THE JOURNAL OF ASIA TEFL
Vol. 16, No. 2, Summer 2019, 608-623
http://dx.doi.org/10.18823/asiatefl.2019.16.2.11.608

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

A Multiword Unit Analysis: COCA Multiword Unit List 20 and *ColloGram*

Dongkwang Shin
Gwangju National University of Education

Yuah V. Chon
Hanyang University

Multiword units receive attention as being an essential part of vocabulary knowledge that will expedite the learning of an L2. However, there is lack of a graded multiword units (MWU) list that can offer direct applications for pedagogy, syllabus design or materials development. The current article aims to report on the development and evaluation of *COCA (Corpus of Contemporary American English) Multiword Unit 20* (COCA_MWU20) where the notion of MWU family is adopted. The compilation involved selecting and grading the MWU by grammatical well-formedness, range, and frequency. The 10,000 MWU families which made up the list were utilized in the development of *ColloGram*, a list-based MWU family analysis program. The *COCA_MWU20* can be expected to help L2 learners increase their knowledge of lexical items beyond single words and *ColloGram* is the first of the kind that can analyze multiword items based on COCA. For researchers, the *ColloGram* can also be a tool for identifying MWU that may appear in a target text.

Keywords: multiword units, grammatical well-formedness, range, frequency, ColloGram

Introduction

Word lists, such as the General Service List (GSL) (West, 1953), have been influential in helping second language (L2) English teachers and materials developers to identify the most frequent words of English. The *Academic Word List* (AWL) (Coxhead, 2000) has also helped to raise awareness on the types of words that should be known for academic support. Recent online word list tools for analysis on vocabulary profiles of texts are now easily accessible to the average practitioner (e.g., *Lextutor*, www.lextutor.ca). However, as pointed out by Martinez and Schmitt (2012), the word lists possess a key deficiency since they are often only the "tips of phraseological icebergs" for representing the recurrent word combinations (p. 302). These multiword items are accepted non-arguably as the prerequisite for proficient language use (Choi, 2019; Lewis, 2000; Nam, 2017; Pawley & Syder, 1983; Schmitt, 2004; Sinclair 1991; Supasiraprapa, 2018) and known for offering a processing advantage (Conklin & Schmitt, 2008; Jiang & Nekrasova, 2007; Vilkaitė & Schmitt, 2017).

Multiword units (MWU) ease the processing overload not only because they offer prefabricated expressions, but also because their salient meanings are easily accessible in online production and processing. For instance, when MWU pre-exist for the learners to retrieve while reading or writing, processing of the items is expedited since learners will not need to unpack the meaning of the individual

words that constitute the items (Martinez & Schmitt, 2012). In the context of the present study, MWU is used in a way to include all lexical items that go beyond single word items (e.g., collocations). However, later sections clarify how our listing of MWU has to satisfy pre-determined criteria for pedagogical purposes.

The previous work on compiling lists of words has helped to establish principles and criteria for compiling MWU lists. There are now pedagogically-oriented listings of high-frequency collocations for spoken English (Shin & Nation, 2008) and academic purposes (Ackermann & Chen, 2013; Biber, Conrad & Cortes, 2004; Durrant, 2009; Simpson-Vlach & Ellis, 2010). Taken together, an evaluation of the studies indicated that the MWU were limited to certain number of words (Biber et al., 2004; Simpson-Vlach & Ellis, 2010), degrees of compositionality (Martinez & Schmitt, 2012), or may lack criteria selected for representing MWU that can be considered pedagogically-viable (Ackermann & Chen 2013; Durrant, 2009).

Biber et al. (2004) investigated the use of multiword sequences in two university registers: classroom teaching and textbooks. They took a frequency-driven approach to the identification of multiword sequences, referred to as 'lexical bundles.' Only twenty percent of their bundles were grammatically well-formed, and their listing of multiword sequences was limited to 4-word items with the exclusion of the many more frequent two item and three item bundles, which have high pedagogical value. Durrant (2009) describes a listing of non-adjacent (positionally-variable) academic collocations and evaluates the extent to which it is likely to be useful to students from across a range of disciplines. Simpson-Vlach and Ellis (2010) also compiled the *Academic Formulas List* (AFL) by being involved in extracting mainly the recurring 2-word grammatical multiword sequences with help from statistical information. Ackermann and Chen (2013) claimed their compilation to be representative of actual use of collocations, but their systematization process, which involved removing the items to make the list more readily accessible for users, was done at the risk of overlooking many of the actual usage features of collocations. The collocations were listed in their base form, such as, by changing adjectives in the comparatives/superlatives to the base form, changing inflected verbs to infinitives, and adding dominant prepositions to collocations. A recent pedagogically-driven multiword item list is the *Phrasal Expressions List* of 505 most frequent multiword expressions (Martinez & Schmitt, 2012), but they were limited to relatively non-compositional collocations intended for receptive use due to the claim that they have higher pedagogical value in comparison to the relatively transparent items.

In particular, the previous studies taking a statistical approach (Ackermann & Chen, 2013; Durrant, 2009; Simpson-Vlach & Ellis, 2010) have uniformly considered mutual information (MI) as a criterion for selecting and ranking collocations. Although MI scores are useful indicators for the statistical measures of cohesiveness, often applied best for 2-word items, the consideration of MI scores will have influenced the way the multiword items were prioritized in the respective lists. The measure tends, in contrast to frequency, to identify rare phrases comprised of rare constituent words, such as many subject-specific phrases. Also, another accepted shortcoming of the MI score is that it will produce abnormal results when the frequencies are very low (Schmitt, 2010). Therefore, phrases selected by MI scores may privilege low-frequency items which will be of less pedagogical value for L2 teachers and students. For the selection of MWU, we chose to adhere to the *bona fide* measures used to rank and order MWU, such as, raw frequency and range.

