A COLLOCATION INVENTORY

FOR BEGINNERS

by

Dongkwang Shin

A thesis
submitted to the Victoria University of Wellington
in fulfillment of the requirements for the degree of
Doctor of Philosophy
in Applied Linguistics

Victoria University of Wellington

2006

ABSTRACT

This study has two goals – (1) to see what criteria are needed to define collocations and (2) to make a list of the high frequency collocations of spoken English that would be useful for guiding teaching, learning and course design. The existing criteria for defining collocations are generally not well defined and have not been applied consistently. Wray and Perkins (2000) identify more than forty terms used for designating multi-word units. To avoid this confusion, three criteria are strictly applied – *frequent co-occurrence*, *grammatical well-formedness* and *predictability in L1*. The ten million word British National Corpus (BNC) spoken corpus is used as the data source, and the 1,000 most frequent spoken word types from that corpus are all investigated as pivot words. It is found that the three criteria can be applied in a systematic way.

The most striking finding is that there are a large number of collocations meeting the first two 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.

There are nine major findings in this study – 1) there is a very large number of grammatically well-formed high frequency collocations, 2) collocations occur in spoken language much more frequently than they occur in written language, 3) the more frequent the pivot word, the greater the number of collocates, 4) a small number of pivot words account for a very large proportion of the tokens of collocations, 5) adjectives tend to have more collocates than other content words, 6) the shorter the collocation, the greater

the frequency, 7) content word plus content word collocations outnumber other patterns of content word collocations, 8) there are more collocates on the left than collocates on the right, but this difference is not striking, 9) a third of the 500 most frequent collocations of English did not have word for word equivalents in Korean (L1).

A balanced approach is needed for the teaching and learning of collocations, employing opportunities for both deliberate and incidental learning, and giving appropriate attention in each of the four skills of listening, speaking, reading and writing.

ACKNOWLEDGEMENTS

I would like to sincerely express my appreciation to the following people for supporting me in many ways throughout my PhD study in the School of Linguistics and Applied Language Studies at Victoria University of Wellington.

- First of all, no words can express my gratitude to my supervisors Professor Paul Nation and Professor Laurie Bauer for their careful supervision including invaluable advice, guidance and encouragement, given with great patience.
- The School of Linguistics and Applied Language Studies, Victoria University of Wellington, for allowing me to use the data source (e.g. the Wellington corpora).
- Fellow PhD colleagues at Victoria University of Wellington who provided critical advice and moral support - in particular, Marianna, Julia, Agnes, Tina, and Laura.
- My former supervisor Professor Duk-ki Kim for encouragement.
- The Scholarships Committee of Research and Postgraduate Studies, Victoria University of Wellington, for supporting me with a PhD Completion Scholarship.
- My flat mate Petre Kusy for supporting me in a variety of ways.
- And my friends and family, in particular Jooyoung Lee, Notae Kim, Taesang Kwon, Youngmin Kwon, Heejin Kim, Bongkyeong Lee and Hyewon Lee for their concern and encouragement, and my parents,

brother and sister for their moral and financial support.

I greatly appreciate the support which has enabled the completion of this study.

CONTENTS

ABSTRACT i
ACKNOWLEDGEMENTS iii
LIST OF TABLES ix
LIST OF FIGURES xiii
CHAPTER 1 1
INTRODUCTION 1
1.1 Justification 1
CHAPTER 2 7
RELATED RESEARCH 7
2.1 What is a collocation and what are the criteria needed to determine a collocation? 7
2.1.1 Frequent co-occurrence 7
2.1.1.1 How frequent is a frequent collocation? 9
2.1.1.2 The frequency of collocations in speech and writing 12
2.1.2 Grammatical well-formedness 13
2.1.3 Mutual information 17
2.1.4 Predictability in L1 22
2.2 What are the sub-categories of collocations and what are the criteria to distinguish them? 26
2.2.1 Compositionality 27
2.2.2 Figurativeness 33
2.3 What sorts of collocation databases are currently available? 35
2.4 Research Questions 46
CHAPTER 3 48

RESEARCH PROCEDURE 48
3.1 The pilot study 48
3.1.1 Instruments 48
3.1.2 Procedure 48
3.2 The Main Study 62
3.2.1 Research procedure 62
3.2.1.1 Data source 62
3.2.1.2 Sample selection 63
3.2.1.4 Searching for collocations 64
3.2.1.5 Choosing a frequency level 65
3.2.1.6 Sorting collocates 69
3.2.1.7 Collocation and colligation 71
CHAPTER 4 74
RESULTS AND DISCUSSION 74
4.1 Number and frequency of collocations 74
4.1.1 Statistical results and comparison of 10 bands of 100 pivot words 75
4.1.2 Zipf's law on frequency distribution for collocations (bi-grams) 78
4.1.3 Inclusion of collocations in a list of high frequency words 87
4.2 Factors affecting the number and frequency of collocations 90
4.2.1 The frequency of the pivot word and the number of collocates 92
4.2.2 What proportion of the words making up the top 4,698 collocations of English are from the high frequency words? 94
4.2.3 Spoken collocations versus written collocations 99
4.2.4 Zipf's law of least effort for 2-word, 3-word, and 4-word collocations 104

4.2.5 Part of speech of the pivot word and number of collocates 108

4.2.6 Collocation patterns and the number of collocations 111

4.2.7 Part of speech of the collocations and the number of collocations 114

4.2.8 Location of the collocates and the number of collocations 119

4.3 Results of predictability in L1 121

4.3.1 The contrastive study 122

4.3.2 The survey 134

4.3.2.1 Participants 136
4.3.2.2 Procedure 136
4.3.2.3 Results and discussion 137

4.4 The transparency of collocations 140

4.4.1 Compositionality, figurativeness, and predictability in L1 140

CHAPTER 5 147
FINDINGS, CAUTIONS, AND FURTHER RESEARCH 147

5.1 Findings 147

5.2 Cautions 153

5.3 Further study 156

CHAPTER 6 159
IMPLICATIONS AND APPLICATIONS 159

6.1 Choosing what to focus on 159

6.1.1 Frequency level 159

6.2 Teaching and learning collocations 162

6.2.1 Deliberate learning and teaching 162

6.2.2 Incidental learning through meaning-focused input 168

6.2.3 Incidental learning through meaning-focused output 170

6.2.4 Fluency development 175

REFERENCES 178

APPENDICES 188

LIST OF TABLES

Table 2.1 Mutual information score versus absolute frequency..... 21

Table 2.2 The comparison of the four collocation studies..... 42

Table 3.1 Collocation span..... 49

Table 3.2 Some examples included and excluded by criterion 1..... 51

Table 3.3 Some examples included and excluded by criterion 2..... 52

Table 3.4 Some examples from the most interesting pairs to the least one by criterion 3..... 53

Table 3.5 The results of each step..... 55

Table 3.6 The collocations of *high* meeting all four criteria..... 56

Table 3.7 The results for the word *field*..... 58

Table 3.8 The comparison of a spoken and a written corpus..... 61

Table 3.9 The number of collocates of some lower frequency words from the spoken word list based on three cut-off points..... 67

Table 3.10 The number of collocates of some high frequency words from the spoken word list based on the three cut-off points. 67

Table 3.11 Some collocates of *up*..... 70

Table 4.1 The number and percentage of the collocations in the 10 frequency ranked bands of 100 pivot words..... 76

Table 4.2 The total tokens and percentage of the collocations of the 10 bands of 100 pivot words..... 77

Table 4.3 Some high and low frequency collocations..... 86

Table 4.4 Cut-off figures from the BNC according to spoken type-based single word figures..... 88

Table 4.5 Cut-off figures from the BNC according to spoken type-based collocation figures..... 88

Table 4.6 The number of collocates of the top 10 pivot words..... 92

Table 4.7 Members of the top 1,000 collocations..... 95

Table 4.8 21 academic words contained in the first 1,000 collocations.. 95

Table 4.9 12 collocations from the first 1,000 collocations list which are not in either the GSL or the AWL..... 96

Table 4.10 Comparison of the 2,000 word list from the BNC spoken corpora and the 2,000 word GSL..... 97

Table 4.11 93 word families contained in the top 4,698 collocations which are not in either the GSL or the AWL..... 98

Table 4.12 Tokens, types, and families of all the word members of the 4,698 collocations..... 99

Table 4.13 The 14 items occurring in both the spoken and written lists.. 101

Table 4.14 Some examples of frequency differences of the top 50 spoken and top 50 written collocations..... 103

Table 4.15 The total number of n-gram collocations..... 105

Table 4.16 The classification of the top 100 collocations based on the number of characters of the longest component of a collocation..... 106

Table 4.17 The classification of the bottom 100 collocations based on the number of characters of the longest component of a collocation..... 107
Table 4.18 Part of speech of the pivot word and the number of collocates..... 109
Table 4.19 Word combination patterns of the collocations of the first 1,000 content pivot words..... 112
Table 4.20 The number of collocations per part of speech..... 114
Table 4.21 Part of speech of pivot words vs. part of speech of collocations..... 116
Table 4.22 The number of collocations in relation to five collocation patterns and part of speech of the collocations..... 117
Table 4.23 Location of the collocates of the top 1,000 content pivot words..... 120
Table 4.24 Some examples for the criterion of *predictability in L1*..... 126
Table 4.25 The results of the analysis of *predictability in L1* of the first 500 collocations..... 130
Table 4.26 Background of the 20 subjects..... 136
Table 4.27 The number of correct answers on predictability in Korean 138
Table 4.28 The results of the three participants who have studied Korean..... 139
Table 4.29 Types of collocational groups..... 144
Table 4.30 The top eight core idioms meeting the frequent occurrence criterion..... 145

Table 6.1 Comparison of the average frequencies per 10,000,000 tokens of a collocation of the 10 bands of 100 pivot words.... 160

Table 6.2 Collocations of some lower frequency pivot words..... 161

Table 6.3 The contrast of the restricted selections of some *dress* verbs between English and Korean..... 166

Table 6.4 Some examples of collocations that are difficult to be predicted in Korean..... 173

LIST OF FIGURES

Figure 2.1 The structure of the phrase *woke your friend up*..... 38

Figure 3.1 A concordance of the pivot word *money*..... 64

Figure 4.1 Zipf curve for the uni-grams (single words) extracted from a 250,000 word token corpus..... 79

Figure 4.2 Zipf curve for collocations of *up*..... 82

Figure 4.3 Zipf curves for the *WSJ87* corpus..... 82

Figure 4.4 Zipf curves for four 1,000 collocation bands..... 83

Figure 4.5 Zipf curves for the most frequent 4,000 collocations..... 85

Figure 4.6 Frequency comparison between single word types and collocations..... 89

Figure 4.7 Frequency comparison between the top 50 spoken and written collocations..... 103

Figure 4.8 The classification and terminology used for multi-word units..... 141

CHAPTER 1

INTRODUCTION

1.1 Justification

The aims of this thesis are to provide a set of reliable and replicable criteria for defining a collocation and to provide a list of the most useful English collocations for beginners learning English. Collocations are sequences of words that go together to make up a sequence such as *you know, I think, pick {smo} up,* and *come back.* 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'. Collocations also provide a useful way of helping learners learn new vocabulary.

Using collocations can develop learners' language fluency. Pawley and Syder (1983), talking about larger units, argue that there are hundreds of thousands of 'lexicalised 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. They looked for 'fluent units' in native speakers' language use. A 'fluent unit' is a stretch of pause-free speech uttered at or faster than the normal speed of articulation (about five syllables per second in English). The typical fluent native speaker makes a pause after every two or three

words. This indicates that there may be memorised sequences in their speech and also raises the possibility of frequent use as memorised sequences. These memorised sequences may be analysed or unanalysed. Sinclair (1991) describes this phenomenon by proposing the idiom principle which suggests that words are 'glued' to other words. Sinclair points out that a large number of semi-preconstructed phrases may constitute single choices. They might be analysable into segments but in fact may be available to a language user as a single unit. Each of these units is thus treated as a kind of idiom and the idiom principle may thus be explained by a natural tendency to gain efficiency in language use, and to deal with the exigencies of real time conversation (Sinclair, 1991, p. 110). Skehan (1996, 1998) argues that when required to perform spontaneously, L2 learners are likely to draw fixed and formulaic expressions from their mental lexicon. The chunked expressions enable learners to reduce cognitive effort, to save processing time and to have language available for immediate use (de Glopper, 2002; Nation, 2001a).

In addition to fluency development, collocations help learners' "native-like selection" (Pawley and Syder, 1983, p. 191). There is usually more than one possible way of saying something but only one or two of these ways sounds 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 English as a Foreign Language (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. This is because the learners have relatively few chances to repeatedly encounter typical English word sequences and so they are more likely to translate from their first language. Using native-like collocations makes learners' speaking and writing seem native-like. Bahns and Eldaw (1993) suggest that learners are more than twice as likely to adopt an unacceptable collocate as they are to select an unacceptable word, and EFL learners' general vocabulary knowledge far surpasses their collocation knowledge. Verstraten (1992) argues that it is far more difficult to produce in a second language than to comprehend in the same language. For production, learners have to be able to select native-like words and to use them according to the rules. If learners lack just one feature, their production may not be native-like. This also partially explains Marton's (1977) findings that even though learners can understand and translate English sentences, they cannot produce those same sentences in English. This is where multi-word units such as collocations can play an important role. Bahns and Eldaw (1993) suggest that learners' collocation knowledge does not expand in parallel with their general vocabulary knowledge, and collocation knowledge is necessary for full communicative mastery of English.

There is evidence that collocations containing known words are easier to learn than new words, making the best use of what is already known is not as difficult as learning completely new items.

Bogaards (2001) compares learning multi-word units with single words. Idioms made up of known words were given to learners (e.g. *du jour au lendemain* consisting of high frequency French words literally corresponds to *from the day to the next* but it can be translated as *unexpectedly*) and they were compared with the other group who was given completely new single words (e.g. *inopinément* is a low frequency French word which has the same meaning as *du jour au lendemain*). He examines the effect of these two types of lexical units in the learning of French as a foreign language by native speakers of Dutch. The result shows that knowledge of form turned out to play a positive role in the learning of lexical units. Completely new single words are harder to learn and retain than multi-word units of the same meaning but with a form that is made up of familiar words. Ellis (2001) also argues that a lot of language learning can be achieved by emphasising associations between sequentially occurring language items. By having collocational knowledge in long-term memory, language reception and production can be made more effective.

It is clear then that there are good arguments for giving attention to collocations. To do this however we need to know what the most useful collocations are.

In the 1970s, Taylor (1983) carried out research on school textbooks to find useful vocabulary for native-speaking learners. The words in the lists were arranged in clusters according to the linguistic notion of collocation (Halliday, McIntosh, & Strevens, 1964). Taylor (1983) worked on the hypothesis that words are not easily learned in isolation. In traditional language learning, vocabulary was memorised from lists, although, as Taylor points out, to some extent, the single words in the lists might be taught with their collocates, the words they typically go with, but the compiling of those word clusters was still based on impressions of usefulness rather than on systematic examination. Taylor wanted his research to provide more reliable data. If selecting useful collocations depends only on teachers' intuition, external reliability will be considerably weakened because different teachers' intuitions are likely to be different. A more reliable method is needed to select useful collocations. Taylor's research at the time did not have access to large corpora and the computer programs that are available today.

The present study assumes that learning collocations is an efficient way to improve the learner's language fluency, native-like selection of language use, and vocabulary retention. In addition to this, 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. This study attempts to discover the most useful collocations in a

way that is both reliable and valid, and is thus capable of being replicated.

CHAPTER 2

RELATED RESEARCH

2.1 What is a collocation and what are the criteria needed to determine a collocation?

Some have used the term 'collocation' and other related terms but often have not provided a clear operational definition. Becker (1975), for example, proposes six categories of multi-word units including 'polywords', 'phrasal constraints' and 'situational builders', but does not provide tests or criteria to distinguish these categories. Lewis (1993) divides the classification into more detailed categories such as 'collocations', 'polywords', 'fixed expressions' and 'semi-fixed expressions'. However, Lewis' definitions include too broad a range of word groups and there are difficulties in reliably assigning items to the categories.

The two criteria of *frequent occurrence* and *grammatical well-formedness* are often mentioned to define 'collocation'. We will now look at these in detail.

2.1.1 Frequent co-occurrence

Palmer (1933) was one of the earliest researchers to use the term

'collocation' and he provided a list of 5,749 collocations. Palmer's notion of a collocation was "a succession of two or more words that must be learned as an integral whole and not pieced together from its component parts" (Palmer, 1933, p. i) and the term is thus used in a phraseological rather than in a frequency-based sense. That is, he mainly focused on types of word combinations, not on the number of co-occurrences of words. This is not surprising as in Palmer's time there was no large corpus or electronic tool to get frequency data. Firth (1957) on the other hand, restricted the notion of collocation to the 'habitual co-occurrence' of lexical items, and collocation is also defined by other researchers as the tendency of a lexical item to co-occur with one or more words (Cruse, 1986; Crystal, 1985; Halliday et al., 1964; Seaton, 1982). The Firthian term "habitual co-occurrence" (Firth, 1957, p. 181) involves the notion of 'frequency' which can be looked at in absolute or relative terms. Absolute frequency is the actual number of occurrences of a collocation in a corpus, for example, suppose the collocation *cause trouble* occurs 10 times in a 1,000,000 word corpus, then its actual frequency is 10 per 1,000,000. On the other hand, relative frequency compares the actual number of occurrences with an expected number of occurrences, in other words, it compares the frequency of co-occurrence of the pivot word and collocate with the frequency of their independent occurrences. Church and Hanks (1990, p. 22) called this measure "mutual information", and it is a measure of the strength of the relationship between the pivot word and its collocate (see 2.1.1.3 below).

Mutual information value is highly affected by the low frequency of one of the items. For example, the frequency of the word *tousled* is very low but the relative proportion of mutual occurrences of *tousled* and *hair* is very large. That is, most of the occurrences of *tousled* are with *hair*. In contrast, in the case of a collocation made of only high frequency words, the mutual information value tends to be low because each item will also occur with many other words. Therefore, when we focus on high frequency collocations whose components are also high frequency words, the mutual information index may not be very revealing.

2.1.1.1 How frequent is a frequent collocation?

If *frequent co-occurrence* is an important criterion for defining collocations, how frequent is frequent? The answer to this question depends partly on the size of the corpus used in the study. Consider one example. Kjellmer (1982, 1984, 1987) decided in his research that "a collocation is a sequence of words that occurs more than once in identical form (in the Brown Corpus)..." (Kjellmer, 1987, p. 133). That is, Kjellmer used a frequency of two occurrences or more as the qualifying frequency level of collocations in the 1,000,000 word Brown Corpus. Kjellmer used frequency level to measure the degree of lexical determination of collocations, that is, he considered 'lexicalised' items are frequently used together regardless of their grammatical well-formedness. For example, *although he* and *hall to* are lexically

determined sequences (that is, they recur in the corpus), by contrast, *yesterday evening* and *green ideas* (which do not recur) are considered as grammatically-determined, in other words, the two items *yesterday evening* and *green ideas* are grammatically well-formed but they do not recur in the Brown Corpus so they are not lexicalised according to Kjellmer. If we want to extract a specific number of collocations, we should determine the qualifying frequency level in proportion to the corpus size used. A collocation is likely to occur more times in a larger corpus. Kjellmer (1982, 1984, 1987) was forced to use the very low frequency level of two co-occurrences because he was working with a relatively small 1,000,000 word corpus. With a larger corpus a higher frequency cut-off could be used because there would be many more occurrences of items. However, Kjellmer used a small-sized corpus, so some potentially recurring sequences were excluded. *Yesterday evening* is easily recognisable as a sequence but it only occurred once in the corpus, so it could not be classified in that study as a recurring sequence. It may be assumed that the most frequent collocations consistently occur regardless of the size of the corpus used. However, if we use a small corpus, we cannot expect the same results (see Table 3.7). In addition, the frequency of a collocation is influenced by topics and registers of the texts used, so we need to use a large corpus including a variety of genres.

Biber, Johansson, Conrad, Leech, and Finegan (1999) used the frequency cut-off point of 400 occurrences or more in a 40,000,000

word corpus (10 times per million). Cortes (2002) using a corpus of 360,704 words set the frequency level at 20 times per million words and set the range criterion of occurrence at occurring in 5 or more texts. Cortes found a total of 93 "lexical bundles" (Biber et al. and Cortes considered these *extended collocations*). Both Biber et al. and Cortes adopted *frequent occurrence* as a criterion. Their frequency cut-off points seem to be very high, but they found a considerable number of lexical bundles. This suggests that it is necessary to use an additional criterion for reducing the number of items to focus on like *grammatical well-formedness*.

However, there are studies that deny the need for frequency as a criterion. Wray (2000) argues that frequency is not a reliable criterion for determining the formulaicity of multi-word units. For this, Wray referred to Moon's (1998a) study. Moon used the 18,000,000 word Hector corpus to examine occurrences of the 6,700 phrases which are listed in *Collins COBUILD English language dictionary,* and found 70% of the phrases have a frequency of less than one in a million or do not occur at all. For example, *bag and baggage, hang fire, kick the bucket, lose your rag*, etc were not found in the corpus. In addition, Wray claimed that formulaic and non-formulaic sequences may look identical, so mechanical frequency counts may not be a reliable way of differentiating them. Moreover, there are a lot of sequences based on "the open choice principle" (Sinclair, 1991, p. 109) where there is a large range of choices and the only constraint is grammaticalness. The

basis of these objections however is that frequency is not necessarily related to formulaicity. If the goal is not to determine formulaicity, but simply to determine usefulness, then we need to consider the cost-benefit advantages that high frequency items provide. High frequency items have more chances to be used and met. Where students learn English as a foreign language in countries such as China, Japan and Korea, students get most of their English input in the classroom, which means there is very limited time for learning and so the best use has to be made of this time. When teachers need to determine what collocations to focus on, frequency can be a useful guide.

2.1.1.2 The frequency of collocations in speech and writing

There are clear frequency differences between speech and writing, especially when lexical use is compared. Altenberg (1994) compares distributions of the high frequency function word *such* in two corpora. The use of *such* in the formal written sections of the Lancaster-Oslo/Bergen (LOB) corpus is more than three times as common as in the more informal spoken texts of the London-Lund Corpus (LLC). However, the word *such* is mainly used as an identifier in the LOB corpus (e.g. *never had such a thing been thought of*), while the use of *such* as an intensifier is dominant in the spoken LLC (e.g. *it's such a bore*). Another example is the use of *pretty*, which occurs predominantly as an intensifier in the spoken corpus (e.g. *pretty horrible weather, pretty*

clearly seen), while almost half of the occurrences of *pretty* are as an adjective in the written corpus (e.g. *pretty girl, pretty picture*). This difference between spoken and written language is the result of a variety of factors including the real time constraints on speech and its interactive nature. However, there are few studies of multi-word units based on spoken corpora. Altenberg (1998) contends that recurrent multi-word units are especially frequent in spoken language. Biber et al. (1999) also find that the number of lexical bundles is greater in conversation than in academic prose. Clearly a study of collocation must decide if the corpus is to be spoken, written or a combination of these and this may have a strong effect on the findings.

2.1.2 Grammatical well-formedness

The second criterion that can be used to identify collocations is *grammatical well-formedness*. If we consider the deliberate teaching and learning of collocations, it makes sense to deal with meaningful units such as Object, Complement, or Preposition+Noun. If this is not done, the very frequent word pairings like *of the,* and *in the* would be classified as collocations because they meet the frequency of co-occurrence criterion. However, in several pieces of research, the criterion of *grammatical well-formedness* was not used. Biber, Conrad, and Cortes' (2004) four-item lexical bundles include a lot of incomplete units such as *what do you think*, *going to be a*, and *I don't know what*. If

there is a focus on teaching and learning, collocations may need to be independent well-formed units.

We also need to consider discontinuous sequences as collocations. This is not incompatible with well-formedness. Palmer (1933) who defined a collocation as "a succession of two or more words" seemed to restrict collocation to continuous sequences. Kjellmer (1984) more explicitly described the criterion of *grammatical well-formedness*. For example, Kjellmer excluded *although he* and *hall to* because even though they recurred in the corpus he used, they were not "grammatically independent" (Kjellmer, 1984 p. 163). However, Kjellmer only dealt with immediately adjacent collocations, so Kjellmer ignored a lot of discontinuous collocational patterns. On the other hand, Renouf and Sinclair (1991) included discontinuous collocational frameworks and also examined replaceable words that occurred in a gap in the discontinuous sequences. For example, the framework *a+ ?+ of* collocates with *man, part, kind* and so on. Moreover, in the written corpus that Sinclair examined, over half of the occurrences of the words *couple, series, pair* and *lot* occurred in that frame. Such discontinuous collocational patterning is common in English. In the present study,"package nouns" (Biber, Conrad, & Leech, 2002, p. 60) including collective nouns (e.g. *a group of, a set of,* etc), unit nouns (e.g. *a bit of, a piece of,* etc), quantifying nouns (e.g. *a cup of, a box of,* etc) and species nouns (e.g. *the sort of, all kinds of,* etc) do not meet the criterion of *grammatical well-formedness,* but these items are included. If we consider these

items as *Determiner+ Noun+ Of* types, all these items are incomplete, so they cannot function as immediate constituents of a clause or sentence. However, if we see these items as a whole unit, they act as quantifiers which modify a following noun.

The issue of discontinuousness is particularly noticeable in relation to phrasal verbs. For example, the two components *pick* and *up* of the phrasal verb *pick up* could be interrupted by other words such as *him, her*, and *you*. Even though the item *pick {smo} up* is a discontinuous form, it could be grammatically and semantically considered as one unit.

