A Corpus-Driven Analysis of Spoken and Written Academic Collocations*

Yuah V. Chon(Hanyang University)
Dongkwang Shin(Korea Institute for Curriculum and Evaluation)

Chon, Yuah V. & Shin, Dongkwang. (2013). A corpus-driven analysis of spoken and written academic collocations. *Multimedia-Assisted Language Learning, 16*(3), 11-38.

As already well-acknowledged in the field, formulaic sequences often make up an important part of lexical knowledge. Within the L2 learning contexts, there is increasing demand for the need to understand and retrieve lexical items as in prefabricated chunks. In tertiary contexts, knowledge of academic collocations is even more required. For individual word forms, there is the General Service List (West, 1953), and the Academic Word List (Coxhead, 2000). However, what the field currently lacks is a list of academic collocations. Identification of collocations was conducted with the British Academic Spoken English and the Academic Corpus for comparison of academic collocations in spoken and written discourse. There was use of 20 node words in each corpus, which produced 934 written and 460 spoken collocations. With more number of written collocations being produced from the corpus, the study provides a comparative view of the top 50 collocations which suggests a concentration of common collocations in the field of economics. Within the guided approach for data-driven learning, there is proposal for the need to facilitate psychological conditions for learning collocations.

I. INTRODUCTION

There is little dispute that the learning of formulaic sequences (e.g., collocations) has proved to be useful for improving achievements in vocabulary learning and reading. However, the question then arises on what may be the collocations worth learning. While there have been word lists used for decades, they have provided information about individual word forms (e.g.,

* This work was supported by the research fund of Hanyang University.

General Service List; West, 1953) and the Academic Word List (Coxhead, 2000). There is also wide-spread acknowledgement of how high-frequency collocations are part of the lexicon which learners need to acquire. This implies—as Coxhead (2008) has recently discussed—that it will be necessary to extend existing wordlists to take account of such items, and some attempts have now been made to generate pedagogically-oriented listings of high-frequency collocations, both in general (Durrant & Schmitt, 2009; Liu, 2003; Shin & Nation, 2008) and in academic English (Biber, Conrad, & Cortes, 2004; Choi & Chon, 2012; Durrant, 2009; Ellis, Simpson-Vlach, & Maynard, 2008; Simpson & Mendis, 2003; Simpson-Vlach & Ellis, 2010). However, the previous listings of collocations have often been analyzed with focus on one of the registers alone (e.g., Michigan Corpus of Academic Spoken English; Simpson & Mendis, 2003) or without distinction for the spoken or written forms of English (e.g., Biber, Conrad, & Cortes, 2004). Also, two prominent academic corpora, the British Academic Spoken English and the Academic Corpus (Coxhead, 2000) have not previously been the subject of analysis for a comparative analysis of collocations. In relation to previous findings, the purpose of the present study is twofold. The researchers would like to add to the list of collocations used for academic purposes in order to potentially provide guidelines on the kinds of collocations that deserve attention for teaching and learning, whereby to sensitize L2 learners to those that deserve noticing. In the Korean context, previous collocation lists have been retrieved from in-house compiled corpora of newspaper articles (Min, S. Kim, & K. Kim, 2010) or from a corpus of Common English I High School textbooks of English (Choi & Chon, 2012), but they lack generalizability. Another purpose of the study is to provide a comparative analysis of the collocations according to the spoken and written forms of discourse. It remains to be seen how the collocational patterns may transpire.

II. BACKGROUND

1. Importance of Collocations in L2 Learning

Research shows that the learning of collocations, which have been studied under more than forty terms, including 'formulaic sequences', 'lexical phrases', 'fixed expressions', 'prefabricated patterns', and 'lexical bundles' (Wray & Perkins, 2000) is important often because they are used repeatedly and are known to make the students' task easier since they can work with ready-made sets of words. It has also been acknowledged in numerous instances that

knowledge of collocations becomes defining markers of being close to a native speaker (Hyland, 2008; Schmitt, 2000; Wray, 2002). Another reason why learners need to pay explicit attention to collocations is that even when learners can produce a considerable number of native-like sequences (Nesselhauf, 2005), there is evidence that learners' restricted collocational repertoires lead them to overuse those sequences with which they feel safe (Granger, 1998). Learners have been found to have a tendency to stick with familiar sequences which they tend to repetitively use, which have also been described as the "islands of reliability" according to Dechert (1984) (Granger, 1998, p. 156). For this type of phenomenon, Durrant and Schmitt (2009) discovered systematicity in this variation, finding that their nonnative writers tended to rely heavily on high-frequency collocations, but that they also tended to underuse less frequent, strongly associated collocations, the type of item which is likely to be highly salient for native speakers. Henceforth, the literature underscores the importance of teaching collocations in L2 learning, and this becomes more crucial when having to address a specific group of audience in specific academic disciplines. That is, learners will need to retrieve the appropriate collocations that fit the expectations of the users who belong to the specific discourse community (e.g., Nesselhauf, 2004). In the same vein, we plan to retrieve and create an academic collocation list that appears in the broad academic disciplinary areas (i.e., Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences) for the spoken and written modes of discourse.

2. Previous Studies on Collocation Lists

Few studies so far exist primarily for the purpose of providing academic collocation lists. Within the Korean context, there have been studies conducted in examining the use of collocations in educational materials. Min, S. Kim, and K. Kim (2010) produced a collocation list from the corpus compiled from newspapers intended for children. In the end, they used 284 words for devising the collocation book, where the main objective of the study was to obtain authentic expressions of English. With secondary school textbooks, Choi and Chon (2012) retrieved 852 lexical collocations by analyzing 16 Common English I High School textbooks of English (1,441,402 running words) after extracting 41 most frequent content words as nodes. Collocations that are closely related to learners' real life and interest (e.g., *volunteer work, good grade, fast food*) were found while the number of collocations combined with adverbs was relatively small comprising less than 5% of the 852 collocations.

In the adjacent areas of research on collocations, Biber et al. (2004) investigated the use of collocations in two university registers: classroom teaching and textbooks. They compared what

they referred to as 'lexical bundles' found in classroom teaching and textbooks to those found in conversation and academic prose. They describe the structural patterns, and then a functional taxonomy of the multi-word sequences. Simpson and Mendis (2003) examined idioms in a specific genre by drawing on a specialized corpus, the Michigan Corpus of Academic Spoken English, where they investigated frequency distributions of idioms in the corpus by academic division, and drew on different pragmatic functions that the idioms perform. Through a corpus-based approach, Liu (2003) investigated spoken American English idioms used most frequently by college and other professional ESOL students learning American English. The study involved a close concordance search and analysis of the idioms used in three contemporary spoken American English corpora: Corpus of Spoken, Professional American English; Michigan Corpus of Academic Spoken English, and Spoken American Media English. According to the search results, four lists of the most frequently used idioms were compiled, with one based on the overall data and the other three on one of the corpora. The study uncovered interesting English idiom use patterns.

