Introducing the First Large-Scale English Collocational Chunk List and Innovative Methods in Which Collocational Fluency Can Be Mastered

James Rogers, Meijo University, Japan

The IAFOR International Conference on Language Learning – Hawaii 2017
Official Conference Proceedings

Abstract
Many researchers agree that knowledge of English collocations and the formulaic chunks they occur in is an important aspect of language fluency. However, a number of issues have led to a lack of research and resources despite the awareness of the importance of these aspects of language knowledge. This study will discuss why this gap in the research exists, and steps that were taken to fill it to create the first large-scale collocational chunk list to help learners master general English. First, an overview of why this gap in the research existed will be provided. Then, the rationale for all steps taken to create the resource will be explained, and a detailed analysis of the usefulness of each of these steps will be provided. Finally, a description of an innovative Leitner algorithm-based smartphone app which students can use to study the contents of the resulting resource will be given to highlight how these chunks can be studied.

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

Introduction

In recent years, more and more researchers have acknowledged the importance of collocational fluency in second language learners. Lewis (2000) believes that mastering such knowledge "should be a top priority in every language course" (p. 8). Hoey (2005) and Hill (2000) agree, stating that collocational knowledge plays a central role in language since much of language itself consists of prefabricated chunks.

Knowledge of these chunks enables learners to speak more naturally (Durrant & Schmitt; 2009, Wray, 2002; Cowie, 1998) and process language more efficiently (Nation, 2001) in comparison to studying isolated vocabulary. In fact, learning vocabulary formulaically can in fact be more efficient in comparison with learning isolated vocabulary items because learners can utilize words in a chunk as mnemonics to help them remember the other words in the chunk (Schmitt, 1997). However, a number of researchers have shown that learners are not obtaining collocational fluency (fluency (DeCock et al., 1998; Kallkvist, 1998; Nesselhauf, 2005). But why are learners failing to acquire this important aspect of language fluency? The main reason is that collocations are not focused on in materials and/or the language classroom (Gitsaki, 1996). This stems from the issue of there being very few studies that identify the common formulaic chunks of English (Durrant & Schmitt, 2009). Until Rogers (2017), which identified over 11,000 formulaic chunks that are frequent in general English, most studied were limited in their size and/or methodology. For instance, Martinez and Schmitt's (2012) phrase list only consisted of 505 items. Shin's (2006) study unfortunately utilized types instead of the more efficient concgramming method with lemma. Furthermore, his study also only examined the most frequent 1,000 types of English, while Rogers (2017) examined items on a much larger scale (the most frequent 5,000 lemma).

This current study will discuss the steps Rogers (2017) took to fill this gap in the research. It will provide a rationale for and a review of the usefulness of each of the steps taken, and then describe an innovative Leitner algorithm-based smartphone app in which the resource Rogers (2017) can be studied with.

Conclusion

Frequency as a criterion

The first step that was taken to fill this gap in the research was to determine a frequency cut-off for collocations of the top 5,000 lemma of the Corpus of Contemporary American English (Davies, 2008), or the *COCA*. At one occurrence per million tokens, the result was approximately 11,000 chunks. This amount may seem impractical in regards to direct study, but these 11,000 chunks actually only consist of around 3,000 word families. For example, "take a break" occurs but also "take a chance" and "chances of winning" and "winning an award". So, it's not 11,000 different items, but rather 3,000 items and the various ways in which they co-occur with each other.

The concgramming approach

At this stage, the list is still just a list of lemma. To go from lemma to chunks, the concgramming (Cheng, Greaves, & Warren, 2006) methodology was utilized. This method is ideal because it considers *constituency variation* (AB, ACB), or when a pair of words not only co-occurring adjacent to one another (*lose weight*) but also with a constituent (*lose some weight*) AND it also considers *positional variation* (AB, BA), or when total occurrences of two or more particular lexical items are counted while 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 chunk concordance search for the lemma *provide* and *support*.

