Beyond single words: the most frequent collocations in spoken English

Dongkwang Shin and Paul Nation

This study presents a list of the highest frequency collocations of spoken English based on carefully applied criteria. In the literature, more than forty terms have been used for designating multi-word units, which are generally not well defined. To avoid this confusion, six criteria are strictly applied. The ten million word BNC spoken section was used as the data source, and the 1,000 most frequent spoken word types from that corpus were all investigated as pivot words. The most striking finding was that there is a large number of collocations meeting the six criteria and a large number of these would qualify for inclusion in the most frequent 2,000 words of English, if no distinction was made between single words and collocations. Many of these collocations could be usefully taught in an elementary speaking course.

Introduction

With the growth of corpus linguistics there has been increasing interest in collocations. In addition, Lewis' influential Lexical Approach (1993) stressed the importance of learning collocations. In this work the author described several categories of multi-word units and a variety of activities for learning and teaching them.

However, there are two critical problems in the Lexical Approach. One is that Lewis did not indicate what to learn and teach first. If we follow Bahns (1993) in arguing that one of the critical problems in teaching lexical collocations is the huge number of collocations, the first thing we need to decide is what to focus on. Another problem with Lewis' suggestions is that his classification of multi-word units was not consistent with existing studies on the multi-word unit. For example, Lewis' classification includes 'collocations', 'polywords', 'fixed expressions' and 'semi-fixed expressions', but these categories refer to a very broad and overlapping range of word groups and there are difficulties in reliably assigning items to the categories.

Thus, the present study has the goal of strictly applying a well-defined set of criteria to arrive at a list of the most useful spoken collocations for elementary learners of English. To date no similar study has been carried out which consistently applies a clearly defined set of criteria to find collocations suitable for teaching in a beginners' spoken language course.

ELT Journal Volume 62/4 October 2008; doi:10.1093/elt/ccm091

Advance Access publication December 6, 2007

Studies by Kjellmer (1994), and Sinclair (1995) have not involved the manual analysis involved in separating different senses of an identical form or in choosing collocations which are meaningful units, which is needed to create a suitable list for teaching. Palmer's (1933) very helpful pioneering work on collocations did not use frequency data and did not involve clearly defined and consistently applied criteria. For example he mentioned non-compositionality as a criterion but did not then consistently apply it as many of his items, such as 'thank you' are clearly compositional.

Why should collocations be taught and learnt?

There are several reasons why teachers and learners should be interested in collocations. One reason is that collocations help learners' language use, both with the development of fluency and native-like selection.

Developing fluency

Pawley and Syder (1983) argue that there are hundreds of thousands of 'lexicalized sentence stems' that adult native speakers have at their disposal, and suggest that the second language learner might need a similar number for native-like fluency. That is, the chunked expressions enable learners to reduce cognitive effort, to save processing time, and to have language available for immediate use.

Native-like selection

Pawley and Syder (op. cit.) also argue that there is usually more than one possible way of saying something but only one or two of these ways sound natural to a native-speaker of the language. For example, 'let me off here' can also be expressed as 'halt the car'. The latter sentence is strictly grammatical, but the problem is that native speakers do not say it in that way. This unnatural language use is problematical for learners in EFL contexts where the focus is on grammar. They may produce grammatically correct sentences, but many of them may not sound native-like. For example, drawing on their first language, Korean students are likely to say 'lying story' for 'tall story', 'artificial teeth' for 'false teeth', 'thick tea' for 'strong tea', etc.

The present study assumes that learning collocations is an efficient way to improve the learner's language fluency and native-like selection of language use. In addition, it is assumed that the most frequent collocations will usually be the most useful because frequent collocations have greater chances of being met and used. However, in using data derived from the analysis of corpora, we need to bear in mind the cautions voiced by Cook (1998) and Widdowson (2000), noting that corpus data is necessarily limited, incomplete, subject to quantitative bias, heavily influenced by the nature of the corpus, and not always easy to interpret. It is only one type of information about language use and needs to be used with these cautions in mind. Despite these caveats, this study is based on the belief that there is value in attempting to discover the most useful collocations in a way that is explicit, and is thus capable of being replicated.

The study

The following research questions are addressed in this study.

1 What are the criteria needed to distinguish collocations from other word groups?

2 What are the most frequent collocations of English?
3 What are the most common collocational patterns?

Criteria used to search for collocations

In this study, as will be clear from the following criteria, *collocation* is used to refer to a group of two or more words that occur frequently together, and it is not restricted to two or three word sequences. A collocation is made up of two parts—a pivot word which is the focal word in the collocation and its collocate(s), the word or words accompanying the pivot word. For example, in the sequences 'high school', 'high court', 'high street', 'so high', and 'too high', 'high' is the pivot word and the other words such as 'school', 'court' and 'so' are the collocates of the pivot word 'high'.