At a more general level regarding studies related to a collocation list, Shin (2007) utilized a spoken corpus from the British National Corpus (10,000,000 running words) and a written corpus (10,000,000 running words) including the Australian Corpus of English (ACE), the Brown corpus, the Lancaster-Oslo/Bergen (LOB) corpus, the Freiburg-Brown (FROWN) and Freiburg-LOB (FLOB) corpora, the Kolhapur corpus, and the Wellington Written (WWC) Corpus, as well as some written text from the British National Corpus. In another study, Shin and Nation (2008) sought to identify the most frequent collocations in spoken English in the British National Corpus, where the researchers report identifying 4,698 collocations. In the process, they established six criteria involving such aspects as frequency and grammatical well-formedness. As a result, the spoken collocations were observed to be typically 50% to 100% more frequent in the spoken corpus. Shin and Nation also considered semantics, particularly individual senses of collocations with the same form (e.g. *looking up* meaning 'to improve' and *looking up* as in to find a word in a dictionary).

Ellis, Simpson-Vlach, and Maynard (2008) focused exclusively on one particular type of collocation. That is, their interest was in three-, four- and five-word sequences, and they found these to be more significant in academic than in non-academic texts. They report dividing their 2.1 million word corpus of academic writing and 2.1 million word corpus of academic speech into five spoken and four written genres and 'grading' bundles according to how well they are spread between genres. However, the results of this procedure are not reported in their later analysis (which focuses instead on the overall frequencies of items), and the inevitably small size

of each genre (an average of 4.2 million/9 = 466,666 words per genre) is again rather limited for work of this kind. In comparison, we were more interested in obtaining a generic list of academic collocations.

Published at a similar time, Durrant (2009) aimed to provide pedagogical descriptions of academic collocations, specifically where he was interested in incorporating positionally-variable expressions and providing a clear account of how well distributed items are across academic disciplines. The results were found to include two-word cross-disciplinary academic collocations (including both fixed and variable items), claiming them to be lists of pedagogical value.

Simpson-Vlach and Ellis (2010) sought to compile a list of the most useful formulaic sequences used in Academic English. The researchers were able to correlate the qualitative judgment data with the quantitative statistics and, through multiple regression to arrive at a metric that could be applied for predicting formulaic sequences that would be worth teaching. However, their method of analysis is different from other collocation lists, which have been ranked by how commonly they occur in discourse (e.g., Choi & Chon, 2012; Liu, 2003; Shin & Nation, 2008; Simpson & Mendis, 2003). In fact, frequency-based information of collocations is bound to be most pedagogically useful for learners and practitioners involved in L2 learning in materials design or syllabus development. Also, at present, there is lack of a frequency-based academic collocation list in contrast to the availability of word lists, particularly with comparison of the spoken and written discourse. Since the present study takes a frequency-based approach to the analysis, there was interest in the top 50 collocations, and a partial list is presented in the Appendix. The present study was conducted with the following questions to guide the inquiry of research:

1. How are the high-frequency single lexical items (i.e., node words) associated with the collocations found in the academic corpora, respectively for the spoken and written corpora?

2. How do the academic collocations in the spoken and written corpora compare to each other, particularly for the high-frequency collocations (i.e., top 50 academic collocations), in potentially providing teaching implications?

III. METHODS

1. The Corpora

To search the academic collocations, two collections of academic corpora were used: the British Academic Spoken English (BASE) and the Academic Corpus (Coxhead, 2000). BASE was used to search for the spoken collocations whereas the Academic Corpus was used to search written collocations.

The BASE Corpus consists of 160 lectures and 40 seminars recorded in a variety of departments (video-recorded at the University of Warwick and audio-recorded at the University of Reading). It contains 1,644,942 tokens in total (lectures and seminars), and is available via the site BASE Files (http://www2.warwick.ac.uk/fac/soc/al/research/collect/base/history). The corpus is distributed across four broad disciplinary groups, each represented by 40 lectures and 10 seminars. These groups are: Arts and Humanities, Life and Medical Sciences, Physical Sciences, and Social Sciences. The text and tagged transcripts of the original BASE corpus were developed as part of the British Academic Spoken English corpus project, 2000–2005.

The Academic Corpus that has previously been used for Coxhead's (2000) Academic Word List was utilized for searching the written academic collocations. The Academic Corpus contains approximately 3,500,000 running words. It is divided into four faculty sections: Arts, Commerce, Law and Science. Each of these faculty sections contains approximately 875,000 running words. Each faculty section is divided into seven subject areas of approximately 125,000 running words. Our choice for the Academic Corpus was due to its common use in the listing of academic vocabulary today (Coxhead, 2000) (See Table 1) (For more information on the corpus, see

〔Table 1〕 Subject Areas in the Faculty Sections of the Academic Corpus

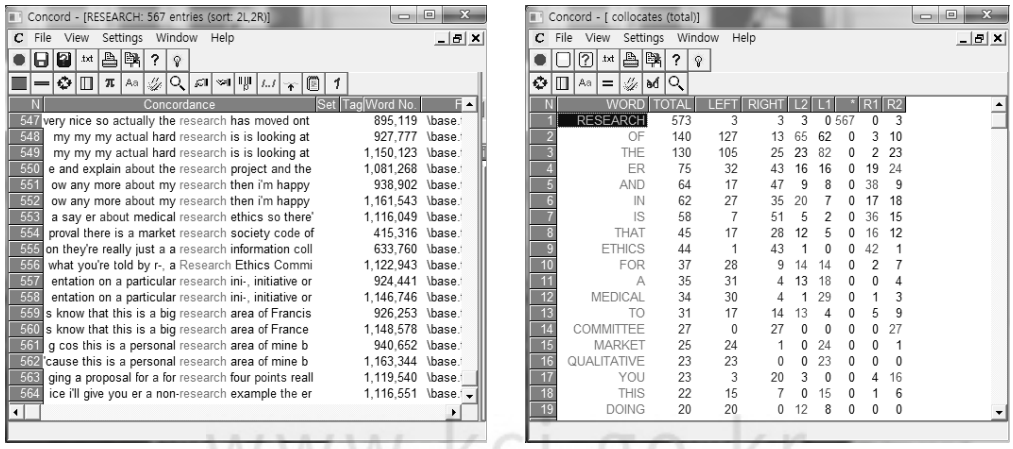
Arts	Commerce	Law	Science
883,214 tokens	879,547 tokens	874,723 tokens	875,846 tokens
122 texts	107 texts	72 texts	113 texts
Education	Accounting	Constitutional Law	Biology
History	Economics	Criminal Law	Chemistry
Linguistics	Finance	Family Law and Medico-Legal	Computer Science
Philosophy	Industrial Relations	International Law	Geography
Politics	Management	Pure Commercial Law	Geology
Psychology	Marketing	Quasi-Commercial Law	Mathematics
Sociology	Public Policy	Rights and Remedies	Physics

http://www.victoria.ac.nz/lals/resources/academicwordlist/information/corpus).

