Furthermore, there are also limitations in how the results of this study can be interpreted due to the fact that only one native speaker gave judgments in regard to which items seemed worthy of direct instruction because of practical time limitations for judging such a large amount of data. If more than one speaker examined the data, an inter-rater reliability analysis could have been conducted to provide more solid data.

This study also acknowledges the limitations of its parameters in regard to frequency, dispersion, and chronological data analysis; no relevant precedents for the corpus used in this study had existed, and thus to an extent its parameter cut-offs were subjective. The present study simply aimed to validate the usefulness of some specific parameters to help identify useful collocations. It acknowledges that results will never be indisputable, but rather offers the best approximation possible within unavoidable constraints. To our knowledge, this was the first study using dispersion and chronological data from the COCA to determine useful collocations; thus, parameters were experimented with that best approximate balanced distribution. Regardless, there are clearly limitations to interpreting the results of this study due to these issues.

Similarly, native speaker intuition judgments on the value of items for learners of general English are subjective. However, the data revealed that such judgments were essential. In addition, this paper acknowledges the limitation of having only one native speaker make judgments on the value of items in regard to the frequency cut-off and parameters. While employing native speakers is ideal, due to time constraints and the large amount of items examined, relying upon a single native speaker was an acceptable expedient.

Despite the above limitations, we believe this paper contributes to collocation research and can inform future works. These limitations should be considered as opportunities for future researchers to improve methodology and resource design. These improvements will hopefully lead to further insights in regard to the identification of useful collocations.

References

Almela, M., & Sanchez, A. (2007). Words as "lexical units" in learning/teaching vocabulary. *International Journal of English Studies*, *7*(2), 21–40. Retrieved from http://revistas.um.es/ijes/issue/view/4811.

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

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.

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

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

Cobb, T. (2015). *Vocabprofile*. Retrieved from http://www.lextutor.ca/vp/bnc/

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

Davies, M. (2008). The Corpus of Contemporary American English: 450 million words, 1990–present. Retrieved from http://corpus.byu.edu/coca/

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

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

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, *47*, 157–177. doi:10.1515/iral.2009.007

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* (Vol. 26, pp. 123–136). Poznan, Poland: Adam Mickiewicz University.

Gitsaki, C. (1996). *The development of ESL collocation knowledge* (Unpublished doctoral dissertation). University of Queensland, Queensland, Australia.

Grant, L., & Bauer, L. (2004). Criteria for re-defining idioms. Are we barking up the wrong tree? *Applied Linguistics*, *25*(1), 38–61. doi:10.1093/applin/25.1.38

Gries, S. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*, 403–437. doi:10.1075/ijcl.13.4.02gri

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, UK: Language Teaching.

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, UK: Language Teaching.

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

Kjellmer, G. (1987). Aspects of English collocations. In W. Meijs (Ed.), *Corpus linguistics and beyond* (pp. 133–140). Amsterdam, the Netherlands: Rodopi.

Liu, D. (2003). The most frequently used spoken American English idioms: A corpus analysis and its implications. *TESOL Quarterly*, *37*, 671–700. doi:10.2307/3588217

Lorenz, G. (1999). *Adjective intensification – Learners versus native speakers: A corpus study of argumentative writing*. Amsterdam, the Netherlands: Rodopi.

Nation, I.S.P. (2001a). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

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, & B. Warvik (Eds.), *Language, learning, literature: Studies presented to Hakan Ringbom* (pp. 167–181). Turku: Abo Akademi University.

Nation, I.S.P., & Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, *23*(1), 35–41. doi:10.1016/0346-251X(94)00050-G

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

Rogers, J., Brizzard, C., Daulton, F., Florescu, C., MacLean, I., Mimura, K., . . . Shimada, Y. (2014). A methodology for identification of the formulaic language most representative of high-frequency collocations. *Vocabulary Learning and Instruction*, *3*(1), 51–65. doi:10.7820/vli.v03.1.2187-2759

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

Shin, D., & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal*, *62*, 339–348. doi:10.1093/elt/ccm091

Snellings, P., van Gelderen, A., & de Glopper, K. (2002). Lexical retrieval: An aspect of fluent second–language production that can be enhanced. *Language Learning*, *52*, 723–754. doi:10.1111/1467-9922.00202

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative methods. *Functions of Language*, *2*(1), 23–55. doi:10.1075/fol.2.1.03stu

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, *21*, 463–489. doi:10.1093/applin/21.4.463

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press.