In this study, six criteria were used to find collocations, and these criteria (as explained further below) largely involve how often two or more words occur together (criteria 3 and 4 below), the ability of the collocational group to stand as a comprehensible unit often as a part of a sentence (grammatical well-formedness) (criterion 5), and distinguishing different meanings of the same group of words (criteria 1 and 6).

Procedure

The corpus

It was decided to use spoken texts because Biber's (1989) study showed striking differences between written and spoken corpora, and there have been suggestions that collocation is likely to play a very important role in spoken language (Shin 2007). Thus, the ten million word spoken section of the British National Corpus (BNC) was used as the data source. The BNC spoken section is the biggest spoken corpus available. There are two almost equally sized parts to the ten million word spoken corpus. One is the demographic part, containing transcriptions of spontaneous natural conversations and the other is the context-governed part, containing transcriptions of recordings of more formal meetings and events.

The computer program

The program used for the search was WordSmith Tools 3.0 (Scott 1999). The program searches for all occurrences of the pivot word and creates a concordance. Because some collocations had several meanings ('come on', 'at the same time') each occurrence was checked manually.

The criteria

Here we will look at how the six criteria used in this study were actually applied.

1 Each pivot word was a word type. That is, the different word forms 'book' and 'books' were treated as different pivot words and investigated separately, rather than treating 'book' and 'books' as one word family. So 'walk', 'walks', 'walking', and 'walked' were each separately examined as different pivot words, as were 'big' and 'bigger'. A major justification for focusing on types rather than lemmas or families was that high frequency collocations need to be used productively as well as receptively and there is evidence that different types of the same word

family have different collocates. The word 'sigh' typically collocates with 'gave', not the other types of 'give'.

2 The pivot word had to be a noun, a verb, an adjective, or an adverb. Adverbial particles like 'up' as in 'get up' were treated as pivot words because they were adverbs. One of the goals of this study is to provide a list of collocations that can be used in teaching and for deliberate learning. These collocations thus need to be meaningful units. The five most frequent two-word groups in the Brown Corpus are 'of the', 'in the', 'to the', 'on the', and 'and the' and would not meet this goal. It was thus decided to search only for the collocates of content words.

3 All the pivot words had to occur in the most frequent 1,000 content words of English according to the spoken word frequency list by Leech, Rayson, and Wilson (2001), available at http://www.comp.lancs.ac.uk/ucrel/bncfreq/flists.html. Thus, in this study it was decided to focus only on high frequency words. The working assumption was that the learning of the collocations should strengthen and enrich words students already know, not add an additional burden by adding unknown, lower frequency vocabulary.

4 Each collocation had to occur at least thirty times in ten-million running words. Pilot testing showed that this frequency cut-off point would include several but not too many collocates for each high frequency pivot word. The results of this study are intended to be used with beginning and low intermediate learners of English whose vocabulary size is around 1,000 words. That is, the collocation should be frequent enough to get into the high frequency words of the language as if it were a single word.

5 Each collocation should not cross an immediate constituent boundary. A sentence can be divided into its principal parts, called 'immediate constituents' (Bloomfield 1933: 161). Immediate constituents are components that immediately make up larger parts of a sentence. To analyse a sentence in terms of its immediate constituents, it is divided into its largest word groups (or phrases), and then each of these parts is progressively divided and subdivided down to the ultimate constituents of the sentence which are morphemes. However, in a study of collocations, the minimal immediate constituent must be a two word group. For example, in sentence A:

$\{I_n\ [(saw_v\ you_n)_{vp}\ (at_{prep}\ (that_{det}\ place_n)_{np})_{pp}]_{pred}\}_s$

there are five immediate collocational constituents:

1 'I saw you at that place',
2 'saw you at that place',
3 'saw you',
4 'at that place', and
5 'that place'

'You at the place' however does not meet this criterion because it crosses an immediate constituent boundary. The single words are also immediate constituents but are of course not collocations.

6 Different senses of collocations with the same form were counted separately. So, 'looking up' meaning 'to improve' was counted separately from 'looking up' meaning 'to search for something, as in a phone book'. The division of entries in the *COBUILD English Dictionary* (1994) was used to distinguish the senses so that this could be done as consistently as possible.

Criterion 3 (frequency of the pivot word), and the high frequency level used in criterion 4 are not essential criteria for defining a collocation, but are directed towards making the resulting list particularly useful for beginning learners. Although a computer did a lot of the work, this study involved a great deal of manual checking and analysis. The *grammatical well-formedness* criterion and the distinguishing of different senses particularly involved much careful checking. In addition, when the data had been gathered, the results were checked against other studies to make sure that no important collocations had been overlooked.

