Vocabulary Learning and Instruction
Volume 9, Issue 2, August 2020
http://vli-journal.org

On Creating a Large-scale Corpus-based Academic Multi-word Unit Resource

James Rogers
Meijo University

Abstract

This study outlines the steps taken to create an academic multi-word unit list derived from corpus data. It gives details on the procedure used and the rationale behind why certain approaches were utilised. It also compares existing resources and makes some suggestions for practical use of the resulting resource.

Keywords: English for specific purposes, academic English, collocation, formulaic language, multi-word units, corpora

1 Introduction

This article outlines a project involving the construction of a corpus-based academic English resource list which focuses on academic multi-word units (MWUs). In recent years, there has been increased awareness of the importance of collocational and MWU fluency for second-language learners, and there is an agreement on the value of such knowledge (Nation, 2013; Siyanova-Chanturia & Pellicer-Sanchez, 2019; Webb & Kagimoto, 2011). Lewis (2000) stated that mastering collocational knowledge should be "a top priority in every language course" (p. 8). Other researchers echoed this sentiment, remarking that much of language is made up of prefabricated chunks, and thus learning them is essential for obtaining language fluency (Hill, Lewis & Lewis, 2000; Hoey, 2005).

1.1 Defining Word Co-Occurrence

MWU is one of the many terms used by researchers to refer to word co-occurrence is referred to by researchers. Some narrowly define the linguistic phenomenon of two words occurring together in high frequency with the term *collocation* (Hoey, 1991). Others define collocations by considering syntactic structures (Gitsaki, 1996), with others utilizing a combination of multiple criteria (Lesniewska & Witalisz, 2007). When the term MWU is used, it can include collocations, but there is a lot of variation of what it specifically refers to. For instance, research has broken down such language into literals, figuratives, and core idioms (Grant & Nation, 2006). Others consider only two-word phrases as collocations and all others as idioms or lexical bundles (Biber, Johansson, Conrad, Leech, & Finegan, 1999). All of these definitions are appropriate, and depend on the type of research, but for this current study, the linguistic phenomenon of word co-occurrence will be defined as two or more words that co-occur in high frequency, and these will be referred to as MWUs.

Please cite this article as: Rogers, J. (2020). On Creating a Large-scale Corpus-based Academic Multi-word Unit Resource. *Vocabulary Learning and Instruction, 9* (2), 17–23. https://doi.org/10.7820/vli.v09.2.rogers

1.2 Currently Available MWU Resources for Non-Native Learners of Academic English

Creating a usable MWU list for language learners is a complex process, and currently available resources are highly limited. The first issue is that they are of small scale. Liu's (2012) study only identified 228 MWUs and Simpson-Vlach and Ellis's (2010) study identified only 207 core items. Chon and Shin (2013) identified slightly more (934 written collocations) but their study was actually small in scope in that their list was derived from only 20 high-frequency academic node words. In comparison, this study begins with 500 node words and ended up with approximately 5,000 MWUs.

1.3 Concgramming as a Method of Identifying High-Frequency MWUs

A *concgram* "constitutes all the permutations of constituency and positional variation generated by the association of two or more words" (Cheng, Greaves, & Warren, 2006, p. 411). When this method is utilised, it allows for consolidation of overlapping items. This makes for more efficient learning in that very similar MWUs are not listed twice but at separate points in a list (e.g., *this study found* and *these studies found*). Concgramming also enables more accurate frequency ratings because consolidation of MWUs that only differ slightly by grammatical inflection are counted as one.

With concgramming, all lemmas of two co-occurring tokens are counted. They are counted while accounting for constituency variation. Therefore, when a corpus search is conducted for the lemma *make* and *money,* MWUs such as *make money* and *make some money* are counted. Concgramming also accounts for positional variation, so a search for the lemma *advice* and *give* will count instances of *advice you give* and *give you advice.*

Previous research has instead utilised the more simplistic n-gram corpus analysis to identify high-frequency MWUs that occur in a linear sequence. However, this method does not have the advantages as that of the concgramming method discussed above. The existence of non-consolidated partial duplicates and/or inaccurate frequency rankings in n-gram method-based studies creates limitations in the resources those studies result in, thus constituting a major gap in the research, since no large-scale studies on identifying academic MWUs have been done using the concgramming method.