Frequency, which has traditionally been an important criterion to extract MWU, was chosen as a selection criterion to indicate how often an item is likely to be met and used. It has also been the main criterion for selecting lexical items (e.g., General Service List, BNC 14,000, BNC-COCA 25,000) and for defining MWU items (Biber et al., 2004; Crystal, 1985; Kjellmer, 1982, 1984, 1987; Liu, 2003). As with the making of word lists (Nation, 2013), lexical bundles (Biber et al., 2004) and idioms (Liu, 2003), *range* was also considered in selecting the MWU items. Range is measured by seeing how many times a particular word or phrase occurs in different texts. The reason for using a range figure would be to ensure that particular MWU are not restricted to corpora of a specific kind, but are generally useful.

Another criterion that has been used by expert-human judgment to select MWU is the criterion of

grammatical well-formedness, which refers to only selecting the MWU that can be treated as a complete cohesive unit (Shin & Nation, 2008). For this purpose, immediate constituents (Bloomfield, 1933), components that immediately make up larger parts of a sentence, are looked for. Martinez and Schmitt (2012) treated this as a criterion of "meaningfulness" in the sense that MWU should be treated as a feasible unit for learning. This entails manual identification of semantically cohesive units.

Taken together, the current study focuses on compiling a list of MWU that can be considered different in three aspects, that is, in comparison to those that have been developed in previous studies for developing multiword lists (Ackermann & Chen, 2013; Biber et al., 2004; Durrant, 2009; Shin & Nation, 2008; Simpson-Vlach & Ellis, 2010). First, as an effort to reflect how L2 learners are likely to experience problems with all sorts of MWU of different degrees of compositionality (transparency), the current study focused on identifying MWU that run on the continuum from non-compositional to the compositional end (cf. Martinez & Schmitt, 2012). The explanation for this is that while some MWU are obviously opaque (e.g., *pass out*) and likely to be noticed for learning, transparent MWU, on the other hand, are also expected to be problematic for L2 learners, for instance, when the learners apply their knowledge of synonymous collocates (e.g., *begin the engine*/ start the engine*). Second, another novel accomplishment was to compile a general MWU list based on the *Corpus of Contemporary American English* (COCA), currently the largest corpus available on modern English, making the study the first of its kind for compiling a list of multiword items at such a scale. A useful feature of having access to a MWU list is that they may provide a number of applications, for instance, in obtaining a sample of the test taker's breadth of lexical knowledge. For this purpose, the MWU needed to be graded systematically based on pre-established criteria so as to improve its pedagogical value. As such, a third distinct feature of the MWU list is that they would be graded in order to guide practitioners and teachers in informing them on the types of MWU that need primary attention for teaching or syllabus design. In the process of grading the MWU, the researchers were also able to classify the items into what the researchers labeled the 'MWU family', being conceptually similar (but not identical) to how a word family may consist of a word in addition to its inflected forms and derived forms. Detailed explanation for this is presented later.

The current study is also a part of project in which we were interested in developing a collocation analysis program, similar to the RANGE program (Heatley & Nation, 2002) that has been used to analyze single word items. For such a program to be developed, a general collocation list needed to be compiled. The program was eventually conceived as *ColloGram* (Shin, Chon, Lee, & Park, 2018) and used for the analysis of collocations in the current study. To the researchers' knowledge, the program that uses multiword items as the unit for analysis to identify target MWU of particular types is also the first of the kind. Awareness has already been raised on the need for this type of program by Martinez and Schmitt (2012) who noticed that the field was in need for a collocation analysis program that can "[flag] up multiword items that may be worth including for explicit instruction or testing" (p. 316). The following section describes how we operationalized and applied the selection criteria for the compilation of the MWU list, *COCA_MWU20*. The following research questions guided the study.

1. What is the composition of *COCA_MWU 20* that varies in terms of non-compositional to compositional MWU of general English?
2. How does *COCA_MWU 20* represent those MWU of general English when validated by another pre-established corpus of English?

Methods

The Corpora

The first step of the compilation process consisted of selecting a large-scale corpus representative of general English. At the time of the analysis, the *Corpus of Contemporary American English* (COCA)

(1990-2012) was found to be the most valid publicly available type of corpus. With COCA being well-balanced and equally divided among spoken, fiction, popular magazines, newspapers, and academic journals, the corpus was expected to provide the most comprehensive list of MWU. TABLE 1 indicates the actual composition of the corpus. As such, each section of the corpus was about 90,000,000 words, and the whole corpus totaled approximately 450 million words.

TABLE 1
Composition of COCA

Academic	Fiction	Magazine	Newspaper	Spoken	Total
87,600,712	85,496,648	92,292,104	88,503,944	94,959,712	448,853,120

Procedure for Compilation

At this stage, there was need to define MWU. As Biber et al. (2004) point out, in spite of their importance, there is little agreement on how to define their characteristics, or how to identify them. While MWU have also been referred to as prefabricated patterns, preassembled units, formulaic sequences, chunks, ready-made utterances, and so forth (Howarth, 1998; Nattinger & DeCarrico, 1992; Wray, 1999), Moon (1997) uses the term, 'multiword items' to refer to formulas that are institutionlized, fixed, and non-compositional. In comparison, we were interested in the extended spectrum of compositionality. We defined MWU as non-compositional to compositional multiword sequences of two to seven widely co-occurring contiguous words. We used a mixed-method, two-step methodology which comprised an exhaustive computer-assisted search for co-occurring words with use of node words, followed by manual vetting of those items with the guidance of pre-determined selection criteria. In terms of what has been referred to as the frequency-based approach (i.e., use of frequency as the main criterion) and the phraseological approach (i.e., consideration of semantic/grammar) (Barfield & Gyllstad, 2009; Nesselhauf, 2005), both were ultimately considered.