A simple example would be the need to identify and rank in frequency the common phrase *take a break. Take/break* occur often together as *take a break,* but they also occur as *take breaks, taking breaks, took a break, take a quick break,* and so on. Essentially, these are all the same collocation. So, with concgramming they are counted together, and then by examining a mini corpus of only *take/break* concordance strings, we can identify *take a break* as the most common chunk. If these items are counted separately, the true frequency of the collocation is not represented. In addition, if such items are counted separately, it does not result in a useable resource for the end users such as teachers and students. For example, Simpson-Vlach and Ellis' (2010) Academic Formula List identified 712 common chunks in academic writing. However, they did not utilize the concgramming method. Thus, in their list there are the following items at different points:

there are a number of
there are a number
are a number of

Such items need to be consolidated for learners, and to improve upon frequency counts as well. When all three of these are added together, the rank of *there are a number of* will go up in the list, and this is important when learners have limited time and only can study a certain amount of items but want to study in the most efficient way possible.

The lack of dedicated software and development thereof

However, to accomplish this step software did not even exist. While it is possible to do it with concordance software such as AntConc (Anthony, 2013), 11,000 files would have to be processed manually, and the results would also have a large amount of noise in each set that would have to be removed manually. For example, if you make a mini corpus in which there are only concordance strings with *take/break* occurring and use AntConc, the top chunks identified will probably be *of the*, *in the*, and so on. Thus, there was a need to create dedicated software for this project. The result was the development of *AntWordPairs* (2013), a one-off software designed specifically for this research project which was able to accomplish the task at hand.

Balanced dispersion and chronological data as criteria

The next step take was to analyze corpus dispersion and chronological data to determine if it was reliable enough to identify items which only had balanced dispersion over a wide range of language genres (because the aim was to create a resource for learners of general English) and also balanced dispersion over time (because it is not appropriate to include dated terms such as *word processor,* and it also is not ideal to include items that only occur during limited time periods such as *saving and loan*). However, corpus data proved unreliable for this step and thus items were examined manually using native speaker intuition. Thus, this step proved extremely time consuming.

Colligation as a criterion

Then *colligation* was considered, or in other words, when a group of words can be substituted by a grammatical marker (such as numbers, days of the week, etc.). Take the example of *early/century* in the table below which analyzed 500 concordance lines from the *COCA* to determine which chunk occurred most often when the two words co-occurred.

Without consideration for colligation		With consideration for colligation (years consolidated)	
10.70%	century earlier	19.20%	early in the [year] century
9.50%	a century earlier	10.70%	century earlier
6.70%	early in this century	9.70%	early [year] century
6.40%	centuries earlier	8.50%	early in this century
5.80%	early in the century	8.30%	early as the [year] century
5.00%	early in the 20th century	8.30%	as early as the [year] century

Figure 1: A comparison between two chunk searches, one with and one without consideration for a specific type of colligation

Without any consideration for colligation, the data analysis results in *century earlier* being the most common chunk occurring when *early* and *century* occur together. However, if years (such as *18th, 19th,* and *20th* and *eighteenth, nineteeth,* and *twentieth*) are counted all as one category, the results drastically change with *early in the [year] century* resulting in double the amount of occurrences in comparison to *century earlier.* The data above highlights how consideration for colligation has the potential to improve upon the accuracy of identifying the most common chunk two words co-occur in. It should be noted that while this step was shown to be useful to a small extent in improving upon the quality of the data, it was extremely complex and time consuming due to a lack of dedicated software.

On extended chunks beyond their cores

Next, an experiment was conducted to determine whether or not native speakers felt it would be beneficial for learners to be exposed to words that commonly occurred to the left and right of the core chunks that were identified initially. For instance, when *close proximity* was identified for the lemma *close/proximity*, slightly lower in rank was *close proximity to*, and then a bit lower *in close proximity to*. In such cases, it was decided by the native speaker to have *in close proximity to* represent the lemma *close/proximity* instead of simply *close proximity*. This step was deemed absolutely essential in that native speaker opted to extend in nearly half of the approximately 11,000 items.

