＜論　文＞

Development of a Basic Spoken Collocation List: A Preliminary Report

Shusaku NAKAYAMA

要旨

本稿では、現在著者が取り組んでいる英語初学者向け話し言葉コロケーションリストの作成に関する研究の一部として、動詞を含むコロケーションに絞り報告を行う。英語の産出において、コロケーション（例：strong tea）は流暢かつ自然な発話を可能にするだけでなく、ワーキングメモリーの節約においても重要な役割を果たす。これまで作成されてきた話し言葉コロケーションリストは包括的なものがほとんどであり、参照資料としては有用であるものの、英語初学者にとってはどのコロケーションを学習するべきであるのか判断が難しいものとなっている。そこで、1,000万語を超える大規模な話し言葉コーパスから、話し言葉の理解に必要な最も基礎的な語彙721語のみで構成されているコロケーションを抽出し、コロケーションリストを作成した。抽出には、コーパス内における発生頻度や語の結びつきの強さなどから重要性が高いと考えられるもののみを抽出し、結果として773種のコロケーションが抽出された。本稿の最後では、今後の研究の方針を示している。

キーワード

英語学習者、コーパス、コロケーション

The development of a corpus, an electronic database of spoken and/or written language texts, is one of the greatest works that brought innovation to “all branches of linguistics including lexicographic and lexical studies, grammatical studies, language variation studies, contrastive and translation studies, diachronic studies, semantics, pragmatics, stylistics, sociolinguistics, discourse analysis, forensic linguistics and language pedagogy (McEnery & Xiao, 2011, p. 383).” Since the release of the first electronic large-scale corpus created by Nelson Francis and Henry Kučera in 1964, the Brown corpus (the Brown University Standard Corpus of Present-Day American English), the field of corpus linguistics and related fields have advanced significantly; researchers, educational institutions, and other research organizations have worked on the creation of such large-scale corpora as the British National Corpus (BNC Consortium, 1994) and the Corpus of Contemporary American English (Davies, 2008-).

Research into multiword units such as idioms, binominals, and collocations is an example of a research field in which the development of electronic corpora moved forward. Many researchers have so far developed lists of multiword units (e.g., Ackermann & Chen, 2013; Lei & Liu, 2018). As researchers now have access to a large size of language data as well as "automated tools and methods with which to extract and explore multiword units (Granger, 2021, p. 1)," it is possible to compare and contrast the frequency of occurrence of word combinations and identify remarkable ones. With a special focus on spoken contexts, I am currently working on creating of a spoken collocation list with the aim of helping learners, especially those at a lower level, improve their communication skills. In this paper, I will report the preliminary results of this research.

Literature Review

What are Collocations Like?

English words do not appear randomly in sentences; rather, they tend to co-occur with specific words more frequently than others. Two or more words that frequently occur together are called multiword units. For instance, when referring to a research article published in conference proceedings, we typically say "conference paper" rather than "conference article."

Depending on their degree of semantic transparency and formulaicity, multiword units, including collocations, can be categorized into various types, ranging from free combinations (e.g., write an essay) through restricted collocations (e.g., strong coffee) and figurative idioms (e.g., do a U-turn), to true idioms (e.g. break a leg). Traditionally, researchers have considered two-word restricted collocations, those whose collocation components cannot be easily replaced with other words, to be so-called collocations (Ackerman & Chen, 2013; Fukuda & Tono, 2022).

Collocation research is largely inspired by the idea proposed by John Sinclair in 1991, the idiom principle. This principle contrasts with the open-choice principle, where people construct sentences word by word. In the idiom principle, on the other hand, people construct sentences using pre-established phrases stored in their mental lexicon, which is a key factor that makes English produced by native speakers sound natural and fluent (Dickinson, 2012; McGuire & Larsen-Hall, 2017).

Why are Collocations Important?

Why do many researchers seek to create lists of multiword units? Put differently, why is then the knowledge of multiword units important for learners? First of all, it can be hard to infer the meanings of collocations from their components, even though one knows the meaning of each collocation component. For example, knowing the meanings of "put" and "off" does not necessarily mean that one knows the meaning of "put off."

Second, collocational knowledge can help one's language processing (McGuire & Larson-Hall, 2017; Schmitt & Carter, 2004). Using collocations, people can speak fluently and naturally. Furthermore, collocations can help save one's working memory, because multiword units including collocations are saved in one's mental lexicon as if they were one word (Kuiper, 2004).