If we follow Bahns (1993, p. 56) in arguing that one of the critical problems in teaching lexical collocations is the huge number of collocations, then the criterion of *grammatical well-formedness* can also be used to reduce considerably the number of collocations to focus on. *Grammatical well-formedness* ensures the multi-word unit is a complete cohesive unit. If multi-word units are required to be immediate constituents of a sentence, then all potential collocations must be able to fill one of the following nine positions - Subject, Predicate, Verb, Object, Adverb, Complement, Conjunction, Preposition and Sentence (or Clause). A sentence can be divided into its principal parts, called "immediate constituents" (Bloomfield, 1933, p.161). Immediate constituents are components that immediately make up larger parts of a sentence. By analysing a sentence in terms of its immediate constituents - word groups (or phrases), each of these parts are then divided and subdivided down to the ultimate constituents of the sentence. In a study of

collocations, the minimal immediate constituent must be a two-word group. The following is an example of immediate constituent analysis.

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

In the sentence A, there are five immediate 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.

Let us look at another example. The phrases such as *{Det-the, a…} high school, {Det-the, a…} high level* can be a subject or an object, and *No. feet high, very high* can be a complement. On the other hand, excluded items like *high in, high from, between high* and *those high* need a noun to meet the above criterion. *Give high* is Verb+ Adjective but the verb *give* is a transitive verb that needs an object, so *give high* will be excluded. If learners are to study collocations as units, they need to make sense as units. The test for *grammatical well-formedness* is to see if the unit acted one or more of the nine constituents listed above. To be well-formed, no sequence of words must cross an immediate constituent boundary unless it continues until the end of that next immediate constituent.

In the following sections, we will look at other possible criteria that have been used to define a collocation.

2.1.3 Mutual information

Another possible criterion to distinguish collocations is *mutual information* which is used to measure the strength of co-occurrence between components of a collocation, that is, *mutual information* is used to measure if the relative proportion of mutual occurrences of some words is large compared with their total frequencies. When this happens, the mutual information value would be high. Let us take *high court* as an example. In a corpus of 4,758,223 tokens (N), there are 35 instances of *high court.* The joint probability of *high* and *court* is 7.4 (calculated by dividing the number of co-occurrences by the size of the corpus - 35/4.7) per million. *Mutual information* compares this probability and chance: the probability of *high* times the probability of *court.* There are 1,784 instances of *high* and 716 instances of *court* in the corpus and so the chance of co-occurrence is (1,784/N)*(716/N) ≈ 0.056 per million. Thus, when we compare the actual co-occurrence (7.4 per million) to chance (0.056), we can see that the actual number is much larger than chance (7.4/0.056 ≈ 132), therefore, *high court* is probably an interesting collocation. The mutual information index of two words, x and y, I(x,y) is defined as a formula:

$$I(x,y)=\log_2[f(x,y)N/(f(x)f(y))]$$

If f(x)=the number of x, f(y)=the number of y, f(x,y)=the number of co-occurrences of x and y, N=the number of the corpus tokens, in general, the observed frequency O, relative to corpus size N, is:

(1) $O = f(x,y) / N$

The expected frequency E of co-occurrence, relative to corpus size N, is:

(2) $E = [f(x)/N] * [f(y)/N] = f(x)f(y) / N^2$

To calculate how much higher than chance the frequency of a collocation is, O/E is calculated by:

(3) $O/E = [f(x,y)/N] / \{[f(x)/N] * [f(y)/N]\}$

$= f(x,y) / [f(x)f(y)/N]$

$= [f(x,y)N] / [f(x)f(y)]$

Fano (1961, p. 28) and Church et al. (1991, p. 120) propose this calculation:

(4) $I(x,y) = \log_2 \{[f(x,y)N] / [f(x)f(y)]\}$

As we have seen, this is simply equivalent to:

$I(x,y) = \log_2 O/E$, from (3)

So, if there is a interesting association between x and y, the restricted occurrence between them, f(x,y)/N will be much larger than f(x)f(y), and then, I(x,y)>0. The logarithmic value is used for historical reasons, which has no real significance here. Its only effect is to reduce, and therefore possibly to disguise, the differences between scores on different collocates. However, SPSS 10.0.5 for Windows (SPSS Inc,

1999) cannot provide the value of base 2 logarithms, so we need to change the formula using the equations $X=log_2Y \rightarrow 2^X=Y$. To change log_2 to *log*, the following procedure is used.

$I(x,y)=log_2[f(x,y)N/(f(x)f(y))]$

$\rightarrow 2^{I(x,y)}= [f(x,y)N/(f(x)f(y))]$

(consider $X=log_2Y \rightarrow 2^X=Y$*)*

$\rightarrow log2^{I(x,y)}=log[f(x,y)N/(f(x)f(y))]$

(consider $X=Y \rightarrow logX=logY$*)*

$\rightarrow I(x,y)log2=log\ [f(x,y)N/(f(x)f(y))]$

(consider $logX^Y=YlogX$*)*

$\rightarrow I(x,y)=log[f(x,y)N/(f(x)f(y))]/log2$

If there is no interesting relationship between x and y, I(x,y) becomes close to zero. If x and y rarely collocate with each other, I(x,y) will be a minus value. However, the problem of *mutual information* is that it only shows relative collocation strength. This means it is difficult to define what an 'interesting' mutual information value is. However, in general, a mutual information score greater than 2 is considered high enough to show an interesting association between two words (Kennedy, 2003).

Kennedy (2003) examines what particular words twenty-four selected amplifiers such as *absolutely, really*, and *very* collocate with using the mutual information measure. Kennedy's study shows *mutual*

information can be used for measuring 'colligation' (collocational frameworks in which units are based on grammaticality and the patterns and words are fixed grammatically or lexically) or 'semantic prosody' (certain words and phrases being associated, through repeated use, with negative, positive, neutral contexts, etc). For example, *perfectly* has exclusively positive associations and it is likely to occur with adjectives ended in *-able* or *-ible* like *perfectly possible*. On other hand, *totally* tends to have mainly negative associations and it mainly occurs with adjectives containing a negative prefix and the suffix *-ed* like *totally unsuited*. However, the mutual information measure does not seem effective in searching for high frequency collocations made of high frequency words. Table 2.1 shows the different effects of using relative frequency (*mutual information*) and absolute frequency.

Table 2.1 uses some data from Kennedy (2003) and additional data from the British National Corpus (BNC). Table 2.1 gives the two sets of the top three collocates of the three amplifiers *completely, entirely* and *perfectly*. One set is based on the mutual information measure, and the other on the number of co-occurrences. The amplifier *completely* has the strongest association with the word *refitted*. The word *refitted* has 35 occurrences in the BNC and the co-occurrences of *completely* and *refitted* are 5. That is, 14% of the total number of occurrences of *refitted* collocate with *completely*.

Table 2.1

Mutual information score versus absolute frequency

Amplifiers	The top 3 collocates based on *Mutual Information* (MI)		The top 3 collocates based solely on the number of co-occurrences	
	Collocates	MI	Collocates	MI
completely	refitted (5)	10.74	different (509)	7.00
	inelastic (8)	10.28	new (261)	4.66
	outclassed (3)	9.11	free (71)	5.38
entirely	blameless (8)	10.53	new (259)	4.95
	coincidental (5)	9.53	different (257)	6.32
	fortuitous (6)	9.52	Clear (76)	5.50
perfectly	contestable (17)	12.75	well (353)	5.84
	proportioned (12)	11.42	clear (116)	6.75
	manicured (6)	10.80	normal (113)	7.74

- The number in brackets shows the number of co-occurrences of each collocation in the BNC.

On the other hand, the most frequent collocate of *completely* is *different* and the two words *completely* and *different* occur together 509 times in the BNC. However, the relative proportion of the co-occurrences of *completely* and *different* is very small. Only 1.07% of the total number (47,607) of occurrences of *different* collocates with *completely* and so the mutual information index of *completely different* (7.00) is lower than *completely refitted* (10.74). However, the collocation *completely different* made of the two high frequency words *completely* and *different* has a high mutual information index like the other high frequency collocations in Table 2.1 (e.g. *completely free* (5.38), *entirely different* (6.32), *perfectly normal* (7.74), etc).

The mutual information value is highly affected by the low frequency of one of the items. So, when we focus only on high frequency collocations whose components are also high frequency words, the mutual information index is not very revealing.

2.1.4 Predictability in L1

A: "Wow, you are really big!"

B: "Huh, what? Big? Am I big?"

A: "Oh, sorry, I mean…tall, you are tall."

B: "I see, well…but I was really upset, you know!"

A: "Sorry, again."

In this conversation, A was a Korean male student who was taking a language course and B was a native New Zealand female student. We can see there was a misunderstanding in their conversation. A used the inappropriate word *big* which conveys the meaning of *being fat* because A confused *tall* with *big*. In Korean the two words *tall* and *big* are translated as the same word 크다 (keuda). Unfortunately, B was a little fat, so A's statement upset B. B could work out what A meant to say through meaning negotiation. Nevertheless, if B had not considered A was a foreign student with limited language proficiency, A could have been considered very impolite.

Some consider that contrastive analysis is "dead meat" (Gregg,

1995, p. 90). However the first language can support or hinder learners in learning and using vocabulary in a second language. The conversation shown above is a good example of how the first language affects language use. The first language also has an effect on what makes it hard or easy for learners to learn L2 vocabulary. There is a variety of factors which could affect learning L2 vocabulary such as word form (spoken and written), word structure (derivations and inflections), syntactic patterns (word patterns in a phrase and sentence), and semantic features of the word. Laufer (1997, pp 149-153) points out the following four semantic features of words:

(1) **Abstractness**: Abstract words are likely to be more difficult to learn than concrete words.

(2) **Specificity and register restriction**: Lexical items frequent in one field or mode of discourse may not be normal in another.

(3) **Idiomaticity**: Idiomatic expressions are much more difficult to understand and learn to use than non-idiomatic meaning equivalents.

(4) **Multiple meanings**: One form can have several meanings and one meaning can be represented by different forms.

Collocational sequences with the same form but different senses may need to be classified as different items.

In a cross-linguistic study, the semantic feature of multiple

meanings could be reinterpreted as a result of comparision between the two languages involved. First, if we compare L2 with L1, we should consider whether a L2 collocation has a L1 equivalent because a L2 collocational combination could be different in L1. For example, the English collocation *strong coffee* is translated as *thick coffee* in Korean. However, in addition, one L2 form can have several meanings in L1 and one L1 meaning can be represented by different forms in L2. For example, the Korean expression 그밖에 누군가 (geubagge nugunga) can be expressed as *anyone else* and *someone else* in English.

Newman (1988) examined the range and register restriction of *dress* and *cooking* collocations between Hebrew (L1) and English (L2). Newman adopted an open/close dichotomy which corresponds to Sinclair's (1991) idiom/open choice principle. For example, Newman classifies four types of *dress* verbs in Hebrew. Type 1 is the general verb *sam* which corresponds to the English *put on*. *Sam* is usually used in an informal register. Type 2 is *lavash* which is close to the English *wear*, but *lavash* can convey the meaning of both action and state, while *wear* only relates to state. Type 3 is a series of three action/state verbs. *Xavash* is restricted to headgear, *na'al* to footwear, and *'anad* to jewellery. These verbs are associated with different body parts. Type 4 is also a series of three verbs whose uses are totally restricted and idiomatic. *Garav* is used for socks, *'anav* for tie, and *xagar* for belt. Newman suggests meaningful exercises for open collocations with the transparent meaning involving comparison of L1 and L2, while it is

suggested directly memorising close collocations which are not transferable to another language. However, all the collocations shown in the study were selected samples aimed at showing a striking difference between Hebrew and English. In addition, Newman did not consider the direction of learning of collocations. According to the direction of learning, "split" or "coalescence" (Prator, 1967) of a collocation between L1 and L2 could be problematic. If focusing on production of L2, *coalescence* in L2 of a L1 collocation would not cause difficulty even if the reverse direction could be a problem. For example, Newman showed English dress verbs such as *wear* and *put on* are split into four types in Hebrew according to different functions (for headgear, footwear, jewellery, etc) and then each type is also subdivided into some different dress verbs. However, if focusing on production of English as a L2, the coalescence of the English dress verbs in Hebrew (L1) would not be too problematic.

Bahns (1993) proposes a more practical method to select some English collocations which do not match with their word for word German translations targeting the German learner of English. Some English collocations were selected and then were translated word for word into German according to the meaning listed in the dictionary. Bahns focused on the aspect of the semantic difference of noun plus verb combinations, however grammatical differences such as word order were not considered. The results show three different types of collocations. Type 1 is a group of English collocations which are directly

transferable into German. For example, the English collocation *seek+ shelter* (=suchen+ Schutz) is translated into *Schutz+ suchen* in German which exactly corresponds to each component of the English collocation. Type 2 is a group of English collocations whose German literal translation does not make sense. For example, the English collocation *(with) draw+ money* is *Geld+ abhaben* in German. However, the English literal translation of *Geld+ abhaben* is *lift+ money* which does not make sense. Type 3 is a group of English collocations whose German literal translation makes sense but the literal translation is different from the meaning of the respective English collocation. For example, the English collocation *lay+ table* should be translated into *Tisch+ decken* in German, but the English literal translation of *Tisch+ decken* is *cover+ table. Cover+ table* also makes sense but the meaning of *cover+ table* in English is different from *lay+ table*. Therefore, types 2 and 3 are unpredictable collocational groups in German. Bahns' study is restricted to a small number of examples and to the verb plus noun combination which shows a striking difference between L1 and L2. However, this study provides a good model to distinguish unpredictable L1 collocations from other collocational groups.

2.2 What are the sub-categories of collocations and what are the criteria to distinguish them?

When considering how the parts of a collocation relate to the

meaning of the whole, collocations can be subdivided into core idioms, figuratives, and literals (Grant & Nation, 2006). These three categories can be distinguished using two criteria, compositionality and figurativeness. Core idioms are non-compositional and non-figurative, for example, *kick the bucket*. Figuratives are non-compositional and figurative. It is possible to see a connection between the literal meaning and its figurative meaning, for example, *jump the gun*. Literals are compositional and non-figurative. That is, the meaning of the parts are related to the meaning of the whole, for example, *thank you*. Let us now look at the criteria of compositionality and figurativeness which are essential for understanding these categories.

2.2.1 Compositionality

Compositionality relates to the degree of semantic opaqueness or transparency of a multi-word unit. If the meaning of a multi-word unit can be deduced from its parts, the multi-word unit is compositional. Grant and Bauer (2004, p. 48) provide a simple way to determine whether a multi-word unit is compositional or non-compositional. In the interest of creating a practical system usable by researchers, if replacing each component in a multi-word unit with its dictionary definition gives the same meaning as the phrase in context, the multi-word unit is compositional. If it does not, the multi-word unit is non-compositional.

Palmer (1933, p. i) defined a collocation as "a succession of two or more words that must be learned as an integral whole and not pieced

together from its component parts". Palmer's definition seems to put non-compositionality as the main criterion. However, if we look at the list Palmer provides, we can see many multi-word units that do not meet this criterion, such as *thank you, to agree with someone* and *in a week*. This failure to stick to criteria has been a major problem in the study of multi-word units. Grant and Bauer (2004), however, are an exception. In their study of idioms, they used non-compositionality and non-figurativeness as two strictly applied criteria to distinguish idioms from other multi-word units.

Grant and Bauer focus first on the criterion of non-compositionality. They assume that if the meaning of a multi-word unit is directly derived from the meanings of its components, it is compositional. If it does not, the multi-word unit is non-compositional. For example, the word *red* of *red paint* means literally the red colour, so it is compositional. On the other hand, in the case of *red herring*, the phrase *red herring* typically does not mean *a soft-finned fish in the red colour range*. It means something that is not important but that takes someone's attention away from the main subject or issue, so it is not related to either the colour *red* or the fish *herring*. Therefore, *red herring* is non-compositional. However, if a multi-word unit has one non-compositional element like *a long face*, the multi-word unit is excluded from core idioms, and instead is classified as a ONCE (one non-compositional element). The next criterion that Grant and Bauer use to distinguish core idioms from non-idioms is non-figurativeness. A

figurative can be interpreted by taking a compositional untruth and extracting probable truth from it by an act of pragmatic interpretation. For example, for the figurative *it takes two to tango* when it is not used to refer to dancing, we can see it is not literally true for that particular situation and thus non-compositional, but we can imagine or visualise *a situation that involves two people and they are both therefore responsible for it* (which conveys the meaning of the phrase) and so it is figurative. According to Grant (2003, pp. 172-174), there turned out to be just 104 core idioms in English which met these two criteria of non-compositionality and non-figurativeness, the most frequent listed by Grant being *by and large* with 487 occurrences in the 100,000,000 word BNC. There is however a very large number of figuratives.

Lin (1999) also measures the non-compositionality of multi-word units. Lin's method is based on the assumption that non-compositional items have markedly different distributional characteristics to expressions derived through synonym substitution over the original word composition. For a multi-word unit, Lin substitutes one of the components with a word with a similar meaning. The list of similar meanings is obtained by taking the ten most similar words according to a corpus-derived thesaurus. The mutual information value is then found for one item produced by this substitution by taking a multi-word unit to consist of three events: the type of dependency relationship, the head lexical item, and the modifier. Put simply, a multi-word-unit a is non-compositional if there does not exist another collocation β such that (1) β

is obtained by substituting the head or the modifier in a with a similar word and (2) there is an overlap between the 95% confidence interval of the mutual information value of a and β (ibid. p. 319). For example, consider *red tape* (5.89 in mutual information value), *yellow tape* (3.75) *orange tape* (2.64), and *black tape* (1.07). Only *red tape* has a quite different mutual information value and the meaning of *red tape* relates to 'bureaucracy', which is not revealed from its components. However, there are critical problems with the underlying assumptions of Lin's method, which is that non-compositional items should have a quite different mutual information value to items formed by replacing component words with semantically similar ones. Let us look at another example such as *economic fallout* (1.66), *economic repercussion* (1.84), *economic potential* (1.24), and *economic risk* (-0.33). According to Lin's assumption, *economic risk* is non-compositional because the mutual information value of *economic risk* is quite different from the other items. Nevertheless, it is clear that *economic risk* is compositional. The whole meaning can be inferred from its parts. The problem comes from the fact that Lin depends on only computer-based data. Another problem is that the results were double-checked with a dictionary of idioms. If an item is in the dictionary, then it is said to be non-compositional. However, as Grant and Bauer (2004) point out, dictionaries of idioms contain very few core idioms.

Biber et al. (1999, p. 989) use the term "lexical bundles", distinguishing them from idioms and collocations. According to their

definition, idioms are phrases which are relatively fixed expressions whose meanings cannot be inferred from their components. Biber et al. also define collocations as two-word phrases which co-occur, and whose meanings are clearly related to their parts. For example, the word *little* prefers collocates such as *baby*, *devil*, and *kitten*. Lexical bundles are regarded as extended collocations such as *do you want me to*, *in the case of the* and *going to be a* even though they are incomplete units. The problem is that Biber et al. include phrasal verbs, prepositional verbs and figurative expressions such as *get up*, *put up with* and *bear in mind* in the category of idioms even though the meanings of some of those expressions can be related to their parts.

Liu (2003) attempts to list the most frequent spoken American English idioms based on three contemporary spoken American English corpora. Liu uses Fernando's (1996) three categories based on 'fixedness' (pure, semi-literal, literal) to distinguish idioms from other multi-word units. Fernando argues that idioms are "indivisible units whose components cannot be varied or varied only within definable limits" (ibid. p.31). However, Liu's list based on Fernando's categories includes a lot of compositional and figurative multi-word units such as *throw away*, *according to* and *use something as a stepping stone* because the criterion of fixedness does not necessarily overlap with the criterion of non-compositionality which is a characteristic of idioms. Grant and Nation (2006) found that at least one-quarter of their 104 core idioms were not frozen. *Pull~leg* also appears as *his leg was pulled*,

stop pulling his leg, pull the other one it's got bells on it, etc.

In contrast, there are studies which take a different view of non-compositionality. Wray (2000) emphasises storage of multi-word units (which Wray calls 'formulaic sequences') rather than non-compositionality. Wray argues that multi-word units are "stored and retrieved whole from memory at the time of use" (2000, p. 465). Ellis (1996, p.111) similarly claims that the words in a formulaic sequence are "glued together" and stored as a single "big word". Wray claims that to encompass the whole range of multi-word units, it is necessary to deal with semantically transparent or syntactically regular items such as *it's lovely to see you, there are three things to consider*, and *firstly...secondly...thirdly...* that are compositional, as well as semantically opaque or syntactically irregular items that are non-compositional including *beat about the bush* and *by and large*. To fully understand Wray's viewpoint on multi-word units, we need to look at Pawley and Syder's (1983) study. Pawley and Syder focus on the storage of word sequences. They compared the sentence *I'm so glad you could bring Harry* with some other paraphrases: *that Harry could be brought by you makes me so glad, that you could bring Harry gladdens me so, your having been able to Harry bring makes me so glad*...etc (ibid. pp. 195-196). The sentence *I'm so glad you could bring Harry* is fully compositional from its smallest vocabulary components. The paraphrases are also grammatical, but they are not likely to be accepted as ordinary, conventionalised usage by native speakers. That is, there are some

lexical or grammatical patterns preferred by native speakers, which are stored as familiar and naturalistic expressions. In the same way, Palmer's (1933) definition of a collocation may not assume non-compositionality of collocations, instead he may be interpreted as referring to lexical sequences in the same way as Wray, and Pawley and Syder do. When storage is the major consideration, compositionality seems less relevant.

Several researchers on collocations thus have not used non-compositionality as a criterion. Kjellmer (1984) is the most notable example. Kjellmer used the criteria *frequent co-occurrence* and grammatically well-formedness to find collocations. He could not use the criterion of non-compositionality because he used only computer-based procedures, and compositionality has to be decided manually.

The criterion of compositionality allows us to distinguish sub-categories of collocation. Primarily it allows us to distinguish core idioms and figuratives which are both non-compositional from literals which are compositional. These distinctions are very relevant when considering the learning of collocations. Let us now look at how core idioms can be distinguished from figuratives.

2.2.2 Figurativeness

Figures of speech such as irony, sarcasm and metaphor are typically non-compositional. Let us look at the following two sentences:

A. To cross the brook we had to use *stepping stones* which I found.

B. Those interns who did plan careers in the political world clearly saw the internships as *stepping stones* to future jobs.

Stepping stones in statement A refers to stones acting as footrests for crossing streams, marshes, etc. On the other hand, *stepping stones* in statement B means a circumstance that assists progress towards some goal. The literal meaning of the first sentence is compositional but the figurative expression of the second sentence is non-compositional - what relationship do stones have to jobs? The figurative use is related to metaphor.

From a teaching perspective, figurative expressions can be interpreted by using general cognitive principles, involving pragmatic competence (Seymour, 2001), while idioms have to be consciously learnt. Even though figuratives do not convey a literal meaning, their underlying meaning is related to the literal meaning and can be inferred from the context. On the other hand, the meaning of core idioms cannot be inferred from their components. Figurativeness can thus be used to distinguish core idioms from figuratives. Figuratives can be understood by visualising their literal meaning and then relating this literal meaning to their figurative meaning in a particular context. The criteria of compositionality and figurativeness can also be applied to distinguish literals. Literals are compositional and non-figurative. The parts clearly

and directly relate to the meaning of the whole.

The existing research suggests we need clear, consistently applied criteria to find the most useful collocations. In the following sections, we will look at how the criteria surveyed can be applied to find these collocations.

2.3 What sorts of collocation databases are currently available?

There are already several substantial sources of collocations. In this section, we will look at six collocation sources. Two of the six sources are in electronic form. The first one in electronic form is the *COBUILD English collocations* (Sinclair, 1995). The collocations listed in this source were derived from the 200,000,000 token Bank of English consisting of a variety of written and spoken sources including newspapers, magazines, and transcriptions of radio broadcasts, everyday conversations and interviews amassed at the University of Birmingham. Most of the texts originated after 1990 and they are primarily British, although approximately 25% are American English and 5% come from other native English varieties. The collocations were found using 10,000 nodes (=pivot words) which were non-lemmatised content words. However, it is not clear what further criteria were used to make the node list. The COBUILD English collocations program provides 140,000 different collocations with 2,600,000 occurrences and also provides corpus-based examples. Each collocation is exemplified by 20 concordance samples. The analysis span was four words before

and four words after the node word. The data presentation focuses on the frequency of the collocations rather than the placement of the collocations. Namely, the database only shows how many times a node occurs with the collocate regardless of the location of a collocate. Let us look at the following examples;

John can *work* **hard** when he wants to
I will *work* very **hard** next week
This is very **hard** *work*
It's **hard** to *work* when you are tired

The locations of the collocate *hard* are all different, but the data only indicates how many times the node *work* co-occurs with the collocate *hard*. According to the location of a collocate, the meaning of the collocation could be different. For example, the meaning of *work hard* and *hard work* is different. The former *hard* means *with a lot of effort* while the latter hard means *very difficult to do*. Separating these different uses has to be done manually.