Selection of node words

The compilation process involved using the most frequent 5,000 lemmas as nodes, downloadable from the COCA website, to search for their MWU entries in COCA. The search was conducted with WordSmith 6.0 (Scott, 2012). It was found more feasible to conduct searches with lemmas rather than by word types since searches otherwise would require an extensive winnowing of repetitive inflectional forms (e.g., *go home, going home, went home, gone home*). The lemmas selected were only content words (i.e., nouns, verbs, adjectives, adverbs). Verbs that are used as auxiliaries—*have, make, take, do, and let*— were also included in the analysis due to their useful functions for expressing meaning with co-occurring words. Pronouns (*he, himself*), exclamations (*wow, hey*), conjunctions (*unless, because*), interrogatives (*what, how*), determiners (*the, those*), prepositions (*over, after*), and numbers (*fifty, sixth*) were excluded. The search for the MWU was conducted with the span of 3 words to the right and to the left from the node, which produced two to seven word MWU. Further details are presented in the following to explain how the selection criteria were operationalized in the development process of the MWU list.

Employment of minimum cut-off point for frequency

A minimum of 20 was used as the cut-off frequency point for the MWU to be selected. Previous studies on MWU (Webb, Newton, & Chang, 2013; Webb, 2007) have shown that a frequency of 10-15 is a useful frequency to reach for the acquisition of MWU. Although acquisition usually increases with the rate of exposure, 20 repetitions seemed sufficient as the minimum cut-off point. Indeed, a much higher

cut-off point might have worked when considering the size of COCA (450 million words), but this would have been at the cost of excluding essential MWU that would be useful for L2 learners. Although the frequency cut-off point that was utilized in the current study can be considered relatively few compared to the size of COCA (i.e., 450 million words), there were selection and inclusion of MWU that would be felt pedagogically meaningful to second language learners, for which information on relatively high frequency figures was utilized foremost.

Refinement by grammatical well-formedness

The next step of the analysis required the researchers to identify MWU by their immediate constituents that make sense as independent units (Bloomfield, 1933), that is, to select items that are grammatically well-formed. For instance, according to Bloomfield, six immediate constituents can be found in the following sentence (Shin & Nation, 2008, p. 342). However, in the context of the present study, MWU that took form as in "the zoo" (i.e., determiner + content word) was excluded due to how the type is expected to have low learning gains.

'I saw that animal at the zoo.'

1 I saw that animal at the zoo
2 saw that animal at the zoo
3 saw that animal
4 at the zoo
5 that animal
6 the zoo.

That animal at the zoo, however, does not meet this criterion because it crosses an immediate constituent boundary. Only including word sequences that meet this criterion was important for making principled decisions on the items to be included. Nonetheless, there was a need to further consider the following rules to make the list more pedagogically useful.

1. Listing only the base form of MWU when the use of articles are considered optional (e.g., *kind of thing*)
2. Adding prepositions even when they are not considered a part of a prepositional verb (e.g., *become involved in*, *take charge of*)
3. Adding interrogatives when they are deemed functionally useful (e.g., *reason why*)
4. Adding *to* of a 'to infinitive' as a part of MWU (e.g., *had a chance to*)

Expert review

The judgement for grammatical well-formedeness involved a panel of four experts with professional backgrounds who knew about the research of MWU. All the researchers had doctorate degrees from the areas of corpus linguistics or vocabulary learning. The experts had a minimum of 15 to a maximum of 18 years experience with having taught L2 learners in the departments of English Language Teaching at the university level. The experts were considered close to being nativelike according to their proficiency level. A fifth expert, a native-speaker instructor of English working at the leading researcher's university was involved in the refinement process of the MWU list to check for identification of MWU that did not seem consistent with our coding scheme.

There was a workshop with a list of benchmark MWU in order for the researchers to share opinions on the types of MWU that should become a part of the list. Since the process of winnowing words to identify immediate constituents, which required qualitative judgement, was not as straightforward as expected, the researchers cross-checked to see that their conceptualization of grammatical well-formedness was being

consistently applied. Thereafter, the researchers were asked to work independently for a period of 3 weeks. While working, the experts were asked to contact the principle researcher for any queries about particular MWU. In fact, the additional rules no. 1~4 listed in the previous section (*Refinement by Grammatical Well-formedness*) evolved as the researchers consulted each other.

The multiword unit (MWU) family

In the process of listing and organizing the MWU, we were able to conceptualize what can be coined as a 'MWU family.' This is alike how a head word may have inflectional or derivational forms of the word. However, when only the node words of MWU are lemmatized, this leaves the accompanying collocates being listed as types rather than lemmas so that lemmatization for the whole MWU is incomplete. As such, it was necessary to lemmatize the co-occurring lexical items as well. FIGURE 1 presents the configuration of *gave birth to* with the total frequency of 4,389. In the configuration of MWU families, the MWU with the highest frequency was listed as the head MWU. The notion of MWU families may not be completely new since it was already discussed by Martinez and Schmitt (2012) when they mentioned the need to lemmatize the MWU item list in the same way that wordlists have been lemmatized. As such, in our classification, head MWU may subsume different inflectional (e.g., inflected verbs, singular/plural forms of nouns) and derivational forms (e.g., different word order, forms with more or less constituents) of the head MWU. For instance, in our list, the inflections for ***go home*** were *goes home*, *going home*, *went home*, and *gone home*. For ***growth rate***, inflections were *growth rates*, *rate of growth*, and *rates of growth*. In comparison, derivations of MWU existed in the form of a word(s) being added to or deleted from the head MWU. An example of derivational forms for ***due process*** are *due process of law* and *without due process of law*.