Given their high prevalence in the real world, mastering multiword units is essential for learners of English. According to Erman and Warren (2000), collocations make up 58.6% and 52.3% of words in spoken and written contexts, respectively, meaning that people are highly likely to encounter collocations in the real world. Multiword units are, therefore, crucial for both productive and receptive language skills.

Previous Studies on the Development of Collocation Lists

Because of their high pedagogic value it would not be surprising that many researchers have worked on the creation of collocation lists. I summarized well-cited research in Table 1.

Table 1:

Key Research Into the Development of Collocation Lists

Research.	Context
Shin & Nation (2008)	General English
Simpson-Vlach & Ellis (2010)	Academic English
Liu (2012)	Academic English
Ackerman & Chen (2013)	Academic English
Lei & Liu (2018)	Academic English
Shin & Chon (2019)	General English
Rogers et al. (2021)	Academic English

As shown in Table 1, most previous studies have focused on academic contexts. These are undoubtedly valuable for not only learners who aim to master academic English, but also teachers and teaching material developers who seek to teach it. In contrast, research into lists for general English seems to be insufficient. Although several collocation dictionaries covering general English exist, they often comprise an overwhelming number of collocations. For example, the Oxford Collocations Dictionary for Students of English (Crowther et al., 2002) comprises more than 250,000 collocations for over 9,000 nouns, verbs, and adjectives, making it unclear which collocations learners should prioritize. The same issue applies to previous studies: the spoken collocation list developed by Shin and Nation (2008) contains 4,698 collocations, while Shin and Chon's (2019) general collocation list includes 31,680 collocations.

According to McLean et al. (2014), Japanese learners of English were not very familiar even with the 2,000 most frequent word families. Given that collocation acquisition often occurs in the later stages of vocabulary learning (Li & Schmitt, 2009), and that a smaller range of words is used in spoken contexts compared to written ones, collocations in spoken contexts would be the first step for such learners with limited knowledge of vocabulary as Japanese. Therefore, I decided to develop a list of very basic collocations.

Methodology

Combined Approach

Traditionally, collocations have been identified based on how well two or more words are semantically and lexically connected. This approach, known as the phraseological approach, has been criticized for its disposition of including subjectivity (Granger, 2021; Xia et al., 2022). Take "error" as an example. We say "make an error" but do not say "do an error," so is it then reasonable to consider the combination to be a collocation? We also say "commit an error." For those who think both "make" and "commit" co-occur with "error" equally, "make an error" would be seen as just a free combination. For others, however, it qualifies as a collocation. This ambiguity highlights the limitations of relying solely on semantic and lexical connections. In reaction to this shortcoming, researchers have turned to a frequency-based approach.

In a frequency-based approach, collocations are identified based on whether each collocation candidate meets pre-set statistical thresholds. That does not mean an exclusive reliance solely on the approach is sufficient for one to identify collocations. As Granger pointed out:

> It would be wrong to assume, however, that the frequency-based approach entirely avoids the fuzziness of the phraseological approach. In the case of statistical collocations, decisions have to be made concerning the size of the span to the left and/or right of the node word, the inclusion of a dispersion criterion and, most importantly, the frequency and statistical thresholds used to establish collocation status. Depending on the options chosen, the sets of collocations can differ quite dramatically. (Granger, 2018, p.4)

Furthermore, an exclusive reliance on a frequency-based approach leads to the inclusion of collocations units that would be of little value for learners, such as "in the" and "am a" (Xia et al., 2022), underscoring the necessity of screening collocation candidates using the phraseological approach, too. In sum, both phraseological and frequency-based approaches comprise pros and cons. To address this issue, researchers have suggested a straightforward yet effective solution: combining the two approaches (Granger, 2018; Szudarski, 2023).

Most researchers have focused on syntactic patterns consisting solely of open-class items (Ackermann & Chen, 2013; Fukuda & Tono, 2022; Gablasoba et al., 2017; Shin & Nation, 2008). However, some researchers have also included patterns involving both open- and closed-class items in their analyses (Lei & Liu, 2018; Xia et al., 2022). As part of a project aiming to develop a spoken collocation list, this study focused on patterns involving verbs. Specifically, the combination patterns analyzed in this study were:

- verb + noun
- noun + verb
- adverb + verb
- verb + adverb
- verb + preposition
- verb + adjective (excluding determiners)

Xia et al. (2022) decided to investigate learners' use of collocations including prepositions because previous studies indicated the difficulty of this type of collocation for learners. Following them, this study also included this collocation pattern in the scope of analysis. As this study also aimed to make a collocation list for learners, this type of collocation was considered essential.

