Vocabulary Learning and Instruction
Volume 3, Issue 1, December 2014
http://vli-journal.org

Article

A Methodology for Identification of the Formulaic Language Most Representative of High-frequency 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.v03.1.rogers.et.al

Abstract

Researchers have stated that learning formulaic language is key to achieving fluency. It has also been stated that studying vocabulary in this way is more efficient than isolated vocabulary learning. However, there is a lack of research in regards to which formulaic language should be taught. There is a further lack of research about how such formulaic language can be identified. This study aimed to evaluate a methodology for identifying the most common formulaic language. It compared multi-word unit identification results from both 500 and 1,000 example sentences and quantified how often native speakers opt to extend multi-word units beyond their core pivot and collocate. This study also identified and quantified colligational issues affecting multi-word unit identification. The results showed no difference in multi-word unit identification between 500 and 1,000 example sentences, that native speakers opted to extend multi-word units more than half of the time, and that colligational issues only affected approximately 3% of the items examined. This study concluded that 500 example sentences are just as reliable as 1,000 when identifying multi-word units. It also found that extending multi-word units beyond their core pivot and collocate is an essential step researchers should take. This study also found that a colligational treatment is necessary if the aim is to achieve the most accurate data; however, the percentage of items that were affected were small and the methodology time-consuming. This finding indicates that there is a need for improved software to better automate the steps taken.

1 Introduction

Researchers agree that knowledge of formulaic language is essential if native-like fluency is to be achieved (Cowie, 1998; Wray, 2002). However, many researchers also state that there is a severe lack of emphasis on formulaic language (Gitsaki, 1996; Nesselhauf, 2005). Thus students fail to develop formulaic knowledge and struggle to obtain native-like fluency. So why is there a lack of emphasis on this important aspect of language fluency despite that researchers have

Please cite this article as: 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.rogers.et.al

agreed on its importance? The reason is connected to the complexity of identifying such items, and the resulting lack of resources to help develop such fluency. Identifying high-frequency collocations and the formulaic language most representative of those collocations is a complex process. In addition, there is a lack of consensus regarding what a collocation is. Furthermore, various different methodologies have been used in the past to identify high-frequency collocations/formulaic language, but all have flaws and/or lack comprehensiveness. Thus despite recent advancements concerning how words and cooccurrence of words can be counted, there is still a lack of research incorporating this knowledge.

In response, this study introduces a methodology that identifies the formulaic language most representative of high-frequency collocations. This study also provides examples of the types of data that can be identified when using such a methodology.

2 Literature Review

Researchers agree that knowledge of formulaic language is central to language fluency and that collocation is a major part of formulaic language. Lewis (2000) believes teaching collocations to be "a top priority in every language course" (p. 8). But what exactly is a collocation? In fact, many researchers still struggle to agree on a comprehensive definition. Traditionally most researchers have defined collocations as the tendency for words to frequently cooccur (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Shin, 2006). Other criteria have been utilized to delimit what can be considered as collocations, such as utilizing mutual information data, but such criteria have been found unreliable (Shin, 2006; Stubbs, 1995).

Other researchers recommend that only semantically opaque words that frequently cooccur be considered collocations (Moon, 1994) and that only such items be taught directly because they have a higher learning burden. However, researchers such as Nesselhauf (2005) and Wray (2000) highlight flaws in this approach. Semantic transparency does not necessarily equate to a lower learning burden since other criteria often affect a collocation's learning burden, such as L1-L2 congruency. In fact, Feyez-Hussein (1990) found that 50% of collocation errors were due to L1 influence. Thus, the most reliable criterion to initially identify collocations worth teaching is still frequency of co-occurrence, and therefore this study will utilize this definition.

If collocations are defined as words that frequently cooccur, how cooccurrence is counted must also be addressed. What exactly is a word and how should we count words? In fact, there are many ways to approach this. For instance, the simplest way to count words would be as *word types*. A *word type* distinguishes all lexical items that have different spellings. For instance, the word *eat* is considered separately to *eats*. However, this method of counting is not ideal for a study of high-frequency collocations because of the sheer number of collocations that exist. Hill (2000) estimates there to be hundreds of thousands of collocations in English. Thus consolidating data is essential.