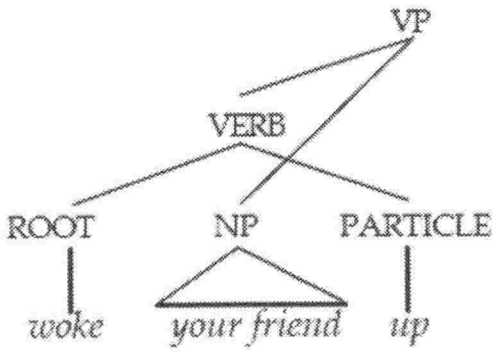
There is an additional category of *stopwords*. These *stopwords* consist of just over one hundred function words such as *to, his* and *will*. These words were omitted from the main list of 10,000 nodes and the database does not provide any example sentences for *stopwords*. This is because the collocation searching project focused predominantly on content words for both nodes and collocates. Even though the Content word+ Content word type is a very useful combination, there is also a

limitation. We need to consider *immediate constituent analysis* (Bloomfield, 1933; Fries, 1952; Gleason, 1955; Hockett, 1958; Wells, 1947), where a sentence is divided into increasingly smaller pieces until every unit is classified as a part of a larger unit. To start, the sentence would be cut into clauses, and each clause cut into Subject and Predicate. The COBUILD database does not limit itself to complete constituents. For example, the database includes items such as *studying the other* ***animals*** *at the zoo…*, *…colourful* ***animals****, zoo keepers…*, and *…killing some zoo* ***animals*** in order to show the collocational relationship between *animals* and *zoo*. The two constituents *animals* and *zoo* of the first two examples are not associated in one constituent. Let us also look at the verb phrase *woke your friend up*. Suppose this phrase contained three constituents *woke*, *your friend*, and *up*. However, if *woke up* is an immediate constituent of this phrase as well, then it would be discontinuous, as (1) *woke* would be a constituent of *woke up*, (2) *up* would be a constituent of *woke up*, (3) *your friend* would not be a constituent of *woke up*, and (4) *your friend* would be linearly ordered between *woke* and *up*. The following tree diagram clearly shows this structure of *woke your friend up*.

Figure 2.1 shows that *woke up* is an immediate constituent even though *your friend* is inserted between *woke* and *up*. This level of analysis has not been applied to the items in the COBUILD database.

Figure 2.1

The structure of the phrase *woke your friend up*



- Ojeda (2005, p. 624)

Another problem is that while searching for collocations, a consistent frequency cut-off point was not used, so according to the node, the frequency cut-off point for its collocates is different. For example, if we look at the first node *abandon* and the last node *zoo* (because the node list is alphabetically ordered), the most frequent collocate of *abandon* is *forced* with 180 occurrences and the lowest frequency collocate is *traditional* with 22 occurrences. On the other hand, the top collocate of *zoo* is *animal* which has a frequency of 76 occurrences and the lowest frequency collocate is *breeding* with 17 occurrences. This is because only the top 20 collocates of a node were entered into the list regardless of their frequencies. The problem is that if a node had 21 or more very frequent collocates, those after 20 were excluded even if they were more frequent than other collocates included for other nodes. On the other hand, if a node had fewer than 20 collocates, some collocates which had very few occurrences were

included. Because of this, it is difficult to use the COBUILD list as a basis for quickly choosing the most useful collocations. So, the COBUILD list is very large but still requires further sorting and analysis to provide data for syllabus design, and teaching and learning.

The second electronic source of collocations is Phrases in English (PIE). PIE was designed by Fletcher (2003/2004) and incorporates a database derived from the second or World Edition of the BNC. It aims to provide a simple yet powerful interface for studying phrases up to eight words long and is useable by both experienced researchers and novice users. Using PIE, we can look up individual words and phrases online.

The PIE system basically consists of four searching functions - n-grams, phrase-frames, POS (Part Of Speech)-grams and chargrams (*n* characters). Here *n-gram* means a sequence of *n* words, where the word span *n* is 1 to 8, and *word* means a token of any lexical entity assigned a BNC POS tag such as AJ0 (adjective (general or positive), e.g. *good, old*), AJC (comparative adjective, e.g. *better, older*), and AJS (superlative adjective, e.g. *best, oldest*). There are 58 such tags.

For searching for n-grams, PIE provides the two choices - Simple Search and Advanced Search. Simple Search is used for searching for individual words and phrases with their POS tags and frequency data, and it also offers a maximum of 50 concordance samples for each item. In addition, there are sub-functions in Simple Search. For example, the symbol + means one word, so *the+ days* includes *the **old** days, the **good** days, the **bad** days,* etc. *The only+ +* means 4-grams beginning with *the*

only. The symbol ~ means an optional word, so *the ~ days* includes *the days, the old days, the good days, the bad days,* etc. *The ~ ~ days* includes *the days, the old days, the good days, the good old days*, etc. * means word variations within one word form, for example, *nation** includes *nation, nations, national, nationalistic, nationalise*, etc. ? is used for searching for word variations with one different character, so *s?ng* includes *sing, sang, sung, song,* etc (see http://pie.usna.edu/simplesearchTab.html for more information). The function of selecting a frequency cut-off point is added in Advanced Search.

Phrase-frames are sets of variants of an n-gram with identical form except for one word, represented here by the symbol *. For example, the most frequent 4-frame is *the * of the,* with 5,652 variants such as *the end of the, the rest of the, the top of the, the nature of the,* etc.

POS-grams are patterns of Part Of Speech tags assigned to word forms without reference to the specific lexical entities. When ordered by types, the most frequent "3-POS-gram" is ART ADJ NOUN like *the other hand.* On the other hand, when ordered by tokens, the 3-POS-gram PREP ART NOUN like *at the end* is more frequent.

Finally, *chargrams* are sequences of *n* letters. Many symbols are used for a lot of different functions. [ ] means word forms with a choice of specified characters in [ ], for example, *t[io]p* includes *tip, top, tipped, stop*, etc. | means alternative groups of chars, for example,

(the/a) cat includes *the cat* and *a cat* (see http://pie.usna.edu/drillGrams.html for more information).

Results can be ordered alphabetically, by frequency or by POS tag. For focused studies, we can filter results for specific word forms and/or word-classes which a query must match or exclude.

Even though PIE has a lot of useful functions, PIE seems closer to the upgraded SARA-32 program which is the word analysis program for the BNC rather than a database of English phrases. The data provided by PIE is still not filtered, That is, it is impossible to apply the criterion of *grammatical well-formedness* or distinguish polysemous uses of an item only by using PIE's computational work. In addition, the concordance samples for each item are restricted to a maximum of fifty samples. The fifty random concordance samples from the BNC cannot cover all the possible different uses of an item. Thus, COBUILD and PIE are useful sources for gathering more instances of useful items, but they require a lot of further manual analysis when compiling a list.

Other collocation sources include Kjellmer's (1994) list, Simpson and Mendis' (2003) list, Liu's (2003) list and Biber et al.'s (2004) list. Each study has its own strengths but leaves some problems to solve. Table 2.2 gives a brief comparison of the three lists.

Kjellmer (1994) found 85,000 collocational types using the 1,000,000 token Brown Corpus. The two criteria *frequent co-occurrence* and *grammatical well-formedness* were used for searching for collocations.

Table 2.2

The comparison of the four collocation studies

	Kjellmer (1994)	Simpson & Mendis (2003)	Liu (2003)	Biber et al. (2004)
Terms	collocation	idiom	Idiom	lexical bundle
Criteria	*grammatical well-formedness, frequent co-occurrence*	*compositionality* or *fixedness, institutionalisation, semantic opaqueness*	*fixedness frequent, co-occurrence*	*frequent co-occurrence, range*
Collocation span				4-word MWUs
Corpus size	1 million tokens	1.7 million tokens	6 million tokens	2 million tokens
Corpus type	written	spoken	Spoken	spoken and written

Kjellmer used a frequency cut-off point of two occurrences because the size of the corpus used was small. As a result the list includes many free combinations which have little collocational relationship and excludes collocations of relatively low frequency. For the second criterion of *grammatical well-formedness*, all collocation entries had to be adjacent and fit one of nineteen grammatical types such as noun phrases, verb plus object, and verb plus verb(s). Discontinuous collocations were excluded and polysemous uses of an identical form were classified as the same item because Kjellmer used a "purely mechanical method" (Kjellmer, 1994, p. xiv). Kjellmer's list included some grammatical categories which are excluded from the present study such as A+Noun (e.g. *an insect*), The+Noun (e.g. *the boat*), Noun+Of (e.g. *father of*), To+Infinitive Verb (e.g. *to examine*),

proper names (e.g. *Bobby Joe*) and non-English expressions (e.g. *per se*).

Simpson and Mendis (2003) found 238 academic spoken idioms occurring in the 1,700,000 token Michigan Corpus of Academic Spoken English (MICASE). Simpson and Mendis used the three criteria of (1) *compositionality* or *fixedness,* (2) *institutionalisation,* and (3) *semantic opaqueness*, which were already noted by Fernando (1996), McCarthy (1988), and Moon (1998b). As Grant and Bauer (2004) pointed out, these criteria could not be consistently applied to distinguish idioms from other multi-word units. In effect, Simpson and Mendis tried to exclude metaphoric expressions (e.g. *a sad showing*), but most of the items in their list were figurative expressions such as *bottom line, the big picture, and draw a line between.* The items were classified into the four academic divisions of 'social sciences & education', 'physical sciences & engineering', 'humanities & arts' and 'biological & health sciences', and the three primary discourse modes of 'monologic/panel', 'interactive' and 'mixed'. The list is useful to focus on academic idiomatic expressions, but the terms used in their study such as 'idioms', 'metaphoric expressions', and 'phrasal verbs' need to be more clearly defined.

Liu's (2003) study focused on spoken idioms, but his criterion of *fixedness* used does not necessarily overlap with the criterion of non-compositionality which is a characteristic of idioms mentioned before. This resulted in the inclusion of a lot of compositional and figurative

multi-word units. Liu identified 9,683 idioms listed in two of four major contemporary English idiom dictionaries and three English phrasal verb dictionaries (Because Liu used the criterion *fixedness*, phrasal verbs were included in the category of idioms as shown in section 2.2.1). Liu found 302 of the 9,683 items identified in three American spoken corpora using a frequency cut-off point of two occurrences per million. However, because Liu's list was restricted to identification based on the existing dictionaries, the category of idioms was not clearly defined and many of the items were grammatically incomplete, for example, *go with, hold on to, knowledge of*, etc.

Biber et al. (2004) identified lexical bundles listed in Biber et al.'s (1999) list in texts from university classroom teaching and textbooks. Biber et al. found 172 lexical bundles in the T2K-SWAL Corpus (TOEFL 2000 Spoken and Written Academic language Corpus; see Biber et al., 2002, 2004) using a frequency cut-off point of forty times per million and a range cut-off point of 20 of 263 different texts. The strength of the list is that register variation in the functional exploitation of lexical bundles was considered. The different registers included conversation, classroom teaching, textbooks, and academic prose. However, the items listed in Biber et al.'s study were restricted to four word items where the items all are adjacent. Here are some of the most frequent bundles they found - *you don't have to, one of the things, what do you think, that's one of the,* etc. About twenty percent of their bundles were grammatically well-formed, for example, *I don't think so, a lot of people,*

at the same time, in the United States.

Biber et al.'s study thus looks at a small set of bundles and excludes the many more frequent two item and three item bundles. It is thus a useful study, but not a sensible starting point for making a list of the highest frequency collocations.

The present study has two goals – (1) to see what criteria are needed to define collocations and (2) to make a list of the high frequency collocations of spoken English that would be useful for guiding teaching, learning and course design. For these purposes, in this chapter we have looked at the criteria of *frequent co-occurrence, grammatical well-formedness, mutual information,* and *predictability in L1. Frequent co-occurrence* could be used to determine usefulness because high frequency items have more chances to be used and met and by using the criterion of *grammatical well-formedness,* we could provide meaningful units for learning and teaching collocations. *Mutual information* is used to measure collocational strength of a multi-word unit but it is highly affected by the low frequency of one of the items. Thus, it may not be effective in distinguishing collocations. However, to ensure this assumption the criterion of *mutual information* needs to be tested. Finally, the criterion of *predictability in L1* would be useful to reduce the number of collocations to focus on. For *predictability in L1,* Bahns' (1993) study is a good model because his method is highly replicable even though using the dictionary definition cannot cover all the sorts of differences between L1 and L2. Thus, these four criteria

will be trialled.

There is clearly a need for a study that uses well-designed criteria that make sense with regard to the teaching and learning of collocations and results in a list of high frequency well-formed collocations.

2.4 Research Questions

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 collocates of high frequency words which are also high frequency words?
(3) What is the relative importance of collocations in spoken and written text?
(4) What are the most common collocational patterns?

There are four assumptions behind this study.

(1) Language use
① Learning collocations can improve the learner's language fluency.
② Learning collocations can improve the native-like selection of a learner's language use.

(2) Language learning

① Learning collocations can improve the learner's vocabulary retention.

(3) Useful collocations

① The most frequent collocations are the most useful in collocation teaching and learning.

These assumptions will not be tested in this study.

CHAPTER 3

RESEARCH PROCEDURE

This chapter describes how the present study was carried out. This involves first of all a description of a pilot study that clarified the criteria that needed to be used. The second part of this chapter illustrates the main study.

3.1 The pilot study

The pilot study was carried out to examine the effectiveness and reliability of the criteria for determining a collocation.

3.1.1 Instruments

WordSmith 3.0 (Scott, 1999) and SPSS 10.0.5 for Windows (SPSS Inc, 1999) were used in the pilot study. Version 3.0 of WordSmith is a very convenient and effective tool for word counts and word concordances. SPSS is very useful for calculating data through mathematical formulae and for drawing graphs based on the data.

3.1.2 Procedure

The pilot study involved the following steps.

1. Choosing a corpus
2. Choosing two words to analyse
3. Using WordSmith Tools to make a concordance
4. Taking all the collocations with a frequency of two or more
5. Applying the criterion of *grammatical well-formedness* to the collocations chosen at step 4
6. Calculating *mutual information* for the collocations chosen at step 5
7. Applying the criterion of *predictability* to the collocations chosen at step 6

The Wellington spoken and written, Brown, and Lancaster-Oslo/Bergen (LOB) corpora containing 4,758,223 tokens and 76,783 types were chosen. Only two words were chosen to analyse in the pilot study. The first word was *high. High* was chosen because *high* has a lot of collocates and there were 1,784 occurrences of *high* in the corpus. The analysis range was from third left collocate to third right collocate. Table 3.1 gives some examples.

Table 3.1
Collocation span

Collocates on the left			Pivot word	Collocates on the right		
3rd	2nd	1st		1st	2nd	3rd
a	little	bit	high			
				and	dry	
				in	the	air

As shown in Table 3.1, the word *high* collocates with the three-word sequence *a little bit* on the left and also collocates with *and dry* on the right. Theoretically in this study a collocation could be, at most, a seven-word sequence including the node and three words to the left and three words to the right. However, it is very unusual to find examples of a seven-word collocation with collocates evenly distributed on both sides.

Using the concordancer in WordSmith Tools, the pilot study found 2,577 multi-word items containing *high.* The number of the initial potential collocations (2,577) was greater than the frequency of *high* (1,784) because the same occurrence of *high* in a sentence can collocate with more than one other word. For example, from the sentence *interest rates have reached very high levels,* the two collocations *very high* and *high levels* are found. Even though they make up one sequence, they are counted separately. So the same occurrence of the word *high* occurs in the two collocations.

2,171 of the 2,576 multi-word groups found by WordSmith did not meet the first criterion, *grammatical well-formedness.* That is, the multi-word groups did not function as one or more of the immediate constituents such as Subject, Predicate, and Verb. As shown in Table 3.1, the word *high* collocates with *and,* but *high and* does not meet the criterion of *grammatical well-formedness.* However *high and dry* meets the criterion. The longest collocation was a 4-gram collocation *high in the air.* Therefore, 405 items remained after the first phase. Table 3.2

contains some items that met the first criterion, and some items that did not meet it and were excluded from potential collocation groups.

Table 3.2
Some examples included and excluded by criterion 1

Included samples	Excluded samples
high school	new high
high level	such high
{No.} feet high	at high
very high	between high
run high	those high
on a high	give high

As shown in Table 3.2, such groups as *high school, high level* and *very high* meet the first criterion and are included. On the other hand, such groups as *new high, such high,* and *give high* are excluded because they do not meet the *grammatical well-formedness* criterion.

The second criterion, *frequent co-occurrence*, uses only absolute frequency. Only the multi-word units that occurred two or more times in the 4,758,223 tokens were included. Table 3.3 shows some of the items that met the criterion and some of the items that were excluded because they occurred only once. In this pilot study, only two co-occurrences are used as a cut-off point to trial the criterion of *frequent co-occurrence*. The low frequency level of two was used to get the maximum number of different collocations. However, in the main study several frequency

levels are compared to choose a suitable frequency cut-off point.

Table 3.3
Some examples included and excluded by criterion 2

The most frequent multi-word units		The multi-word units that occurred once	
high school	154	on a high	1
very high	56	high wardrobe	1
high level	41	usually high	1
too high	41	high vacancies	1
high court	35	high velocities	1
high standard	28	high statue	1
high speed	23	high voltages	1
high up	22	unacceptably high	1
high quality	21	unfairly high	1
high degree	20	unnecessarily high	1

This criterion was applied to the 405 multi-word groups remaining after the application of criterion 1 and 234 items were excluded, leaving 171 word groups. In Table 3.3, we see that such word groups as *high school*, *very high* and *high level* are included as collocations because they meet the frequency cut-off point of two or more. On the other hand, such word groups as *high voltages, unfairly high,* and *high statue* are excluded because they all occur only once in the corpus, which does not meet the criterion of *frequent co-occurrence*. However, there is a

problem with this criterion as some very acceptable low frequency collocations such as *on a high, high tide* and *high tea* are excluded by the absolute frequency criterion. Clearly a larger corpus is needed.

The third criterion is *mutual information* which shows the strength of the relationship between the pivot word and its collocate. If f(x)=the number of x, f(y)=the number of y, f(x,y)=the number of co-occurrences of x and y, and N=the number of the corpus tokens, the mutual information value I(x,y) can be calculated by using the following formula:

$$I(x,y) = \log_2 \{[f(x,y)N] / [f(x)f(y)]\}$$

Table 3.4 shows the statistical data and some examples from the most interesting pairs to the least.

Table 3.4

Some examples from the most interesting pairs to the least one by criterion 3

	I(x,y)	From high I(x,y) to low I(x,y)		-value(I<0)	
N	172	high heeled (N)	10.70	high one	-1.54
Max	10.70	high birthrate	10.38	get high	-6.20
Min	-6.20	high coercivity	10.38		
M	5.7572	.	.		
SD	2.2190	.	.		
		high point	1.67		
		high place	1.01		

Table 3.4 shows that this criterion was applied to 171 items and the average of the mutual information value was about 5.76. The collocation with the highest mutual information value was *high heeled (N)* with a mutual information value of 10.7 because *heeled* is a low frequency word, but it typically collocates with the word *high*. Using the mutual information criterion, there were some items consisting of three words such as *a little high, a bit high,* and *high and dry* which were difficult to classify because the measurement of *mutual information* is based on the relationship between two words. Therefore, multi-word collocates such as *a little, a bit* and *and dry* were counted as a chunked unit because counting these collocates as one unit is more valid than counting single collocates such as *little, bit,* and *dry.*

Table 3.4 also shows some problems with the criterion of *mutual information. Mutual information* is highly affected by the low frequency of one component of the items. For example, *high coercivity* is an expression with a very restricted use, but its mutual information value of 10.38 is very high. The total number of tokens of the word *coercivity* is just 4 which all occur in the Wellington Written Corpus (WWC) and 2 of the 4 cases collocate with *high.* It is used in physical science as in *it must have high coercivity at room temperature, a low Curie point, and a high Kerr or Faraday rotation.* This means technical terms could have a higher mutual information value because of their very restricted use. *Mutual information* is usually used to check if one collocation has a more interesting relationship between its components than other collocations,

so it is difficult to set a cut-off point for *mutual information*. However, in general, a mutual information score greater than 2 is considered high enough to show an interesting association between two words (Kennedy, 2003). In the present study, thus a score of 2 was used as a cut-off point for *mutual information*.

By applying the criterion of *mutual information* using a cut-off point of 2, only seven items including the two collocations *high one* and *get high* with a minus value were excluded. The result was that 164 word groups remained from 171 items. This criterion thus did not have much effect on the results.

At the fourth step, *predictability in L1/* or *semantic opaqueness,* 148 items from 164 items were classified as predictable collocations because they are directly translatable word for word into Korean and only 16 items are included as unpredictable collocations. Tables 3.5 and 3.6 summarise the results.

Table 3.5
The results of each step

	Step 1	Step 2	Step 3	Step 4
Included	405	171	164	16
Excluded	2,171	234	7	148
Total	2,576	405	171	164

As shown in Table 3.5, three of the four criteria were efficient tools to determine a collocation. However, although the third criterion,

mutual information or *collocational specialisation,* showed the most specialised information of a word sequence, it was not influential in including or excluding items.

Table 3.6
The collocations of *high* meeting all four criteria

1.	high street	13
2.	high water	13
3.	high rise	7
4.	high spirits	7
5.	high minded	5
6.	high tech	5
7.	high flyer	4
8.	high seas	4
9.	high tension	4
10.	high and dry	4
11.	high road	3
12.	run high	3
13.	high sounding	3
14.	high and low	2
15.	high society	2
16.	high flying	2

Table 3.6 shows the sixteen collocations resulting from applying the four criteria, and the frequency of occurrence of each collocation. So, *high street* met all four criteria and occurred 13 times in the

4,758,223 word corpus. *High and low* has 5 occurrences but 3 of them are used with a literal meaning, and the other 2 cases are used with the meaning of "everywhere" as in *I have searched high and low for it. High society* occurs 3 times, but one occurrence is a movie title. This is why the total number of occurrences of the two collocations is reduced. These tokens were excluded after calculating *mutual information*, but it did not affect the results of *mutual information*.

In addition to the analysis of *high*, another word, *field* was examined. *Field* was chosen because it is a much less frequent word than *high* and is a different part of speech. The same corpus consisting of the Wellington spoken and written, Brown, and Lancaster-Oslo/Bergen corpora containing 4,758,223 tokens was used, using different sampling sizes of collocations selected randomly to see the effect of sample size on the result. This analysis was carried out under the assumption that the most frequent collocations would be found regardless of different sampling sizes. If the same result could be reached by extracting only a part of the total instances, then time could be saved in the search for collocations. On the concordance option of WordSmith Tools, the number of entries required can be set and the entries can be randomly chosen. Table 3.7 gives the results.

As shown in Table 3.7, the total number of occurrences of *field* in the corpus was 809 (tokens), and the total number of collocation entries of *field* (e.g. *medical field, field again, magnetic field*) was 388 (types). This is less than 809 because the same collocation occurred several

times.

Table 3.7

The results for the word *field*

No. of tokens of 'field'	No. of collocation entries of 'field'	Criteria 1~2	Criterion 3	Criterion 4
809 (1 in 1) =(whole corpus)	388	68	62 (62/809≈0.08)	5
401 (1 in 2) =(1/2 size)	209	36	33 (33/401≈0.08)	5
264 (1 in 3) =(1/3 size)	120	19	16 (16/264≈0.06)	1

● (1 in 3) means that one of three random samples is kept and the other two samples are ignored.

After applying criteria 1 to 3, 62 of the 388 items remained (e.g. *medical field, magnetic field*), which make up 8% of the total number of tokens of *field.* An interesting point is that in the case of the word *high*, after applying the first three criteria, 9% of the total number of tokens of *high* remained, which is almost the same proportion as for *field.* After applying criterion 4, only 5 items were included as unpredictable collocations (e.g. *field marshal, field service, visual field*). Subsequently, 401 of the total number of occurrences of *field* were randomly selected, that is, a half of the total number of 809 tokens of *field* were randomly selected. 209 items were found as potential collocation entries. After applying criteria 1 to 3, 33 items remained (e.g. *ski field, field service, field work*), which is 8% of the randomly selected 401 tokens of *field.*

After applying the last criterion, 5 items were included as collocations. Table 3.7 shows that the different sampling sizes (809 and 401) include substantially different numbers of collocations except for criterion 4. Although criterion 4 was restricted to Korean, criteria 1 to 3 aimed at general usefulness for learners of English regardless of L1. This means that the corpus used was too small. However, this result is already expected from Table 3.6. The collocations found have a small number of occurrences, so when we examined a smaller sample size, many of those collocations did not occur at all. It is thus clear that a much bigger corpus than 4,788,223 tokens is needed. Using a large corpus with a high frequency cut-off point seems more reliable than using a small corpus with a low frequency cut-off point.

Let us return to the starting point. The goal of this study is to make a collocation inventory for beginners learning English. For this purpose, we decided to use four criteria: 1. *grammatical well-formedness*, 2. *frequent co-occurrence*, 3. *mutual information*, 4. *predictability in L1 (e.g. Korean)/ or semantic opaqueness*. However, through the pilot studies, it was found that the third criterion, *mutual information,* was strongly related to the second criterion, so it has no additional discriminating influence on distinguishing a collocation from other sequences. The fourth criterion *predictability in L1* is restricted to Korean, so the findings cannot be generalised to other languages. Therefore, the first and second criteria are particularly important in this study. The criterion, *frequent co-occurrence* involves two critical

decisions. One is, as Richards (1974) points out, that frequency depends greatly on the nature of the corpora used and there are differences, too, between written and spoken text. It is necessary to decide which corpus to use. The other is what frequency level we should use for the second criterion. To investigate these questions, firstly, we compared the Wellington spoken (WSC) and written (WWC), Brown, and LOB corpora with the BNC. The WSC, WWC, LOB, and Brown corpora together have 4,758,223 tokens and the BNC spoken section has 10,000,000 tokens. The word *high* was the sample word analysed in the comparison. There are 1,784 occurrences of *high* in the combined WSC, WWC, LOB, and Brown corpora, and the 10,000,000 word BNC spoken section has 2,068 occurrences of *high*. The tokens of *high* do not differ greatly between the two corpora even though the BNC spoken section is almost twice as large as the combined written corpus. However, the whole 100,000,000 word BNC has 38,184 occurrences of *high*, which is more than twenty times that of the combined WSC, WWC, LOB, and Brown corpora.