According to the information available at the site for the Academic Word List, the Corpus consists of journal articles, book chapters, course workbooks, laboratory manuals, and course notes. Where possible, Coxhead kept a balance between the number of short texts (2,000-5,000 running words), medium length texts (5,000-10,000 running words) and long texts (over 10,000 running words) among the four faculty areas. The texts are representative of the academic genre for an academic audience, and there are 414 texts by more than 400 authors in the Academic Corpus. It is also stated at the source site that the majority of the texts are written for an international audience; sixty-four percent were sourced in New Zealand, 20% in Britain, 13% in the United States, 2% in Canada, and 1% in Australia.

2. The Computer Program

The program used for the search was *WordSmith Tools 3.0* (Scott, 1999). WordSmith Tools is an integrated suite of programs for looking at how words behave in texts, and it is one of the most commonly used programs for extracting collocations from a corpus. Of the programs, the WordList tool lets you see a list of all the words or word-clusters in a text, set out in alphab etical or frequency order. The concordancer, Concord, gives you a chance to see any word or phrase in context so that you can see what sort of company it keeps. Concord was the primary means of analysis in the study for extracting the collocations. The program creates a concordance and provides numerical information of co-occurrences of components making up a collocation (See Figure 1). The next section elaborates on the specific criteria that were



〔Figure 1〕 Collocation-Search Process on WordSmith Tools 3.0

considered in the process.

3. Criteria and Procedure for Identification and Coding of Collocations

As previously mentioned, there are more than forty terms used for designating multi-word units such as 'collocations', 'polywords', 'fixed expressions' and 'semi-fixed expressions' (Wray & Perkins, 2000). In an attempt to categorize the multi-word items, Grant and Bauer (2004), Howarth (1998) have favored a continuum description. For instance, Grant and Bauer would consider *at all times* as compositional, since its meaning is still retained when each lexical word is replaced with its own definition (p. 52). On the other hand, the phrase *at all costs* is less transparent and makes it potentially difficult for L2 learners to decode the item. However, deciding where to place the multi-word items (e.g., collocations) on the continuum has not always been a straightforward job. To avoid this confusion in this study, criteria proposed by Shin and Nation (2008) were adopted for the identification of collocations, and their study incorporates criteria from previous studies (Kjellmer, 1994; Stubbs, 2000).

The spoken and written academic corpora were coded and counted for the collocations using the following four set of criteria for the purpose of coding high-frequency academic collocations that are grammatically well-formed and make sense as of itself.

1) The first step of the analysis involved making decisions on the type of words that would be used as node words so that collocates could be found to form a collocation. Also, since our interest was in the listing of academic collocations, the top 20 ranking academic node words were retrieved from each of the academic corpus—BASE and the Academic Corpus—based on those that correspond to the words on the Academic Word List(Coxhead, 2000). We limited our research to 20 node words since we realized in a preliminary analysis that several hundred collocations are deemed sufficient for obtaining information on the high-frequency collocations (See later Table 2 for the list of academic words).

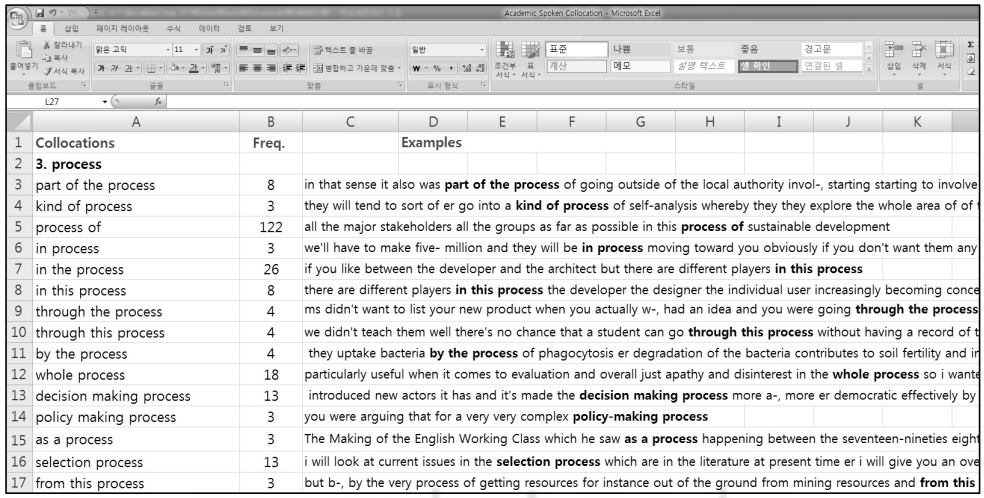
2) Another criterion that had to be reached is that the node word had to be a noun, a verb, an adjective, or an adverb. When two different node words share a collocation, the overlapping collocation was counted once. For example, the two node words *research* and *data* share *research data* as a collocation. Each node word was a word type. That is, the different word forms *involve* and *involved* were treated as different node words and investigated separately. As claimed by Stubbs (2000), a major justification for focusing on types rather than lemmas or families was

that different types of the same word family have different collocates. For instance, this is because words such as *break* and *broken* even when belonging to the same word family will have different high-frequency collocates.

3) To satisfy as a collocation, each collocation had to occur at least 3 times in the spoken corpus (1,600,000 running words), and 6 times in the written corpus (3,500,000 running words). This is because the sizes of the spoken and written corpora were different, and accordingly different frequency cut-off points were set according to the sizes of spoken and written corpus. This cut-off point was derived from Kjellmer (1994) who used the cut-off point of 2 for every 1,000,000 running words.

4) To be coded as a collocation, it also had to be grammatically well-formed. That is, the collocation had to be complete in itself in that it can act as the constituent of a sentence, however, with exclusion of articles and demonstrative adjectives (e.g., *this, those*). This criterion was needed so as to satisfy the criterion of 'replicability' which was needed to be able to code collocations consistently as 'meaningful units.'

To sum up, (1) the node word in a collocation must be an academic word, (2) the node word in a collocation must be a content word, (3) the collocation should occur frequently in the academic corpus, and (4) the collocation should be grammatically well-formed.



	A	B	C	D	E	F	G	H	I	J	K
1	**Collocations**	**Freq.**		**Examples**							
2	**3. process**										
3	part of the process	8	in that sense it also was **part of the process** of going outside of the local authority invol-, starting starting to involve								
4	kind of process	3	they will tend to sort of er go into a **kind of process** of self-analysis whereby they they explore the whole area of of t								
5	process of	122	all the major stakeholders all the groups as far as possible in this **process of** sustainable development								
6	in process	3	we'll have to make five- million and they will be **in process** moving toward you obviously if you don't want them any								
7	in the process	26	if you like between the developer and the architect but there are different players **in this process**								
8	in this process	8	there are different players **in this process** the developer the designer the individual user increasingly becoming conce								
9	through the process	4	ms didn't want to list your new product when you actually w-, had an idea and you were going **through the process**								
10	through this process	4	we didn't teach them well there's no chance that a student can go **through this process** without having a record of t								
11	by the process	4	they uptake bacteria **by the process** of phagocytosis er degradation of the bacteria contributes to soil fertility and in								
12	whole process	18	particularly useful when it comes to evaluation and overall just apathy and disinterest in the **whole process** so i wante								
13	decision making process	13	introduced new actors it has and it's made the **decision making process** more a-, more er democratic effectively by								
14	policy making process	3	you were arguing that for a very very complex **policy-making process**								
15	as a process	3	The Making of the English Working Class which he saw **as a process** happening between the seventeen-nineties eight								
16	selection process	13	i will look at current issues in the **selection process** which are in the literature at present time er i will give you an ove								
17	from this process	3	but b-, by the very process of getting resources for instance out of the ground from mining resources and **from this**								

〔Figure 2〕 Spreadsheet for Identification and Coding of Collocations with '*process*'

After the criteria needed for satisfying the conditions for collocations were identified, the collocations had to be organized and entered into spreadsheets for each of the node words (e.g., *obviously, data, process, research, lecture, structure*) and the related collocations. The collocations were also sorted according to the form and frequency of collocations (See Figure 2 for how the collocations were organized according to each node word, in the following case with *process*).