Data consolidation can be accomplished in a number of ways. For example, it would be preferable to count the formulaic sequences *eat dinner* and *eats dinner*

together because the learning burden of learning one after the other is very low. With general affix knowledge, a learner can handle such differences on their own without both items needing to be taught at separate times. This can be accomplished by counting words as *lemma.* A lemma is a "set of related words consisting of the stem and inflected forms that are all the same part of speech" (Nation & Meara, 2002, p. 36). For example, the verbs *run, runs, running* and *ran*, are all counted together when lemmas are used, while the noun *run* would be counted separately. Counting word as *word families* consolidates data even further. A *word family* is "a headword, its inflected forms, and its closely related derived forms" (Nation, 2001, p. 8). When counting words with word families, the verb and noun forms of *run* are counted together and listed as a singular entry under the headword *run.*

Counting words as word families certainly has advantages in specific types of research, however, such a methodology can be problematic as well. This is because the headword that represents a word family is not always the most frequent lexical item that the family includes. For example, in Table 1, it is clear how the headword *depress* can be misleading. The word family is represented by the verb *depress,* despite the fact that the noun *depression* has significantly higher frequency. Furthermore, it is erroneous to make the assumption that learners can simply extend their affixed knowledge, thus equating to knowledge of the verb *depress* extending to knowledge of the whole word family. Both Schmitt and Meara (1997) and Daulton (2008) found Japanese learners to struggle with this task. So if the goal of a study is to identify a specific example of formulaic language to teach directly to learners, then breaking down word families into smaller groups of words, such as *lemma*, would be preferable. Lemma grouping can be broken down even further into word *types*, but this does not take advantage of a learner's ability to extend simple affix knowledge, such as the difference between the nouns *dog* and *dogs*, which is more plausible than, for instance, extending knowledge of a noun to an adjective.

For the above reasons, lemmatized collocation pairs are preferable when identifying high-frequency collocation, and therefore, this method of counting

Table 1. Frequency Counts in the Corpus of Contemporary American English (COCA; Davies, 2008) for Word Types in the Word Family *depress*

Word type	Frequency in the corpus
depression	19,176
depressed	6,715
depressing	2,032
depressive	1,598
anti-depressants	758
anti-depressant	533
depress	411
depressingly	152
depresses	144
depressant	58
depressives	31
depressants	18

words will be utilized in this study. The way this is achieved is by counting cooccurrence of words as *concgrams*.

A *concgram*, as defined by Cheng, Greaves, and Warren (2006), "constitutes 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 cooccurring 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 includes 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 formulaic language concordance search for the lemma *provide* and *support*. Table 2 shows the first five results of an actual concgram search for the lemma *provide* and *support*. These data are sourced from the Corpus of Contemporary American English (COCA)'s online interface, which allows for lemma concgram searches and provides snippets of the sentences these concgrams are occurring in.

Then, this concordance data could be processed to identify the formulaic language, or *multi-word unit*, most representative for the lemma *provide* and *support*. When 500 such snippets from the COCA are processed, it is revealed that *provide support* is the most common multi-word unit. Table 3 shows the top three multi-word units for this lemma pair.

Concgramming has significant advantages when the goal is to identify formulaic language most representative of high-frequency collocations. Cheng et al. (2006) state that "searches which focus on contiguous collocations present an incomplete picture of the word associations that exist" (p. 431). In other words, attempts to identify formulaic language that are not done as concgram searches are not reliable. However, much of the previous research that aimed to identify high-frequency formulaic language was actually conducted in this way (Biber, Conrad, & Cortes, 2004; Shin, 2006; Simpson & Mendis, 2003). Therefore, there is a severe gap in the research that this study aims to fill.

Table 2. A Sample of Data from the COCA for a Concgram Search for the Lemma *provide* and *support*

. . . low-cost measures, the United States can extend the same lifesaving **support** that it has **provided** to the little boy in a rural, dusty village to the working-age woman living . . .
. . . psychiatrists, nurses, addiction and employment counselors, and peer **support** specialists. PHF **provides** community-based services, and a service coordinator is always on call to help clients address . . .
. . . it, then provide technical support to assist them. This **support** can usually be **provided** through a single phone call or demonstration. If needed, seek assistance from school . . .
. . . losing those aid dollars that we need in order to get **support** when Pakistan does **provide** it, which is real and does help us in the case of drones to . . .
. . . for low-income adults in occupational programs as well as financial **support** to colleges to **provide** support services for such students. States and colleges interested in adopting a model similar . . .