Another issue that transpired while refining the list for the head MWU and its types was frequency. As the items were being identified for their immediate constituents in trying to extract word sequences that satisfied grammatical well-formedness, there was the need to record the frequency figures, that is, for the head MWU and its types (i.e., MWU family). This involved what has been referred to as the *subtractive method* (Martinez & Schmitt, 2012), which was needed to arrive at a more accurate frequency figure of the MWU in developing *COCA_MWU20*. However, in order to record the exact frequency, the frequency for each MWU had to be subtracted from that of the head MWU to eliminate any instances of the frequencies being counted more than once. For example, ***not so easy to*** can be subsumed under ***so easy to***, a shorter type of the same MWU family. However, in order to obtain the exact frequency for ***so easy to***, there was need to subtract the number of occurrences of the string ***not so easy to*** (178) from the number of times the trigram ***so easy to*** appears in the corpus (1,046). As a result, the true frequency of ***so easy to*** is 868. Also, if any of the MWU did not reach the minimum cut-off point of 20 after having calculated the frequency by the method described above, then the item was excluded from the list.

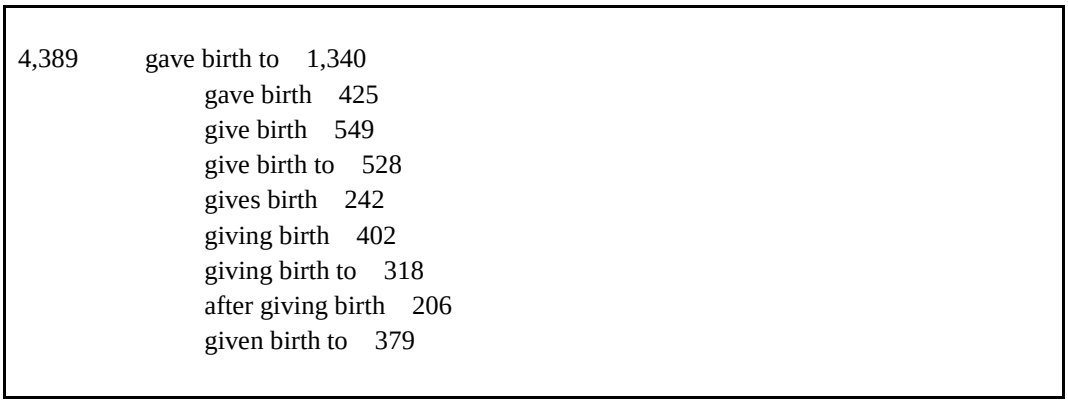

Figure 1. COCA_MWU family.

As a result of 'familizing' the MWU, the list consisted of 10,299 MWU families. In the MWU list, the shortest form of the MWU was not necessarily always the base form of the MWU. The decision for selecting a head MWU was made based on which of the MWU type was most frequent and pedagogically useful among the items.

Employment of range for MWU

Once the MWU had passed the grammatical well-formedness judgement test with the minimum threshold of 20 occurrences in COCA, the MWU needed to be validated by *range* based on the distribution of the five types of corpora in COCA. This was to indicate how widely a particular MWU family is distributed in our target corpus, similarly to the way the British National Corpus (BNC) word family list was created. The MWU master list (i.e., 10,299 MWU families) was uploaded to *ColloGram* at this stage. MWU families that received ranges of 4-5 from the 5 domains of COCA were considered valid for their coverage and representativeness of the MWU list. A minimum range of 4 produced 10,214 MWU families. Previously, Liu (2003) and Biber et al. (2004) also used range as a measure to respectively compile a list of idioms and lexical bundles. As such, it was found valid to use range figures to compile the MUW list and the methodology enabled the researchers to include high coverage MWU items occurring in general native-speaker English.

Division of MWU bands

The master list of 10,214 MWU families needed to be graded for wider applications in learning and teaching. Word lists have traditionally been graded by every 1,000 word (Nation 2001). However, when results are multiplied by a 100, this may actually be an overestimation of the non-native speakers' vocabulary size (Gyllstad Vilkaité & Schmitt, 2015). Since the non-native L2 learners' opportunities for exposure to the second/foreign language would be far less compared to native-speakers, we felt that the bands needed to be smaller in size. That is, 500 MWU families was considered a more valid unit to divide the bands for assessing L2 learners' knowledge of MWU. This is line with how Kremmel (2016) has found smaller 500-item bands to be more informative for vocabulary test development since we were also considered about the pedagogical utility of the list. Graded as such, 20 bands were produced, which had coverage for 10,000 MWU families (i.e., 20 bands * 500 MWU families). The final list of MWU families was referred to as *COCA_MWU20* where 20 indicates the number of graded bands. The composition of the list as a whole is explained in the *Results and Discussion* Section.

The ColloGram: A multiword family analysis program

As mentioned previously, the field on MWU is in need of a program that can analyze words beyond single words. Realizing both local (the current study) and global needs (within the academic community), we developed *ColloGram*, named from the compound Collocation and N-gram or Program.

The functions of *ColloGram* are similar to those of RANGE, the vocabulary analysis program (Heatley & Nation, 2002). This program provides MWU lists (Basecollo1.txt, Basecollo2.txt, Basecollo3.txt…Basecollo21) and an execution file (32bit, 64bit), available without an installation process. As for the functions of the program, it can count the frequency of MWU types, families, and MWU family members (derivation and inflectional forms). The program can also extract head MWU, and remove duplicate MWU in the list. FIGURE 2 illustrates the screenshot of *ColloGram*.