In the process, the data for the written and spoken corpora were stored separately. Although the computer did most of the work, the present study involved manual checking and analysis for the process of identifying the collocations that satisfied the four criteria. Also, in the process, extracting sample sentences containing collocations in addition to seeing that it satisfied the grammatical well-formedness criterion involved constant checking.

IV. RESULTS AND DISCUSSION

1. Frequency of Node Words and Number of Academic Collocations

The first research question of our study was to examine high-frequency node words that are used by speakers and writers as seen in academic corpora and to see if any relationships exist regarding academic collocations. There was first analysis of the top 20 academic spoken and top 20 academic written node words. Considering the size of the two corpora, use of the 20 most frequent node words was deemed fair. This is based on the estimate presented by Shin (2009) where he found in the analysis of the spoken, ten million British National Corpus that with the use of 100 node words, this produced 60% coverage of the spoken collocations found in the listing of collocations. As such, the use of 20 node words to retrieve the academic collocations from the smaller BASE and the Academic Corpus in the present study was sufficient to have coverage of the frequently occurring collocations. All in all, when 20 academic node words were compared between the two types of corpus, ten of them were found to occur in both the spoken and written word lists as highlighted in Table 2.

〔Table 2〕 Frequency of the Top 20 Academic Node Words and No. of Collocations

	Spoken Academic Corpus			Written Academic Corpus		
Rank	Nodes	Freq of Nodes	No. of Cllctns	Nodes	Freq of Nodes	No. of Cllctns
1	OBVIOUSLY	774	9	INCOME	3,413	95
2	DATA	691	39	SECTION	2,948	56
3	PROCESS	582	43	RESEARCH	2,567	68
4	RESEARCH	566	36	POLICY	2,508	78
5	LECTURE	493	27	DATA	2,482	89
6	STRUCTURE	483	26	PROCESS	1,967	54
7	MEDICAL	481	32	ECONOMIC	1,806	50
8	AREA	476	19	ANALYSIS	1,780	45
9	PERIOD	473	21	EXPORT	1,643	46
10	ECONOMIC	434	39	STRUCTURE	1,557	38
11	THEORY	421	23	APPROACH	1,514	32
12	ISSUE	360	21	EVIDENCE	1,439	47
13	EVIDENCE	360	19	PERIOD	1,408	38
14	INDIVIDUAL	359	20	LABOUR	1,399	36
15	NORMAL	358	17	AREA	1,387	32
16	ISSUES	356	22	REQUIRED	1,378	20
17	ANALYSIS	350	19	INDIVIDUAL	1,374	27
18	INVOLVED	329	6	SIMILAR	1,361	17
19	FUNCTION	319	18	SIGNIFICANT	1,327	30
20	CONTEXT	313	8	LEGAL	1,299	47
Total		**8,978**	**464**		**36,557**	**945**

Note: Highlighted blocks indicate common node words in the spoken and written corpus

Information presented in Table 2 also indicates that the frequent 20 spoken academic words yielded 464 collocations. However, as mentioned above, since there were overlaps of collocations, such as in when the two words *data* and *analysis* share the collocation *data analysis*, the exclusion produced 460 collocations. In the written corpus, 945 collocations were found, and 934 collocations were found without overlaps. The results indicate that the number of

written collocations are roughly two times more frequent than the number of spoken collocations even though we have applied different frequency cut-off points for leveling the different sizes of the two academic corpora (See Appendix for a sample of the collocation list). This suggests that there is a greater variety of academic collocations in the written area.

This finding contrasts to how previous research (Biber et al., 2004) has found fixed expressions to be more frequent in the spoken register (i.e., conversation, classroom teaching) than those in the corpora of written registers (i.e., textbooks, academic prose). Biber et al. (2004) demonstrated how they found a greater range of these different 'lexical bundles' in the spoken register. We attribute this difference to the way the fixed sequences of words were defined, specifically in Biber et al.'s case, as four-word sequences (e.g., *what do you think*, *something like that*) whereas our coding system was based on two-word sequences. When word sequences are broken down into smaller units, it seems the collocational units in the written corpora become more readily noticeable. This pattern of results seems to have occurred when writers (in contrast to real-time constraints of face-to-face communication) tend to use fuller expressions, contemplating over what needs to be stated.

In relation to the node words, an evolving query was to examine the relationship between the frequency of the node words and the frequency of collocations. That is, we examined whether it was the high frequency node words that produced more collocates than the lower frequency node words. For this, the correlation between the raw frequency of the node words and the number of collocates was measured with Pearson *r*. The correlation between the frequency of the written node words and the number of their collocations was significant at $r = .865$, $p < 0.01$, with effect size of $R^2 = 74.8$, indicating a highly reliable correlation; for the spoken node words, the correlation did not show a strong relationship ($r = .414$, $p = .07$). Increases in the frequency of the high-frequency node words seem to have contributed accordingly to the production of collocations in the written corpus, but not as strongly in the spoken corpus. The analysis for the relationship between nodes and the collocates demonstrates that while the frequency of collocations is relatively larger in the written academic corpus, there is a tendency of the higher frequency node words to prompt the production of written collocations. For instance, referring back to Table 2 for the written corpus indicates that the first ranking node word *income* (95) produced more number of words than *legal* (47).

2. High-Frequency Academic Spoken and Written Collocations

This section examines the high-frequency academic collocations themselves, and differences or similarities between the two registers. As highlighted in Table 3, there were 11 collocations that occurred in both the top 50 spoken and top 50 written collocation lists. They are *unit area, this area, this period, period of time, in the process, economic policy, economic development, economic growth, in the area, of the data,* and *this analysis*. The overlapping collocations indicated *area, period, process, economic, data,* and *analysis* to be the most commonly occurring node words. Likewise, the related collocations can be regarded as the lexical items most potentially worth learning since there is high chance that the related collocations will appear in lectures in the classroom or in academic texts.