To examine the different nature of spoken and written texts, two corpora were made from the BNC: one was a spoken corpus that has 4,960,744 tokens, and the other was a written corpus that has 4,849,337 tokens. That is, the size of the two corpora was almost the same. The word *high* was also analysed, using a cut-off frequency of 5 occurrences. Table 3.8 shows the results.

Table 3.8

The comparison of a spoken and a written corpus

	Spoken	Written
Frequency	1,275	2,387
No. of collocations	47	82

As shown in Table 3.8, the written corpus has 2,387 occurrences of *high* while there are 1,275 occurrences of *high* in the spoken corpus. That is, the word *high* is much more frequent in the written corpus. The written corpus has almost twice as many collocations containing *high* as the spoken corpus. 35 of the 47 spoken collocations also occur in the written corpus such as *very high, high level*, and *too high*. Some items occur in only one of the corpora, for example, *quite high* and *fairly high* only occur in the spoken corpus. Looking at the spoken/written distinction will clearly reveal interesting data. However, because spoken proficiency is seen as the main goal of beginning courses, the main data will be gathered from a spoken corpus.

The pilot study showed that (1) a large corpus was needed, (2) the mutual information criterion was not effective for finding high frequency collocations, and (3) a frequency cut-off point would need to be worked out that fitted with the size of the corpus.

3.2 The Main Study

In this thesis, a collocation will be defined as a sequence of words, usually but not necessarily adjacent, which occur together frequently and which can act as a complete functional part of a sentence such as a subject, an object, or an adjective.

3.2.1 Research procedure

3.2.1.1 Data source

The 10 million-word spoken section of the BNC was used as the data source. The BNC spoken section is the biggest spoken corpus available. The pilot study indicated that this may be big enough to get plenty of occurrences of the collocations. There are two almost equally sized parts to the 10-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 Demographic part is made up of 153 texts sampled from 124 male and female volunteers of a wide range of ages, living at 38 different locations across the UK. The context-governed part is made up of 757 texts with the following four broad categories: (1) **Educational** and **informative** events, such as lectures, news broadcasts, and classroom discussion, (2) **Business** events such as sales demonstrations,

trade union meetings, and consultations, (3) **Institutional** and **public** events, such as sermons, political speeches, and council meetings, and (4) **Leisure** events, such as sports commentaries, after-dinner speeches, and club meetings. To ensure representativeness of spoken language for the BNC, a broad range of sampling in both the target population and the types of spoken text was carried out.

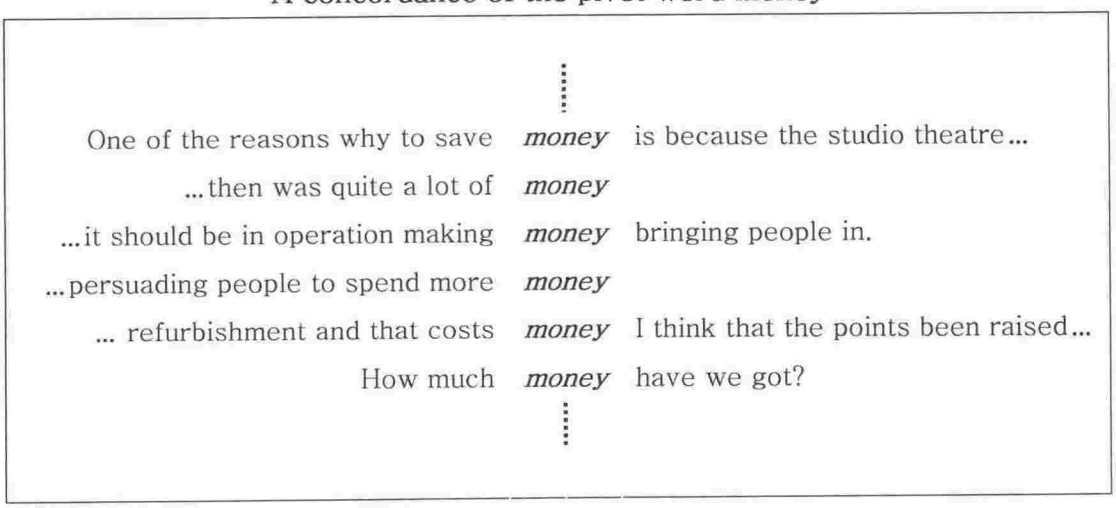
3.2.1.2 Sample selection

The analysis of this corpus focused on the collocates of the most frequent 1,000 content words from the spoken word frequency list by Leech, Rayson, and Wilson (2001), available at http://www.comp.lancs.ac.uk/ucrel/bncfreq/. The first 1,000 content words were the pivot words, that is the focal points of the collocations. For example, the pivot word of the collocations *sex education* and *child sex abuse* is *sex*. What is the pivot word is simply determined by what the researcher has decided to focus on. The word list by Leech et al. is the one based on the BNC spoken section. It consists of word types and is thus non-lemmatised, for example, the different inflected forms of *look, looks, looked* and *looking* are listed as different words. The word type, rather than lemma or word family, was chosen because different inflected forms may take different collocates. For example, only *looking* collocates with *by* as in *by looking at {sth}*.

3.2.1.4 Searching for collocations

The word analysis program, WordSmith Tools was used, particularly its two functions, concordance and collocate-sort. The concordance option finds the target word and provides a context for that word on each side of the word. Figure 3.1 gives an example.

Figure 3.1
A concordance of the pivot word *money*



Collocate-sort in WordSmith Tools sorts the contexts so that the same and similar collocations occur together to the right or left of the node. The analysis range was from third left collocate to third right collocate, in other words, it includes the first to the left, the second to the left, the third to the left, the first to the right, the second to the right, and the third to the right. For example, the collocates of *go* include ***in one go, let go, go outside***, and ***go in there.*** However, there are some issues to consider when counting collocations. Firstly, when collocations

consist of more than two words such as the collocation of *up, up and down*, the two collocates *and* (right 1) and *down* (right 2) are separately counted by WordSmith, not as one whole unit. Therefore, the number of the initial potential collocation entries can be bigger than the total occurrences of the pivot word. Secondly, some collocations could overlap when counting the total number of collocations. For example, the collocation *very good* is shared by the two pivot words *very* and *good*. When the pivot word *very* is analysed, *very* has the collocate *good*. When the pivot word *good* is analysed, *good* has the collocate *very*. Thirdly, the range for collocates was restricted to the third left collocate to the third right collocate. Therefore, if a word sequence is beyond that range, a collocation might be shown as an incomplete form, so it could be excluded. For example, the pivot word *day* has the collocate *end of the*, but *end of the day* is an incomplete form. It needs more complementary words such as ***at the*** *end of the day*. Fortunately, in this case the word *end* is included in the first 1,000 pivot words, so when the pivot word *end* is analysed, *at the end of the day* is included as a collocation.

3.2.1.5 Choosing a frequency level

In the present study, it was necessary to set a frequency level (cut-off point) for the collocations considering the size of the corpus and the size of the resulting collocation inventory. If the cut-off point is

too low, then the total number of collocations is too large. If the cut-off point is too high, many useful collocations may be excluded. Therefore, several cut-off points were trialled to find an appropriate frequency level which can be applied across all the data, that is for nodes at the top of the frequency scale and for nodes at the end of the most frequent 1,000 types. When a frequency of 30 occurrences per 10,000,000 tokens for each collocation is applied as a cut-off point, the average number of collocates of the most frequent 50 node words is about 28. When a frequency of 20 occurrences per 10,000,000 tokens for each collocation is used as a cut-off point, the average number of collocates is about 37. When a frequency higher than 30 is used as a cut-off point, no collocates are discovered in the lower frequency pivot words which are still in the most frequent 1,000 words of English or at best very few collocates are found. For example, some lower frequency words such as *fund* (634 in frequency), *previous* (542) and *passed* (521) have no collocates when a cut-off point of 60 is applied and the two pivot words *watching* and *dad* have only one collocate each, namely *watching it* and *mum and dad*. Similarly, *woman* has two collocates - *this woman* and *that woman* - that meet the two criteria. This means if we use a low cut-off point like 20 occurrences per 10,000,000 tokens, there are too many collocates meeting the criterion, whereas if a high cut-off point like 60 occurrences is used, many lower frequency pivot words may have few or no collocates that meet that frequency criterion. Even if a few collocates are found, they are likely to be function words such as

this, that, and *it.* Consider Tables 3.9 and 3.10.

Table 3.9

The number of collocates of some lower frequency words from the spoken word list based on three cut-off points

Cut-off point / Pivot word	20	30	60
green (88/million)	5	1	0
tell (668)	12	1	0
give (845)	8	4	0
great (363)	9	7	1
sale (46)	6	2	1
worry (94)	2	1	1

- The pivot words have a frequency below 1,000 occurrences per million in the 10,000,000 word corpus.

Table 3.10

The number of collocates of some high frequency words from the spoken word list based on the three cut-off points

Cut-off point / Pivot word	20	30	60
up (2,891/million)	169	156	65
very (2,373)	109	60	42
down (1,472)	98	71	28
good (1,566)	95	65	32
other (1,313)	85	42	22
go (2,885)	77	43	26
here (1,640)	71	39	21

- The number in brackets shows the total frequency per million of each word.

Table 3.9 shows that the pivot word *green* has 88 total occurrences per million in the 10,000,000 token corpus. When we use a cut-off point of 20 or more occurrences for each collocation which

meets the two criteria, we end up with only five collocates meeting those criteria. When the frequency cut-off point is raised to 30 or more occurrences, this drops to 1, and when it is 60 or more, to zero. Table 3.10 is interpreted in the same way as Table 3.9, but Table 3.10 contains only very high frequency words. *Very* has a total occurrence of 2,373 tokens per million in the 10,000,000 word corpus. *Tell* in Table 3.9, by contrast, has only 668 per million.

As shown in Table 3.9, the number of collocates for words at a cut-off frequency of 20 occurrences for each collocation is small (mostly less than 10). When a frequency cut-off of 60 occurrences per collocation is applied, no collocates occur or there are very few collocates. When we consider only Table 3.9, a frequency of 20 occurrences seems to be an appropriate cut-off point. However, we also need to consider high frequency words to determine a generalisable cut-off point. Table 3.10 shows the number of collocates of some high frequency words based on three cut-off points (20, 30 and 60). In Table 3.10, the number of collocates using a frequency of 20 occurrences as a cut-off point is too many (even 71 different collocates which is the smallest number of occurrences in Table 3.10 is too many). Besides, if a frequency of 20 occurrences is applied as a cut-off point, many "open choice" collocations where "a large range of choice opens up and the only restraint is grammaticalness" (Sinclair, 1991, p. 109) are more likely to be included. When we consider Table 3.10, a frequency of 60 occurrences seems appropriate for a cut-off point because the average

number of collocations per pivot word is about 34. Nevertheless, to make a consistent list, we need to use the same frequency cut-off point for all collocations whether their pivot word (node) is a very high frequency word or not. This is why eventually a cut-off point of 30 occurrences per 10,000,000 words (3 per 1,000,000) was decided as an appropriate cut-off point for collocations in this study. This cut-off point provides an adequate number of collocations for most of the pivot words in the first 1,000 word types of English.

3.2.1.6 Sorting collocates

When a collocation has different parts of speech or different meanings, each meaning or part of speech is listed and sorted as a different item. The distinction in the meaning relies on the *Collins English dictionary* (1994). For example, the collocation *get up* is defined as - (1) *to wake up and rise from one's bed,* (2) *to rise to one's feet; stand up,* (3) *to ascend or cause to ascend,* (4) *to mount or help to mount,* etc, in *Collins English dictionary.* Even though all the uses of *get up* convey the core meaning of *to rise, raise oneself or come to somewhere or some posture*, foreign learners are likely to recognise those different uses as different expressions, and thus the distinction in the meaning is separated in the lists. When collocations are listed, we sort collocations according to frequency. If a collocation has different parts of speech or different meanings, those are subdivided into separate collocation items

according to frequency order. Each collocation is classified with the part of speech of the collocation, and a sample sentence or phrase as shown in Table 3.11. So in the final lists in Appendices 1 and 2, polysemous uses of a collocation have been separated.

Table 3.11

Some collocates of *up*

Rank	Collocation	Part of speech	Frequency	Context
5	*(be-verb) set up (sth)*	VP	500	And we will set up mailboxes for you as individuals.
65	*(be-verb) grown up (sth)*	VP	60	…at least until the children are grown up and probably well beyond?
70	*getting up*	VP	56	Aha, that's what I do, it's like getting up in the morning.
128	*getting up (swe)*	VP	21	Well, it's getting up there.
137	*getting up*	VP	12	…why am I getting up and speaking.
139	*{Det-a [8]} grown up*	NP	9	…you are a grown up now aren't you?

- The frequency of some polysemes does not meet the frequent occurrence criterion, but the total frequency of their identical forms meets the criterion.

The three example collocations in Table 3.11, *set up, getting up,* and *grown up,* are ordered by frequency, and as shown in the case of *grown up,* when a collocation has more than one part of speech, the collocation is divided into different collocations which are also arranged by frequency. In addition, when a collocation has more than one meaning,

the item is also divided into different items. For example, *getting up* (89 in frequency) is classified into three items by frequency order in Table 3.11 The first meaning of *getting up* (56) is *waking up*, the second meaning (21) is *coming up* and the last (12) has the meaning of *standing up*. In this study, only content words such as common nouns, verbs, and adjectives from the word frequency list are analysed. Interjections (e.g. *god!, well!*) are excluded from the list of pivot words, but a collocation made of content words can be used as an interjection such as *good gracious!* and is classified as such. Table 3.11 shows some collocational groups of the pivot word *up*, however, in Appendix 2, all the collocations are re-sorted by frequency order regardless of the pivot words.

3.2.1.7 Collocation and colligation

We need to distinguish *collocation* and *colligation*. In the present study, a collocation is a well-formed sequence of words that frequently go together, such as *you know, I think*, and *come back*. The parts of a collocation are usually, but not necessarily, adjacent to each other.

Some collocations listed in this study such as *check* ***sth*** *out*, *brought* ***sth*** *to* ***smo,*** and *No. years* however may be included in the notion of *colligation*. Colligation refers to collocational frameworks (Renouf & Sinclair, 1991) in which some units are based on a grammatical class rather than on a particular word. Fillmore (1997) gives a good example to explain colligation. The phrase *ripe old age*

frequently occurs with some collocates like *live to* or *reach*, but also with other semantically similar verbs such as *attain, survive to* and *go on to*. However, the preferred grammatical pattern is *verb+ preposition + a ripe old age - live to a ripe old age, survive to a ripe old age* etc. Because the pattern does not consist solely of particular words but includes grammatical categories (e.g. Verb+Preposition), it is a colligation rather than a collocation.

In a simpler example, it is easily seen that the verb *want* prefers the colligation *want+ to INF* (e.g. *want to go*) to being followed by an *-ing* form. *Smo* as in *brought sth to smo* represents personal pronouns or personal proper nouns such as *him, her,* and *John*. It is a grammatical frame where similar sorts of words are substitutable, so the pattern containing the non-lexical items such as *sth* (something), *smo* (someone), and *swe* (somewhere) is regarded as a colligation. Colligates are added to show the items occurring between non-adjacent collocates but they are not part of the collocation such as *smo* as in *pick {smo} up*. The colligates which are included as part of a collocation in this study are:

a, the as in *{Det-the [16], a [12]} bus stop*

N as in *nearly every {N}*

INF as in *managed to {INF}*

No. as in *{No} years*

make-verb as in *{make-verb} {smo's} mind up*

be-verb as in *{be-verb} very surprised*

In the present study, { } signals an obligatory colligation, ( ) signals an optional but a possible part of the collocation, and [ ] brackets the 'frequency figure'.

The research procedure described in this chapter used the computer but also involved a great deal of manual analysis and checking. Although *frequent occurrence* can be determined by the computer, *grammatical well-formedness* and *predictability in the L1* can only be determined manually. Because frequency of occurrence involved the separation of different senses of collocations, this also involved a great amount of manual analysis. The results of all this counting and analysis are reported in the next chapter.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter looks at the results of the corpus search for collocations that meet the criteria of frequency and grammatical well-formedness. Later in the chapter we look at the effects of applying the criterion of predictability in the L1.

4.1 Number and frequency of collocations

In section 4.1, we will first look at the number and frequency of collocations. There are two major questions - 1) how many collocations meet the two criteria of *frequent co-occurrence* and *grammatical well-formedness*? and 2) is the frequency distribution of these collocations similar to that of single words? The answers to these questions could be used for choosing how many collocations we need to teach and learn.

The frequency of a collocation will always be less than the frequency of its least frequent member, thus it is possible that collocations made of high frequency words could in fact be low frequency items. For this reason, we did not expect many collocations to occur with very high frequencies, so we will check to see how many of the collocations would occur among the high frequency words of the language, if no distinction was made between words and collocations. If we consider the cost-benefit advantages that high frequency items

provide, more frequent collocations are more useful because high frequency items have more chances to be used and met. That is, the collocation should be frequent enough to get into the high frequency words of the language.

Next, we will examine whether Zipf's law on frequency distribution for single words is applicable to collocations.

4.1.1 Statistical results and comparison of 10 bands of 100 pivot words

Under the assumption that more frequent collocations are more useful collocations, it is important to know how many high frequency collocations meet the criteria used in the present study.

5,894 collocations were found for the 1,000 most frequent content pivot word types of English. We searched for collocations according to the pivot words, so sometimes different pivot words have some collocations in common. For example the pivot word *tired* has the two collocations *very tired* and *so tired*, but these collocations are also included in the collocations of *very* and *so*. There are 1,196 such overlaps in the list, so when the list is re-sorted alphabetically to get rid of these overlaps, the new list contains 4,698 collocations. If overlaps are not removed, this is an average of over six collocations per pivot word, but as can be seen in Table 4.1 the range across the 1,000 pivot words is very wide. Each one of the 5,894 collocations was examined

individually in this study.

To look at the data in more detail the 1,000 most frequent pivot words were divided into the equally sized groups of 100 according to the frequency of the pivot words. Tables 4.1 and 4.2 show how many collocations each 100 pivot word band includes and how much they account for the total frequencies of the collocations of the first 1,000 pivot words.

Table 4.1
The number and percentage of the collocations
in the 10 frequency ranked bands of 100 pivot words

10 bands of 100 pivot words	No. of collocations	Percent (%)	Cumulative percent (%)
1st 100 words	2,052	34.82	34.82
2nd 100 words	843	14.30	49.12
3rd 100 words	704	11.94	61.06
4th 100 words	564	9.57	70.63
5th 100 words	481	8.16	78.79
6th 100 words	373	6.33	85.12
7th 100 words	281	4.77	89.89
8th 100 words	233	3.95	93.84
9th 100 words	187	3.17	97.01
10th 100 words	176	2.99	100
Total	5,894	100	100

In Table 4.1, we can see that the most frequent 100 pivot words have a total of 2,052 collocations which make up almost 35% of the total

number of the collocations of the first 1,000 pivot words. The first 300 pivot words include more than half of the total number (about 61%). The number and percentages very roughly follow Zipf's law on frequency distribution where the rank of the band times the frequency (or in this case, number of collocations) equals a constant figure (1 times 2,052 = 2,052; 3 times 704 = 2,112 and so on). This means that collocation use is heavily concentrated on the most frequent words. This phenomenon can be more obviously seen in Table 4.2 which gives the number of tokens rather than types.

Table 4.2
The total tokens and percentage
of the collocations of the 10 bands of 100 pivot words

10 bands of 100 pivot words	Total tokens of collocations	Percent (%)	Cumulative percent (%)
1st 100 words	387,634	52.66	52.66
2nd 100 words	111,468	15.15	67.81
3rd 100 words	64,962	8.82	76.63
4th 100 words	44,896	6.10	82.73
5th 100 words	36,903	5.01	87.74
6th 100 words	27,846	3.78	91.52
7th 100 words	22,433	3.05	94.57
8th 100 words	16,872	2.29	96.86
9th 100 words	12,615	1.71	98.57
10th 100 words	10,515	1.43	100
Total	736,144	100	100

As shown in Table 4.2, the total number of tokens of the collocations of the first 100 pivot words is 387,634 which account for just over 52% of the total number (736,144) of tokens of the collocations found in this study. The first 200 pivot words make up a little over 67% of the total tokens. The top 200 pivot words account for a very large proportion of the collocations.

4.1.2 Zipf's law on frequency distribution for collocations (bi-grams)

One way to look at the frequency distribution of collocations is to see how Zipf's law on frequency distribution applies to the collocation data in this study.

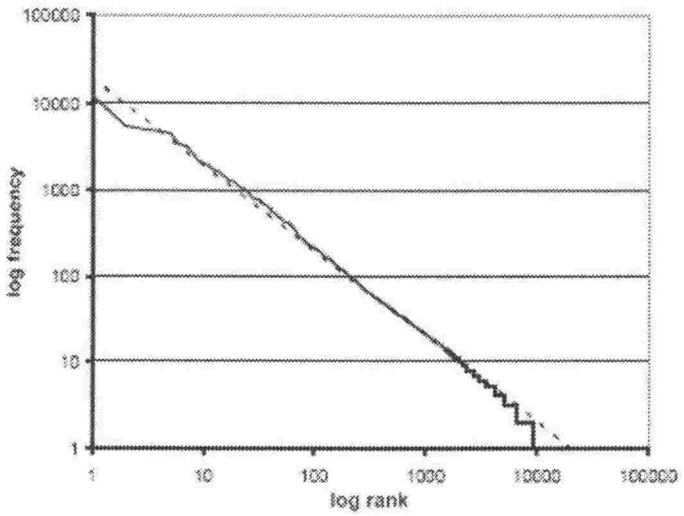
Zipf's law on frequency distribution for individual words is expressed in the following equation (Zipf, 1945, 1949):

$$f * r = C$$

If r is the rank of a word in a list of all the words in a sample in order of decreasing frequency and f is the frequency of the word in the sample, then C is a constant figure (always the same number). For example, Zipf (1935, 1949) found that *the* is rank 1 occurring 1,631 times, *and* is rank 2 occurring 866 times in *Alice in Wonderland.*

According to Zipf's law on frequency distribution, the result of the calculation for *the* which is 1,631 (1,631*1=1,631) should be the same as that for *and* which is 1,732 (866*2=1,732). Both figures are close to each other suggesting that they are near enough to a constant figure. When *log(f)* is drawn against *log(r)* in a graph (which is often called a Zipf curve), a dotted line is drawn with a slope of −1 as in Figure 4.1.

Figure 4.1
Zipf curve for the uni-grams (single words) extracted from a 250,000 word token corpus



(From Ha, Sicilia, Ming, and Smith, 2002, p. 315)

If Zipf's law on frequency distribution applied perfectly to a set of data (it almost never does!), the top ten items would follow a frequency pattern like this:

Rank	Frequency	Rank * Frequency
1	240	240
2	120	240
3	80	240
4	60	240
5	48	240
6	40	240
7	34	240
8	30	240
9	27	240
10	24	240

This concocted example, multiplying rank by frequency always results in a constant figure, in this case 240.

Zipf's law on frequency distribution is typically tested on frequency data involving the frequency of words, but language is made of phrases as well as single words. Phrases of 2, 3 and more words are called n-grams, for n=2, 3, etc. Ha et al. (2002) looked at differences among the occurrences of n-grams, using several English corpora made from the *Wall Street Journal* (Paul & Baker, 1992) from issues published in 1987, 1988, 1989, with corpus sizes of approximately 19 million, 16 million and 6 million tokens respectively. They found some differences with the results of the n-grams using Zipf's law on frequency distribution and suggested a modified Zipf's law on frequency distribution, $f * r^{\beta} = C$. In the three *Wall Street Journal* corpora *(WSJ)*, the average β for bi-grams is 0.65. In the present study, most collocations consist of two-word phrases, so the formula, $f * r^{.65} = C$, was used to examine the constant

figures. As a sample test, the word *up* was examined. 122 of 156 collocations of the pivot word *up* were examined. If a collocation has different parts of speech or different meanings, those are subdivided into separate collocation items, and they are called polysemes. However, 34 polysemes containing *up* were excluded because they did not meet the 30 occurrences of the frequency cut-off point. For example, *follow up* occurs 39 times as a verb phrase and it also occurs 6 times as an adjective. So, the adjective *follow up* does not meet the frequency cut-off point, and thus it was excluded from the analysis. The 122 collocations of *up* consist of 99 bi-grams, 21 tri-grams and 2 four-grams. Most of the collocations are bi-grams, so the formula, $f * r^{.65}=C$ was used. Figure 4.2 shows a Zipf curve for the collocations of *up*, and Figure 4.3 shows Zipf curves for n-grams in the *WSJ87* corpus *(Wall Street Journal 1987* corpus*).*

In Figure 4.3, the dotted line represents the formula, $f * r^{.65}=C$. Note that the curves for shorter units (such as a 1-gram, that is, a single-word unit) have a steeper slope, and the bi-gram curve is more closely parallel to the dotted line with a slope -1. Even though Figures 4.1 and 4.2 are a little different from Figure 4.3 in slope, comparing Figures 4.1 (uni-grams) and 4.2 (bi-grams) confirms that the single words curve has a steeper slope.