Table 3. Top Three Multi-word Units for the Lemma *provide* and *support* Found after Examining 500 Concordance Strings in the COCA

Multi-word unit	Frequency
provide support	55
support provided	39
support provided by	32

However, simply identifying lemma pairs that co-occur frequently is insufficient to provide learners with specific items to study. For instance, *take/walk* collocate, but it is not enough to simply expose students to this lemma pair. Rather, a more specific example of how the two collocate as a multi-word unit needs to be identified. Is it *taking walks, took walks, take a walk,* etc.? Thus steps are required to identify the multi-word unit most representative of that lemmatized concgram. This is accomplished via concordance software. However, working with concgrams is not simple, and thus this paper provides guidance on how this can be accomplished. Furthermore, another pertinent question is whether a multi-word unit identified as most representative of a lemmatized concgram should go beyond the pivot and collocate. For instance, should an identification method stop at *take a walk* or should it extend beyond this to identify *take a walk to*?

Colligation, or the counting various lexical items that can easily substitute for one another as grammatical categories (Gitsaki, 1996; Renouf & Sinclair, 1991), is another important criterion for formulaic sequence identification about which there is a lack of research. An example of colligation is counting the collocates *early* and *century* as *early* [*year*] *century* when they occur with years, which would account for instances, such as *early twentieth century*, *early nineteenth century*, etc., together. Table 4 shows the advantage of processing corpus data with consideration for colligation. One thousand example sentences were collected from the COCA (Davies, 2008), and a concordance search identified the multi-word unit most representative of how *century* and *earlier* occur together. One search was done with consideration for colligation, replacing every instance of a year with the marker *[year]*. By considering colligation, the top multi-word unit identified was shown to have nearly double the frequency in comparison with the top multi-word unit identified without consideration for colligation.

However, depending on the goal of the research, colligation also has the potential to create more problems than it solves. For instance, when major content word categories, such as nouns or verbs, are replaced with colligational markers, the limitations of how a multi-word unit can be formulated may not be conveyed to the learner. Take the colligational framework *[adjective] tea,* for instance. Typical examples such as *hot tea, brown tea,* or *strong tea* are perfectly logical, but it becomes very difficult to explain why *powerful tea* is not an option. Due to this idiosyncratic way collocations occur, grammar alone is not sufficient to determine which lexical items cooccur (Lewis, 2000). Regardless, colligation may be an important criterion to consider when identifying formulaic language. Yet how this criterion can be implemented and the extent of its value remains to be seen. Thus

Table 4. A Comparison between Two Multi-word Unit Searches, One with and One without Consideration for a Specific Type of Colligation

% of occurrences in 1,000 example sentences	Multi-word unit with cooccurrence of *century* and *early*	% of occurrences in 1,000 example sentences	Multi-word unit with cooccurrence of *century* and *early*
Without consideration for colligation		With consideration for colligation	
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*

this paper aims to clarify the value of specific types of colligational searches and provide examples of the types of data that result from such consideration.

3 Research Questions

1. What percentage of the most common multi-word units is affected when specific types of colligations are considered?
2. Compared with the results of multi-word unit searches without consideration for colligation, what percentage of items identifies a different multi-word unit as being most representative of the lemmatized concgram?
3. Should multi-word units be extended beyond the pivot and collocates, at the beginning and end of a multi-word unit, to provide learners with more information about how the target items commonly occur formulaically?

4 Materials

This study will begin by utilizing Rogers et al.'s (in press) list of 12,604 high-frequency lemmatized concgrams. This list was originally derived from Davies' (2010) *Word List Plus Collocates,* a list of collocations that occur with the most frequent 5,000 lemmas of the COCA. To distinguish only items from this list that are useful for learners of general English, Rogers et al. (in press) delimited the list by frequency (approximately one occurrence per million tokens), and only included items with balanced range and chronological data.