Results

All the criteria used in the present study had to be easily replicable. The steps in applying the criteria had to be explicitly described and as much as possible had to involve a minimum of intuitive judgement. As a result, it is hoped that the findings of the present study provide reliable data that can be added to by further research, and provide a procedure that can be applied to other corpora. The appendix includes the top 100 spoken collocations.

In the present study, there are four major findings.

1 There are a very large number of grammatically well-formed high frequency collocations.
5,894 collocations were found using the first 1,000 content pivot words of English. There are 1,196 overlaps in the list. For example, the two different pivot words 'keep' and 'going' share 'keep going' as their collocation. So when the list is re-sorted by frequency removing duplicated items, the new list contains 4,698 collocations.

2 The more frequent the pivot word, the greater the number of collocates.
The most frequent 100 pivot words have 2,052 collocations which make up about 35 per cent of the total number of the collocations of the first 1,000 pivot words. The first 300 pivot words cover more than half of the total number (about 61 per cent). The first 100 pivot words have an average of 20.5 collocations, while the second 100 words have 8.4. After the second 100, the number of collocates gradually decreases as the frequency of the pivot words reduces. These results show that most frequent pivot words are likely to have more collocations and the majority of collocations are concentrated on the first 200 pivot words. However, this general rule has many individual exceptions.

3 A small number of pivot words account for a very large proportion of the tokens of collocations.
When the frequencies of all the collocations are added together, the total number of occurrences of the collocations of the first 100 pivot words is 387,634 which cover about 53 per cent of the total number of tokens of the

collocations (736,144) found in the present study. The first 200 pivot words make up about 68 per cent. This means that collocation use is heavily concentrated on the most frequent pivot words.

4 The shorter the collocation, the greater the frequency.
Two-word collocations make up 77 per cent of the total number of collocations. In addition, when analysing the top 100 collocations and 100 collocations from the bottom of the frequency list, the collocations containing a short word are more frequent than collocations containing a long word. The number of collocations decreases in inverse proportion to the number of characters of the longest component making up a collocation. However, in the bottom items, there are more collocations containing a relatively longer word compared with the top 100 collocations.

What is the nature of the most frequent collocations?

The collocation list (see Appendix) gives interesting insights into the nature of spoken language. The most frequent collocation is 'you know', which occurs 27,348 times in the 10 million running words. It is extremely frequent if we consider the third most frequent spoken collocation, 'a bit', has 7,766 occurrences. Collocations such as 'you know', 'a bit', and 'come on' are to be expected in the spoken English making up the corpus, because of its interactional nature. We can see that these sorts of interjections and amplifiers are much more frequent in the list than other collocations.

Shin (2007) shows two striking differences between written and spoken corpora. One is that the usage between spoken collocations and written collocations is considerably different. Only fifteen collocations occur in both the top 50 spoken and top 50 written lists. The here-and-now nature of spoken language is reflected in items like 'this morning', 'at the moment', and 'over there', and the personal and interactional nature is reflected in items like 'thank you', 'you know', and 'come in'. The written list contains items that can act as conjunctions like 'as well as', 'even if', and 'even though'. The other is that collocations occur in spoken language much more frequently than they occur in written language. Without exception, each collocation which appeared in both the top 50 spoken and top 50 written collocations has a much higher frequency in the spoken list than in the written list. See Shin (op. cit.) For most items at the same rank in the two lists, spoken items are 50 per cent to 100 per cent more frequent than the items at the same rank in the written list. For example, the most frequent spoken collocation 'you know' has 27,348 occurrences in the 10 million running words, while the most frequent written collocation 'of course' has 2,698. Thus, collocations are particularly important in spoken language and courses focusing on spoken language should give particular emphasis to them, perhaps more so than in written English.

Figure 1 is a frequency comparison between single word types from the BNC spoken section (Leech *et al.*'s list (2001)) and collocations to show how many collocations would meet the frequency cut-off points to get into the first four thousand words of English.

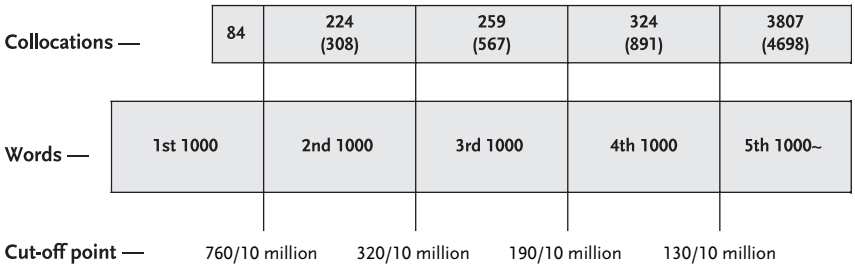

FIGURE 1
Frequency comparison between single word types and collocations

* The number in brackets shows the cumulative number of collocations.