Figure 4.2

Zipf curve for collocations of *up*

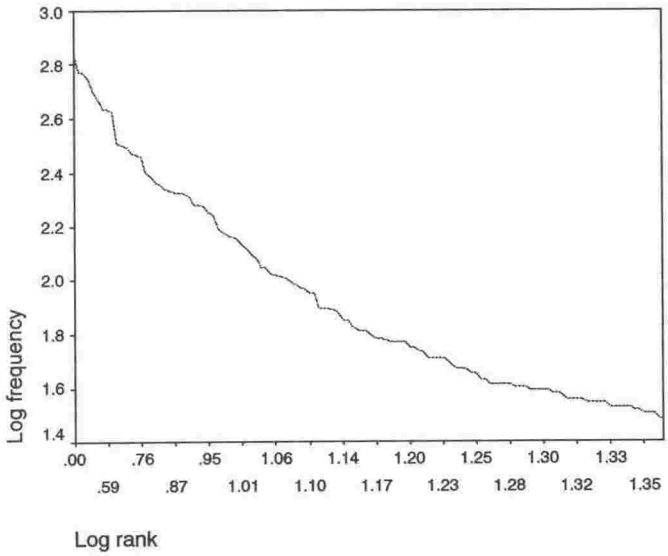
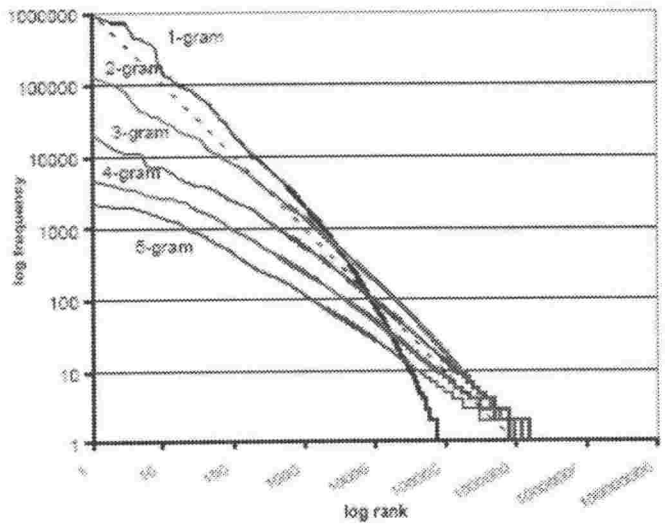


Figure 4.3

Zipf curves for the *WSJ87* corpus



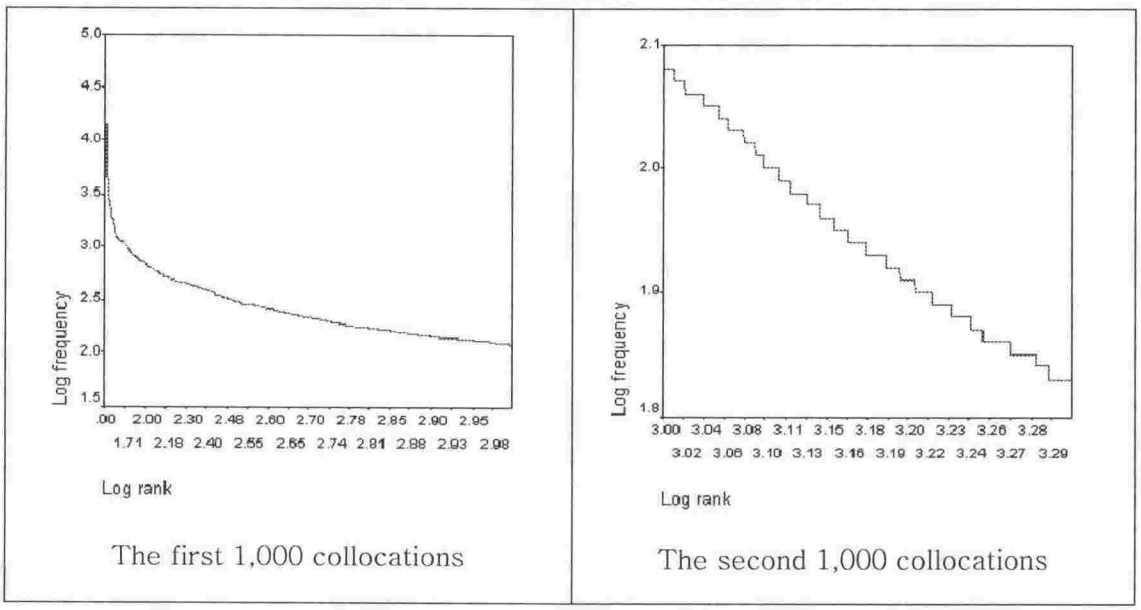
In the analysis of Zipf's law on frequency distribution for collocations in the present study, even though the Zipf curve of each

pivot word does not strictly meet a slope of -1, there is a series of figures that are close to the constant. To get a valid result using Zipf's law on frequency distribution requires many more cases. Therefore, the collocations list re-sorted by frequency regardless of pivot words was used to check this. The list includes 4,698 collocations and four 1,000 collocation bands were examined. The fifth band was excluded from the analysis because it includes a lot of polysemes having less than 30 occurrences. Figure 4.4 gives the results.

Figure 4.4 shows Zipf curves for the four 1,000 collocation bands. The first 1,000 collocation band shows a slope near -1.5 when the two extremely frequent collocations *you know* and *I think* are excluded. The other three bands show slopes close to -0.9, -0.8, and -0.7 each.

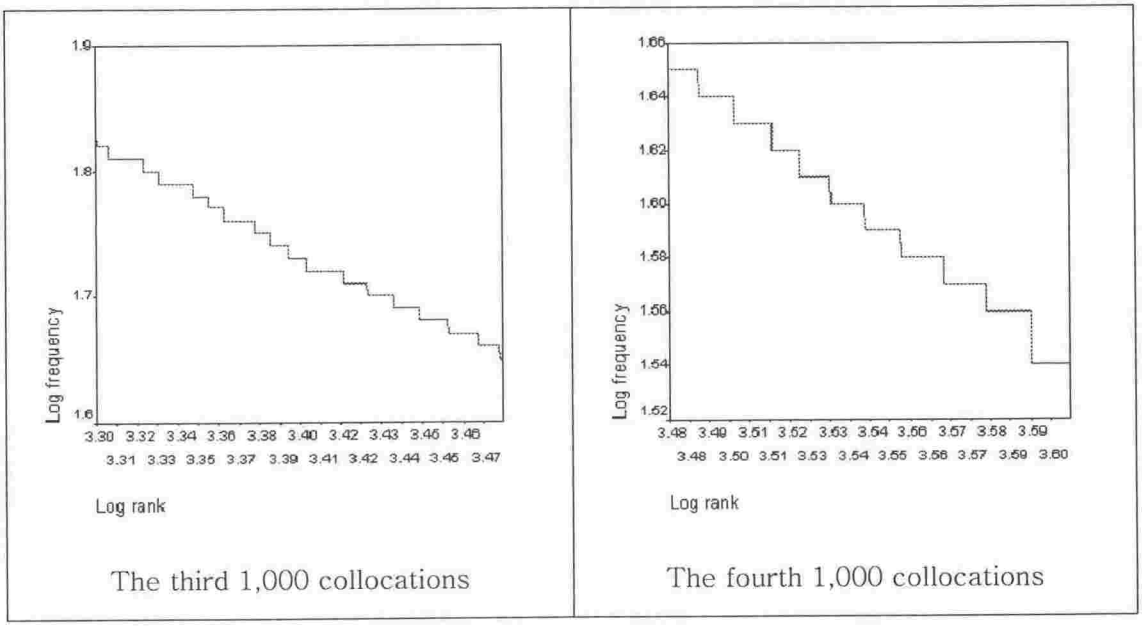
Figure 4.4

Zipf curves for four 1,000 collocation bands



The first 1,000 collocations

The second 1,000 collocations



The third 1,000 collocations

The fourth 1,000 collocations

As the rank goes down, the slope levels out. This means that compared with the predictions by Zipf's law on frequency distribution, more frequent collocations are more frequently used compared with their ranks and at the same time there is a more rapid drop in the frequency of the first 1,000 collocations than Zipf's law on frequency distribution would predict. If the four collocation bands are joined together, the result is Figure 4.5.

Figure 4.5 shows that after the steep section on the left, the slope gradually becomes more gradual. The steep section includes the two extremely frequent collocations *you know* and *I think* and after those two items, the curve representing the other 3,998 collocations shows a slope around -0.9. This means that Zipf's law on frequency distribution fits for a certain section of the collocation list as well as fitting for uni-grams (see Figure 4.1).

Figure 4.5

Zipf curves for the most frequent 4,000 collocations

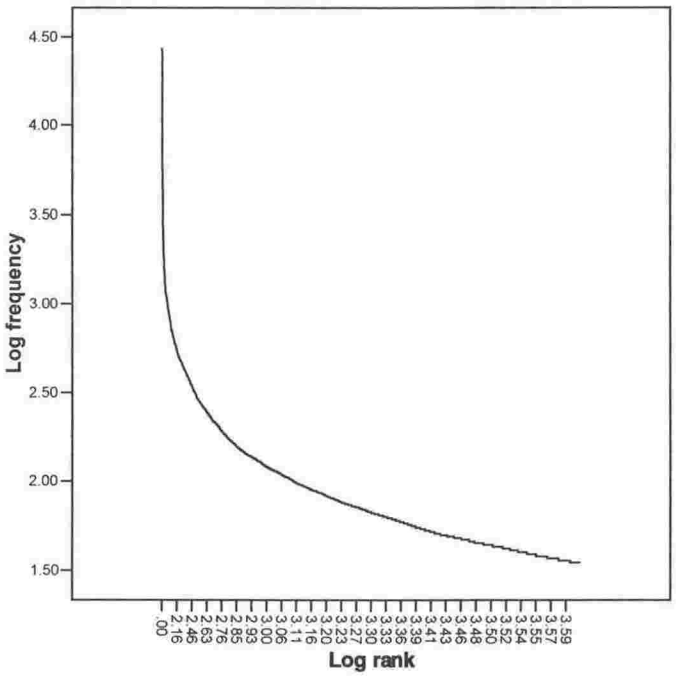


Table 4.3 also confirms the results of Figures 4.4 and 4.5.

As mentioned above, the frequencies of the top two collocations, *you know* (27,348 tokens) and *I think* (25,862 tokens) are well outside the frequency range of the other items, even compared with the other top collocations such as *as well* (5,754), *thank you* (4,789 tokens), and *in fact* (3,009). *You know* and *I think* are interactional markers and are clearly typical of spoken language.

Table 4.3

Some high and low frequency collocations

Rank	Collocation	Frequency
1	*you know*	27,348
2	*I think (that)*	25,862
3	*a bit*	7,766
4	*(always [155], never [87]) used to {INF}*	7,663
5	*as well*	5,754
6	*a lot of {N}*	5,750
7	*{No.} pounds*	5,598
8	*thank you*	4,789
9	*{No.} years*	4,237
10	*in fact*	3,009
4411	*say about that*	30
4412	*saying anything*	30
4413	*social housing*	30
4414	*stand back*	30
4415	*stand here*	30
4416	*start up (sth)*	30
4417	*starting off (with sth [8])*	30
4418	*sure about that (S V)*	30
4419	*take {sth} on board*	30
4420	*that ball*	30

The lists of collocations show the same rapid drop that is found in frequency ranked lists of single words.

Zipf's law on frequency distribution (f*r=c) seems to apply as well to collocations as it does to single words. In practical terms, this law says that in a frequency ranked list, a relatively small number of highly frequent types will account for the majority of the tokens. A very large

number of infrequent types account for the remaining small number of tokens. From a teaching and learning perspective, greater value will be gained from focusing on the small number of high frequency items than on the large number of low frequency items.

4.1.3 Inclusion of collocations in a list of high frequency words

To see how the high frequency collocations were ranked compared to word types, the collocation frequency list was compared with a frequency list which contains the BNC spoken frequency figures by Leech et al. (2001), available at http://www.comp.lancs.ac.uk/ucrel/bncfreq/. The frequencies in the list show occurrences of a word type per 1,000,000 running words. However, the list only provides data for the first 4,000-odd non-lemmatised words. As mentioned in previous sections of this thesis, the collocation list is not lemmatised, that is, we used word types to search for collocations using the 10,000,000 BNC spoken corpora. Let us look at Tables 4.4 and 4.5.

Table 4.4 shows cut-off figures from the BNC according to spoken type-based single word figures. The last word of the 1st 1,000 list is the noun *terms* which has 76 occurrences in one million running words, or 760 per ten million running words.

Table 4.4

Cut-off figures from the BNC

according to spoken type-based single word figures

	Spoken type-based single word frequency (/million)
1st 1,000	76 (≈ 760/10 million) - (*terms* (Noun))
2nd 1,000	32 (≈ 320/10 million) - (*beat* (Verb))
3rd 1,000	19 (≈ 190/10 million) - (*trading* (Noun))
4th 1,000	13 (≈ 130/10 million) - (*hurts* (Verb))

Table 4.5

Cut-off figures from the BNC

according to spoken type-based collocation figures

	Spoken type-based collocation frequency (/10 millions)
1st 100	689 (*(be-verb) interested in {sth}*)
2nd 100	454 (*went down*)
3rd 100	325 (*get out of {sth}*)
4th 100	257 (*found that (N, S V)*)
5th 100	214 (*very high*)
1st 1,000	118 (*used it*)
2nd 1,000	64 (*support that (N)*)
3rd 1,000	43 (*(just [11]) one example*)
4th 1,000	33 (*full up*)

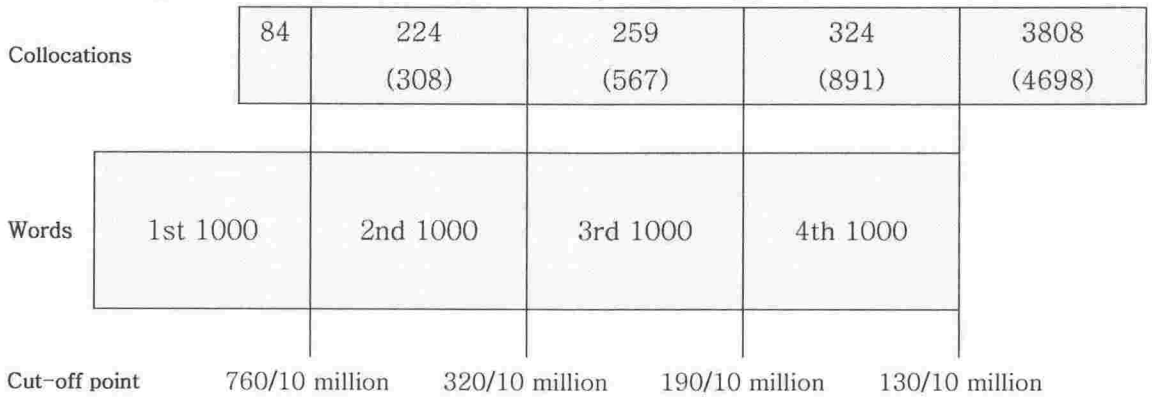
Table 4.5 shows cut-off figures from the BNC according to spoken type-based collocation figures. The last collocation of the 1st 100 list is *(be-verb) interested in {sth}* with 689 occurrences from the ten million running words.

Some argue that collocations should be treated as if they were like single words, that is, units to be learned as a whole item. If we accept this idea, it is interesting to see how the frequencies of these collocational units compare with the frequencies of single words.

If we compare Table 4.5 with Table 4.4, the last collocation of the 3rd 100 list of collocations is *get out of {sth}* with 325 occurrences and this would include it in the 2nd 1,000 list of frequency ranked word types in Table 4.4. The frequency cut-off point of the 1st 1,000 list of the collocations is the frequency of 118 occurrences, which almost matches the cut-off figure of the fourth 1,000 spoken type-based single words (≈ 130/10 millions). This means that the top 1,000 collocations would all qualify for entry into the top 4,000 words of English using a frequency criterion. Figure 4.6 gives a diagrammatic representation of the occurrence of single word types and collocations.

Figure 4.6

Frequency comparison between single word types and collocations



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

Figure 4.6 shows that 84 collocations meet the cut-off point of 760 occurrences per 10 million and thus these are included in the level of the first 1,000 word types. 308 collocations are included in the first 2,000 corresponding to the cut-off point of 320 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, any more, last year, very nice*, etc. The 224 collocations of the second collocation band include *I see, I bet, in a minute, go away, at the end of the day, up here*, etc. The 324 of the third collocation band are *manage to {INF}, in the world, 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, piss off, stand back*, 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.

4.2 Factors affecting the number and frequency of collocations

In section 4.2, we will examine factors affecting the number of collocations. There are four factors - 1) the frequency of the pivot words, 2) the length of the collocations, 3) part of speech, and 4) the location of the collocates. It is useful to look at these factors because they may allow us to predict what items are likely to have a large

number of collocates without having to count them.

The first factor of the frequency of the pivot words has two sub-divisions - (1) the number of collocates, and (2) the frequency of collocations. We will first look at whether high frequency pivot words have more collocates than lower frequency pivot words, and then we will check whether high frequency pivot words occur in more frequent collocations. In addition, we will examine how many collocations are made up of high frequency words and the differences between the frequencies of spoken and written collocations. The second factor of the length of collocations is related to Zipf's law of least effort (shorter words are more common than longer words). This Zipf's law is usually applied to the number of phonemes or syllables of a single word, but we will check whether Zipf's law of least effort is applicable to the number of words making up a collocation.

The factor of part of speech is concerned with three questions - (1) pivot words of which part of speech have more collocates?, (2) which combination of parts of speech of collocates is the most common?, and (3) which part of speech of a whole collocation is the most common?

For the final factor of the location of collocates, we will look at the number of collocates on the right, on the left, and on both sides of the pivot word.

4.2.1 The frequency of the pivot word and the number of collocates

First we will examine the relationship between the frequency of the pivot word and the number of collocates.

This study focuses on the top 1,000 word types as these are the most problematical and time-consuming to analyse. Each band of 100 words (10 bands) in the first 1000 types was examined and compared. When we consider some randomly selected sequential words from the pivot word list, there seems to be no obvious relationship between the frequency of a pivot word and the number of its collocates (see Table 4.6). However, from a broader view, when each band of 100 pivot words is compared, we can more readily see a decrease in the number of collocates as pivot word frequency goes down (see Tables 4.1 and 4.2).

Table 4.6
The number of collocates of the top 10 pivot words

Rank	Pivot word	Frequency (/million)	No. of collocates
1	well	5,912	12
2	know	5,550	11
3	so	5,151	16
4	think	3,977	8
5	just	3,820	11
6	right	3,356	17
7	up	2,891	156
8	go	2,885	43
9	now	2,864	12
10	said	2,685	7

As shown in Table 4.6, the top 10 pivot words ordered by frequency do not show any systematic relationship between frequency and the number of collocates. *Think* which occurs 3,977 times per million in the 10,000,000 word corpus has 8 collocates, while *up* occurs 2,891 times per million but it has 156 collocates. One reason why the relatively lower frequency word *up* has more collocates than *think* which is a higher frequency word is that such words as *up*, *out*, *over* and *in* occur in a lot of phrasal verbs (e.g. *pick up*, *carry out* and *take over*).

To make sure that phrasal verbs were not overlooked in this study, the phrasal verbs were double-checked by using Fletcher's (2003/2004) database of Phrases in English which includes all phrases with a length between 1 and 8 words which occur more than 3 times in the BNC. Even though Fletcher's database (http://pie.usna.edu) includes a lot of grammatically ill-formed phrases because he collected the data by using only a computational process, we found 14 phrasal verbs which we had missed, and these were added to the study. The most frequent of these were *woke up* (93 tokens), *picks up* (59 tokens) and *stood up* (40 tokens).

When we look at another pivot word *go* in Table 4.6, there seems to be no obvious relationship between the frequency of a pivot word and the number of its collocates. The word *go* also has more collocates than *think* even though *think* is a more frequent word. However, as seen in Tables 4.1 and 4.2, frequency can have an influence on the number of collocates, but there are clearly other important factors that affect the number of collocations a pivot word can occur in. Before looking at

these, let us look further at the frequency distribution of the collocations found.

4.2.2 What proportion of the words making up the top 4,698 collocations of English are from the high frequency words?

One of the purposes of the present study is making the best use of already known words, that is, not adding an additional burden by adding unknown, lower frequency vocabulary. Thus, we examined how many collocations are made up of high frequency words.

The frequency level of all the individual words which make up 4,698 collocations were examined by using the RANGE program (Heatley, Nation and Coxhead, 2002). Three ready made base word lists are available with the RANGE program. The first base word list includes the 1st 1,000 most frequent word families of English. The second base word list includes the 2nd 1,000 most frequent word families. The sources of these two lists are the General Service List (GSL) by West (1953). The third base word list includes 570 word families not in the first 2,000 words of English but which are in the Academic Word List (AWL) by Coxhead (1998, 2000).

As a result of this analysis, it was found that the most frequent collocations are largely made up of high frequency words (see Table 4.12).

Table 4.7

Members of the top 1,000 collocations

	No. of word families	Cumulative (%)
1st 1,000	392	392 (82%)
2nd 1,000	55	447 (93%)
AWL	21	468 (98%)
Not in the lists	11	479 (100%)

Table 4.7shows that the top 1,000 collocations are made up of 479 word families and about 93% (=447) of them are included in the 2,000 word GSL. Very high frequency members include *the, in, of* and *to*. If the first two thousand word families of English are considered the high frequency words of the language (West, 1953; Nation, 2001b), then 956 of the top 1,000 collocations are made up of these 2,000 word families. A further 32 collocations contain twenty-one academic words and the remaining 12 collocations contain some words which are from neither the GSL nor the AWL. Table 4.8 lists the twenty-one academic words.

Table 4.8

21 academic words contained in the first 1,000 collocations

access, area, authority, available, aware, benefit
couple, involve, issue, job, labour, odd
paragraph, percent, period, prime, role, sector
similar, structure, transport

As shown in Table 4.8, only 21 academic words such as *couple, period* and *issue* were found in the top 1,000 collocations. Table 4.9 lists the 12 collocations containing 11 words (the word *county* was used in

two collocations in the second column) which were not included in either the 2,000 word family GSL or the 570 word family AWL. This is largely because the GSL has some limitations. These mainly occur for the following reasons. First, the GSL is based on written texts, so it is not sufficiently relevant to spoken texts.

Table 4.9
12 collocations from the first 1,000 collocations list
which are not in either the GSL or the AWL

Rank in the first 1,000 collocation list	12 words not from the first 2,000 high frequency words		Collocations
30	county	(166)	*county* council
86	bet	(498)	I *bet*
204	reckon	(546)	I *reckon*
225	hell	(395)	bloody *hell*
270	awful	(531)	an *awful* lot (of sth)
452	poll		*poll* tax
525	reform		land *reform*
527	fed		*fed* up
722	guy	(757)	this *guy*
773	county	(166)	this *county*
795	fucking		*fucking* hell
900	television	(622)	on *television*

- The number in brackets shows the frequency rank of each word in the first 1,000 pivot word list from the BNC spoken section.

The collocation list is from the BNC spoken section and 7 words such as *bet, awful,* and *guy* of the 11 words were included in the first 1,000 content words from the BNC spoken section. In addition, the word *fucking* was intentionally excluded from the 1,000 content pivot word

list for educational reasons, but it came in as a collocate of *hell*. Second, the GSL was made several decades ago, so it does not include new words reflecting recent changes such as *television*. Third, the BNC is also strongly British, which brings in words like *county, pence, poll* and *steward*. Fourth, the definition of word family used in the GSL is a little narrower than the criteria used in the BNC word families lists (Nation, 2004). Table 4.10 shows a comparison of the 2,000 word family list from the BNC spoken corpora and the 2,000 word GSL.

Table 4.10
Comparison of the 2,000 word list
from the BNC spoken corpora and the 2,000 word GSL

	Word types from the 2,000 word family list of the BNC spoken corpora	Word types from the word family GSL
Overlapping types	5,864 (e.g. *ability, ask, burn*, etc)	5,864 (e.g. *ability, ask, burn*, etc)
Unique types	6,091 (e.g. *abuse, income, beer*, etc)	1,987 (e.g. *abroad, yield, widow*, etc)
Total	11,955	7,851

As shown in Table 4.10, the first 2,000 BNC word family list contains 11,955 types, while the 2,000 GSL word family list has 7,851 types. That is, the BNC word list contains 4,104 more types than the GSL. The two lists have 5,864 word families in common. The third row of Table 4.10 is a good example of the difference between the two lists. The first 2,000 BNC word family list contains 6,091 word types which are not included in the 2,000 GSL word family list. In the total 4,698

collocations there were only 93 word families not in either the GSL or the AWL. Table 4.11 shows the 93 word families.

Table 4.11 shows that, in the 4,698 collocations, there were 93 word families not in either the GSL or the AWL.