〔Table 3〕 Top 50 Academic Spoken Collocations vs. Top 50 academic Written Collocations

Rank	Academic Spoken Collocations Top 50	Freq	Academic Written Collocations Top 50	Freq
1	[in (26)] **this lecture**	85	**income tax** [act (168), purposes (16)]	572
2	[per (19)] **unit area**	75	[use of (50), types of (11), awareness of (10), acquisition of (8), literature on (7), sources of (6)] **export information**	372
3	[end (14), part (10), beginning (5)] **of the lecture**	60	[from (13), other (10)] **assessable income**	339
4	[in (30)] **this area**	60	[export (121), international (25)] **marketing research** acquisition(6), sources(6)]	294
5	[in (17), during (12), at (4)] **this period**	54	**in section** [No.]	277
6	[at the (11)] **medical school**	43	**of income**	261
7	[a long (7), a certain (6), a short (4), for a (5)] **period of time**	42	[in (112)] **this section**	245
8	[medical (4)] **research ethics**	38	[active (32)] **labour market** [policy(38)]	231
9	**in the lecture**	32	[each (6)] **income year**	204
10	**medical research** [ethics (4)]	29	**of policy**	196
11	**research ethics committee** [committees (6)]	28	**of the income**	188
12	[of (5)] **medical students**	27	**of labour**	187

Rank	Academic Spoken Collocations Top 50	Freq	Academic Written Collocations Top 50	Freq
13	**medical student**	27	[with (16)] **this approach**	160
14	**get involved**	26	**of research**	145
15	**in the process**	26	[in (111)] **this area**	142
16	**in theory**	26	**export performance**	135
17	[No.] **year period**	25	[of (51), on (10)] **policy advice**	129
18	**last lecture**	25	**in the area**	123
19	**market research**	24	**export marketing research** [information (20)]	120
20	**qualitative research**	23	**this process**	115
21	[between (4)] **economic policy**[and social policy(4)]	22	**of data**	112
22	[in (3)] **this theory**	22	[pursuant (37), contrary (18)] **to section** [No.]	104
23	[quality (5), some (4)] **of the research**	22	[during (32), in (22), over (15), of (12)] **this period**	102
24	**medical history**	22	**export market intelligence**	102
25	[other (3), British (6), addicted to (3)] **medical dramas**	20	**monetary policy**	100
26	**economic development**	20	**public policy**	97
27	**economic growth**	20	[the (28), duplex (8), this (6)] **average structure**	90
28	**a long period**	18	**under section** [No.]	86
29	**in the area**	18	**labour party**	84
30	**in the structure**	18	**economic growth**	82
31	**lecture notes**	18	**in the process**	82
32	**of the data**	18	**of the data**	82
33	**past medical history**	18	**market research**	81
34	**the whole issue**	18	**economic policy**	79
35	**whole process**	18	**social policy**	78
36	**observed data**	17	[qualitative (6)] **data analysis**[methods(6)]	77
37	**over a period**	17	[total (12), daily (6)] **unit area** [loadings (50)]	75

Rank	Academic Spoken Collocations Top 50	Freq	Academic Written Collocations Top 50	Freq
38	**some data**	17	[on (6), from (6)] **this analysis**	74
39	**this analysis**	17	**data collection**	73
40	**next lecture**	16	**fiscal policy**	73
41	**research project**	16	[the (67)] **next section**	71
42	**this structure**	16	[a (18)] **period of time**	68
43	**medical drama**	15	[the (51)] **section heading**	68
44	**social and economic** [history (3)]	15	**policy making**	68
45	[from a (4)] **normal distribution**	14	**section headings**	68
46	**at the data**	14	**immigration policy**	67
47	**that structure**	14	**low income**	67
48	**today's lecture**	14	**export information acquisition** [modes (19), mode (6)]	66
49	[from (8)] **one individual**	13	**of the labour**	66
50	**decision making process**	13	**economic development**	62

Note: Bold letters indicate the core collocation; highlighted blocks indicate common collocations in the spoken and written corpus; [ ] indicates the additional component of the collocation that was listed when the collocate(s) occurred at least 3 times in the spoken corpus and 6 times in the written corpus; ( ) indicates the frequency of the collocate(s); [No.] indicate instances where numbers occurred

When the collocations were analyzed for the different structural types, the academic collocations existed in three forms. They were 'referent + academic word' (e.g., *this period, this analysis*), 'noun phrases' (NP) (e.g., *unit area, period of time, economic policy, economic development, economic growth*), and 'prepositional phrases' (PP) (e.g., *in the process, in the area, of the data*). All the top 50 collocations in the collocation list for both the spoken and written corpora could be analyzed within this framework so that there were more similarities rather than differences in the structural types. There was one exception in the spoken collocation list where there was identification of a 'verb phrase (with passive verb)' as in *get involved*. As found in Biber et al. (2004), identification of lexical bundles in classroom teaching, textbooks, and academic prose produced a higher percentage of NP/PP-based bundles in comparison to conversation (spoken register). Almost 70 per cent of the common lexical bundles in academic prose consisted of noun phrase expressions (e.g., the nature of the) or a sequence that bridges across two

prepositional phrases (e.g., as a result of). In a similar vein, Durrant (2009) who was interested in identifying two-word, positionally-variable collocations used across disciplines in academic writing, was able obtain top listing of collocations for: *this study*, *this paper*, *present study, associated with*, *based on*, *due to*, *and respectively*, *consistent with*, *related to*, *compared to*, *was performed*, *was used*, and *number of*.

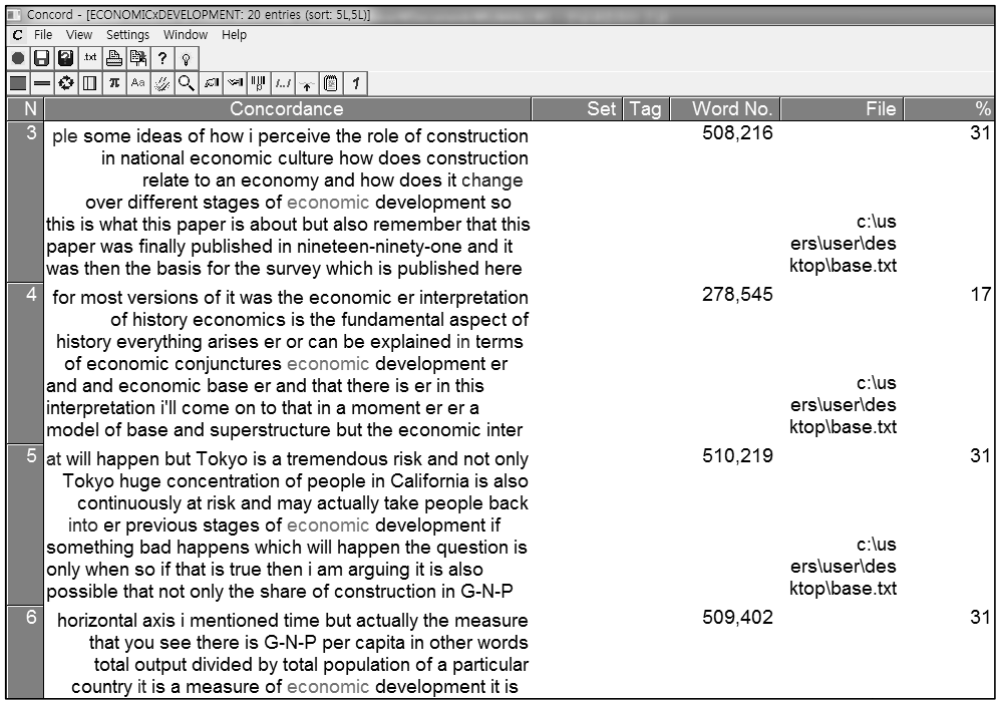
With our focus on producing a discipline-encompassing collocation list (rather than one for the comparison between the sub-disciplines), the array of academic collocations shows that there is prevalence of economic terms throughout both the spoken and written corpus, such as in *economic development*, *economic growth*, and *economic policy*. The results indicate the fixedness of collocations in the academic fields. Other academic collocations, such as, *income tax*, *assessable income*, *low income, labour market, social policy, fiscal policy, policy making,* and *immigration policy* were also found in the written corpus within the sub-disciplines of social sciences. In the spoken corpus, there were also numerous instances of collocations with *research* (i.e., *research ethics*, *medical research*, *research ethics committee*, *market research*, *qualitative research*, *of the research*, *research project*, and *market research)*. The spoken corpus also has a concentration of collocation listings in the field of medical sciences, such as in *medical research ethics*, *research ethics committee*, *medical student(s)*, *medical history*, *medical dramas*, *past medical history*, and *medical drama,* but it may need to be noted that this was due to how the written corpus did not include corpus from the medical field. All in all, the commonly listed academic collocations in both the spoken and written corpora may be able to give practitioners involved in materials development or syllabus design a sense of direction on the kinds of collocations that deserve teaching.