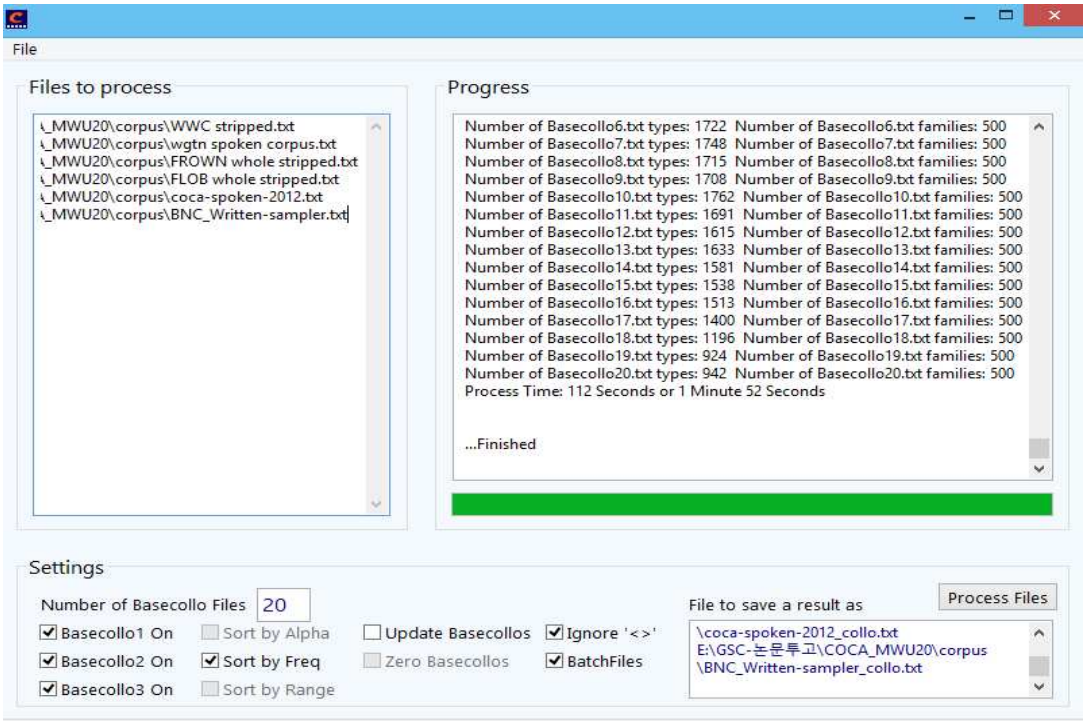

Figure 2. A screenshot of *ColloGram.*

The analysis can be conducted by going to 'File → Open' to upload the target corpus. The same function can be conducted by a 'Drag & Drop' function by which the target file can be made available on the left side of the platform. After having ticked the intended options, the analysis can be run by clicking 'Process Files.' As default, 3 Basecollo Files are ticked. When needing to analyze all 20 bands of *COCA_MWU20*, 20 can be typed in for 'Number of basecollo Files.' In addition, 'Ignore < >' can be checked when needing to remove all POS tagging from the target corpus. When 'Batchfiles' are checked, the results of individual files can be produced as a single file. The *ColloGram* is optimized to analyze a corpus of a million words, but can analyze up to a hundred million words. The program identifies MWU where all the words are immediately adjacent to each other for a maximum of 10 words.

The program can automatically analyze the occurrences of MWU in a target text according to *COCA_MWU20*. This is in contrast to how previous programs (e.g., *Wordsmith, Concgrams, AntConc*) have needed to use keywords (nodes) to search words individually to find the constituents of MWU. *ColloGram* can also provide analysis for a list of particular MWU and allow a graded list of MWU to be utilized.

Results and Discussion

Composition of COCA_MWU20

The final *COCA_ MWU20* list of 10,000 MWU families consisted of 31,680 MWU types where 9,791 items fell in the range of 5 and 209 items in the range of 4. The top 100 MWU items are listed in TABLE 2. Analysis with Range BNC-COCA 25,000 (Nation & Webb, 2011) indicated that they were composed of mostly the top 4,000 words in English (98.85%); 89.94% comprised of words from the first 2,000 word families, and 97.72% from the top 3,000 word families. The vocabulary profile accords with the claim that a majority of MWU are formed by the use of high-frequency words (Kim, 2016; Martinez & Murphy,

2011). This provides an explanation for how MWU may become a comprehension problem when familiar words that make up MWU will often go unnoticed or misunderstood, which indeed provides a rationale for the need of a MWU list and explicit attention to it.

TABLE 2
Top 100 Multiword Unit Items in COCA

Rank	MWU	Frequency	Range	Rank	MWU	Frequency	Range
1	as well	111,979	5	51	young man	18,467	5
2	years old	93,506	5	52	on the other hand	18,045	5
3	years ago	79,557	5	53	take on	17,835	5
4	all right	60,485	5	54	sit down	17,615	5
5	so much	53,298	5	55	a lot of people	17,471	5
6	come up with	46,693	5	56	take care of	17,454	5
7	right now	46,250	5	57	good morning	17,321	5
8	come back	45,207	5	58	web site	17,317	5
9	come out	45,004	5	59	time when	17,207	5
10	focus on	42,650	5	60	turned off	17,190	5
11	high school	42,481	5	61	stood up	17,163	5
12	come in	42,228	5	62	supreme court	17,153	5
13	in order	39,839	5	63	show up	16,704	5
14	in addition	38,908	5	64	other people	16,628	5
15	no longer	38,085	5	65	go down	16,362	5
16	pick up	37,833	5	66	most important	15,931	5
17	go back to	35,608	5	67	trying to get	15,786	5
18	very much	34,961	5	68	very good	15,761	5
19	pointed out	32,581	5	69	once again	15,616	5
20	know how	32,447	5	70	want to do	15,562	5
21	much too	32,075	5	71	more likely to	15,495	5
22	grew up in	31,516	5	72	put on	15,381	5
23	feel like	31,233	5	73	take over	14,940	5
24	out there	30,622	5	74	let go	14,822	5
25	go through	28,204	5	75	get up	14,786	5
26	last week	28,180	5	76	old man	14,737	5
27	find out	27,882	5	77	never seen	14,476	5
28	health care	27,814	5	78	get back to	14,408	5
29	go out	27,436	5	79	in particular	14,388	5
30	make sure	26,570	5	80	work out	14,198	5
31	long term	25,934	5	81	other things	14,147	5
32	so far	25,447	5	82	in general	14,107	5
33	all over	25,001	5	83	just want to	14,050	5
34	vice president	24,904	5	84	wake up	13,955	5
35	little bit	24,360	5	85	on the other side	13,850	5
36	end up	23,885	5	86	in other words	13,800	5
37	set up	23,656	5	87	prime minister	13,618	5
38	much more	23,284	5	88	came down	13,546	5
39	get out	22,154	5	89	figure out	13,504	5
40	for a long time	21,985	5	90	real estate	13,194	5
41	even more	21,648	5	91	in part	13,179	5
42	willing to	21,290	5	92	federal government	13,081	5
43	come on	20,892	5	93	worry about	12,902	5
44	looked up	20,792	5	94	go ahead	12,892	5
45	take place	20,683	5	95	one more	12,887	5
46	make up	20,535	5	96	want to know	12,884	5
47	years later	20,241	5	97	last month	12,791	5
48	give up	20,038	5	98	very well	12,718	5
49	just like	19,281	5	99	living room	12,717	5
50	engage in	18,996	5	100	high level	12,566	5