1.4 The Need to Extend a Core MWU Beyond the Initial Identified Unit

When the concgramming method is utilised, concordance software will provide the most common MWUs in the corpus data, ranked by frequency. It is important to realise that the most frequent MWUs identified could be improved upon if they are extended beyond the core items. Often, extending MWUs in this way provides valuable information for learners on their extended contextual usage. A clear example of this is how a concgram search for the lemma *numerous* and *study* identifies *numerous studies*

Table 1. Identification of "Numerous Studies Have Shown" as the Exemplary Representative for the Lemmatised Concgram "Numerous" (Adjective) and "Study" (Noun) Using Data from Davies' (2008) Corpus of Contemporary American English's Academic Section

Rank	MWU
1.	*numerous studies*
2.	*numerous studies have*
3.	Study numerous
4.	are numerous studies
5.	studies numerous
6.	**numerous studies have shown**
7.	there are numerous studies
8.	in numerous studies
9.	numerous studies have been
10	numerous studies have found

In this table, italics is used to highlight the initial string and any other strings that contain it. Bold and italics are used to highlight the longest italicized string, which is considered as the exemplary MWU for the collocation *numerous/study*.

as the most frequent, but that when *numerous studies* occurred, the majority of the time the next word was *have*, and when *numerous studies have* occurred, the majority of the time the next word was *numerous studies have shown* (see Table 1). In Rogers (2017), experienced English as a foreign language (EFL) practitioners manually examined such data and decided whether or not to extend a MWU in this way with learners in mind. In that study, 53% of the items identified were extended, and the fact that the majority of the items were extended indicates that this may be an important criterion to consider.

2 Procedure

2.1 Search

This study began by using the most frequent 500 lemmas in Gardner and Davies' (2013) high-frequency academic vocabulary list as pivot words to search for lemmatised collocates in the academic section of the Corpus of Contemporary American English COCA (Davies, 2008), a corpus consisting of over one billion tokens of American English sourced from material from 1990-2017, which is evenly divided into the following five genres: spoken, fiction, magazines, newspapers, and academic journals. For each collocate found, a file was created with 500 concordance lines from the academic section of the COCA containing both the pivot and the collocate. These files were analysed using the custom concordance software *AntWordPairs* (Anthony, 2013) to identify the most frequent MWUs that both the pivot and collocate occurred in.

Following the parameters set by Rogers' (2017) study, which accomplished the same task set in this study but for general English, one occurrence per million tokens was frequency cut-off. Previous research has also implemented a parameter using the statistical measure of mutual information (MI), and Stubbs (1995) and Hunston (2002) recommend an MI cut-off score of 3 or higher for collocates, and therefore this study also only considered collocates with such a score.

2.2 Manual Removal of Noise

High-frequency and an M.I. score of 3 or higher does not always result in useful collocations being identified. The results of Rogers' (2017) study indicated that manual checking of data was essential, and therefore, six experienced EFL practitioners' intuition was used to identify and remove items that were not useful for learners. Such items included proper nouns, noise in the data, and also MWUs which only occurred in a particular genre of academic English, and therefore, had little value for learners of general academic English. Each practitioner had their own list to review. After they reviewed it, a second practitioner also went through the entire list. Any items flagged as not being useful for learners were reviewed by a different practitioner, and if that practitioner agreed, the item was removed from the list. The practitioners were given a protocol to follow, which listed potential reasons why an item would be flagged. The protocol was to flag any items if they were:

Proper Nouns: Items such as organisation titles, journal titles, descriptions of particular ethnicities, etc. Some examples that were already flagged include *Center for School Counseling Outcome Research*, *Census of Population and Housing*, *Native American population* and *Muslim population*.

Grammatical Formulations: Items that are devoid of meaning as a whole but occur frequently due to grammar. An example that was already flagged is *positive negative.* When further data were analysed, this item was revealed to be part of a list, with the two words separated by a comma, such as in *positive*, *negative* and *neutral.*

Too Specific to Particular Academic Fields: Items which only tend to occur in particular kinds of academic fields and therefore do not have general value for learners of academic English. Some examples that were already flagged include *the primary tumour*, *God's presence* and *reading fluency.*

Too Technical: Items which occur frequently due to the particular types of journal in the corpus, but which do not general value for learners of academic English. An example that was already flagged is *stance phase.*

Not Particular to Academic English: Items which occur more often in general English and are not particular to academic English. An example that was already flagged is *obstacle course.*

Unnatural Formulations due to Formatting: Items that are the result of the way a journal formats their papers with titles which are not actually MWUs. An example that was already flagged is *Results Preliminary Analysis.*

Others: Any items deemed to have little or no value for learners who aim to improve their academic English fluency based on your teaching experience and native-like speaker intuition.