Concordance data for each of the 12,604 concgrams was collected from the COCA. This study's approach necessitated the writing of custom concordance software to identify the most common multi-word units. Using normal concordance software, such as Anthony's (2011) *AntConc,* was not an option because

this study aimed to identify only multi-word units in which both lemma occurred, a function not possible with *AntConc* or other concordance software. Furthermore, the large amount of data (over 12,000 pairs) required a batch processing option, another feature not possible with current concordance software. Thus this study used the custom concordance software *AntWordPairs* (Anthony, 2013), a program written specifically for this study. It utilizes Someya's (1998) *E-lemma list*. For coding purposes, Someya's lemma list could not contain duplicate entries, and thus was modified to remove homonyms. For part of speech tagging, the software *GoTagger Version 0.7* (Goto, 2005) was utilized, and for colligational marker substitution, the software *Textcrawler* (Digital Volcano, 2011) was utilized.

5 Procedure

The first step was to collect concordance data (example sentences) for each of the 12,604 lemma pairs. Lemmatized concordance searches were conducted, using the COCA's online concordance interface, to identify instances when the collocate occurred either three words to the left or right of the node word. The rationale for this length (seven words) was influenced from findings on typical human memory limitations (Miller, 1956). The COCA's interface provides options for 100, 200, 500, or 1,000 example sentences to be extracted. Since more data provide the most reliable results, this study began by collecting 1,000 example sentences for each pair. However, because of COCA download limits, and the time required for sentences to load, 1,000 sentences was deemed impractical. However, to ensure that 500 example sentences provided as reliable data as 1,000 sentences would, results from 10 random lemma pairs were compared using both 500 and 1,000 example sentences. Starting with pairs which had frequency counts of 1,000 or more, every 500th pair was selected from the list which was sorted by frequency. Extracting 500 example sentences per lemma pair essentially created a mini corpus for each pair consisting of approximately 13,000 words per pair.

The next step was to identify specific categories of lexical items that occur in high frequency that could be substituted with colligational markers. Essentially, the goal was to experiment with a number of items that could be substituted with a marker that does not impede the meaning of the multi-word unit as a whole, while providing more accurate frequency counts. Table 2 is a perfect example of such an item. However, since no previous research existed, a number of items needed to be chosen and experimented with. A multi-word unit search was conducted on all 12,604 lemma pairs without consideration for collocation. A scan of the full data by a native English speaker revealed that particular categories of words (pronouns) occurred quite often in the multi-word units identified and could easily be substituted without disruption of the meaning of the multi-word units as a whole. In addition, a number of other word categories were used in the colligation treatment: *months, days of the week, oridinal numbers,* and *cardinal numbers.*

To use the colligational categories, adjustments for homonyms in the corpus data were necessary. This was done by part of speech tagging using *GoTagger* and making replacements using *Textcrawler.* First, all instances of the pre-nominal possessive pronoun *her* were changed to *his* as to not interfere with the object pronoun *her.* Then, instances of the ordinal number *second* were changed to 2nd as

to not interfere with the noun *second.* Next, instances of the nominal possessive personal pronoun *his* were changed to *hers* to not interfere with the pre-nominal possessive pronoun *his.* Then, the nominal possessive personal pronoun *mine* was replaced with *yours* to not interfere with the noun *mine.* Furthermore, instances of the month *May* and *March* were replaced with *January* to not interfere with the auxiliary verb *may* and the verb *march,* respectively. In addition, the day of the week abbreviations *Sun, Wed,* and *Sat* were replaced with *Mon* to not interfere with the noun *sun* and the verbs *wed* and *sat*, respectively.

Then, *Textcrawler* was used to replace all the pronouns, months, days of the week, ordinal and cardinal numbers with distinct colligational markers in each mini-corpus. The data were then processed with *AntWordPairs* to identify the most common multi-word units each lemma pair occur in. Because the amount of resulting data was excessive, only multi-word units occurring in 5% or more of the corpora were collected. Furthermore, a limit of seven words was set for the length of the multi-word units.

Next, five native English speakers examined the data to not only extract the most frequent multi-word unit but to also extend the multi-word unit beyond the most frequent item to its left or right when the native speaker judged any additions to be part of the natural unit.