In line with how the list was compiled from the latest corpus of general English, the items represent MWU that need primary attention for teachers and learners of English. Having a list as such is expected to increase the usefulness and applicability for materials development and syllabus design. For instance, in addition to word lists that have been utilized for the development of graded readers (Nation & Ming-Tzu, 1999; Waring, 2003; Wodinsky & Nation, 1988), further incorporation of the graded MWU list will allow learners to be exposed to large quantities of MWU items through reading in spite of the L2 learners' limited vocabulary. On the other hand, for teaching MWU, the whole MWU list is not meant to be taught exhaustively. The value of the list lies in selecting MWU items from different MWU bands appropriate to learners' level of language proficiency or by selecting the idiomatic expressions for which the learners will need more explicit attention.

The top 10 MWU were *as well, years old, years ago, all right, so much, come up with, right now, come back, come out*, and *focus on.* It can be noted that *years* and *come* are frequent single node items that become the core component for MWU. Although an exhaustive analysis of *COCA_MWU20* by their part-of-speech (POS) was beyond the scope of the current study, the top 100 MWU indicated that there was a high proportion of verbal MWU (47%) as in *stood up, take over*, and *take place*. Research in the future may need to conducted to seek whether verbal MWU (e.g., verb + noun/adj, verb + adv) are likely to be more challenging for L2 learners in comparison to other part-of-speech combinations, but the list implies that verbal MWU may deserve more attention than other POS combinations. Scrutiny of the list beyond the 1,000th MWU indicated decreasing number of verb combinations whereas a higher portion of noun combinations and binomials could be found (*positive attitude, upper and lower, message boards*).

Validation

In order to determine whether the *COCA_MWU 20* is representative of general English, a validation study was conducted with an aim to investigate the list's overall coverage of its source corpus in another well-established corpus. This was conducted with use of *ColloGram* in a comparable corpus of general English—The *Wellington Corpus of Spoken New Zealand English* (WSC) (1 million) and the *Wellington Corpus of Written New Zealand English* (WWC) (1 million), both compiled at a similar time. The corpora were deemed logical choices for comparing the distributional features of *COCA_MWU20* respectively in the spoken and written language. Validating the results of previous studies (Biber & Conrad, 1999; Erman & Warren 2000, Shin, 2007) that provides information on the occurrences of MWU in the spoken and written language were used for validation. A technical validation was conducted concurrently with the performance of *ColloGram* for its efficiency and accuracy.

The coverage configuration obtained by *ColloGram* for the 20 bands in WSC and WWC (Figures 3 & 4) indicated that there were generally natural falls in the coverage of MWU towards the lower frequency bands, and this verified the acceptability of *COCA_MWU20*. The number of MWU families indicated that the COCA_MWU families had occurred 1,479 times more in the written language (5,363) than in the spoken language (3,626). However, calculation of tokens and families indicated that each MWU had been repeated 31 times (13,925/451) in the spoken language and 18 times (8,588/474) in the written language. This indicates that spoken language makes much more frequent use of its common MWU than written language does (Biber et al., 1999; Erman & Warren, 2000; Leech, 2000; Nation, 2016, Shin, 2007). Erman and Warren (2000) found that the density of MWU (i.e., prefabs) is somewhat greater in spoken than in written language (59 vs. 52 percent), which lent support to our validation of *COCA_MWU20*.

WSC

COLLOCATION LIST	TOKENS/%	TYPES/%	FAMILIES/%
one	13925/54.80	1172/20.58	451/12.44
two	3535/13.91	763/13.40	396/10.92
three	1731/ 6.81	547/ 9.61	329/ 9.07
four	975/ 3.84	438/ 7.69	292/ 8.05
five	836/ 3.29	343/ 6.02	237/ 6.54
six	753/ 2.96	346/ 6.08	245/ 6.76
seven	653/ 2.57	295/ 5.18	210/ 5.79
eight	456/ 1.79	246/ 4.32	181/ 4.99
nine	399/ 1.57	216/ 3.79	170/ 4.69
ten	318/ 1.25	202/ 3.55	169/ 4.66
11	293/ 1.15	176/ 3.09	148/ 4.08
12	275/ 1.08	184/ 3.23	142/ 3.92
13	207/ 0.81	134/ 2.35	112/ 3.09
14	299/ 1.18	122/ 2.14	97/ 2.68
15	226/ 0.89	133/ 2.34	107/ 2.95
16	184/ 0.72	117/ 2.05	96/ 2.65
17	129/ 0.51	100/ 1.76	90/ 2.48
18	97/ 0.38	83/ 1.46	81/ 2.23
19	64/ 0.25	44/ 0.77	43/ 1.19
20	57/ 0.22	33/ 0.58	30/ 0.83
Total	25412	5694	3626

