

# From Words to Phrases: High-Frequency Spoken Collocations

Shusaku Nakayama (shusaku-nakayama@outlook.jp)  
Osaka University of Economics and Law, Osaka, Japan

## 1. Why are Collocations Important?

**HARD to infer meanings from components**

Ex.) put off / of course

**HELP language processing**

Speaking fluently and accurately (McGuire & Larson-Hall, 2017)

Saving “working memory” (Kuiper, 2004)

**PREVALENCE**

58.6% in spoken contexts

52.3% in written contexts (Erman & Warren, 2000)

## 3. Methodology

Table 1. Collocation patterns under scrutiny

NODE WORD	PLACE OF COLLOCATE	
	LEFT	RIGHT
NOUN	NOUN	NOUN
	VERB	VERB
	ADJECTIVE	
		PREPOSITION
VERB	NOUN	NOUN
		ADJECTIVE
	ADVERB	ADVERB
		PREPOSITION
Adjective		NOUN
	VERB	
	ADVERB	
		PREPOSITION
ADVERB	VEBR	VERB
		ADJECTIVE
	ADVERB	ADVERB

### Collocation Identification Process

**721 NGSL-S words**  
(Browne & Culligan, 2017)

LogDice score = **5 or higher**  
(Frankenberg-Garcia et al., 2019)

Frequency  
of  
occurrence

Word  
association  
strength

**30** times per million  
words (Shin & Nation, 2008)

Directionality

DeltaP of the node word  $\geq$  collocate

**Spoken module of  
the BNC2014**  
(Love et al., 2017)

**Online OXFORD Collocation  
Dictionary of English**

## 2. Literature Review

### Development of collocation lists

◆ **Shin & Nation (2008): 4,698 entries**

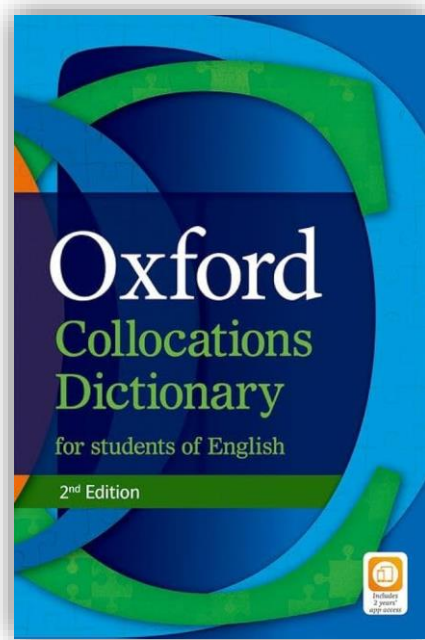
- Node words = the most frequent 1,000 content words in the BNC corpus.
- A frequency threshold = 30 times of occurrences per 10 million words

◆ **Shin & Chon (2018): 31,680 entries**

- Node words = the most frequent 5,000 content words in the COCA corpus
- A frequency threshold = 20 times of occurrences in the corpus (and occurring in 4-5 sub-corpora out of the 5 sub-corpora in the corpus)

◆ **Oxford Collocations Dictionary for Students of English:**

250,000 word combinations for over 9,000 nouns, verbs, and adjectives



### Aim of the Study:

To develop a basic list of spoken collocations for learners with limited knowledge of collocations.

## 4. Results

Table 2. Number and percentage of combination types

Combination pattern	Number of collocations	Percentage
Verb-noun	301	26.4%
Verb-preposition	179	15.7%
Noun-preposition	126	11.1%
Adjective-noun	120	10.5%
Adverb-adjective	96	8.4%
Verb-adverb	87	7.6%
Verb-adjective	58	5.1%
Noun-noun	53	4.7%
Adverb-verb	45	4.0%
Noun-verb	40	3.5%
Adjective-preposition	34	3.0%
Adverb-adverb	0	0%
Total	1,139	100%

- In total, 1,139 different collocations were identified.
- Collocations involving prepositions, which previous studies often excluded from the scope of analysis, were relatively prevalent.
- Some collocation patterns were more common than others.

**RESULTS ARE ACCESSIBLE FROM HERE ⇒**



## 5. Conclusion

- ◆ The completed list comprises a relatively smaller number of collocations than others (**31,680** (Shin & Nation, 2008) vs. **4,698** (Shin & Chon, 2018) ) ⇒ Depending on their learning goals and teaching targets, learners and teachers can choose the most suitable one.
- ◆ Directionality evaluated by DeltaP reasonably identified the distribution of the collocation patterns ⇒ Symmetric measures might misleadingly over-/under-estimate some collocation patterns.

### NEXT STEP:

To investigate the validity of the list ⇒ how well my list covers 2-word collocations occurring in general spoken texts?

### Selected References:

- Browne, C., & Culligan, B. (2017). The New General Service List – Spoken version 1.2. <https://www.newgeneralservicelist.com/faqs-4>
- Erman, B., & Warren, B. C. (2000). The idiom principle and the open choice principle. *Text - Interdisciplinary Journal for the Study of Discourse*, 20(1), 29-62.
- 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.
- Kuiper, K. (2004). Formulaic performance in conventionalised varieties of speech. In Schmitt, N. (Ed.), *Formulaic sequences* (pp. 37–54). John Benjamins.
- 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), 319-344.
- McGuire, M., & Larson-Hall, J. (2017). Teaching formulaic sequences in the classroom: Effects on spoken fluency. *TESL Canada Journal*, 34(3), 1-25.
- Shin, D., & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal*, 62(4), 339-348.
- 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.