Frequency-based Approach

The frequency-based approach in this study incorporated three different parameters: frequency of occurrence, strength of word associations, and directionality. In their creation of a spoken collocation list using the British National Corpus, Shin and Nation (2008) found that a frequency threshold of 30 occurrences per 10 million words in the corpus was a reasonable cut-off point to "include several but not too many collocates (p. 342)" for each node word. This study also adopted the same threshold.

For measuring the strength of word associations, researchers have used various association measures to suit their research objectives. Among these, logDice scores, which utilize Dice coefficients, are widely considered to be a reliable option, especially for their capability of evaluating the strength of word associations using relative frequency. One can thus rule out the likelihood of results being skewed by corpus size and can compare research outcomes between different language data (Rychlý, 2008). In contrast, other traditional association measures such as t-scores and MI scores have been questioned in terms of their validity. More specifically, t-scores are inclined to be affected by corpus sizes, making it difficult to interpret research outcomes (Gablasoba et al., 2017); MI scores (Schmitt, 2012) tend to highlight too many infrequent items. Dice coefficients are calculated as follows:

$$\text{Dice coefficient} = \frac{\text{2*freq.of cooccurrence of node word and collocate}}{\text{freq.of node word + freq.of collocate}}$$

LogDice scores are computed by converting Dice scores into logarithmic ones:

$$\text{LogDice score} = 14+\log_2 \frac{\text{2*freq.of cooccurrence of node word and collocate}}{\text{freq.of node word + freq.of collocate}}$$

Seemingly, logDice scores are a simple transformation of Dice scores by applying logarithmic conversion and adding 14; however, these two steps can not only address a shortcoming of Dice scores that "the values of the Dice score are usually very small numbers (Rychlý, 2008, p. 9)," but also make logDice scores easy to interpret (Tsunekawa, 2020).

Previous studies have adopted different cut-off points for this parameter. Frankenberg-Frankenberg-Garcia et al. (2019) employed a threshold of logDice ≥ 5, and Kim and Oh (2020) followed their threshold level. Cao and Deignan (2019) adopted a threshold of logDice ≥ 4. To my knowledge, unfortunately, there seems to be no definitive threshold. My preliminary testing with logDice ≥ 4 indicated that it might be too strict, as many collocations were unreasonably left out. For the preliminary analysis in this study, I adopted a threshold of logDice ≥ 5 combined with a minimum co-occurrence frequency of 30 times per 10 million words.

In extracting collocates for each node word, researchers have suggested taking into account directionality (Gries, 2013; Szudarski, 2023), that is to say, which word of a collocation more strongly attracts the other. This suggestion stems from the understanding that each component of a collocation does not necessarily attract the other component at the same level of strength. Nevertheless, "nearly all measures that have been used are bidirectional, or symmetric (Gries, 2013, p. 141)," indicating the need for a different parameter. In the field of corpus linguistics, the

Δ*P* measure is a suitable option for evaluating directionality. Take "win the lottery" as an example. Calculating the Δ*P*s occurring in the British National Corpus, the Δ*P* was 0.0009 when making "win" its node word; when making "lottery" its node word, the Δ*P* was 0.03. Thus, "win the lottery" is more likely to occur when "lottery" appears than when "win" does. In other words, we would expect "win the lottery" more readily upon encountering "lottery" than upon encountering "win." I extracted only collocates whose ΔP values were the same as or less than those of the node words of the collocations.

Phraseological Approach

To determine whether each collocation candidate identified through the frequency-based approach should be included in my collocation list, I consulted the Online Oxford Collocation Dictionary, an online dictionary developed based on the British National Corpus. This dictionary comprises over 150,000 collocations for approximately 9,000 node words. Extracting only those listed in the dictionary made it possible to filter out non-collocations without relying on arbitrary judgment.

Word list Under Analysis

Unlike previous studies where researchers selected node words from large-scale corpora, I selected them from an existing word list, namely, the spoken module of the New General Service List-Spoken 1.2 (NGSL-S; Browne & Culligan, 2017). This word list consists of the 721 most frequent words in general spoken English, providing up to 90% coverage of general spoken texts. I extracted collocations consisting only of the NGSL-S words.

It is unlikely that learners will understand the meaning of a collocation as a whole without knowing the meanings of its individual components, suggesting that learners should move on to learning collocations only after acquiring knowledge of individual word meanings of target collocations. Hence, a collocation list based on the foundational word list can provide learners with a natural step forward for vocabulary learning, aligning with my research aim of creating a list of basic spoken collocations rather than a comprehensive one.