Typically, once a list of collocations has been produced, one of the next logical steps of the process would be to devise a method for teaching the lexical items. Foremost, the teaching of the forms would have to be coupled with the context-specific meaning of the collocations, that is, via data-driven learning (DDL). Although the approach has been the foundation and inspiration behind pedagogic applications over the past two decades (Chambers, 2010), the approach has been underestimated for the learning of collocations themselves. According to Johansson (2009), DDL is usually associated with an inductive, discovery-based approach to learning in which students work out rules or probabilities from the examples provided, whereby there is emphasis on gaining insight rather than establishing habits. However, in the approach of teaching collocations with DDL, there is criticism that solely an inductive approach would make high demands on the students in terms of language proficiency, observation and inductive reasoning (Kennedy & Miceli, 2010). That said, Flowerdew (2012) has pointed out how a guided approach is needed so as to provide L2 learners with the competence to be able to work

with the corpus, and gently familiarize themselves with corpus methodologies such as the inductive approach.

Within the guided approach, we make some practical suggestions on teaching academic collocations by presenting the psychological conditions that need to be met for optimum vocabulary learning (Nation, 2003), which has been applied more for the teaching and learning of single lexical items. However, the framework can be expanded for establishing learning conditions of multi-word items, in our case, academic collocations. For any vocabulary acquisition to occur, the psychological conditions that have to be reached are 'noticing', retrieval', and elaborating.'

'Noticing' involves paying attention to a word as a language feature. To create conditions for noticing, the target collocations can be made salient to the learners via screenshots of the concordance output in the form of a worksheet or in the actual concordancer, in our case at *WordSmith*. The teacher would try to draw the learners' attention on how often the collocations have appeared in the corpus and point out the learning benefits. See Figure 3 for examples on *economic development* from BASE.


Concord - [ECONOMICxDEVELOPMENT: 20 entries (sort: 5L,5L)]

File View Settings Window Help

N	Concordance	Set	Tag	Word No.	File	%
3	ple some ideas of how i perceive the role of construction in national economic culture how does construction relate to an economy and how does it change over different stages of economic development so this is what this paper is about but also remember that this paper was finally published in nineteen-ninety-one and it was then the basis for the survey which is published here			508,216	c:\users\user\desktop\base.txt	31
4	for most versions of it was the economic er interpretation of history economics is the fundamental aspect of history everything arises er or can be explained in terms of economic conjunctures economic development er and and economic base er and that there is er in this interpretation i'll come on to that in a moment er er a model of base and superstructure but the economic inter			278,545	c:\users\user\desktop\base.txt	17
5	at will happen but Tokyo is a tremendous risk and not only Tokyo huge concentration of people in California is also continuously at risk and may actually take people back into er previous stages of economic development if something bad happens which will happen the question is only when so if that is true then i am arguing it is also possible that not only the share of construction in G-N-P			510,219	c:\users\user\desktop\base.txt	31
6	horizontal axis i mentioned time but actually the measure that you see there is G-N-P per capita in other words total output divided by total population of a particular country it is a measure of economic development it is			509,402		31

〔Figure 3〕 Concordancer Output for *economic development* from BASE

Noticing has in fact been recognized as a key concept in the lexical approach as it plays the role of transforming input into intake. Lewis (1997) insists that "exercises and activities which help the learner observe or notice L2 more accurately ensure quicker and more carefully-formulated hypothesis about L2, and so aid acquisition" (p. 52). Other ways of eliciting learners to notice academic collocations would be to ask them to find any recurring collocational items in academic journals within specific disciplines.

Another condition for learning collocations is 'retrieval' which can be receptive or productive; it may involve recalling the meaning or part of the meaning of a form when the spoken and written form is met (receptive retrieval), or recalling the spoken or written form in order to express a meaning (productive retrieval). To operationalize conditions for 'receptive retrieval', learners may be asked to discuss, for instance, what *economic development* entails or encompasses in each of the lines that can be found in the corpus output at the concordancer. Figure 4 illustrates a line from BASE where the speaker is trying to inform the listener(s) as to what may have been the inhibiting factors (i.e., Hinduism and the caste system) of *economic development*.

N	Concordance
2	into the eighteenth century before they were pushed aside especially by the British er the great merchants of the south er the South er China Seas the er the whole of the Indian Ocean area that whole area between the Mediterranean and China er were that whole area was dominated by Indian merchants trading across this va–, vast area so certainly it didn't er mean that the Indians were not not good a–, good at this but he argued the affect of Hinduism and the caste system inhibited **economic development** compared to the West in China he noted high er levels of evolution but with Confucianism

〔Figure 4〕 Concordancer output for 'economic development' from BASE

After having consolidated the learners' knowledge of collocations by 'receptive retrieval', the next step of learning would involve 'productive retrieval.' One of the activities that can be incorporated are tasks where the learners are asked to paraphrase or summarize the main idea orally or in writing by use of the target collocation, preferably without looking back at the concordance passage.

The final step of creating conditions for learning collocations is 'elaborating.' Elaborating is a more effective process and enriches the memory for an item as well as strengthening it. According to Nation (2003), examples of elaboration include meeting a known word in

listening or reading where it is used in a way that stretches its meaning for the learner (receptive generative use), and using a known word in contexts that the learner has not used it in before (productive generative use). For 'receptive generative use', the learners may be asked to electronically look up collocations in the academic corpus for other collocates of a particular node word of interest. For instance, according to the previous [Table 3] Top 50 Academic Spoken Collocations vs. Top 50 academic Written Collocations, other academic collocations that can be drawn to the learners' attention with *economic* as the node word are *economic policy, economic development,* and *economic growth*. The learners at this stage can be led to scrutinize the different usages of academic collocations with *economic*.