On semantic transparency as a criterion

After that, the extent of semantic transparency of the items were determined by native speakers. Only 14 percent were considered to be semi-figurative, figurative, or core idioms. With 86 percent of the high-frequency collocations of English being literal formulations, to say we shouldn't teach literal formulations directly would highly limit exposure to the vast amount of high-frequency collocations for learners, and thus this criterion was deemed to be problematic in regard to the high-frequency chunks identified in this study.

Literal	ONCE	Figurative	Core Idiom	Outlier
9,641/*86.06*	76/*6.01*	93/*1.7*	179/*1.65*	19/*4.7*

Figure 2: Sematic transparency ratings of the collocations (percentage of total items in italics)

On L1-L2 congruency as a criterion

In regards to L1-L2 congruency, this criterion in fact trumps semantic transparency because it does not matter if a chunk is a literal chunk or not since if it is said in a different way in the learner's L1, they will have a high chance of making an error with it and thus it needs to be taught directly. For instance, in English we say *get credits* for a class, but in Japanese the way it is said can literally be translated into *take credits,* and thus students will often make this error by directly translating. Thus, the literalness of *get credits* becomes moot. Therefore, to take this criterion into consideration it was necessary to translate all 11,000 items into Japanese and to give each an L1-L2 congruency rating on a scale from 0-12.

0-3	4-6	7-9	10-12 (12)
996	2,419	2,905	4,888 (4,146)

Figure 3: L1-L2 congruency ratings of high-frequency English MWUs with Japanese translations

When items that received a rating of six or less were kept, the 11,208 items becomes only 3,414 items that have a higher chance of learners making an error with them. Such a resource, at 50 items per week, could be mastered in approximately a year and a half, exposing learners to a large majority of the way high-frequency vocabulary collocate that they have the highest chance of making an error with.

On the reliability of native speakers to create high-frequency context

However, just having the chunks themselves is not enough for the end users (students/teachers). Having example sentences for each item is ideal because learners can then see the proper context in which these chunks are normally used. A team of native speakers were thus given the task to create an example sentence for each of the approximately 11,000 chunks. They were instructed to try to only use high-frequency vocabulary when they created the surrounding context for the chunk's sentence. So, the next research question became whether or not native speaker intuition could be relied upon to select only high-frequency items for surrounding context. The example sentences added 160,000 words of content to the list. The resulting resource not only covered 90 percent of the top 3,000 word families of English, but in addition, 97 percent of the words in the sentences created fell within the top 3,000 word families. Thus, the answer was clearly affirmative that native speaker intuition is highly reliable for this task.

On Japanese university students' knowledge of high-frequency chunks

The final step in this study was to confirm that this resource constitutes knowledge that all native speakers have, but that learners do not. Thus a 50 question cloze productive test was created with a balanced selection from the 11,000 chunks in regards to the following criteria:

1. Frequency
2. Semantic transparency
3. L1-L2 congruency

Test questions were then created for each, such as the following:

I doubt my son is going to follow t _ _ _ _ _ on his promise to cut the grass.

Pilot tests with native speakers showed that all items could be answered correctly. 549 Japanese university freshmen with an average TOEFL score of 421, and a wide range of proficiency took the test and the average score was 23 percent correct. Such a low average score in comparison with native speakers perfect scores showed that this resource consists of knowledge that native speakers possess but the learners in questions lack, and thus confirmed its value.