The next step was a random sample of the multi-word units that were affected by the colligational treatment, and a concordance search with the original data not treated for colligational to judge whether a different multi-word unit was identified.

The final step taken in this study was to examine a random sample of 100 multi-word units identified and determine which percentage native speakers extended beyond the pivot and collocate.

6 Results

Data from 10 random concordance searches were examined for differences between using 500 and 1,000 example sentences.

Between the two amounts, the same top multi-word unit was identified for every pair examined, regardless of whether 500 or 1,000 example sentences were used. The data also show that the frequency counts varied very little when comparisons were made. Table 5 shows the top multi-word unit identified for each of the 10 pairs examined.

After the initial concordance search, distinct categories of words were found to occur frequently in the multi-word units identified. The vast majority of these was pronouns. Thus colligational markers were created for the following types of pronouns:

1. Pre-nominal possessive pronouns (your, his, her, their, my, our, its)
2. Subject pronouns (I, you, he, she, they, we, it)
3. Object pronouns (me, us, him, her, them)
4. Nominal possessive personal pronouns (theirs, his, hers, yours, mine)
5. Singular reflexive personal pronouns (myself, yourself, himself, herself, itself, yourselves, themselves, ourselves)

Table 5. The Top Multi-word Unit Identified When 500 and 1,000 Example Sentences Were Utilized

Lemma	POS	Lemma	POS	Multi-word unit identified	% out of 500 sentences	% out of 1,000 sentences
announce	verb	week	noun	announced last week	21.6	20.0
trade	noun	deficit	noun	trade deficit	85.6	84.7
body	adj.	upper	adj.	upper body	87.2	86.2
up	adv.	high	adv.	high up	70.0	66.5
little	adv.	better	adv.	little better	100	97.5
court	noun	hold	verb	court held	40.2	42.5
take	verb	charge	noun	take charge	46.4	38.7
care	verb	people	noun	people who care	15.4	10.8
get	verb	look	noun	get a look	23.2	15.7
too	adv.	often	adv.	too often	57.4	33.4

It was also determined that four other additional colligational categories should be replaced with colligational markers since they were seen occurring in the original concordance search, did not disrupt the meaning of the multi-word unit as a whole, and could potentially provide more accurate frequency counts. There were:

1. Months (January, Jan, February, Feb, Mar, April, Apr, May, June, Jun, July, July, August, Aug, September, Sept, October, Oct, November, Nov, December, Dec)
2. Days of the week (Sunday, Sun, Monday, Mon, Tuesday, Tue, Wednesday, Wed, Thursday, Thurs, Friday, Fri, Saturday, Sat)
3. Ordinal numbers (1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th, 9th, 10th, 11th, 12th, 13th,14th, 15th, 16th, 17th, 18th, 19th, 20th, 21st, 30th, 40th, 50th, 60th, 70th, 80th, 90th, 100th, first, second, third, fourth, fifth, sixth, seventh, eighth, ninth, tenth, eleventh, twelfth, thirteenth, fourteenth, fifteenth, sixteenth, seventeenth, eighteenth, nineteenth, twentieth, twenty-first, thirtieth, fortieth, fiftieth, sixtieth, seventieth, eightieth, ninetieth, one-hundredth)
4. Cardinal numbers (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38,39,40,41,42,43,44,45,46,47,48,49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000, 8,000, 9,000, 10,000, 20,000, 30,000, 40,000, 50,000, 60,000, 70,000, 80,000, 90,000, 100,000, 200,000, 300,000, 400,000, 500,000, 600,000, 700,000, 800,000, 900,000, 1,000,000, one, two, three, four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fourteen, fifteen, sixteen, seventeen, eighteen, nineteen, twenty, thirty, forty, fifty, sixty, seventy, eighty, ninety, one-hundred, one-thousand, ten-thousand, one-hundred thousand, one-million)

It should be noted that these selections are not all-encompassing and other potentially useful colligational patterns may certainly be present in the data.

However, due to practical time and computing limitations this paper could only deal with the above colligational categories and the items listed within them.