For a single word type to get into the top 1,000 word types of the spoken corpus, it needs to have a frequency of 760 occurrences per 10 million. Figure 1 shows that 84 collocations meet this cut-off point of 760 occurrences per 10 million and thus these could be included in the level of the first 1,000 word types. Similarly, 308 collocations could be included in the most frequent 2,000 types, which have a frequency higher than 320 occurrences per 10 million. More than 500 collocations meet the frequency level of the first 3,000 word types. The 84 collocations of the first collocation band include 'you know', 'I think', 'come back', etc. The 224 collocations of the second collocation band include 'I see', 'I bet', 'at the end of the day', etc. The 324 of the third collocation band are 'on earth', 'nothing else', 'give up', etc. Relatively infrequent collocations beyond the frequency level of the fourth 1,000 word types are 'particularly good', 'go to church', 'from the bottom', etc. A large number of collocations meet the criteria used, and a reasonably large number of these qualify for inclusion in the most frequent 2,000 items in English if no distinction was made between single words and collocations. This is the most striking finding in this study because a collocation is always less frequent than the frequency of its less frequent member, so we would not expect many collocations to occur among the high frequency words of the language. There are in fact many, and these deserve the same kind of attention given to high frequency words.

Conclusions

There are some limitations that we should note. We determined usefulness considering the cost-benefit advantages that high frequency items provide even though frequency has to be balanced with other factors when choosing items to focus on in language teaching. Care needs to be taken when choosing items from this list, bearing in mind that the corpus it is derived from is spoken, British, largely adult, and a mixture of very colloquial and rather formal speech. Items 7('{No.} pounds'), 12 ('{No.} pound'), and 30 ('County Council') indicate the strongly British nature of the corpus, and these items can be ignored where English is taught outside the United Kingdom. Items 1 ('you know'), 75 ('mind you'), 85 ('I see'), 86 ('I bet') and 98 ('hang on') are strongly colloquial and may be best approached through identifying them in stretches of connected speech rather than teaching them for use. Other than these, the list is not very surprising, except perhaps that greetings like 'good morning', 'good afternoon', 'good evening', and 'How are you?' do not occur in the top 100, which indicates frequency is not everything. Although frequency in the

language is an important criterion for selecting what to focus on, it is only one of several important criteria like learner need, range of use (for example in both spoken and written use), difficulty, teachability, and suitability for the age and background of the learners. However having a list of the most frequent collocations in spoken English to choose from is a useful starting point for syllabus design.

Final revised version received June 2007

References

Bahns, J. 1993. 'Lexical collocations: a contrastive view'. *ELT Journal* 47/1: 56–63.
Biber, D. 1989. 'A typology of English texts'. *Linguistics* 27: 3–43.
Bloomfield, L. 1933. *Language*. London: George Allen and Unwin.
Collins COBUILD English Dictionary. 1994. (Third edition). Glasgow: Harper Collins Publishers.
Cook, G. 1998. 'The uses of reality: a reply to Ronald Carter'. *ELT Journal* 52/1: 57–63.
Kjellmer, G. 1994. *A Dictionary of English Collocations: Based on the Brown Corpus*. Oxford: Clarendon Press.
Leech, G., P. Rayson, and **A. Wilson.** 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Longman.
Lewis, M. 1993. *The Lexical Approach*. Hove: Language Teaching Publications.
Palmer, H. E. 1933. *Second Interim Report on English Collocations*. Tokyo: Kaitakusha.
Pawley, A. and **F. Syder.** 1983. 'Two puzzles for linguistic theory' in J. Richards and R. Schmidt (eds.). *Language and Communication*. London: Longman.
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.
Sinclair, J. M. (ed.). 1995. *COBUILD English Collocations* Version 1.1. London: Harper Collins Publishers.
Widdowson, H. G. 2000. 'On the limitations of linguistics applied'. *Applied Linguistics* 21: 3–25.