On the creation of a smartphone app to study the resource with

A number of researchers have cited the potential of flashcard programs to improve education (Burston, 2007; Ishikawa, 2004). Goodwin-Jones (2010) specifically points out how software which features spaced repetition of items can help a learner to better commit information to long-term memory. An even more advanced method would be software which utilizes an *Leitner* algorithm to determine which items are studied next. With such an algorithm, not only is time considered (as is with spaced repetition), but also the difficulty of the item because the *Leitner* algorithm enables a user to mark an item as something that they know or don't know. When users mark items in such a way, the algorithm takes that into consideration when it calculates which item is to come next. Thus, learners are exposed to more difficult items more often and therefore get the extra exposure needed to master such items. Rogers and Reid (2015) found that studying chunks with a smartphone app that featured such an algorithm resulted in an average score of 57 percent, while when learners studied the same amount of similar chunks but on paper, the average score was only 41 percent.

Therefore, a smartphone app series was developed to make the chunk resource available to learners in. A customized version of the app *Flashcards Deluxe* was created, an app which has over 100,000 pay downloads and has been listed as an iTunes Store bestseller. It not only features a unique and advanced Leitner algorithm, but new features were also added to it, such as the ability to take quizzes and to submit them online and teachers can also able to see how many minutes students are studying with the apps for. The approximately 11,000 chunks were released in an app called 英語マスター1 万 (English Master 10,000) (Rogers, et al., 2015) and the approximately 3,000 chunks that differ with Japanese learners L1 was released as an app called 英語マスター3 千 (English Master 3,000) (Rogers, et al., 2016) for iOS and Android smartphones and tablets.

Summary

In conclusion, this study discussed the importance of collocational fluency and the reasons why students lack such knowledge and why no large-scale resources are available. Then it discussed the rationale for and results of a number of steps that were taken to create such a resource. Finally, a description of a smartphone app that students can use to study these items with was given. Although much more research still needs to be done in regard to helping students attain collocational fluency, this study can still be regarded as a significant step in addressing this gap in the research and it is hoped that students and researchers alike can use it to help improve upon the efficacy of second language acquisition.

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, Available on request.

Burston, J. (2007). CALICO software review: WordChamp. *CALICO Journal*, *24*(2*), 473-486.*

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

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

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

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

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics, 47*, *157-177.*

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

Goodwin-Jones, R. (2010). Emerging technologies. From memory palaces to spacing algorithms: Approaches to second-language vocabulary learning. *Language Learning and Technology, 14*(2), *4-11.*

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

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

Ishikawa, S. (2004). Vocabulary instruction at college level using JACET 8000 and educational software. In JACET Basic Words Revision Committee (Ed.), *How to make the best of JACET 8000: For educational and research application* (pp. 7- 14). Tokyo: JACET.

Kallkvist, M. (1998). Lexical infelicity in English: the case of nouns and verbs. In K.

Haastrup and A. Viberg (Eds.), *Perspectives on lexical acquisition in a second language* (pp. 149-174). Lund: Lund University Press.

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

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

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

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

Rogers, J. (2017). *What are the collocational exemplars of high-frequency English vocabulary?On identifying MWUs most representative of high-frequency, lemmatized concgrams.* (Unpublished doctoral dissertation). University of Southern Queensland, Australia.

Rogers, J., Brizzard, C., Daulton, F., Florescu, C., MacLean, I., Mimura, K., … Shimada, Y. (2015). 英語マスター1 万 [English Master 10,000] (Version 1.0) [Mobile application software]. Retrieved from http://itunes.apple.com

Rogers, J., Brizzard, C., Daulton, F., Florescu, C., MacLean, I., Mimura, K., … Shimada, Y. (2016). 英語マスター3 千 [English Master 3,000] (Version 1.0) [Mobile application software]. Retrieved from http://itunes.apple.com

Rogers, J., & Reid, G. (2015). How effective are smartphone flashcard applications for learning a second language? *The IRI 1st Research Forum, 24-33.*

Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt and M. McCarthy (Eds). *Vocabulary: Description, Acquisition and pedagogy* (pp. 199-227). Cambridge: Cambridge University Press.

Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics 31(4), 487-512.*

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

Contact email: jrogers@meijo-u.ac.jp