In the end, 'elaborating' is finalized when the learning of academic collocations is generated for 'productive generative use.' At this stage, one of the options may be to ask learners to write or discuss a particular topic of academic interest (e.g., *What are the stages in economic development? What does it require to develop a nation economically? Assessing the economic performance of South Korea*) so that learners can be offered opportunities to use the collocations acquired during the previous stages. This will give learners the opportunity to consolidate the form and meaning of the academic collocations that have been exposed to them (i.e., *economic development, economic growth, economic policy)* in different context.

All in all, the activities are intended to provide learners with a more learner-centered, corpus-driven approach, allowing them to browse the corpus for usages of common collocations, or in noticing particular verbs, nouns or prepositions that may be used as a collocate of an academic node word or with an academic collocation. One of the roles of the teacher at the intermediary stages would be to seek that the learners are able to notice the common patterns of language use. Dictionaries at this stage may be used to look up lexico-grammatical queries, but the corpus will still have the advantage of providing users with many examples of the search item (O'Keeffe, McCarthy & Carter, 2007).

V. CONCLUSION AND LIMITATIONS

It has been well documented in the literature that the knowledge of collocations is often equated with native speaker fluency (Hyland, 2008; Schmitt, 2000; Wray, 2002), but it is often difficult for materials writers and teachers to be able to make guided decisions about the collocations that need to be included in their materials or teaching. This is due to the vast array of collocations that may be used, and in the present study, the present study drew on

complementary types of discourse, BASE and the Academic Corpus (Coxhead, 2000) for analysis.

In the study, there was focus of the 50 high-ranking collocations, and they took the structural form of 'referent + academic word', 'noun phrases', and 'prepositional phrases.' The common collocations appearing in both the spoken and written discourse were within the field of Economics, and related to collocations dealing with various components and types of research.

To elaborate on the findings of academic collocations, a guided approach in applying DDL was proposed. That is, with some of the high-frequency collocations, this study has proposed a way of teaching collocations in a step-by-step fashion for 'noticing', 'retrieval', and 'elaborating.' While the DDL approach itself may be relatively overwhelming to the learners, instruction coupled with the guided approach can be felt more manageable to the learners. The researchers believe that when there is facilitation of the psychological conditions for learning collocations, the learning burden can be reduced to help learners become more fluent and accurate in retrieving the form and meaning of collocations for the receptive and productive tasks of academic discourse.

This study is not without its limitations. For depth of analysis, this research was limited to analyzing primarily the high-frequency academic collocations, that is, those that could be retrieved from the top 20 academic node words. As such, further research needs to be conducted for the lower frequency collocations.

Another area worth researching is to retrieve a list of collocations for the separate sub-disciplines (e.g., Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences) in the academic discourse so as to derive a collocation list that may be of improved practical value. Also, in future research, a comparative identification of academic collocations may be more systematically derived when an academic collocation list is created from two complementary academic corpora, BASE and the British Academic Written English Corpus (BAWE), which have been compiled by the identical research group. Both corpora considers the distribution across four broad disciplinary groups—Arts and Humanities, Life Sciences, Physical Sciences, and Social Sciences.

Last but not least, now that we have quite a number of lectures conducted in English at the Korean universities of tertiary level, another area of research that would need further investigation is in examining how academic collocations are utilized among L2 learners in the classrooms or in their assignments, via compiling a learner corpus. The learners' repertoire of academic collocations can also be compared to those frequently found in native academic

discourse, for instance, English textbooks, or journals written for academic purposes. A comparative study of the sort may be able to sensitize teachers, materials developers and L2 learners on what and how learners need to be exposed to academic collocations via their textbooks or lectures; future studies may also need to seek any explanatory variables that may account for the L2 learners' use of academic collocations.

REFERENCES

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371−405.

Chambers, A. (2010). What is data-driven learning? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 345-358). London: Routledge.

Choi, H., & Chon, Y. V. (2012). A corpus-based analysis of collocations in tenth grade high school English textbooks. *Multimedia-Assisted Language Learning, 15*(2), 41-73.

Coxhead, A. (2000). A new academic wordlist. *TESOL Quarterly, 34*(2), 213−238.

Coxhead, A. (2008). Phraseology and English for academic purposes: Challenges and opportunities. In F. Meunier & S. Granger (Eds.), *Phraseology in language learning and teaching* (pp. 149-161). Amsterdam: John Benjamins.

Dechert, H. (1984). Second language production: Six hypotheses. In H. Dechert, D. Mohle, & M. Raupach (Eds.), *Second language productions* (pp. 211-230). Tubingen: Gunter Narr Verlag.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes, 28*(3), 157-179.

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.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 41*(3), 375−396.

Flowerdew, L. (2012). *Corpora and language education.* New York: Palgrave Macmillan.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 145−160). Oxford: Oxford University Press.

Grant, L., & Bauer, L. (2004). Criteria for re-defining idioms: Are we barking up the wrong tree? *Applied Linguistics, 25*(1), 38-61.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics,* 19(1), 24-44.

Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41−62.

Johansson, S. (2009). Some thoughts on corpora and second language acquisition. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 33-44). Amsterdam: John Benjamins.

Kennedy, C., & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference source. *Language Learning and Technology, 14*(1), 28-44.

Kjellmer, G. (1994). *A dictionary of English collocations: Based on the Brown corpus*. Oxford: Clarendon Press.

Lewis, M. (1997). *Implementing the lexical approach: Putting theory into practice*. Hove: Language Teaching Publications.

Liu, D. (2003). The most frequently used spoken American English idioms: A corpus analysis and its implications. *TESOL Quarterly, 37*(4), 671-700.

Min, D., Kim S., & Kim K. (2010). A study on the development of a corpus-based collocation book using children's English newspapers. *Multimedia-Assisted Language Learning, 13*(1), 173-200.

Nation, I. S. P. (2003). Materials for teaching vocabulary. In B. Tomlinson (Ed.), *Developing materials for language teaching* (pp. 394 - 405). London: Continuum.

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

Nesselhauf, N. (2004). Learner corpora and their potential in language teaching. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125-152). Amsterdam: John Benjamins.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Schmitt, N. (2000). *Vocabulary in language teaching.* Cambridge: Cambridge University Press.

Scott, M. (1999). *Wordsmith tools version 3*. Oxford: Oxford University Press.

Shin, D. (2007). The high frequency collocations of spoken and written English. *English Teaching, 62*(1), 199-218.

Shin, D. (2007). *A collocation inventory for beginners*. Unpublished doctoral dissertation, Victoria University, Wellington.

Shin, D. (2009). *A collocation inventory for beginners: Spoken collocations of English*. Köln: Lambert Academic Publishing.

Shin, D., & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken

English. *ELT Journal, 62*(4), 339–348.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list (AFL). *Applied Linguistics*, *31,* 487-512.