Table 4.11

93 word families contained in the top 4,698 collocations which are not in either the GSL or the AWL

abuse, agenda, alright, anthem, auditor, awful
aye, bet, bitch, bloke, booklet, boot
borough, bother, boxing, brigade, brilliant, budget
bun, careers, cash, chap, chip, client,
congress, conservation, contingency, county, curriculum, daddy
darling, executive, fed, forecast, fortnight, frankly
fuck, funnily, glazing, golf, gracious, greenbelt
grief, guy, haircut, hell, horrible, jolly
keen, kettle, kids, lad, lousy, madam
mayor, mummy, obviously, okay, olds, opt
parish, parliament, payers, peasant, penalty, pence
piss, plastic, poll, premier, pub, pussy
quid, reckon, recorder, reform, rural, server
session, silly, steward, surveyor, switch, television
tiny, traffic, tremendous, unionists, urban, vast
video, wee, yeah

- The words in Table 4.11 are word types but represent word families which the word types belong to. For example, the two types *booklet* and *booklets* were found in the collocations, but only one type *booklet* was included in Table 4.11.

In the top 4,698 collocations these 93 word families had 108 word types. 92.38% of the types (100 minus 7.62) and 90.63% of the families (100 minus 9.37) in the 4,698 collocations were included in the GSL or

the AWL. These results show that not only the 1,000 pivot words but most of their collocates are also high frequency words.

Table 4.12

Tokens, types, and families of all the word members of the 4,698 collocations

Word list	Tokens	Types	Families
GSL 1st 1,000	11,875 (91.34%)	998 (70.43%)	642 (64.65%)
GSL 2nd 1,000	653 (5.03%)	205 (14.47%)	168 (16.92%)
AWL	288 (2.23%)	106 (7.48%)	90 (9.06%)
Not in the lists	182 (1.40%)	108 (7.62%)	93 (9.37%)
Total	13,002 (100%)	1,417 (100%)	993 (100%)

This interest in the words that make up the collocations is important because it is an effective way of seeing how easy the learning of collocations could be. If the collocations are made of known items, learning will be much easier.

4.2.3 Spoken collocations versus written collocations

There is plenty of evidence that there are significant differences between spoken language and written language (Biber, 1989; Halliday, 1985). Are these differences also found in collocations? To find this out, a comparison was made between two equally sized corpora, the BNC spoken section and a similar sized written corpus.

The collocations were found by searching in two collections of modern English text. These two corpora were (1) the spoken corpus of

around 10,000,000 running words from the BNC used in other parts of the research, and (2) a written corpus of around 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 BNC. The written corpora include newspaper editorials, reviews and news items, magazines, articles, learned and scientific texts, and fiction.

The comparison was made by comparing the 50 most frequent collocations in each of the two corpora.

The most striking finding of the spoken/written comparison was the very big difference between the results from the spoken corpus and the results from the written corpus. The difference is of two kinds, (1) the different items in the lists, showing that spoken use and written use favour different collocations, and (2) the very high frequency of the collocations in the spoken corpus.

Only fifteen collocations occur in both the top 50 spoken and top 50 written lists. They are listed in Table 4.13.

A few of these overlapping collocations (e.g. *you know, I think,* etc) seem much more typical of spoken rather than written English, and the overlap may be caused by the fact that about one fifth of the written corpus was from novels and newspapers which both included a lot of direct speech. It is clearly difficult to make an inclusive written corpus that does not include representations of spoken language. The 35

spoken items and the 35 written items that did not overlap in the two top 50 lists are not uniquely spoken or written. All 70 of them did occur in both corpora.

Table 4.13
The 15 items occurring in both the spoken and written lists

Rank	Spoken collocations	Freq	Rank	Written collocations	Freq
1	*you know*	27,348	15	*you know*	1,074
2	*I think (that)*	25,862	4	*I think (that)*	1,565
4	*as well*	5,754	22	*as well*	917
8	*(for) {No.} years*	4,237	9	*(for) {No.} years*	1,287
9	*in fact*	3,009	3	*in fact*	1,679
10	*very much*	2,818	29	*very much*	823
23	*(Det-the [86]) last year*	1,347	17	*(Det-the [58]) last year*	1,027
24	*so much*	1,334	10	*so much*	1,238
25	*{No.} years ago*	1,314	11	*{No.} years ago*	1,233
27	*this year*	1,255	13	*this year*	1,156
29	*last night*	1,244	36	*last night*	755
30	*{Det-the [1178], a [30]} fact that {S V}*	1,208	5	*{Det-the [1351], a [28]} fact that {S V}*	1,482
36	*at the end (of sth [737])*	1,122	23	*at the end (of sth [797])*	877
44	*too much (N)*	1,034	35	*too much (N)*	764
50	*(Det-the [47]) last week*	956	50	*(Det-the [18]) last week*	564

However, the here-and-now nature of spoken language is reflected in items like *this morning, at the moment, last night,* and *over there*, and the personal and interactional nature is reflected in items like *thank you, thank you very much, you know, I think*, and *come in*. The

written list contains items that can act as conjunctions like *as well as, just as, even if,* and *even though.*

The data shows that among the most frequent collocations there are substantial differences between written and spoken English.

The overwhelming feature however is the enormous difference in the frequency of the items. There are several ways of showing this. Although the total number of items meeting the various criteria were virtually the same in both corpora (2,261 in the spoken corpus and 2,266 in the written corpus), the top 50 spoken collocations occurred 147,217 times, while the top 50 written collocations occurred only 48,782 times. That is, the top 50 spoken collocations occurred almost three times as often as the top 50 written collocations. Spoken language makes much more frequent use of its common collocations than written language does.

Without exception, each collocation in the frequency ranked spoken list in Appendix 3 is much more frequent than the collocation at the same rank in the written list. Table 4.14 lists some points of comparison. As Table 4.14 shows, items in the spoken list are 50% to 100% more frequent than the corresponding items in the written list. Table 4.13 shows that where a collocation occurs in both lists, it is always more frequent in the spoken corpus.

Table 4.14

Some examples of frequency differences

of the top 50 spoken and top 50 written collocations

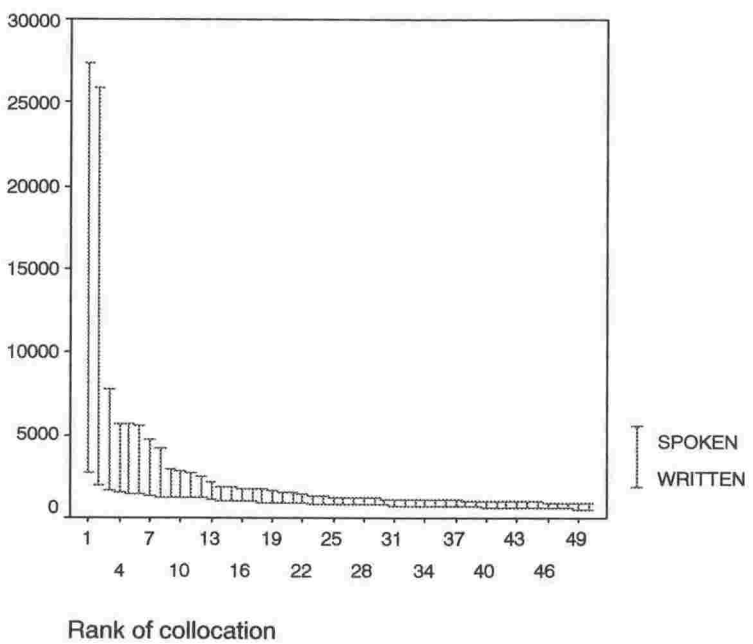
Rank	Spoken		Written	
1	*you know*	27,348	*of course*	2,698
10	*very much*	2818	*so much*	1,238
20	*come in (swe)*	1571	*say that (N, S V)*	932
30	*{Det-the [1178], a [30]} fact that {S V}*	1,208	*(the) only one (N)*	796
40	*{FREQUENCY, QUANTITY) a week*	1,056	*any other {N}*	672
50	*(Det-the [47]) last week*	956	*(Det-the [18]) last week*	564

Figure 4.7 makes this comparison in a graphic form.

Figure 4.7

Frequency comparison

between the top 50 spoken and written collocations



The figure is made up of fifty vertical lines with a cross bar at the top of each line showing the frequency in the spoken corpus, and the cross bar at the bottom showing the frequency in the written corpus of the collocation at the same ranked position. So the vertical line at the extreme left represents the top ranked collocations in both corpora, *you know* in the spoken corpus and *of course* in the written corpus. The gap between the frequencies of these two collocations (27,348 and 2,698) is very big, so the vertical line is long. As we move down to rankings (moving from left to right in Figure 4.7), the vertical line becomes shorter but there is always a gap.

The final way we will use to show the striking frequency difference between spoken and written collocations is to consider what would happen if the two 50 item lists were mixed together to make one list and then ranked according to frequency. If this was done, the top eleven collocations would all be from the spoken list. The top written collocation would come in at rank 12 on the list. The final thirty-two items on the combined list would all be from the written corpus. The high frequency spoken collocations occur much more frequently than the high frequency written collocations.

4.2.4 Zipf's law of least effort for 2-word, 3-word, and 4-word collocations

Zipf (1935) measured word length by syllables (for Chinese and

Latin) and phonemes (for English newspapers) and he found short words are much more common than long words. This is called "the law of least effort" because it says that we try to reduce the effort of language production by making sure the items we use often are short and thus easy to produce. This law is not restricted to the syllables or phonemes of a word. Zipf's law of least effort may also be applied to collocations which consist of two or more words. It is consistent with the findings about collocational frameworks by Renouf and Sinclair (1991). Renouf and Sinclair showed that short collocations are much more frequent in text than longer ones. This is the same for the present study. The total numbers of 2-word, 3-word, 4-word, 5-word, and 6-word collocations in this study are listed in Table 4.15.

Table 4.15
The total number of n-gram collocations

N-gram	No. of collocations	Percent
2 word collocations	3,616	76.97
3 word collocations	956	20.35
4 word collocations	112	2.38
5 word collocations	12	0.26
6 word collocations	2	0.04
Total	4,698	100

- Optional collocates were not counted for collocation length in Table 4.15. For example, *(in [12]) different places* was regarded as a two word collocation.

The results show that the number of two-word collocations is 3,616 and that makes up about 77% of the total number of 4,698. Two and three word collocations make up about 97% of the total number of

collocations.

In addition, we examined if collocations containing a short word are more frequent than collocations containing a long word. For this, the top 100 collocations and the 100 collocations from the bottom of the frequency list were compared. When choosing the 100 collocations from the bottom of the 4,698 collocation list, 261 polysemes were excluded to avoid overlaps, so the 100 items frequency ranked 4,339 to 4,438 were chosen. The number of collocations was counted according to the number of characters making up the longest component of each collocation. Tables 4.16 and 4.17 show the results.

Table 4.16
The classification of the top 100 collocations based on the number of characters of the longest component of a collocation

No. of characters making up the longest component	No. of collocations	
3	9	(e.g. *a **bit***)
4	49	(e.g. *come on*)
5	19	(e.g. *last **night***)
6	10	(e.g. *not **really***)
7	12	(e.g. *for **example***)
10	1	(e.g. *interested in sth*)

Table 4.16 shows six different types of the top 100 collocations based on the number of characters making the longest member of a collocation. As shown in Table 4.16, collocations containing a short word are more frequent than collocations containing a long word except for three-character collocations. After nine collocations containing a

three-character word, the number of collocations decreases in inverse proportion to the number of characters of the longest component making up a collocation. The collocations containing a four-character word are the most common (49) from the top 100 collocations.

Table 4.17
The classification of the bottom 100 collocations based on the number of characters of the longest component of a collocation

No. of characters making the longest component	No. of collocations	
3	2	(e.g. ***pay*** *up*)
4	25	(e.g. ***very*** *fair*)
5	27	(e.g. ***stand*** *back*)
6	19	(e.g. *from the* ***bottom***)
7	15	(e.g. ***running*** *round*)
8	8	(e.g. ***starting*** *off*)
9	2	(e.g. *very* ***dangerous***)
10	2	(e.g. *great* ***difficulty***)

In Table 4.17, for the 100 items from the bottom, after the collocations containing a four-character word, the number of collocations also decreases in inverse proportion to the number of characters of the longest component making up a collocation. However, in the bottom items, we can see more collocations containing a relatively longer word compared with the top 100 collocations shown in Table 4.16. While almost a half (49) of the top 100 is concentrated on the collocations containing a four-character word like ***come*** *on*, the collocations containing a five to ten-character word increase in the 100 bottom items, especially, the collocations containing a five-character

word (27) are more common than the collocations containing a four-character word (25). Collocations containing a short word are more common than collocations containing a long word.

Zipf's law of least effort (short items are more frequent than long items) takes us back to looking at the range of factors that affect the frequency and number of collocations. The number of words in the collocations is clearly a very important factor.

4.2.5 Part of speech of the pivot word and number of collocates

In the previous section, we saw that frequency is one of the important factors affecting the number of collocates. Now we will look at the influence of part of speech on the number of collocates. To examine whether part of speech of a pivot word affects the number of collocates, the number of collocates of four different parts of speech was compared.

Table 4.18 shows the relationship between part of speech and the number of collocates.

Table 4.18

Part of speech of the pivot word and the number of collocates

Frequency range (/million)	Part of speech	Pivot word	Frequency (/million)	No. of collocates	No. of collocates / frequency	Pivot word	Frequency (/million)	No. of collocates	No. of collocates / frequency
1252-1819	Noun	time	1,819	59	0.032	way	1,252	36	0.030
	Verb	come	1,737	22	0.013	put	1,640	17	0.010
	Adj	good	1,566	65	0.043	other	1,313	42	0.032
	Adv	down	1,472	71	0.048	in	1,307	16	0.012
603-668	Noun	week	631	19	0.03				.
	Verb	tell	668	1	0.001				.
	Adj	new	603	15	0.025				.
	Adv	too	629	30	0.048				.
529-577	Noun	pounds	572	5	0.009	number	531	15	0.028
	Verb	saying	577	7	0.012	find	529	8	0.015
	Adj	big	549	14	0.026	old	529	21	0.040
	Adv	also	556	1	0.002	over	540	43	0.080
431-479	Noun	night	465	12	0.026	house	460	8	0.020
	Verb	came	473	11	0.023	keep	449	7	0.016
	Adj	sure	479	11	0.023	sorry	431	6	0.014
	Adv	else	475	13	0.027	perhaps	444	0	0
301-367	Noun	children	367	12	0.033	area	356	9	0.025
	Verb	feel	364	8	0.022	told	362	1	0.003
	Adj	great	363	7	0.019	able	339	3	0.009
	Adv	please	361	0	0	maybe	301	2	0.007
280-299	Noun	name	299	3	0.01	government	295	6	0.020
	Verb	leave	289	5	0.017	working	282	10	0.035
	Adj	bad	296	8	0.03	important	280	12	0.043
	Adv	certainly	299	2	0.007	obviously	296	1	0.003
179-194	Noun	system	189	4	0.021	shop	188	7	0.037
	Verb	bought	194	8	0.041	seems	186	6	0.032
	Adj	fine	187	1	0.005	high	185	10	0.059
	Adv	absolutely	183	4	0.022	later	179	7	0.039

	Noun	Verb	Adjective	Adverb

No. of collocates/part of speech	195 (2)	111 (4)	215 (1)	190 (3)
No. of the top rank items in no. of collocates	3 (3)	1 (4)	5 (1)	4 (2)
No. of the top rank items in no. of collocates/frequency	3 (3)	1 (4)	5 (1)	4 (2)

● The numbers in brackets in each row show the rank in each category.

As shown in Table 4.18, we analysed 52 pivot words making 13 groups. Each group consists of four words representing four different parts of speech; Noun, Verb, Adjective, and Adverb. To control for the effect of the frequency, the frequency of the four words of each group was matched. That is, the four words in each group have a similar frequency. Where possible, each band is represented by two sets of four words. In the first band, the range of frequency is 1,256 to 1,819 (/million). This means the least frequent word in this band is *way* with a frequency of 1,256 and the most frequent word is *time* with a frequency of 1,819, which is a big interval compared with the 179 to 194 range of the last band. This is because it is not possible to find a four word set with a closer frequency from each part of speech. In addition, the frequency interval becomes rapidly larger as we move up the frequency scale, so the number of collocates per frequency band was also examined to avoid the effect of the frequency interval in the last column. The reason for controlling for frequency was so that the effect of part of speech could be isolated. The bottom section of Table 4.18 gives the results. The 13 adjectives examined have 215 collocates, which shows adjectives have more collocates than other parts of speech. The number

of collocates of verbs is 111, which is the smallest number among the four parts of speech. The results for adjectives, nouns and adverbs are rather similar. To make sure, the number of collocates and the number of the collocates per frequency level of each part of speech were counted. Both results are exactly the same. Five adjectives of the thirteen word sets occupy the top rank, four adverbs, three nouns, and one verb. The only difference from the number of collocates per part of speech is that adverbs have one more top item than nouns. As a result, adjectives are likely to have more collocates, while verbs have fewer collocates. Another finding is that the adverbs *down* and *over* have more collocates than other adverbs because a lot of phrasal verbs include those adverbs, and sentence adverbs such as *please, perhaps*, and *maybe* have few or no collocates. Part of speech of the pivot word thus is a factor affecting the number of collocates, but the major contrast is between adjectives, nouns and adverbs in one group with verbs in the other group. We have to be cautious in generalizing from this small set of data, but this data set tries to control the important variable of frequency range when looking at relationship between the part of speech and number of collocates.

4.2.6 Collocation patterns and the number of collocations

Another aspect of part of speech is how part of speech affects the type of collocation in terms of possible content/function word

constructions. For this, the types of combinations that make up the collocations of the first 1,000 content words are examined to see their effect on the number of collocations. 29 patterns were found from 4,698 collocations. Two item combinations include *content word+ content word* (C+C), *content word+ function word* (C+F), and *function word+ content word* (F+C). Three item combinations include *function word+ function word+ content word* (F+F+C), *content word+ function word+ content word* (C+F+C), and *function word+ content word+ content word* (F+C+C). The other combinations account for only a small number of the collocations found. Table 4.19 shows some major combinations of collocations based on the content word/function word distinction.

Table 4.19
Word combination patterns
of the collocations of the first 1,000 content pivot words

N-gram	No. of collocations	Word combination type
2 word collocation	2,039	C+C
	835	F+C
	742	C+F
Total	3,616	
3 word collocation	355	F+F+C
	250	C+F+C
	147	F+C+C
	84	F+C+F
	50	C+F+F
	42	C+C+F
	28	C+C+C
Total	956	

- Most of the collocations consist of 2 and 3 word collocations, so Table 4.19 focused on the two types of collocations.

As shown in Table 4.19, the combination of content word+ content word (C+ C) is the most dominant collocation pattern. This pattern includes collocations like *last year, really nice,* and *go back.* That pattern makes up 2,039 (56%) of the 3,616 two word collocations and 43% of the total (4,698) collocations in the present study.

In the three item collocations, the F+ F+ C and C+ F+ C patterns are the most common. These include for example, *for a minute, in the case,* and *sorts of things.* These two types make up 605 (63%) of the 956 three word collocations. The number of 4-6 item collocations is very small, and the pattern F+ C+ F+ C as in *a pair of shoes* is most frequent of the 4 item collocations. To a large degree, the decision to count only content word pivot words has influenced the results in Table 4.19. If *the,* for example, was treated as a pivot word, the F+ C category would have a very large number of collocations.

Content word plus content word collocations outnumber other patterns of content word collocations. This is a positive finding in the present study because we aimed at searching for meaningful units and the content word plus content word pattern is arguably the most meaningful unit.

4.2.7 Part of speech of the collocations and the number of collocations

In the previous section, we found that adjectives are likely to have more collocates. Now we will look at part of speech of the collocations as whole units and we will check whether this result matches with the results from the previous section. The number of collocations classified by part of speech of the whole unit was examined. Table 4.20 gives the results.

Table 4.20
The number of collocations per part of speech

Part of speech	No. of collocations
VP	1,734
NP	1,498
PP	649
AP	330
AVP	190
CA	114
AVP/NP	60
INT	46
AP/AVP	36
C	15
C/P	10
AP/NP	9
AP/AVP/NP	3
AP/NP/VP	1
AP/VP	1
AVP/PP	1
INT/NP	1
Total	4,698

- VP=Verb Phrase (or Predicate), NP=Noun Phrase, PP=Prepositional Phrase, AP=Adjective Phrase, AVP=Adverb Phrase, CA=Clause (or Sentence), C=Conjunction, P=Preposition, and INT=Interjection. Slash (/) categories like AP/AVP show a collocation which is used as more than one part of speech. For example, *much more* can modify both nouns and adjectives (or verbs), so it can be used both as an adjective and as an adverb (e.g. *I think* ***much more*** *discourse is required, I suppose it's always* ***much more*** *expensive to do*, etc)

As shown in Table 4.20, verb phrase collocations were the most common type in the present study with 1,734 occurrences. This contrasts with the figures for pivot words where verbs had the fewest collocates. Next were noun phrases with 1,498 items. These two types of collocations make up 3,233 (69%) of the total of 4,698 collocations. Table 4.21 shows why the frequency of the pivot word categories in Table 4.20 does not match the frequency of the part of speech of the whole collocations. Five example words are examined covering the four content word parts of speech.

As shown in Table 4.21, the verb *went* has 19 collocates and the 19 collocations containing the word *went* are all verb phrases. The adverb *up* has 156 collocates, but 131 of the 156 collocations are phrasal verbs containing the word *up*. So the 131 collocations are classified as verb phrases. The adjective *good* has 65 collocates, but 35 of the 65 collocations are noun phrases. Only 14 collocations remain as adjective phrases. The noun *way* has 36 collocates and 27 of the 36 collocations are noun phrases. Another adverb *now* which is not related to phrasal verbs was examined. *Now* has 12 collocates and 5

collocations are adverb phrases, 3 prepositional phrases, and 2 verb phrases. The results show the two parts of speech of pivot words, verb and noun, keep their part of speech in their collocations while many of the adjectives and adverbs occur in noun and verb phrases.

Table 4.21

Part of speech of pivot words vs. part of speech of collocations

Pivot words	No. of collocates	Part of speech of collocations	
went (Verb)	19	VP	19
up (Adverb)	156	VP	131
		NP	13
		AVP	10
		AP	2
good (Adjective)	65	NP	35
		AP	14
		INT	11
		VP	4
		INT/NP	1
way (Noun)	36	NP	27
		AVP	3
		PP	3
		VP	2
		AP/AVP/NP	1
now (Adverb)	12	AVP	5
		PP	3
		VP	2
		NP	1
		INT	1

The number of adjective and adverb phrases is heavily concentrated on a small number of collocates such as *very* and *good*. Clauses (or sentences) like *how much is it* were relatively few because

longer collocational frames are likely to be less frequent than shorter collocational frames (Renouf and Sinclair, 1991). Only fifty-two (46%) of the total of 114 clause collocations consist of more than 3 words.

The structure of collocations was further examined by looking at the relationship between the part of speech of the collocations and the patterns of combinations of content and function words. Table 4.22 shows the results.

Table 4.22
The number of collocations in relation to
five collocation patterns and part of speech of the collocations

	1. C+C	2. F+C	3. C+F	4. F+F+C	5. C+F+C
VP	817	61	621	11	111
NP	747	367	60	19	78
AP	251	12	29	2	12
AVP	113	21	10	2	24
AP/NP	43	0	0	0	2
INT	28	16	2	0	0
AP/AVP	23	4	1	0	2
CA	7	54	0	14	1
PP	6	274	2	303	18
AP/AVP/NP	2	0	0	0	0
AP/VP	1	0	0	0	0
INT/NP	1	0	0	0	0
C/P	0	1	8	0	0
C	0	3	8	4	0
AVP/PP	0	1	0	0	0
AP/NP/VP	0	0	1	0	0
AVP/NP	0	21	0	0	2
Total	2,039	835	742	355	250
	4,221				

Twenty-nine collocation patterns were found such as C+C (e.g. *much more*), C+C+C (e.g. *thanks very much*), and F+F+C+C (e.g. *to a certain extent*), and just five of these make up 4,222 (90%) of the total 4,698 collocations.

If we look at Table 4.22 which is linked to these five patterns, we can see that the C+C type functioning as a verb phrase accounts for the largest number of different collocations (817). A large number of verb phrases consist of a verb and an adverb and these include a lot of phrasal verbs such as *pick up, turn out*, and *set off*. The second dominant grammatical function (747) is noun phrases which usually consist of an adjective and a noun, or a noun and a noun such as *{Det-a [19]} good reason* and *course work*.

The second major type is the F+C pattern. In this pattern, noun phrases are the most common (367). Most of them consist of a determiner and a noun such as *this man, those people,* and *that case*. The next most frequent (274) sub-type is prepositional phrases which are made of a preposition and a noun such as *in work, in control* and *at college*. For this type, verb phrases are not the dominant part of speech compared with the former two types C+C and C+F. Most of the verb phrases are made of a delexicalised verb which loses its core meaning and a lexical verb such as *get started, make sense*, and *get married* where the delexicalised verbs *get* and *make* do not convey any strong concrete meaning.

The third dominant type is the C+F pattern. For this pattern, verb

phrases were also the most frequent (621). However, the components of the C+F pattern are different from those of the C+C pattern. Most of the verb phrases in the C+F pattern consist of a verb and a pronoun or a verb and a conjunction such as *try it, want one* and *knew whether {S V}*.

There are two dominant patterns for three word collocations. The first one is the F+F+C pattern. Prepositional phrases account for about 85% of the F+F+C pattern. The F+F+C pattern mainly consists of a preposition, and a determiner and a noun such as *in the world, in a sense*, and *on the phone*. The other dominant type of the three word collocations is the C+F+C pattern. The most frequent part of speech for this pattern is a verb phrase consisting of (1) a verb, (2) a determiner and (3) a noun, or (1) a verb, (2) a pronoun (or noun) and (3) an adverb, the so-called 'phrasal verbs', for example, *want a drink, say a word, send it back,* etc.