3 Results

3.1 Item Analysis

Initial results identified a total of 10,190 collocations. After these were analysed by the EFL practitioners, 5,057 of them (49.6%) were judged to be useful

Table 2. Every 500th Academic MWU Identified in This Project (with Core MWUs Identified in Bold)

Rank in list	Frequency	MWU identified
1	36,167	the results of this study **suggest that**
500	779	**gain skills and knowledge**
1000	477	**described in detail**
1500	330	**highly effective**
2000	253	**samples were analysed**
2500	209	**situational factors**
3000	176	**consistent with the literature**
3500	151	**design features**
4000	130	**potential implications for**
4500	113	**requires detailed**
5000	100	**widely disseminated**

for learners who aim to improve their general academic English fluency. The vast majority of the items judged to have little value only occurred frequently in particular genres, such as medical academic English. Table 2 lists every 500th item to highlight the type of MWUs that were deemed useful.

3.2 On Extending Core MWUs Beyond the Initial Identified Unit

A majority of the 5,057 items (53.9%) were extended beyond their initial identified unit. This percentage was similar to that found by Rogers (2017), in which intuition was utilised in comparison to this study's usage of frequency data.

3.3 Accessing the Results

The resulting resource can be accessed via a custom webpage at https://www.smartsmart.org/academic-english. This website enables users to view the entire list ranked by frequency, along with a custom example sentence written with EFL learners for each multi-word unit. The site also allows users to search for any MWUs that contain or start with a particular word or words.

4 Discussion

4.1 Findings

This study filled gaps in the literature of scope and methodology that previous studies that attempted to identify academic MWUs had. The experienced EFL practitioners who participated considered the approximately 5,000 MWUs identified to all be useful for learners intending to improve their general academic English fluency. This study also confirmed a previous study's finding that manual analysis of data is essential for creating useable learning materials in that approximately half of the items initially identified were manually removed because they

were deemed to be not useful for learners despite falling within the parameters set. It should be noted that such manual checking of data is extremely time-consuming, and therefore, future software development should consider any possible ways to automatise this task and/or corpus compilers should consider ways in which this issue could be avoided when developing and organizing a corpus.

5 Conclusion

Much more research is still needed with regard to methodology of identifying MWUs. However, this study is a first step towards confirming the previous study's findings. It also provides teachers, students and researchers with a large-scale resource that can be immediately used, which previously constituted a major gap in the research. Future research is also needed to discover how such a resource could be best studied or utilised in language courses which aim to improve learners' academic English fluency.

References

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* London: Pearson Education. doi: 10.1017/s0022226702211627

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

Chon, Y., & Shin, D. (2013). A corpus-drive analysis of spoken and written academic collocations. *Multimedia-Assisted Language Learning, 16*(3), 11–38. doi: 10.15702/mall.2013.16.3.11

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

Gardner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327. doi: 10.1093/applin/amt015

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

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

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. 88–116). Hove, England: Language Teaching Publications.

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

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

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

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

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

Liu, D. (2012). The most frequently-used multiword constructions in academic written English: A multi-corpus study. *English for Specific Purposes, 31*(1), 25–35. doi: 10.1016/j.esp.2011.07.002

Nation, P. (2013). *Learning vocabulary in another language* (2nd edn.). Cambridge: Cambridge University Press. doi: 10.1017/cbo9781139524759

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

Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487–512. doi: 10.1093/applin/amp058

Siyanova-Chanturia, A., & Pellicer-Sanchez, A. (Eds.). (2019). *Understanding formulaic language: A second language acquisition perspective.* New York: Routledge.

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative methods. *Function of Language 2*(1), 1–33. doi: 10.1075/fol.2.1.03stu

Webb, S., & Kagimoto, E. (2011). Learning collocations: Do the number of collocates, position of the node word, and synonymy affect learning? *Applied Linguistics, 32*(3), 259–276. doi: 10.1093/applin/amq051