Corpus Analysis Software

The extraction of collocation candidates and all the statistical analyses were performed using LancsBox X 3.0.0 (Brezina & Platt, 2023). In this software, users can analyze both their own language data as well as built-in corpora. For this study, I adopted the informal spoken module of the BNC 2014 (Love et al., 2017), a large language data of contemporary British English comprising 11.5 million words. I extracted collocates occurring within a ±5-word window containing the node word and performed lemma queries on node words.

Results and Discussion

Using the combined approach, I identified a total of 773 verb-collocations (all the collocations are available from https://1drv.ms/x/s!AuEnnnXgDEv2h5E-VeLRDrmAKcYErw?e=5xhtef). Table 2 summarizes the number of collocations identified for each combination pattern, along with the percentage each pattern represents.

First and most importantly, verb-noun collocations were the most prevalent among the six types, accounting for 41% of all collocations, whereas its inverted pattern, noun-verb collocation, was less common, making up only 7% of all collocations. The high prevalence of verb-noun collocations was also the case in previous studies (Ackerman & Chen, 2013; Lei & Liu, 2018). Table 2 also shows the relatively low prevalence of collocations involving adjectives and adverbs, again matching what were found in previous studies. Ackermann and Chen (2013) reported that verb-adjective, adverb-verb, and verb-adverb combinations respectively accounted for 1.2%, 0.6%, and 1.2% of their identified collocations. Similarly, Lei and Liu (2018) found verb-adjective combinations to account for only 0.19% of their identified collocations.

Please note that this study focused solely on verb-collocations and cannot definitively generalize the distribution of all collocation patterns in the real world. Nevertheless, these observed similaritise to previous studies focusing on academic contexts suggest that regardless of language contexts, not all collocation patterns appear with equal frequency; some patterns are more common than others. From a pedagogical perspective, these findings emphasize the importance of prioritizing high-frequency collocation patterns in language learning, as they are what learners are highly likely to encounter in the real world.

Table 2:
Summary of Analysis Results

Combination patterns	Number of types	Proportion
verb-noun	314	41%
verb-preposition	178	23%
verb-adjective	95	12%
verb-adverb	86	11%
noun-verb	54	7%
adverb-verb	46	6%
Total	773	100%

Unlike most previous studies (Ackermann & Chen, 2013; Fukuda & Tono, 2022; Gablasoba et al., 2017), this study included collocations involving prepositions in the scope of analysis, finding a relatively high prevalence of this collocation pattern. This finding suggests that including this collocation pattern could provide collocation lists more valuable for learners.

Suggestions for Future Research

This study yielded several findings that align with those of previous studies. However, the similarities might be attributed to the current study's exclusive focus on verb-collocations. It would be, therefore, indispensable to expand the investigation to explore other collocation patterns beyond verb-collocations.

In extracting collocations, I consulted the Oxford Collocation Dictionary as a phraseological approach in order to mitigate subjective judgments. Nevertheless, the resulting list included several free combinations, such as “drink tea.” To remove such free combinations and enhance the practical value of the collocation list, consulting English experts to assess the pedagogical relevance of collocation candidates, as done in previous studies (Ackermann & Chen, 2013; Shin & Chon, 2019), could be a beneficial strategy.

While beyond the scope of this study, it is crucial to assess the validity of the collocation list. To this end, one potential method is to evaluate how well the list covers general spoken English texts. Without such validation, it would be almost impossible to determine whether or not the list is worthwhile for learners.

References

恒川元（2020).「logDice 係数はどのような共起指標か」『言語文化論究』45巻, pp. 35-44. https://doi. org/10.15017/4104141

Ackermann, K., & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL): A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*(4), 235–247. https://doi. org/10.1016/j. jeap. 2013.08.002

BNC Consortium. (2007). *The British National Corpus, XML edition*. Oxford Text Archive. http://hdl.handle.net/20.500.14106/2554.

Brezina, V., & Platt, W. (2023). *#LancsBox X 3.0.0* [Computer software]. http://lancsbox.lancs.ac.uk.

Browne, C., & Culligan, B. (2017). *The New General Service List – Spoken version 1.2*. https://www.newgeneralservicelist.com/faqs-4

Cao, D., & Deignan, A. (2019). Using an online collocation dictionary to support learners’ L2 writing. In C. Wright, L. Harvey & J. Simpson (Eds.), *Voices and practices in applied linguistics: Diversifying a discipline* (pp. 233–249). White Rose University Press. https://doi.org/10.22599/BAAL1.n

Crowther, J., Dignen, S., & Lea, D. (Eds.). (2002). *Oxford collocations dictionary for students of English*. Oxford University Press.