Figure 3. Coverage of *COCA_MWU20* in the Wellington Spoken Corpus (WSC) with *ColloGram.*

WWC

COLLOCATION LIST	TOKENS/%	TYPES/%	FAMILIES/%
one	8588/38.74	1219/14.61	474/ 8.84
two	2780/12.54	911/10.92	438/ 8.17
three	1795/ 8.10	734/ 8.80	399/ 7.44
four	1316/ 5.94	632/ 7.58	383/ 7.14
five	955/ 4.31	523/ 6.27	346/ 6.45
six	912/ 4.11	483/ 5.79	327/ 6.10
seven	808/ 3.65	450/ 5.39	322/ 6.00
eight	679/ 3.06	399/ 4.78	288/ 5.37
nine	626/ 2.82	399/ 4.78	283/ 5.28
ten	560/ 2.53	353/ 4.23	271/ 5.05
11	506/ 2.28	328/ 3.93	249/ 4.64
12	547/ 2.47	360/ 4.31	270/ 5.03
13	440/ 1.98	285/ 3.42	229/ 4.27
14	352/ 1.59	269/ 3.22	214/ 3.99
15	363/ 1.64	262/ 3.14	222/ 4.14
16	320/ 1.44	237/ 2.84	196/ 3.65
17	197/ 0.89	160/ 1.92	142/ 2.65
18	175/ 0.79	140/ 1.68	131/ 2.44
19	134/ 0.60	115/ 1.38	105/ 1.96
20	114/ 0.51	84/ 1.01	74/ 1.38
Total	22167	8343	5363

Figure 4. Coverage of *COCA_MWU20* in the Wellington Written Corpus (WWC) with *ColloGram.*

All in all, the coverage configuration of the *COCA_MWU20* was able to show how the list was compiled to represent the most common expressions used in general English, and the *ColloGram* was able to facilitate this validation procedure. Ultimately, *COCA_MWU20* and *ColloGram* are intended to be used as a complement to existing lists and instruments of second language instruction that use them.

Conclusion, Implications and Limitations

The purpose of the current study was to develop a MWU list that can have pedagogical value for learning contemporary English. The compilation of *COCA_MWU20* involved utilizing the criterion of grammatical well-formedness to include only sequences with discrete meaning. This was further refined by range and frequency, the two most objective measures to represent native-speakers' use of English (Nation, 2004; Nation & Webb, 2011). The *COCA_MWU20* adds to the few comprehensive lists of MWU, for instance, the *Phrasal Expression List* of non-transparent multiword expressions (Martinez & Schmitt, 2012). However, to address the need to encompass the whole continuum of MWU (Wray, 2000), *COCA_MWU20* ranged from semantically transparent to opaque MWU.

While there is a lack of MWU lists for general purposes, our project for developing *COCA_MWU20* has tried to fill this gap by adopting a corpus-driven approach. The MWUs that constitute *COCA_MWU20* were systematically compiled with conceptualization of a MWU family which adopts a head MWU as the main unit to derive variants of MWU within its family. The configuration of the list is expected to offer pedagogical value for both teachers and learners since the different MWU derivatives and inflections are listed by each MWU family.

For instance, although a finite number of phrases can be assembled by the application of rules to words (i.e., grammar), our compilation of multiword items by MWU families demonstrates how only a few of these are actually used by speakers of any language. For instance, when learners are asked to recall the variants for 'give a call', grammatically permissible forms of the MWU would be 'gives a call', 'gives a call', 'giving a call', 'gave a call' and 'will give a call.' However, actual identification of its MWU family members indicates that only variants as in "Please, give me a call" or "I will give you a call" are permissible while 'gave a call' would not be used as in when meaning to say "I called/rang you yesterday." As such, utilizing MWU family members jointly with carefully built concordance samples can show both the extent of the formulaic phenomenon as well as the usage characteristics for particular MWU (Cobb, 2018).

In relation to the developments of CALL for MWU research, a noteworthy contribution of the present study is that we have provided access to an MWU analysis program, the *ColloGram*, in response to the realization that the field of vocabulary research is urgently in need of a program that can analyze MWU. Although the main unit of analysis is MWU families, MWU types may also be uploaded to the program for analysis according to different research aims, and different frequency cut-off points may be used to examine certain MWU.

The availability of the list now allows us to develop a MWU size test (cf. Phrasal VST, https://www.lextutor.ca/tests/pvst/) similarly to how there have been attempts to assess the vocabulary size of L2 learners (Nation, 2001; Nation & Beglar, 2007; Schmitt, Schmitt, & Clapham, 2001). MWU size tests of the kind could be developed for both meaning recognition and form recognition, however, in reduced band sizes, that is, by 500 MWU families, which may be the more legitimate size for measuring phrasal knowledge of ESL/EFL learners.

A possible limitation for the pedagogical use of *COCA_MWU20* is that due to its large collection of MWU and its family members varying in terms of compositionality, instructors of L2 learners may be in the most propitious position to be able to make judgments on the types of MWU for which learners need to have their attention drawn for explicit instruction. Instructors may find that it is the MWU with low degrees of transparency (i.e., non-compositional MWU) that deserve most attention. In common with other types of MWU, they will often be composed of high-frequency words, but the meaning of them cannot be easily guessed merely by knowing the single words that constitute them (Martinez & Murphy, 2011). Moreover, MWU will not be easily acquired particularly when the MWU go unnoticed during the learners' encounters with them during reading or listening. As such, the more non-compositional types will deserve attention when learners can only spend a limited amount of time to study them.

The Authors

Dongkwang Shin is an associate professor at Gwangju National University of Education, South Korea. He had his PhD from Victoria University of Wellington in 2007. His expertise and interest are in vocabulary research and applied corpus linguistics.