After all the mini-corpora were adjusted for homonyms and processed with *AntWordPairs* to identify the multi-word units, and native speakers extracted the multi-word units most representative of how each lemma pair cooccurs, and the amount of multi-word units identified that were affected by the colligational treatment were counted. The results are shown in Table 6.

The colligational treatment for prenominal possessive pronouns was shown to be the most common. As much as 2.1% of the lemma pairs' top multi-word units were affected by this colligational treatment. Treatments for subject pronouns and cardinal numbers also resulted in a significant amount of items being affected. In total, 5.8% of all of the top multi-word units (727 items) identified were affected by all the colligational treatments conducted.

Ten random samples were then taken from the top three types of colligation treatment found to affect the top multi-word unit identification. These were then compared to a top multi-word unit search with untreated data. Out of the 30 items selected, only 13 (43.3%) resulted in different multi-word units being identified. For items affected by the prenominal possessive pronoun treatment, only 4 out of 10 top multi-word units differed. With the subject pronoun treatment, only 3 out of 10 top multi-word units differed. With the cardinal number treatment, 6 out of 10 of the top multi-word units differed. These results are summarized in Tables 7, 8, and 9.

Native speakers opted to extend multi-word units beyond the core pivot and collocate in 53 percent of the 100 random multi-word units sampled. For instance, the most frequent multi-word unit for the lemma pair *come* and *term* was found to be *come to terms*, at 243 occurrences (see Table 10). However, the next most common string in the data beyond *come to terms* was *come to terms with* (229 occurrences), and beyond that, *to come to terms with* (129 occurrences). Thus a native speaker judged *to* **come to terms** *with* as being the multi-word unit most representative of the lemma pair *come* and *term.* Core multi-word units were

Table 6. Amount of Top Multi-word Units That Were Affected by Each of the Colligational Treatments

Colligational treatment	Number of top multi-word units affected	Percentage of total lemma pairs (%)
Pre-nominal possessive pronouns	259	2.1
Subject pronouns	208	1.7
Cardinal numbers	171	1.4
Object pronouns	74	0.6
Ordinal numbers	14	0.1
Singular reflexive personal pronouns	1	0.007
Nominal possessive personal pronouns	0	0
Months	0	0
Days of the week	0	0
Grand totals	727	5.8

Table 7. Comparison between 10 Random Samples of Top Multi-Word Units Affected by the Colligational Treatment for Prenominal Possessive Pronouns and the Results That Would Have Occurred without the Treatment

Lemmatized concgram pair	Multi-word unit identified w/ colligational treatment	Multi-word unit identified w/o colligational treatment
***hand* (noun) *wave* (verb)**	***waved * hand***	***waved a hand***
***live* (verb) *life* (noun)**	***live * life***	***live life***
***base* (verb) *experience* (noun)**	***based on * experience***	***based on experience***
***attention* (noun) *focus* (verb)**	***focus * attention***	***focus attention***
head (noun) *gun* (noun)	*gun to * head*	*gun to his head*
hand (noun) *extend* (verb)	*extended * hand*	*extended his hand*
eye (noun) *wipe* (verb)	*wiped * eye*	*wiped her eye*
life (noun) *ruin* (verb)	*ruin * life*	*ruin your life*
put (verb) *hand* (noun)	*put * hand*	*put her hand*
sit (verb) *desk* (noun)	*sitting at * desk*	*sitting at his desk*

Note: Items in bold indicate those that showed differences in the top multi-word unit identified, and instances of a slot in which a pre-nominal possessive pronoun exists are represented with "*".

identified in bold and any strings present in the data and also judged to be typically cooccurring with the multi-word unit were added in italics. To accomplish this, native speakers relied on their intuition to not only add strings that truly represented common usage, but that also provided learners with useful information.