Davies, M. (2008-). *The Corpus of Contemporary American English (COCA)*. https://www.english-corpora.org/coca/

Dickinson, P. (2012). Improving second language academic presentations with formulaic sequences. *Bulletin of Niigata University of International and Information Studies Department of Information Culture. 15*, 25–36. https://nuis.repo.nii.ac.jp/records/2608

Erman, B., & Warren, B. C. (2000). The idiom principle and the open choice principle. *Interdisciplinary Journal for the Study of Discourse, 20*(1), 29-62. http://dx.doi.org/10.1515/text.1.2000.20.1.29

Francis, W. N., & Kučera, H. (1964). *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers*. Providence.

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P., & Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL, 31*(1), 23–39.

http://dx.doi.org/10.1017/S0958344018000150

Fukuda, K., & Tono, Y. (2022). Constructing a collocation database for the CEFR-J wordlist. *Proceedings of Language Resources Workshop, 1*, 133-146. https://doi.org/10.15084/00003733

Kim, Y. S., & Oh, S. Y. (2020). A corpus-based analysis of collocations in Korean middle and high school English textbooks. *Language Research, 56*(3), 437-461. https://doi.org/10.30961/lr.2020.56.3.437

Kuiper, K. (2004). Formulaic performance in conventionalised varieties of speech. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 37–54). John Benjamins.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning, 67*, 155-179. https://doi.org/10.1111/lang.12225

Granger, S. (2018). Formulaic sequences in learner corpora: Collocations and lexical bundles. In A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 228-247). Routledge

Granger, S. (2021). Phraseology, corpora and L2 research. In S. Granger (Ed.), *Perspectives on the L2 phrasicon: The view from learner corpora* (pp. 3-21). Multilingual Matters. http://dx.doi.org/10.21832/9781788924863-002

Gries, S. Th. (2013). 50-something years of work on collocations: What is or should be next …. *International Journal of Corpus Linguistics, 18*(1), 137-166. https://doi.org/10.1075/ijcl.18.1.09gri

Lei, L., & Liu, D. (2018). The academic English collocation list: A corpus-driven study. *International Journal of Corpus Linguistics, 23*(2), 216-243. https://doi.org/10.1075/ijcl.16135.lei

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing, 18*, 85-102. https://doi.org/10.1016/j.jslw.2009.02.001

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes, 31*, 25-35.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3), pp. 319-344.

McEnery, A. M., & Xiao, R. Z. (2005). Help or help to: What do corpora have to say? *English Studies, 86*(2), 161-187. http://dx.doi.org/10.1080/0013838042000339880

McGuire, M., & Larson-Hall, J. (2017). Teaching formulaic sequences in the classroom: Effects on spoken fluency. *TESL Canada Journal, 34*(3), 1-25. http://dx.doi.org/10.18806/tesl.v34i3.1266

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the vocabulary size test. *Vocabulary Learning and Instruction, 3*(2), 47-55. http://dx.doi.org/10.7820/vli.v03.2.mclean.et.al

Rychlý, P. (2008). A lexicographer-friendly association score. In P. Sojka & A. Horak (Eds.), *Proceedings of recent advances in Slavonic natural language processing, RASLAN 2008* (pp. 6–9). Masaryk University.

Rogers, J., Müller, A., Daulton, F. E., Dickinson, P., Florescu, C., Reid, G., & Stoeckel, T. (2021). The creation and application of a large-scale corpus-based academic multi-word unit list. *English for Specific Purposes, 62*, 142–157.
https://doi.org/10.1016/j.esp.2021.01.001

Schmitt, N. (2012). Formulaic language and collocation. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–10). Blackwell.
https://doi.org/10.1002/9781405198431.wbeal0433

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing, and use* (pp. 1-22). John Benjamins Press.

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

Shin, D., & Chon, Y. V. (2019). A multiword unit analysis: COCA multiword unit list 20 and ColloGram. *The Journal of Asia TEFL, 16*(2), 608-623.
http://dx.doi.org/10.18823/asiatefl.2019.16.2.11.608

Shin, D., & Nation, I. S. P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal, 62*(4), 339-348.
https://doi.org/10.1093/elt/ccm091

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

Szudarski, P. (2023). *Collocations, corpora and language learning*. Cambridge University Press.

Xia, D., Chen, Y., & Pae, H. K. (2022). Lexical and grammatical collocations in beginning and intermediate L2 argumentative essays: A bigram study. *International Review of Applied Linguistics in Language Teaching, 61*(4), 1421-1453.
http://dx.doi.org/10.1515/iral-2021-0188