Department of English Education
Gwangju National University of Education
55 Pilmundae-Ro, Buk-Gu
Gwangju 61204, South Korea
Mobile: + 82 105492-2232
Email: sdhera@gmail.com

Yuah V. Chon (corresponding author) is an associate professor at Hanyang University in the Department of English Education, South Korea. She has a PhD in ELT from University of Essex. Her research has covered a wide range of topics, including vocabulary, pedagogical use of corpus and learner strategies with a focus on dictionaries.

Department of English Education
Hanyang University
222 Wangshimli-Ro, Seongdong-Gu
Seoul 04763, South Korea
Mobile: + 82 105397-3451
Email: vylee52@hanyang.ac.kr

References

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.

Barfield, A., & Gyllstad, H. (2009). *Researching collocations in another language: Multiple interpretations*. New York: Palgrave Macmillan.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora* (pp. 181-190). Amsterdam: Rodopi.

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London, UK: Longman.

Bloomfield, L. (1933). *Language*. London: George Allen & Unwin.

Choi. W. (2019). A corpus-based study on "Delexical Verb + Noun" collocations made by Korean learners of English. *The Journal of Asia TEFL*, *16*(1), 279-293.

Cobb, T. (2018). From corpus to CALL: The use of technology in teaching and learning formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 192-211). New York: Taylor & Francis.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, *29*(1), 72-89.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

Crystal, D. (1985). How many millions? The statistics of English today. *English Today*, *1*(1), 7-9.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, *28*(3), 157-169.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse*, *20*(1), 29-62.

Gyllstad, H., Vilkaité, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL-International Journal of Applied Linguistics*, *166*(2), 278-306.

Heatley, A., & Nation, I. S. P. (2002). *RANGE and FREQUENCY programs* [Software]. Retrieved August 20, 2016 from the World Wide Web: http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx

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

Jiang, N. A., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, *91*(3), 433-445.

Kim C. (2016). L2 learners' recognition of unfamiliar idioms composed of familiar words. *Language Awareness*, *25*(1-2), 89-109.

Kjellmer, G. (1982). Some problems relating to the study of collocations in the Brown corpus. In S. Johansson (Ed.), *Computer corpora in English language research* (pp. 25-33). Bergen: Norwegian Computing Centre for the Humanities.

Kjellmer, G. (1984). Some thoughts on collocational distinctiveness. In J. Aarts & W. Meijs (Eds.), *Computer corpora in English language research* (pp. 163-171). Bergen: Norwegian Computing Centre for the Humanities.

Kjellmer, G. (1987). Aspects of English collocations. In *Proceedings of the International Conference on English Language Research on Computerised Corpora* (pp. 133-140). Amsterdam: Rodopi.

Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, *50*(4), 976-987.

Leech, L. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning*, *50*(4), 675-724.

Lewis, M. (2000). *Teaching collocation: Further developments in the lexical approach.* Hove, England: 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.

Martinez, R., & Murphy, V. A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly*, *45*(2), 267-290.

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

Moon, R. (1997). Vocabulary connections: Multi-word items in English. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 40-63). Cambridge: Cambridge University Press.

Nam. D. (2017). Functional distribution of lexical bundle in native and non-native students' argumentative writing. *The Journal of Asia TEFL*, *14*(4), 703-716.

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

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In B. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 3-13). Amsterdam: John Benjamins Publishing Co.

Nation, I. S. P. (2013). *Teaching & learning vocabulary*. Boston: Heinle Cengage Learning.

Nation, I. S. P. (Ed.). (2016). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins Publishing Co.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9-13.

Nation, P., & Ming-Tzu, K. W. (1999). Graded readers and vocabulary. *Reading in a Foreign Language*, *12*(2), 355-380.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston: Heinle Cengage Learning.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

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

Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191-226). London: Longman.

Schmitt, N. (Ed.). (2004). *Formulaic sequences: Acquisition, processing, and use* (Vol. 9). John Benjamins Publishing.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. London: Palgrave Macmillan.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the vocabulary levels test. *Language Testing*, *18*(1), 55-88.

Scott, M. (2012). *WordSmith Tools version 6* [Software]. Liverpool: Lexical Analysis Software.

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

Shin, D., Chon, Y. V., Lee, S., & Park, M. (2018). *COCA_MWU20 ColloGram* [Computer Software]. Seoul, South Korea: e-future. Retrieved from http://cfile281.uf.daum.net/attach/99EBA0495A80F110304D74

Shin, D., & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal*, *62*(4), 339-357.

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

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

Supasiraprapa, S. (2018). Second language collocation acquisition: Challenges for learners and pedagogical insights from empirical research. *The Journal of Asia TEFL*, *15*(3), 797-804.

Vilkaitė, L., & Schmitt, N. (2017). Reading collocations in an L2: Do collocation processing benefits extend to non-adjacent collocations? *Applied Linguistics*, *40*(2), 329-354.

Waring, R. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*(2), 130-163.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46-65.

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning*, *63*(1), 91–120.

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

Wodinsky, M., & Nation, P. (1988). Learning from graded readers. *Reading in a Foreign Language*, *5*(1), 155-161.

Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching*, *32*(4), 213-231.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, *21*(4), 463-489.

Appendix

Sample MUW Families

Basecollo1

35924 health care 29475
in health care 1416
health care system 2171
mental health care 403
health care providers 818
health care costs 1288
health care services 353

12138 hold on 5059
on hold 1559
hold on to 1969
holding on 460
holding on to 833
held on 2258

Basecollo3

2211 brown hair 1535
dark brown hair 194
light brown hair 217
curly brown hair 133
long brown hair 132

2199 minimum wage 2011
raise the minimum wage 83
minimum wages 105

Basecollo5

1352 just a matter of 790
just a matter of time 252
not just a matter of 195
just a matter of time before 115

1261 center stage 632
center stage in 82
take center stage 177
takes center stage 120
took center stage 97
to center stage 77
taking center stage 76