The authors

Dongkwang Shin received his PhD in Applied Linguistics in 2007 from Victoria University of Wellington, New Zealand. His expertise is in vocabulary learning and teaching, and corpus linguistics. He is currently working for the Korean Institute of Curriculum and Evaluation.
Email: sdhera@hotmail.com
Paul Nation is a professor of Applied Linguistics in the School of Linguistics and Applied Language Studies, at Victoria University of Wellington, New Zealand. He has taught in Indonesia, Thailand, the United States, Finland, and Japan. His specialist interests are language teaching methodology and vocabulary learning. His book *Learning Vocabulary in Another Language* was published by Cambridge University Press (2001) and there is a book called *Vocabulary Teaching: Strategies and Techniques* appearing in 2007 from Thomson Heinle publishers.
Email: paul.nation@vuw.ac.nz

Appendix
The first 100 collocations

RK	Collocations	FRE
1	*you know*	*27348*
2	*I think (that)*	*25862*
3	*a bit*	*7766*
4	*(always [155], never [87]) used to {INF}*	*7663*
5	*as well*	*5754*
6	*a lot of {N}*	*5750*
7	*{No.} pounds*	*5598*
8	*thank you*	*4789*
9	*{No.} years*	*4237*

10	*in fact*	*3009*
11	*very much*	*2818*
12	*{No.} pound*	*2719*
13	*talking about {sth}*	*2489*
14	*(about [91]) {No.} percent (of sth [580], in sth [54], on sth [44], for sth [38])*	*2312*
15	*I suppose (that)*	*2281*
16	*at the moment*	*2176*
17	*a little bit*	*1935*
18	*looking at {sth}*	*1849*
19	*this morning*	*1846*
20	*(not) any more*	*1793*
21	*come on*	*1778*
22	*number {No.}*	*1661*
23	*come in (swe, sth)*	*1571*
24	*come back*	*1547*
25	*have a look*	*1471*
26	*in terms of {sth}*	*1463*
27	*last year*	*1347*
28	*so much*	*1334*
29	*{No.} years ago*	*1314*
30	*{Det-the [879], this [39], a [21]} county council*	*1273*
31	*this year*	*1255*
32	*go back*	*1250*
33	*last night*	*1244*
34	*rather than*	*1243*
35	*come out*	*1163*
36	*very good*	*1160*
37	*I hope (that [455]) {N, S V}*	*1155*
38	*{No.} times*	*1147*
39	*that way*	*1145*
40	*said well (that, what) {S V}*	*1135*
41	*at the end (of sth [737])*	*1122*
42	*{Det-that [425], this [146], the [142]} sort of thing*	*1113*
43	*for example (if S V [30])*	*1107*
44	*as far as*	*1079*
45	*said to {smo}*	*1076*
46	*mean (that) {S V}*	*1066*
47	*come on (to swe, smo [65])*	*1059*
48	*{FREQUENCY, QUANTITY} a week*	*1056*
49	*all the time*	*1044*
50	*thank you very much*	*1041*
51	*too much*	*1034*
52	*over there*	*1017*
53	*that sort (of sth [953])*	*1016*
54	*looking for {sth}*	*990*
55	*make sure (that [394]) {S V}*	*990*
56	*very well*	*987*
57	*{Det-the [47]} last week*	*956*
58	*in the morning*	*952*
59	*it seems {N, A, to INF, that S V}*	*945*
60	*next week*	*940*
61	*a number of {sth}*	*929*
62	*out there*	*929*
63	*what I mean*	*929*

64	*get in (swe, sth)*	*912*
65	*find out {sth}*	*908*
66	*know that (S V)*	*889*
67	*leave it*	*886*
68	*at home*	*884*
69	*and so on*	*872*
70	*(about [226]) {No.} minutes*	*867*
71	*(do) n't mind (sth)*	*862*
72	*other people*	*839*
73	*not really*	*837*
74	*talking to {smo}*	*829*
75	*mind you*	*822*
76	*want it*	819
77	*much more*	816
78	*looked at {sth}*	805
79	*the other one*	805
80	*(at [207], about [110], till [50], by [24]) half past {No.1∼12}*	798
81	*some people*	797
82	*this week*	794
83	*this time*	787
84	*very nice*	784
85	*I see*	756
86	*I bet (S V)*	746
87	*these things*	742
88	*call it (A, N)*	737
89	*(be-verb) not sure*	721
90	*at the time*	717
91	*thought that {S V}*	714
92	*going out*	712
93	*it comes*	712
94	*go out*	711
95	*quite a lot*	711
96	*even if*	707
97	*last time*	704
98	*hang on*	701
99	*believe that (S V, N)*	696
100	*(be-verb, become-verb) interested in {sth}*	689

* { } signals an obligatory type of word that needs to occur in the collocation, ( ) signals an optional but a possible part of the collocation, and [ ] brackets the 'frequency figure'. RK refers to Rank and FRE to Frequency (the number of occurrences in the corpus).