The most common items are (1) C+C as a verb phrase (e.g. *pick up {smo, sth}*), (2) C+C as a noun phrase (e.g. *(a [435]) good idea*), (3) C+F as a verb phrase (e.g. *managed to {INF}*), and (4) F+C as a noun phrase (e.g. *on holiday*).

4.2.8 Location of the collocates and the number of collocations

An additional analysis involves the location of the collocates in relation to the node, namely, on the left side of the pivot word, on the

right side, or on both sides. Table 4.23 provides this data.

Table 4.23

Location of the collocates of the top 1,000 content pivot words

Location / Number	Collocates on the left	Collocates on the right	Collocates on both sides	Others	Total
No. of collocations	3,161	2,408	319	6	5,894
Percent	53.63	40.86	5.41	0.1	100
e.g.	*very well, all the time*	*right hand, put it down*	*on the way home, on the other hand*	*now now, from time to time*	

- The analysis for Table 4.23 was based on the collocation list ordered by the top 1, 000 content pivot words of English (see Appendix 1).

As shown in Table 4.23, there are more collocates on the left (3,161) than on the right (2,408). There are only 319 collocations with collocates on both sides. However, because the location of the collocates is influenced by the part of speech of the pivot word, small differences in the number of collocates are not meaningful. If a collocation is made up of more than three words, the other words with the exception of the pivot word are counted as one collocate to avoid overlaps while counting. Look at the collocation *put it down*, if *put* is a pivot word, *it down* is a right collocate of *put*. There are six cases that are problems to classify into the three categories, because their pivot words occur twice in each item such as the collocations *now now* and *from time to time*. Not surprisingly, the location of the collocates is closely related to the part of speech of the pivot words as shown in Table 4.22. For example, if the pivot word is a noun, for collocates on

the left the three dominant parts of speech are adjectives, verbs and prepositions, as in *more time, had time* and *in time.*

We have now looked at a range of factors affecting the number of collocates a pivot word has. These have included the size of the collocations (two word collocations are by far the most common), the part of speech of the pivot word (adjectives have the most collocates), the grammatical function of the collocation (verb phrases are marginally the most common), and the pattern of the collocates (C+C is the most common). The most frequent collocations then are those like *make sure, go back* and *show up*, which combine most of these features.

4.3 Results of predictability in L1

Section 4.3 involves the test of the criterion of *predictability in L1.* In the present study, the criterion of *predictability in L1* is restricted to Korean, but the procedure of the test may be applicable to another language. This criterion could be effective in reducing the number of collocations to focus on. If there is a parallel L1 construction of a collocation of English, it would not be problematic even if learning single words and then combining these words in a grammatical frame. By using the criterion of *predictability in L1,* it could be possible to exclude predictable items from the collocation list or at least mark them as items requiring little learning effort. We will look at whether this criterion is effective and then if so, we will investigate how many collocations are

unpredictable in Korean. In addition, we will examine if native or near-native speakers of English intuitively recognise the difficulty of English collocations for learners of English. That is, we will look at whether native or near native speakers could see regularity or the lack of it in English collocations. If they can, then it may not be necessary to keep checking this through translation because it could be done by intuition.

4.3.1 The contrastive study

A contrastive study was done to investigate the effect of the criterion of *predictability in L1* using Korean as the L1.

In this study, the first 500 items of the total number of 4,698 collocations found by using the two criteria *frequent co-occurrence* and *grammatical well-formedness* were examined to analyse their predictability in Korean. As shown in Figure 4.6, the first 300 collocations are all within the frequency band of the most frequent 2,000 words of English, so these 500 items would be more than enough to test the effect of the criterion of *predictability in L1.*

To examine the criterion of *predictability in L1* the following four step procedure was used. The goal of the procedure is to see if a word by word translation of the Korean equivalent of the English collocation would result in the English collocation. That is, if learners want to express the Korean phrase *매우 많이* (maewoo mani) in English, would the learner be likely to produce the English equivalent correctly if they

did a word by word translation? The steps are made as mechanical as possible so that they are as reliable and replicable as they could be. Someone else following the same procedure and using the same dictionaries should get the same result. The procedure could also be used for other languages.

Step 1. To decide the primary meaning of a constituent of a collocation, the two English dictionaries the *Oxford advanced learner's English dictionary* (2000) and the *Collins COBUILD advanced learner's English dictionary* (2003) are used. The first entry in the two dictionaries is examined to see if they are the same and either entry whose meaning is closer to the meaning used in the whole Korean translation of the collocation is selected as a primary meaning. For example, each component of the collocation *very much* was checked by referring to the two English dictionaries. O/C means that the primary meaning of a component appears in not only the *Oxford advanced learner's English dictionary* but also in the *Collins COBUILD advanced learner's English dictionary*. The *Oxford advanced learner's English dictionary* describes the word *very* as "in a high degree" and the *Collins COBUILD advanced learner's English dictionary* explains it as "to give emphasis to an adjective or adverb". The former is based on the semantic aspect while the latter is from the syntactic view point. However, basically both show the word

very is used as an amplifier, so both definitions are considered the same. In addition, the meaning of *much* is the same in the two English dictionaries "to a great degree". Step 1 asks, what is the first listed meaning of the parts of the collocation? This is another way of seeing whether there is an easily agreed fundamental meaning. It was decided that the ideal situation would be if both dictionaries listed the meaning used in the collocation first. Because dictionaries have different policies about the order of the listing of meaning, it was less desirable but still acceptable if just one of the dictionaries listed the meaning used in the collocation first. If the first meaning listed in the dictionary was the one used in the collocations, then the collocation was easily predictable in English.

Step 2. If the collocation is predictable from the viewpoint of an English speaker how predictable would it be if a Korean speaker wanted to produce it. Production rather than reception was chosen as the test because the focus of this thesis is on learning collocations for productive spoken use. If the favoured Korean translation of each of the English components matches the translation of the whole collocation, then it is highly predictable. To examine the Korean translation of each component, the first entry which was decided as an English primary meaning was checked to see if the primary meaning also matches the first

entry in the two English-Korean dictionaries - the *Essence English-Korean dictionary* (2002) and the *Prime English-Korean dictionary* (2004). E in Table 4.24 stands for the *Essence English-Korean dictionary* and P stands for the *Prime English-Korean dictionary*, and E/P means that the primary meaning of a component appears in both dictionaries. For example, the word *very* of the collocation *very much* is translated into 매우 (maewoo) and *much* has 많이 (mani) as its Korean equivalent in the two English-Korean dictionaries.

Step 3. To examine if the primary meaning of each component is used with the same meaning in the collocation, the collocation is translated into Korean as a whole unit. For example, the collocation *very much* is translated into 매우 많이 (maewoo mani).

Step 4. The Korean translation of each individual component is compared with the whole Korean translation of the collocation. For example, the whole Korean translation of *very much* 매우 많이 (maewoo mani) matches the translation of the individual words; 매우 and 많이. Therefore, the English collocation *very much* is fully predictable from the Korean equivalent 매우 많이 (maewoo mani). There are four possibilities; (1) none of the components match, (2) only some of the components match the

whole Korean translation, (3) sometimes it is not clear-cut to decide if individual components match the whole Korean translation, and (4) all the components match the whole Korean translation. The first three cases are likely to be unpredictable in Korean, and only the last one is predictable in Korean.

Table 4.24 gives some examples of applying the four steps. As the first step, each component of the collocation *you know* was checked by referring to the two English dictionaries to see if their first entries for *you* and *know* were the same.

Table 4.24

Some examples for the criterion of *predictability in L1*

Collocations	The primary meaning according to the two English dictionaries		The primary meaning according to the two Korean dictionaries		Korean translation of the first constituent	Korean translation of the second constituent	Korean translation of the whole meaning of the collocation	Do all the Korean translations match with the whole translation?
you know	O/C	O/C	E/P	E/P	너 (당신-deferential)	알지(아시죠-deferential)	(너 (당신)) 알지(아시죠)	grammatical
too good	_ /C	O_C	_ /_	E/P	또한	좋은	너무 좋은	1 matches
follow me	O/C	O/C	E/_	E/P	따르다	나를	이해하지 (이해하시죠)	different
give way	O/C	_ /C	E/P	E/P	주다	길	양보하다	different
report back	O/C	O/C	E/P	E/P	보고하다	뒤로	돌아와 보고하다	1 matches
very much	O/C	O/C	E/P	E/P	매우	많이	매우 많이	match

The second column in Table 4.24 is subdivided into two columns. The first sub-column is for the first component *you*, and the second is for the word *know*. The *Oxford advanced learner's English dictionary* describes the word *you* in this way "the subject or object of a verb after a preposition refers to the person or people being spoken or written to", and the *Collins COBUILD advanced learner's English dictionary* explains it as "the second pronoun which refers to one or more people". The former is based on the syntactic aspect, while the latter is from the semantic viewpoint. However, both show the word *you* is the second person pronoun, so they agree and thus O/C is entered in column two. They also agreed on *know*, so O/C is entered in column three.

In the fourth and fifth columns, E/P means that the primary meaning of a component appears in both dictionaries. The primary meaning of *you* is 너 (neo) or 당신 (dangshin) as shown in the sixth column. 너 is usually used for younger people, and 당신 is a deferential expression. The primary meaning of *know* is 알지 (alji) or 아시죠 (asijyo) as a question form (see column seven). Like 당신, 아시죠 conveys the deferential meaning. The eighth column gives a whole Korean translation of the collocation. *You know* is translated into 알지 (아시죠). The word *you* is frequently omitted in Korean. However, if adding the word *you*, it also makes sense, so it is not clear-cut to decide if *you know* is fully predictable or partially predictable in Korean. To make a clear distinction, the item *you know* is classified as a

different category because the difficulty only comes from the grammatical difference between the two languages. In this study, we focus on the semantic aspect of collocations assuming grammatical features should be looked at separately. Therefore, grammatical is entered in column nine.

Let us look at the collocation *too good*. The primary meaning of the first word *too* is "to give emphasis to an adjective or adverb", which only appears in the *COBUILD advanced learner's English dictionary* as the first listed meaning and is the same meaning used in the whole Korean translation. It is translated into 너무 (neomu) in Korean. The other dictionary the *Oxford advanced learner's English dictionary* gives the meaning of "as well" or "also" for the word *too* as the first entry, which is 또한 (ddohan) in Korean. Moreover, in both the *Essence English-Korean dictionary* and the *Prime English-Korean dictionary,* the first listed meaning of the English word *too* is "as well" or "also" as well. On the other hand, the first listed meaning of *good* is "high quality or an acceptable standard" in the *Oxford advanced learner's English dictionary* and "pleasant or enjoyable", but the two meanings are translated into the same Korean equivalent 좋은 (joeun), so it could be considered predictable in Korean. It is not that common to find one English primary meaning as the first entry in both the English dictionaries, but either one is acceptable from the Korean translation. This is why O_C is entered in column three instead of O/C. O_C means the first listed entries in the two dictionaries are different but either is

acceptable. Therefore, the word *too* is unpredictable but *good* is predictable in Korean.

The primary meaning of the two components *report* and *back* of *report back* is the same in all the dictionaries. The primary meaning of *report* is "to give people information about something that you have heard, seen, done, etc" whose Korean parallel construction is 보고하다 (bogohahda). The word *back* conveys the meaning of "away from the front or centre or behind you" which is translated into 뒤로 (dwiro). However, the meaning of *back* of the collocation *report back* is different from the primary meaning. The word *back* of *report back* means "after returning or again" whose Korean equivalent is 돌아와 (dorawa) or 다시 (dasi), so the second component *back* is unpredictable.

The collocation *follow me* is used as either a figurative expression or a literal meaning. If considering a literal use, it is fully predictable from the Korean equivalent. However, the collocation *follow me* in Table 4.24 is used in a figurative use which means "understand", so it is unpredictable in Korean.

From the analysis of *predictability in L1* of the first 500 collocations, we found 267 predictables, 174 unpredictables, and 59 others. Table 4.25 shows the results and some examples.

Table 4.25

The results of the analysis of *predictability in L1* of the first 500 collocations

Predictables	Unpredictables	Grammaticals	Undecided	Total
267	174	58	1	500
e.g. *next week, very nice, get out of (sth, swe)*	e.g. *as far as, this evening, *as well *as well as*	e.g. *in fact, in the morning, somebody else, not necessarily, thanks very much*	*good morning*	

- See Appendix 4 for more data.

The first column in Table 4.25 shows some examples of the 267 predictable collocations. *Next week* is translated into the Korean equivalent 다음 주 (daum ju), *very nice* is 아주 좋은 (aju joeun), and *somebody else* is 그밖의 어떤사람 (geubakke eoddeonsaram). *Very nice* has the same Korean equivalent as *very good*, but *nice* and *good* can be interchangeable in English, so it would not be a problem. *Get out of (sth, swe)* means "move out from somewhere" (e.g. *I've got to get out of here*), but it can also convey the meaning of "gain sth from sth/swe" (e.g. *that was all they were going to get out of her*). However, only the first meaning of *get out of (sth, swe)* was considered when the collocation was searched for, so the collocation *get out of (sth, swe)* needs to be subdivided into its different uses. Nevertheless, because this study focuses on the production of English, it would not cause a difficulty in the case that two or more Korean expressions are just one item in English even if the reverse direction could be a problem. It has a word for word equivalent in Korean, so it is included in *predictables*.

The second column gives the examples of unpredictable collocations. 174 items were found to be unpredictable. *As far as* is a figurative expression which conveys the meaning of "to a limited degree". It is not predictable in Korean. *This evening* is translated into *오늘 저녁*. However, *오늘 저녁* is word by word translated into *today evening* in English, so it is not predictable, either. Only two core idioms whose meanings cannot be inferred from their components were found. *As well* is a core idiom which has the meaning of "in addition or too" as an adverbial, and *as well as* which conveys the meaning of "in addition to" as a conjunction or a preposition is also a core idiom. The two items have no word for word equivalent in Korean, so they are included in *unpredictables*.

The third column shows that 58 items are affected by grammatical differences between the English and Korean languages. Five different sorts of grammatical features were found in those 58 items.

The first category is *ellipsis*. For example, *in fact* is translated into the single word *사실* (sasil). *사실* corresponds to the English word *fact* without *in*. The Korean word *사실* can be used either as a noun or an adverb. Such ellipsis of a function word makes up 19 items of the total number of 58 *grammaticals*. Ten items involve ellipsis of a plural suffix. *{No.) years ago* is *{No.) 년 전* (nyeon jeon) in Korean. Each component has a Korean equivalent but *년* (*year* in English) does not contain any plural suffix because the Korean plural suffix *들* (deul) is almost only used for people, plural personal pronouns and plural demonstrative

pronouns such as 사람들 (*people*-saramdeul), 이것들 (*these*-igeotdeul), and 그들 (*they*-geudeul). In addition, there were three items where a content word or both content and function words were omitted during translation into Korean such as *one* ***thing****, (for) the first* ***time*** and *at the end of the* ***day****. Thing, time*, and *of the day* of those three collocations do not convey any meaning in each Korean translation.

The second category is *preposition use*. There were 19 items in this category. For example, *in the morning* is 아침에 in Korean. The English preposition *in* is used to indicate a place according to the fundamental meaning of the two English dictionaries the *Oxford advanced learner's English dictionary* (2000) and the *Collins COBUILD advanced learner's English dictionary* (2003) corresponding to the Korean suffix -에. However, -에 is also used for time. This is why -에 is added to 아침 (*morning* in English). That is, the two uses of the suffix -에 are grammatically different but they look identical. *In the morning* is included in *grammaticals* because it is related to the issue of grammar.

The third category is *split*. As we mentioned above, because this study focuses on English production, if one Korean expression was split into two or more expressions in English, it would be difficult for Korean students to choose the right one of those different English expressions. There were five items related to the issue *split*. *Somebody else* is *else somebody* in Korean word order, but in this study the word order is not seen as a problem. However, the word *somebody* is not distinguishable from *anybody* in Korean. In general, the English word *anybody* is used in

negative or interrogative sentences. Even though *somebody* and *else* each have a Korean equivalent, it is impossible to distinguish the different uses of *somebody* and *anybody* only with a semantic comparison. *Somebody else* requires a grammatical explanation, so it is included in *grammaticals*.

The fourth category is *partial negation*. There was one item included in *partial negation*. *Not necessarily* is a partial negative expression, but it is not distinguishable from a complete negative expression in Korean. In Korean, *not necessarily* is the same as *mustn't be* which conveys the meaning of a complete negation. To express partial negation, the Korean suffixes *-하지는* (hajineun) or *~한건* (hangeon) should be added to the negative form.

The fifth category is *conversion of part of speech*. There was one item included in this category. *Thanks very much* is *감사합니다* (gamsahamnida) or *고마워* (gomaweo). *고마워* is an informal expression, so it is closer to the use of *thanks*. However, *thanks* is a noun, but *고마워* is a verb. In Korean it is unusual that a noun is used like a verb. Even though *thanks* conveys the meaning of *고마워*, the noun form is not a grammatical expression in Korean.

If we look at the fourth column, one item was found hard to classify, so it was designated *undecided*. That was *good morning*. *Good morning, good afternoon*, and *good evening* are coalesced into *안녕하세요* (annyeonghaseyo) in Korean, and so *good morning* would be completely unpredictable from Korean considering only the semantic aspect.

However, the Korean literal translation 좋은 아침 (joeun achim) of *good morning* is also used although *good afternoon* and *good evening* are not translated because *good morning* is much more frequently used in Korean. Even though 좋은 아침 is not a standard Korean expression, it is allowed in informal conversation.

Table 4.25 shows that one third of the collocations were unpredictable from Korean. Fifty-eight *grammaticals* were not included in *unpredictables* because if learners know some English grammar, they would be predictable. This suggests that quite a lot of learning would be required to make the collocations part of a learner's repertoire for spoken production. Even when a collocation is predictable, however learners may be hesitant in applying Korean models to make English collocations.

4.3.2 The survey

In addition to the contrastive study, a survey was carried out. The aim of this part of this study was to examine if native or near-native speakers of English recognise the difficulty of English collocations for learners of English. That is, it investigated whether participants could predict what English collocations would be transparent to users of another language even though they had no knowledge of that language. In effect, the study looked at whether native or near native speakers could see regularity or the lack of it in English. Regular features may be more

likely to parallel use in another language while irregular patterns would be much less likely to have parallels.

Native speakers tend to look at the whole meaning of a familiar collocation rather than to combine the meaning of each component of that collocation when they transfer that into another language (familiarity is closely related to frequency). For example, Spöttl and McCarthy (2004) found most of their twenty participants transferred holistically when asked to transfer English (L2) collocations into other languages (L1, L3, or L4). Only three participants used a word for word translation strategy. This suggests learners are likely to recognise collocations as such and take account of the unit when they transfer them into another language.

On the other hand, if participants have a sophisticated knowledge of the structure of English, they might adopt a word for word analysis strategy. In Kellerman's (1983) experiment, Dutch participants were asked to judge the acceptability of English expressions involving *break*. Advanced learners who were sophisticated metalinguistically rejected forms where causative and non-causative functions were not made distinct because they tried to linguistically analyse these two expressions, while lower level learners simply accepted the two uses of *break* because both English functions correspond to Dutch expressions. Native speakers might transfer either holistically or word for word, but neither method is necessarily successful. Therefore, this survey provides an opportunity to look at what strategies native speakers use when handling collocations and how well native speakers can guess the

difficulty of the collocations.

4.3.2.1 Participants

Twenty subjects who were all postgraduate students or academic staff in the School of Linguistics and Applied Language Studies participated in the survey. The 20 subjects had no problem with English language proficiency and had a sophisticated knowledge of the structure of English. The 20 subjects consisted of eight different ethnic groups. Table 4.26 shows the 20 subjects' background.

Table 4.26
Background of the 20 subjects

New Zealander	Chinese	Malay	Australian	German	Greek	Indian	Russian
9	3	3	1	1	1	1	1

Half of the participants (nine New Zealanders and one Australian) were native speakers of English.

4.3.2.2 Procedure

All the participants were required to guess which items in a set of 50 collocations might have a parallel construction in Korean or not. The 50 collocations were randomly selected from the most frequent 500 collocations of English and *grammaticals* were excluded from the survey. So the 50 collocations consisted of 25 *predictables* and 25 *unpredictables*.

Predictables have a word for word Korean translation. For example, *next year* in Korean is translated by the Korean word for *next* and the Korean word for *year*. It is not necessary for the English and Korean expression to have exactly the same word order. For example, the verb is located at the end of the sentence in Korean, and thus *go home* has the Korean equivalent 집에 가다 (home go), so it is predictable in Korean. On the other hand, *unpredictatbles* have no parallel Korean translation. For example, *strong coffee* is translated into *thick coffee* in Korean, so it is unpredictable in Korean. The participants were also required to answer the following three questions:

1) What languages do you know?

2) Have you studied Korean at all?

3) Are you a native speaker of English?

In addition, the participants were asked for the reasons why they chose *predictable* or *unpredictable* for some specific items (see Appendix 5).

4.3.2.3 Results and discussion

It was found that an average of about 32 of the total 50 items the 20 participants had chosen matched with the results of the 50 items by the criterion of *predictability in L1*. Table 4.27 gives more information on the results.

Table 4.27

The number of correct answers on predictability in Korean

	From the total 50 items	From the 25 predictable items	From the 25 unpredictable items
Minimum	27	8	8
Maximum	37	22	23
Mean	32.20	15.95	16.25
Standard Deviation	3.090	3.99	4.25

- The numbers in columns 2-4 were based on the number of correct answers.

As shown in column 2 of Table 4.27, the minimum score was 27 and the maximum was 37. The standard deviation was 3.09 which is not a large number. The 20 participants chose an average of about 32 correct answers of the total 50 items, which means about 64% were predictable by native or near native speakers. The 50 collocations consisted of 25 *predictables* and 25 *unpredictables*, so each area was also separately examined. Columns 3 and 4 show the results. There was little difference between *predictables* and *unpredictables.* The 20 participants had an average of about 16 correct answers on each area.

18 of the total 20 participants have studied another language in addition to English. Three participants (two New Zealanders and one Australian) have studied Korean before. Most participants were likely to think from the perspective of their L1 or another language, but not English because they were required to guess predictability of English collocations in Korean even though they did not know about Korean. The three participants who have studied Korean gave quite different results for *predictables* and *unpredictables.*

Table 4.28

The results of the three participants who have studied Korean

	From the total 50 items	From the 25 predictable items	From the 25 unpredictable items
A	27	8	19
B	29	9	20
C	36	13	23

- The numbers in Table 4.28 show the number of correct answers.

Table 4.28 shows that A had 27 correct answers from the total 50 items. 8 of the 27 correct items came from predictable items and the other 19 from predictable items.

C was the most proficient in Korean. This is why *C* had a higher score than the other two participants though not much above the mean for the group of twenty participants. *A* and *B* have learned Korean, but their Korean language proficiency was at a novice level, which might interfere with their guessing. So their scores (27 and 29 each) were lower rather than the average number of 32. Nevertheless, all three participants' guessing was more accurate on *unpredictables*. *A* had 8 correct answers on *predictables* but 19 answers were correct on *unpredictables*. *B* had 20 correct answers and *C* had 23 on *unpredictables*. Most participants expected prepositions or adverbs such as *up, of,* and *on* would be difficult for foreign learners because in many cases those words are not used with their literal meaning (e.g. *up to you, thinking of sth*, and *hang on*).

To sum up, 64% of the choices the 20 participants made were correct and the three participants who have studied the Korean language

made better guesses on *unpredictables*. However, the results were not striking. Only two choices, *predictable* and *unpredictable,* were used, so the probability of a correct answer was 50%. 64% is larger than 50%, but it is not a striking difference. Nevertheless, this survey suggests that even native or near-native speakers of English do not make strikingly accurate guesses of predictability. This suggests a systematic process is needed to find out which collocations are predictable or not in another language rather than relying on intuition.