Simpson, R., & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly, 37*(3), 419-441.

Stubbs, M. (2000). Using very large text collections to study semantic schemas: A research note. In C. Heffer & H. Sauntson (Eds.), *Words in context: A tribute to John Sinclair on his retirement* (ELR Monograph 18, CD-ROM). Birmingham: University of Birmingham.

West, M. (1953). *A general service list of English words*. London: Longman, Green and Co.

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

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language and Communication, 20*(1), 1-28.

APPENDIX

Rank	Spoken Collocations (460 Collocation Types)	Freq.	Written Collocations (934 Collocation Types)	Freq.
51	lot of research	13	labour force	62
52	selection process	13	this research	61
53	economic power	12	of analysis	60
54	educational research	12	during the period	58
55	first lecture	12	of the process	58
56	medical market place	12	taxable income	56
57	sorts of issues	12	for research	55
58	surface area	12	research and/& development	55
59	the normal range	12	in the income	53
60	[get(3)] normal ranges	11	in the period [No.(22)]	53
61	[of(8)] medical schools	11	legal system	53
62	after the lecture	11	that income	53
63	every individual	11	labour relations	52
64	in the data	11	these data	52
65	lecture theatre	11	[strongly(2)] significant difference	51
66	the deception theory	11	foreign income	51
67	three dimensional structure	11	by section [No.]	50
68	utility unction	11	in the analysis	50
69	economic base	10	personal income [tax(42)]	50
70	international structure	10	[household(11), net(5), real(9)] disposable income	49
71	more research	10	some evidence	49
72	post-war period	10	the most significant	49
73	this data	10	[strongly (2)] significant differences	47
74	time period	10	economic activity	47
75	whole area	10	[the(10), decision making process	46
76	a certain period	9	time period	46
77	about the structure	9	[active labour(28)] market policy	45

Rank	Spoken Collocations (460 Collocation Types)	Freq.	Written Collocations (934 Collocation Types)	Freq.
78	data register direct	9	export market orientation	44
79	economic theory	9	for the period	44
80	ethical issues	9	government policy	44
81	experimental data	9	of evidence	44
82	I mean obviously	9	of the period	44
83	kidney function	9	[No.] year period	43
84	normal form	9	[the(21)] crystal structure	42
85	normal people	9	for income [tax(26)]	42
86	particular area	9	of the research	42
87	pest analysis	9	[by(17)] gradual process	41
88	sort of research	9	[in the(27)] previous section	41
89	survival analysis	9	in the labour	41
90	those issues	9	such an approach	41
91	current issues	8	empirical evidence	40
92	data collection	8	little evidence	40
93	doing research	8	policy development	40
94	economic and political	8	[common(7)] factor analysis	39
96	empirical evidence	8	legal systems	39
97	general medical council	8	this data	39
98	in this process	8	[non(33)] resident withholding income	38
99	market structure	8	[the(18)] evidence suggests [that S V(27)]	38
100	medical practice	8	of the structure	38
101	part of the process	8	foreign source income	37
102	perfectly normal	8	[export marketing(20)] research information	36
103	production process	8	from income [tax(30)]	36
104	protein structure	8	policy makers	36
105	quantitative research	8	legal title	35
106	to the structure	8	on income	35
107	[particular(2)] research area	7	sterling area	35

Rank	Spoken Collocations (460 Collocation Types)	Freq.	Written Collocations (934 Collocation Types)	Freq.
108	a longer period [of time(5)]	7	[for(15)] future research	34
109	become involved [in(4)]	7	business income	34
110	discourse analysis	7	economic conditions	34
111	do the research	7	statistically significant	34
112	enter the data	7	[of(7)] scientific research	33
113	environmental issues	7	economic efficiency	33
114	good evidence	7	other income	33
115	important issue	7	over a period [of time(13)]	33
116	kinds of issues	7	over the period	33
117	medical culture	7	export sales	33
118	medical services	7	for a period	32
119	ordinal function	7	in the data	32
120	phase function	7	of the area	32
121	research area	7	[for(8)] further research	31
122	research projects	7	department of labour	31
123	set theory	7	evidence of change	31
124	sine function	7	labour markets	31
125	statistical analysis	7	legal aid	31
126	undergraduate medical education	7	survey data	31
127	with real data	7	any income [year(11), tax(8)]	30
128	[Quantifier(6)] amount of data	6	as income	30
129	[these(3)] kind of issues	6	export information sources	30
130	a lot of evidence	6	for analysis	30
131	all the data	6	section [No.] provides	30
132	area based approach	6	data points	28
133	assessment process	6	employment income	28
134	complexity theory	6	[exercise (6)] significant influence [over(6)]	27
135	continuing medical education	6	[higher(3), minimum(3), low(3), lower(2), moderate(2)] income levels	26
136	development issues	6	export decisions	26

Rank	Spoken Collocations (460 Collocation Types)	Freq.	Written Collocations (934 Collocation Types)	Freq.
137	different context	6	in area	26
138	economic activity	6	policy ministries	26
139	economic organization	6	short period	26
140	economic system	6	[in an(23)] export setting	25
141	enough data	6	accounting period	25
142	every other individual	6	by virtue of section [No.]	25
143	historical context	6	data set	25
144	input-output analysis	6	of the analysis	25
145	lecture one	6	significant changes	25
146	market analysis	6	through the process	25
147	medical staff	6	[subsequent(5), preceding(3), earlier(2), future(2), succeeding(2)] income years	24
148	more evidence	6	[the(18)] design process	24
149	much involved [in]	6	cross section	24
150	normal way	6	in similar [circumstances (4), situations (3)]	24
440	serious issue	3	legal research	24
441	simplify analysis	3	policy factors	11
442	sort of context	3	policy outputs	11
443	sort of function	3	research paper [No.]	11
445	sort of individual	3	research suggests	11
446	source of evidence	3	sample period	11
447	specific issues	3	skills required	11
448	surface process	3	state policy	11
449	survival function	3	study area	11
450	system analysis	3	subject area	11
451	task structure	3	syntactic analysis	11
452	tension structure	3	the whole process	11
453	the interwar period	3	to the income	11
454	the other issue	3	with data	11

Rank	Spoken Collocations (460 Collocation Types)	Freq.	Written Collocations (934 Collocation Types)	Freq.
455	training issue	3	with income [tax(2)]	11
456	transformation process	3	with the data	11
457	types of data	3	[further(4)] research is needed	10
458	ventricular function	3	[soild(7)] state structure	10
459	wave-length theory	3	[the(2), a(3)] data file	10
460	whole period	3	[the(7)] decision process	10

Key words: academic collocations, corpus analysis, frequency
Applicable levels: tertiary education

Author(s): Chon, Yuah V. (Hanyang University, 1st author); vylee52@hanyang.ac.kr
Shin, Dongkwang (Korea Institute for Curriculum and Evaluation, corresponding author); sdhera@kice.re.kr

Received: July 31, 2013
Reviewed: August 20, 2013
Accepted: September 15, 2013