7 Discussion

Regarding the amount of data collected to create each mini-corpus used in this study, 500 example sentences were deemed as reliable as 1,000 example sentences when concordance data were compared. The example shown in Table 3

Table 8. Comparison between 10 Samples of Top Multi-Word Units Affected by the Colligational Treatment for Subject Pronouns and the Results That Would Have Occurred without the Treatment

Lemmatized concgram pair	Multi-word unit identified w/ colligational treatment	Multi-word unit identified w/o colligational treatment
***see* (verb) *mirror* (noun)**	***mirror * saw***	***mirror and saw***
***wear* (verb) *dress* (noun)**	***dress * wore***	***wearing a dress***
***take* (verb) *back* (adverb)**	***take it back***	***take back***
how (adverb) *interact* (verb)	*how * interact*	*how they interact*
get (verb) *when* (adverb)	*when * got*	*when I got*
make (verb) *hard* (adverb)	*makes * hard*	*makes it hard*
could (verb) *suppose* (verb)	*suppose * could*	*suppose you could*
belong (verb) *where* (adverb)	*where * belong*	*where I belong*
think (verb) *pretty* (adverb)	*think * is pretty*	*think she is pretty*
want (verb) *whenever* (adverb)	*whenever * want*	*whenever you want*

Note: Items in bold indicate those that showed differences in the top multi-word unit identified, and instances of a slot in which a subject pronoun exists are represented with "*".

Table 9. Comparison between 10 Random Samples of Top Multi-word Units Affected by the Colligational Treatment for Cardinal Numbers and the Results That Would Have Occurred without the Treatment.

Lemmatized concgram pair	Multi-word unit identified w/ colligational treatment	Multi-word unit identified w/o colligational treatment
***get* (verb) *second* (noun)**	***got * seconds***	***seconds to get***
***nearly* (adverb) *decade* (noun)**	***nearly * decades***	***nearly a decade***
***just* (adverb) *year* (noun)**	***just * years***	***just a few years***
***live* (verb) *mile* (noun)**	***live * miles***	***live within 50 miles***
***nearly* (adverb) *mile* (noun)**	***nearly * miles***	***nearly a mile***
***minute* (noun) *second* (noun)**	***minutes * seconds***	***seconds to one minute***
estimate (verb) *percent* (noun)	*estimates that * percent*	*estimates that 80 percent*
divide (verb) *group* (noun)	*divided into * groups*	*divided into two groups*
over (adverb) *month* (noun)	*over * months*	*over six months*
roughly (adverb) *percent* (noun)	*roughly * percent*	*roughly 10 percent*

Note: Items in bold indicate those that showed differences in the top multi-word unit identified, and instances of a slot in which a cardinal number exists are represented with "*".

demonstrates that collection of 500 versus 1,000 example sentences for each lemma pair made no difference in identifying the most common multi-word unit. However, collecting the data was a manual process of copy and pasting from the COCA's interface, something it was not designed for. Thus through the process unnecessary data were also copied, and therefore, a multi-step process of pasting into an Excel file, then copying only the sentences and pasting again into a Word file, and then saving the file, was necessary to remove this data. Being a cumbersome, time-consuming process, corpus computer interface designers may want to consider this for future design.

When the initial concordance data were examined after processing the compiled mini-corpora, various types of pronouns occurred quite often within the multi-word units identified. Other categories of words, such as cardinal numbers, also frequently occurred. Thus such word categories became the focus of this study's colligation experiment. However, because of a lack of previous research, other categories were experimented with as well. Not all of these proved

Table 10. Multi-word Units Identified from 500 Example Sentences in Which the Lemma Pair *come* and *term* Both Occur

Multi-word unit	Occurrences in 500 sentences
come to terms	243
come to terms *with*	229
to **come to terms**	133
to **come to terms** *with*	129
coming to terms	96
coming to terms with the	86
to **come to terms** *with the*	44
come to terms *with* [pre-nominal possessive pronoun]	28
coming to terms with the	26

fruitful, however, the resulting data did provide an insight as to specific types of colligation that, when addressed, can improve upon the reliability of multi-word unit identification.

The colligational treatment for prenominal possessive pronouns was shown to be the most useful. Treatments for singular reflexive personal pronouns, nominal possessive personal pronouns, months, and days of the week did not prove useful; only one item was affected in the entire list by all of these treatments. At first glance, the colligational treatment was shown to be an important step in the identification of the most frequent multi-word units, most representative of lemmatized concgrams, in that 727 (5.8%) of the total concgrams examined had their most common multi-word unit change. However, when a sample of the multi-word units was compared to the multi-word units that would have been identified without a treatment for colligation, only 43.3% of the items actually had differing results. Therefore, while frequent counts were always improved upon, the treatments did not always end with improved results.

Yet before the colligational treatment could be conducted, homonym interference in the data had to be dealt with. The process was complex, cumbersome, and very time-consuming due to the lack of dedicated software to conduct such a task. It would be useful if software developers considered such functionality and ways to improve the efficacy of conducting such data modification.

In regards to the value of extending multi-word units beyond the core pivot and collocate, the data suggest that this is an important criterion to consider when attempting to identify multi-word units most representative of lemmatized concgrams. Native speakers opted to extend multi-word units in more than half of the items examined. Corpus data and software alone cannot accurately identify such extensions, and thus this aspect of the study highlighted the importance of native speaker intuition and intervention in multi-word unit identification.

8 Conclusion

This study discussed a methodology to identify multi-word units most representative of lemmatized concgrams. It highlighted the value of counting words as lemma to identify formulaic language most representative of high-frequency collocations, compared results from different sized pivot word/collocate corpora, provided hard data as to the type of results one can expect when conducting specific colligational treatments on data, and highlighted the value of extending multi-word units beyond the core pivot and collocate.

This study showed how 500 example sentences that contain a target pivot word and collocate are just as reliable as 1,000 example sentences. This study also showed that multi-word unit searches for 5.8% of the lemma pairs examined were affected by the steps taken in this study. However, when a sample of these items was examined more deeply, it was found that nearly half showed no difference in the top multi-word unit identified. While results did improve for approximately 3% of the items examined, the steps needed to achieve these improvements were time-consuming and complex. Therefore, this study indicated the need for a more efficient methodology for such colligation treatments. Software designers should

thus consider ways to automate some of the steps taken in this study. This study also highlighted the importance of extending multi-word units beyond the core pivots and collocates, as over half of the items examined benefited from this procedure.

This study had its limitations. Due to the lack of previous research and no standard on how to conduct such a data analysis, choices for the types of colligation examined were subjective. Quite possibly other types of colligation exist in the data that could also prove fruitful if treated. Thus more research is needed in regards to other types of colligation that may improve results if treated. Despite these limitations, this study did provide new insights into a previously unexplored area of linguistic analysis that certainly has the potential for creating improved resources that help learners achieve fluency in a second language.

References

Anthony, L. (2011). *AntConc* (Version 3.2.2) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2013). *AntWordPairs* (Version 1.0.2) [Computer Software]. Tokyo, Japan: Waseda University.

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., 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

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

Daulton, F. (2008). *Japan's built in lexicon of English-based loanwords*. Clevedon, Oh: Multilingual Matters.

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

Digital Volcano. (2011). *Textcrawler* (Version 2.5) [Computer Software]. Retrieved from http://www.digitalvolcano.co.uk/content/textcrawler

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: Adam Mickiewicz University.

Gitsaki, C. (1996). The development of ESL collocational knowledge (Unpublished doctoral dissertation). University of Queensland, Brisbane.

Goto, K. (2005). *GoTagger* (Version 0.7) [Computer Software]. Retrieved from http://web4u.setsunan.ac.jp/Website/GoTagger.htm

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: Thomson Heinle Language Teaching.

Lewis, M. (2000). Language in the lexical approach. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 8–10). Hove, UK: Thomson Heinle Language Teaching.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97. doi:10.1037/h0043158

Moon, R. (1994). The analysis of fixed expressions in text. In M. Coulthard (Ed.), *Advances in written text analysis* (pp. 117–135). London, UK: Routledge.

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

Nation, P., & Meara, P. (2002). Vocabulary. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 35–54). London, UK: Edward Arnold.

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

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

Rogers, J., Daulton, F., Brizzard, C., Florescu, C., MacLean, I., Mimura, K., … Shimada, Y. (in press). On using corpus frequency, dispersion and chronological data to help identify useful collocations. *Research in Corpus Linguistics*.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework. *Studies in Second Language Acquisition*, *19* (1), 17–36. doi:10.1017/S0272263197001022

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

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

Someya, Y. (1998). *E-lemma list*. Retrieved from http://www.antlab.sci.waseda.ac.jp/software/resources/e_lemma.zip

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.