4.4 The transparency of collocations

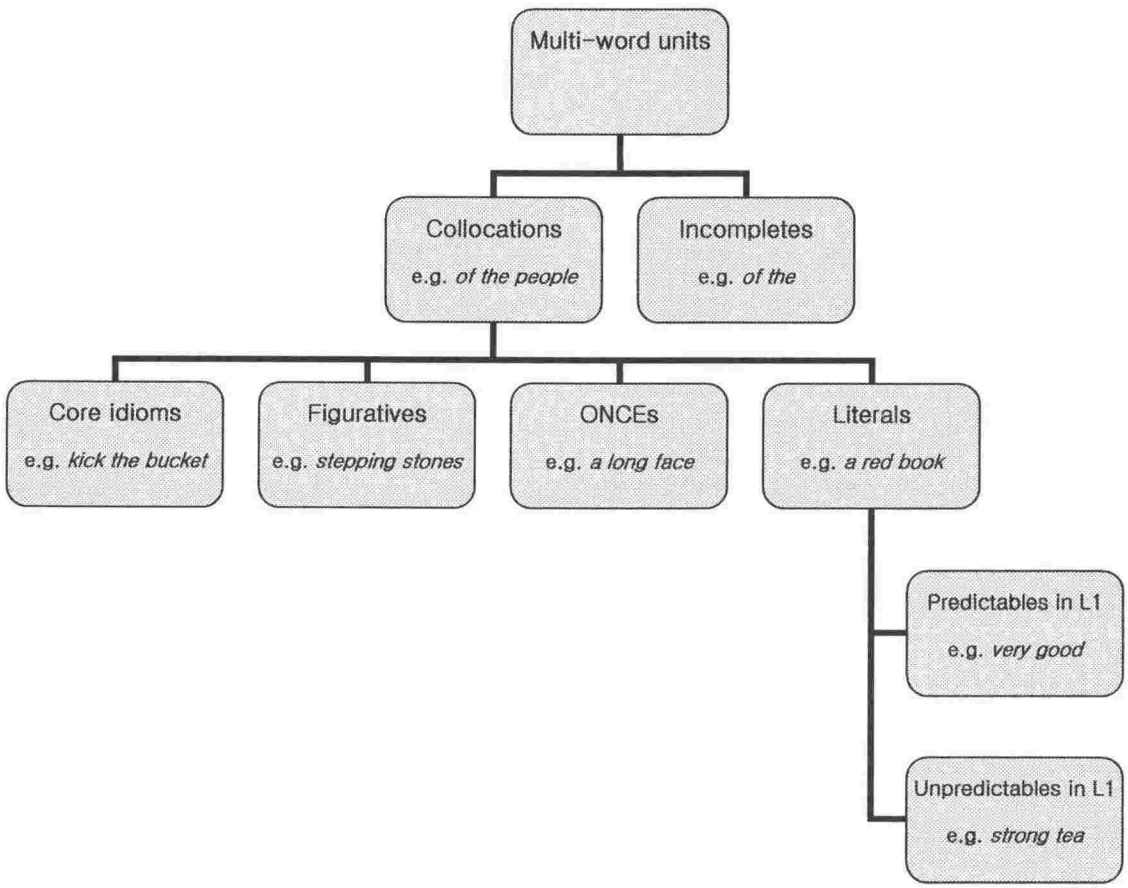
Multi-word units can differ in the way the meanings of the parts of the multi-word unit relate to the meaning of the whole. By using the three criteria of *compositionality, figurativeness,* and *predictability in L1,* we will see if is is possible to classify multi-word units into meaning-based categories. It is worth classifying these items into different categories because different categories of multi-word units need to be treated in a different way when they are taught and learned.

4.4.1 Compositionality, figurativeness, and predictability in L1

Even though the vast majority of collocations in the present study are literal collocations, by applying the two criteria of *compositionality* and *figurativeness* considered above, we can divide collocations into four

types: core idioms, figuratives, ONCEs and literal sequences. These four types are based on the relationship of the meaning of the parts of the collocation to the meaning of the whole collocation. In this study, the term collocation is used as a broad notion using the criteria of *frequent co-occurrence* and *grammatical well-formedness*. Figure 4.8 illustrates the classification system and terminology used for multi-word units.

Figure 4.8
The classification and terminology used for multi-word units



A multi-word unit consists of two or more frequently co-occurring words. As shown in Figure 4.8, multi-word units are divided into

collocations and incompletes using the criterion *grammatical well-formedness.* For example, *of the* is typically the most frequent pair of co-occurring words, but it does not meet the *grammatical well-formedness* criterion, so it is classified as an incomplete. On the other hand, the multi-word unit *of the people* meets the *grammatical well-formedness* criterion and so is classified as a collocation.

The next step then is to use Grant and Bauer's (2004) criteria of compositionality and figurativeness to distinguish the various types of collocations. These criteria distinguish collocations on the basis of how the meanings of the collocates relate to the meaning of the whole collocation. Grant and Bauer (2004) suggest three steps to distinguish core idioms from other types of collocations.

(1) Is the meaning of the collocational group retained if you replace each lexical word in the MWU with its own definition?

YES= compositional NO= non-compositional

(2) Is it possible to understand the meaning of the collocational group by recognising the untruth and pragmatically reinterpreting it in a way that correctly explains the collocational group?

YES= figurative NO= non-figurative

(3) Is there only one word in the collocational group which is either not literal or non-compositional?

YES= ONCE (a one non-compositional element)

NO= More than one element is non-compositional

- If we answer NO to these three questions, the collocational group is a 'core idiom'.

Grant and Bauer (2004) presented the following examples applying the three steps.

(1) Compositional

a red book → YES= ∴ a literal

a red herring → NO= ∴an idiom or figurative

(2) Figurative

as good as gold → YES= ∴ a figurative

a red herring → NO= ∴ an idiom

(3) ONCE, as only one part is 'not-literal' or 'non-compositional'

a long face → YES= ∴ an ONCE

a red herring → NO= ∴ a core idiom

Only the collocation *a red herring* (which conveys the meaning of "anything that diverts attention from a topic or line of inquiry") is classified as a core idiom. These criteria can be used to distinguish the various collocational groups into core idoms, figuratives, ONCEs and literals. Each of these categories can be subdivided into two subtypes: predictable and unpredictable sequences from L1. For this, we can apply the criterion of *predictability in L1*. Core idioms, figuratives and ONCEs can also be subdivided into predictables and unpredictables, although it is very unlikely though not impossible that core idioms will have word-for-word equivalents in the L1. Table 4.29 summarises the types of collocational groups discussed above.

Table 4.29
Types of collocational groups

Types		Transparency
Core idioms		*Cannot be predicted or analysed*
Figuratives		*Cannot be predicted and need to be interpreted*
ONCEs		*Only one element cannot be predicted*
Literals	No L1 equivalent	*Cannot be predicted but can be analysed*
	L1 equivalent	*Can be predicted and analysed*

Only six core idioms from the core idiom list of Grant and Bauer's

study (2004) meet the criterion of *frequent co-occurrence*. However, three more core idioms *as well, of course,* and *as well as* were found in the present study. Table 4.30 lists those nine items.

Table 4.30

The top eight core idioms meeting the frequent occurrence criterion

Rank	Idioms	Part of speech	Frequency	Context
1	*as well*	AVP	5,754	Can I just say something else as well?
2	*of course*	AVP	5,661	And then of course we sign them up later.
3	*as well as*	C/P	620	…other schools in the area used to use this facility as well as we did.
4	*so (-) and (-) so*	NP	192	Well, we've come to deliver this furniture for Mr. so-and-so at number twenty five.
5	*and what have you*	CA	124	They'd be setting their own goals, targets and what have you.
6	*by and large*	AVP	89	By and large, it's according to people's occupations.
7	*such (-) and (-) such*	NP/AP	57	God just decides that such and such is correct or such and such is wrong.
8	*take the piss*	VP	53	It's not funny, you can't take the piss out of all those poor people that died.
9	*out (-) of (-) hand*	AP	48	So I do think it's an industry which is very dangerous now and which could well get out of hand.

In this section, we have looked at how the meanings of the parts of the multi-word unit relate to the meaning of the whole. Grant and Bauer's two criteria, *compositionality* and *figurativeness,* were effective in distinguishing core idioms from the more transparent collocations. In their study, only 104 core idioms were found. Furthermore, by applying the criterion of *predictability in L1* we could distinguish unpredictable literal collocations from predictable literal collocations in the L1. These criteria could give insight into ways of reducing the number of collocations to focus on.

The most striking findings of this research are (1) the large number of collocations found that are as frequent as the high frequency words of English (the first 300 collocations are included in the first 2,000 frequency ranked word types), (2) the small amount of overlap between the high frequency spoken and written collocations (only 15 collocations of the top 50 collocations of the first 150 spoken and written content pivot words occur in both lists), (3) the very high frequency of spoken compared to written collocations (the top 50 spoken collocations are 2 to 10 times more frequent than the top 50 written collocations), and (4) the unpredictable nature of a large proportion of the collocations (one third (174) of the first 500 collocations are unpredictable in Korean). The implications of these findings will be looked at in the following chapters.

CHAPTER 5

FINDINGS, CAUTIONS, AND FURTHER RESEARCH

5.1 Findings

This study has attempted to show what criteria can be used to define collocations and has resulted in a list of the high frequency spoken collocations of English. It was found that the existing criteria for defining collocations do not work well, largely because the existing criteria have not been applied consistently. This resulted in vague categories and a lot of overlap among different categories of multi-word units. To avoid this confusion, we strictly applied the three criteria of *frequent co-occurrence*, *grammatical well-formedness*, and *predictability in L1*.

Firstly, *frequent co-occurrence* is a very simple criterion but it is the most effective way to select the most useful collocations for beginning and low intermediate learners of English. An appropriate frequency level can be used to get a suitable number of collocations, considering the goals and needs of learners. This criterion embodies the classic Firthian definition of a collocation. We found *mutual information* not effective in this study because the mutual information value is strongly affected by having a low frequency word as one of the components of a collocation. However, 92.38% of the types and 90.63% of the word families making up the components of the 4,698 collocations

found in this study were high frequency words occurring in the GSL or the AWL. This means that not only the 1,000 pivot words but most of their collocates are high frequency words.

The criterion of *grammatical well-formedness* facilitates learners' understanding and memory for collocations. While there is value in searching for and describing grammatically incomplete "lexical bundles" (Biber et al., 1999), the goals of this study led us to search for grammatically complete units that could act as well-formed immediate constituents of sentences. The list resulting from applying these first two criteria can be found in Appendices 1 and 2. Thirdly, we found the criterion of *predictability in L1* can be used effectively to select the more difficult collocations for a specific group of foreign language learners. The values of applying this criterion from a Korean perspective are (1) it provides a useful list for people working and studying in Korea, (2) it gives some indications of the general predictability and regularity of high frequency English collocations.

The most striking finding was the large number of collocations meeting the criteria, and the large number of these that would qualify for inclusion in the most frequent 2,000 words of English if no distinction was made between single words and collocations. 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 were in fact many.

Finally, we applied the criteria of compositionality and figurativeness to distinguish core idioms from more transparent collocations.

All of the criteria used in this study had to be reliable and 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 this study provide reliable data that can be added to by further research, and a procedure that can be applied to other corpora.

In this study, there are nine major findings.

1. **There is 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 a collocation. So when the list is re-sorted by frequency without distinguishing the pivot words, the new list contains 4,698 collocations. There are plenty of very frequent collocations made from very frequent words both in spoken and written English.
2. **Collocations occur in spoken language much more frequently than they occur in written language.** The frequency cut-off point for the 100 most frequent spoken word types of English is around a frequency of 16,999 occurrences per 1,000,000 tokens and for the

100 most frequent written word types around 9,595 occurrences, that is, the first 100 spoken collocations are almost twice as frequent as the first 100 written collocations. However, word frequency decreases more rapidly in the spoken corpus, so after the first 400, the written word types are more frequent than the spoken word types. There are 218 collocations in the spoken corpus which would get into the top 2,000 words of spoken English, 56 of these would be in the first 1,000. There are 54 collocations which would get into the top 2,000 words of written English, 14 of these would be in the first 1,000.

3. **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% 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%). 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.

4. **A small number of pivot words account for a very large proportion of the tokens of collocations.** The total number of tokens of the

collocations of the first 100 pivot words is 387,634 which cover about 53% of the total number of tokens of the collocations (736,617) found in this study. The first 200 pivot words make up about 68%. The first 200 spoken collocations are much more frequent than the first 200 written collocations. Research on collocations provides support for Zipf's law on frequency distribution (frequency * rank = a constant figure). The Zipf curve of the first 1,000 collocation band shows a slope of around -1.5 when the two most frequent collocations *you know* and *I think* are excluded. The other three 1,000 collocation bands show slopes of around -0.9, -0.8, and -0.7 each. As the rank goes down, the slope is more gradual. This means that very frequent collocations are more frequently used compared with their ranks.

5. **Adjectives tend to have more collocates than other content words.** Some randomly selected pivot words with different parts of speech were compared, controlling for their frequency. This comparison shows that adjectives are likely to have more collocates, while verbs have fewer collocates. There was very little difference between nouns and adverbs. When we examined the part of speech of collocations as a whole unit, verb phrases were most common in the present study. 1,734 verb phrases were found. Next were noun phrases with 1,498 items. These two types of collocations make up 3,233 (69%) of the total of 4,698 collocations.

6. **The shorter the collocation, the greater the frequency.** Two word

collocations make up 77% 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 infrequent items, there are more collocations containing a relatively longer word compared with the top 100 collocations.

7. **Content word plus content word collocations outnumber other patterns of content word collocations.** There are three dominant collocation patterns found in this study. The combination of content word+ content word (C+ C) such as *fund raising* is the most dominant collocation pattern which makes up 2,039 (56%) of the 3,616 two word collocations. The second dominant combination is the pattern of content word+ function word (C+ F) such as *try it* and the third dominant combination is the pattern of function word+ content word (F+ C) such as *in use*. Collocations consisting solely of function words were not examined in this study.

8. **There are more collocates on the left than collocates on the right, but this difference is not striking.** There are 3,161 collocates on the left, and 2,408 on the right, and there are 319 collocations with collocates on both sides.

9. **A third of the first 500 collocations of English analysed did not**

have word for word equivalents in Korean. Through analysis based on a dictionary definition of each component of a collocation, we found 267 predictable items such as *next week*. 174 items were unpredictable such as *as far as*. The other 59 items could not be classified by this method of analysis which focused only on the semantic difference between the English and Korean languages.

There are few surprises in the resulting lists, just as there are few surprises in lists of the most common English words. Collocations like *you know, a bit, thank you, very much* are to be expected in spoken English.

5.2 Cautions

In this study, a 10 million word corpus consisting of spoken texts from the BNC was used, and the most frequent 1,000 content words from the frequency-ordered spoken word list by Leech et al. (2001) were used for searching for collocates. However, it is clear that there would be big differences if a written rather than spoken corpus was used. Frequency is also influenced by topics and discourse modes (narrative, descriptive, expository and argumentative). For example, in a scientific text sample, *data, figure, formula*, and *example* occur frequently. In a commercial text sample, *information, market, value,* and *export* are frequently occurring words. Therefore, high frequency in one

corpus cannot guarantee wider usefulness. From this perspective, range might be a more valid criterion than frequency. This however would have resulted in a list which could have been misleading, encouraging teachers and learners to use colloquial collocations in formal written language, and to use formal language in colloquial situations. Further research could usefully add to the fifty written collocations looked at in part of this study.

Another issue is that the spoken list is not lemmatised. This is probably an advantage. In the COBUILD English collocations (Sinclair (Editor-in-chief), 1995), Stubbs (2000) found different forms of the lemma *seek* were used quite differently. The collocates of *seeks* rarely overlap at all with the collocates of *seek, seeking* and *sought*. The collocates of *seeks* largely come from political and legal contexts, in the semantic field of "help and support", only six collocates of *seeks* were shared by the three different forms: *asylum, court, government, help, political*, and *support*. The non-lemmatised list was used to make sure that all the possible collocations were picked up. Searching using only the base form would have missed many frequent collocations.

A ONCE (One Non-Compositional Element) is one type of multi-word unit based on the classification used by Grant and Bauer (2004). However, the use of the category of ONCEs could be unnecessary because the one non-compositional element that ONCEs contain could be considered a polysemous or homonymous use of a word. For example, the word *long* of *long face* is used with the meaning of *gloomy*

or *worried* which is not related to the notion of length. So it was considered non-compositional but some might argue that use of *long* comes from its polysemous use. That is, the word *long* could be used in more than one sense. Nevertheless, we decided to leave the category of ONCEs in this study because ONCEs are a marginal part of multi-word units, and those expressions have little influence on the results of this thesis.

About 170 Verb+Preposition types were found in this study. Most of the 170 items were prepositional verbs like *wait for*, *deal with*, and *look forward to*. Some of them however are close to Verb+Adverbial constructions, the so called 'free combinations', for example, *go with {sth, smo}*, *live in {swe}*, and *walk to {swe}*. The most common criterion used for distinguishing the two types is *wh-question formation* (Biber et al. (2002); Greenbaum & Quirk (1990)). When a sentence containing these word sequences is changed to a wh-question form, and the preposition could be omitted, then the prepositional phrase would be an adverbial. For example, the preposition *about*, when in *what are talking about?*, cannot be omitted, so *talking about* is a prepositional verb. Let us look at another example. *I am walking to that place* could be transformed to *where are you walking?* without the preposition *to*, however, consider *what place are you walking to?* In this case, the criterion *wh-question formation* is not effective, that is, it needs to be investigated in further research. In this study, all these types were included because the inclusion of these items may be more useful than

the exclusion of these types.

To check the results of this thesis, the collocations in the Brown Corpus were compared with the data from the present study. The Brown Corpus collocations were listed using a frequency cut-off point of 50 occurrences. The list contains 2-word to 5-word collocations. There were several Brown Corpus collocations not listed in our collocation list based on the BNC spoken corpora. It was found that we missed the three collocations *no more* (382 occurrences in the ten million BNC spoken section), *each other* (639), and *the fact that {S V}* (1,178). The other missing collocations such as *of course, no doubt, in addition,* and *with respect to {N}* were not listed in our list because the components of those collocations were not included in the first 1,000 content pivot words. There are undoubtedly useful function word plus function word collocations that would meet the criteria used in this study. These deserve future investigation.

5.3 Further study

There are four recommendations for future research:

1. Further research is needed on written collocations. As a part of the present study, a small number of spoken and written collocations were compared using the 150 most frequent spoken content pivot words and the 150 most

frequent written content pivot words. In this analysis, 2,261 spoken collocational types and 2,266 written collocational types were found. The comparison focused only on the top 50 spoken collocations and the top 50 written collocations and the results showed big differences between spoken and written language. One hundred and fifty written pivot words however is not enough to make a definitive list of the most frequent written collocations. So, further research in this area would be useful.

2. The collocation list made in the present study may be misleading, encouraging teachers and learners to use colloquial collocations in formal written language, and to use formal language in colloquial situations. The use of collocations is likely to be different in different discourse modes. Even though the present study was based on spoken language, 'formality' was not considered. Biber's (1989) research indicates that friendly letters are similar to friendly conversation. Making a formality distinction would be interesting and revealing.

3. In the present study, collocations were subdivided into four types: core idioms, figuratives, ONCEs and literal sequences. Because the present study focused on finding

the core idioms, the distinction between figuratives and literal sequences was not made. Some phrasal verbs like *get up* may be included in the category of figuratives. Further research on figuratives is needed.

4. When the learner learns a language, members of some word groups can interfere with each other semantically (Nation, 2000) or formally, so there is a need to examine if teaching the various related collocations of different types of a lemma or semantically associated collocations causes any interference effects making learning more difficult.

CHAPTER 6

IMPLICATIONS AND APPLICATIONS

6.1 Choosing what to focus on

Bahns (1993, p. 56) argues that one of the main obstacles to teaching lexical collocations systematically is their number, which amounts to tens of thousands. However, this enormous teaching and learning load can be reduced in three ways - (1) by only giving attention to high frequency items, (2) by taking different approaches to core idioms, figuratives, and literals, and (3) by taking a contrastive approach to the concept of lexical collocation, involving comparison of the L1 and the target language.

6.1.1 Frequency level

Because there is a relationship between the frequency of the pivot word and the number and frequency of the collocations involving that word, there is likely to be a point in a frequency level list of pivot words where there are very few or no collocations which meet the frequency criterion, or where the collocations are not very interesting. The most interesting collocations are usually where content words collocate with content words. The least interesting collocations with content words are where content words collocate with very high frequency function words

like *the*, *that* or *those*. An appropriate number of pivot words to focus on should be decided by considering the usefulness or meaningfulness of the collocations. In Table 6.1, the average frequencies of collocations in ten bands of 100 pivot words are compared. The data comes from the 10,000,000 token BNC spoken corpus.

Table 6.1
Comparison of the average frequencies
per 10,000,000 tokens of a collocation of the 10 bands of 100 pivot words

10 bands of 100 pivot words	Total frequency of collocations (token)	No. of different collocations (types)	Average frequency per collocation
1st 100 words	387,634	2,052	188.91
2nd 100 words	111,468	843	132.23
3rd 100 words	64,962	704	92.28
4th 100 words	44,896	564	79.60
5th 100 words	36,903	481	76.72
6th 100 words	27,846	373	74.65
7th 100 words	22,433	281	79.83
8th 100 words	16,872	233	72.41
9th 100 words	12,615	187	67.46
10th 100 words	10,515	176	59.74
Total	736,144	5,894	

Table 6.1 shows that the total frequency of the collocations of the first 100 pivot words is very high. After the first 100 pivot word band, the total frequencies of the collocations and the number of the

collocations initially drop rapidly. The first band has more than twice the number of collocations compared to the second band and the total frequency of the collocations is over three times higher. After the first two bands, all the figures then decrease more slowly. Even though the number of the collocations drops, there is no big change in the average frequency per collocation in 10,000,000 running words. The average frequency of the 7th band in fact is higher than the 6th band. This means that there may be still some quite frequent collocations for lower frequency pivot words. Table 6.2 gives some examples.

Table 6.2
Collocations of some lower frequency pivot words

Rank	Pivot word	Collocation	Frequency
979	bothered	*(not [82]) bothered about {sth}*	109
		(not [72]) bothered to {INF}	85
986	managed	*managed to {INF}*	319
999	due	*(be-verb) due to {INF}*	272
		due to {sth}	97
		in due course	72

- In the present study, thirty occurrences were used as a frequency cut-off point.

Table 6.2 shows the collocations of the three pivot words *bothered, managed* and *due* which are all words near the end of the first 1,000 pivot words. The pivot word *bothered* has just two collocations, *managed* has one, and *due* three in 10,000,000 running words. However, as shown in Table 6.2 those collocations have a high frequency, for example, *managed to {INF}* has 319 occurrences per 10,000,000 tokens. This shows that it is still worth searching for collocations in the 10th

band of pivot words or in other words at the 1,000 word level.

6.2 Teaching and learning collocations

The time available for learners to get their English input in most EFL classroom situations is short, so deciding what to focus on is one key to a successful language programme. From this point of view, the collocation list could be used as a source for a deliberate learning approach to the most frequent collocations of English. However, some might argue that focusing on collocations out of context is not helpful to learning, especially in an EFL situation. This negative attitude toward deliberate learning might come from a reaction against traditional methods of language learning which were heavily dependent on rote learning. The problem, however, is not in deliberate learning itself, but lies in a lack of balance with other ways of learning, particularly incidental learning. There is value in learning collocations in a range of ways, through direct study and teaching and through incidental learning in context. Let us look at each of these ways.

6.2.1 Deliberate learning and teaching

The most common and efficient way of dealing with collocations is direct learning and teaching. Learners or teachers need to choose high frequency collocations or topic-related collocations and deliberately

focus on their form, meaning, and use. Some vocabulary strategies can be taught for learning collocations such as using word cards, using a dictionary and comparing L2 with L1. Boers, Eyckmans, Kappel, Stengers, and Demecheleer (2006) set up a small scale experiment involving 32 college students majoring in English. The participants were exposed to considerable input of authentic listening and reading material for 22 teaching hours spread over an eight month period. The seventeen students in the experimental group were directed to pay attention to particular standardised formulaic sequences (collocations) during listening exercises which included gap-filling activities targeting words making up formulaic sequences, and reading exercises, where they identified useful word combinations and compared their selection with their peers in small groups or with the teacher's recommendations. The fifteen control group students' attention was given to individual words or grammar patterns, rather than collocations. Afterwards, all the participants' oral proficiency was tested in an interview by two raters, not otherwise involved in the study.

The results showed the experimental group made significantly better progress in overall oral proficiency, fluency, and range of expression compared with the control group. However, no significant difference was found in accuracy. To see if deliberate learning of formulaic sequences was useful to help improve oral proficiency, correlations between the experimental students' overall oral proficiency and the number of formulaic sequences these students used were

calculated. The range of the correlations was .325 to .609. The results came from the two formulaic sequence counting judges and the two different oral proficiency raters. The correlation of .325 is not high but .609 is a very impressive figure. This suggests "the use of formulaic sequences can indeed play a part in students' coming across as proficient speakers" (Boers et al., 2006, p. 14) and formulaic sequences can be deliberately focused on and learned with measurable effects on oral use.

The deliberate learning of collocations will be greatly helped by meeting the collocations in meaningful contexts. There are four ways of providing such contexts: (1) collocations embedded in sentences, (2) collocations associated with visual aids, (3) collocations taught in comparison with the learner's first language (L1), and (4) collocations used during problem solving. Presenting collocations embedded in sentences is the simplest way. For example, the pivot word *provide* collocates with *service* 30 times in the spoken section of the BNC. However, the phrase *provide (a/the) service* is not enough context. Introducing the collocation embedded in a sentence such as *they don't clearly* ***provide a service*** *to all our customers* makes it easier to understand and learn the collocation. Secondly, using visual aids to present collocations reinforces the association between collocations and a visual image. Rommetveit et al. (1971) provided evidence that pictorial materials can help students learn vocabulary easily. Thus a picture of a train with its rear coach beyond the end of a station platform, and a

sad-faced woman looking at it from the platform, was accompanied by the utterance, *it's too bad*[1] *she didn't get there*[2]*in time*[3]. The collocation *too bad* has a total occurrence of 354 tokens in the 10,000,000 word corpus, *get there* has 332, and *in time* has 263. These high frequency collocations can be learned by associating them with visual images which will enrich learners' long-term memory of the collocations. Thirdly, teaching collocations by comparing the learner's L1 with the target language can be useful, especially, in an EFL classroom. This is likely to be helpful even when there is not a word for word match with the L1. Some L2 collocations that match the L1 can be learned easily, so this association technique should be encouraged. However, some collocations cannot be translated into the L1. For example, one of the collocations of the verb *turn*, *turn up* has two main meanings. One is "show up", or "appear" (e.g. ... *where maybe we'd sit here and nobody would turn up*) and the other is "increase the amount of sound, heat, or power" (e.g. *let's turn up the sound then*…). The use of the first meaning has a total occurrence of 210 tokens in the 10,000,000 word corpus and the second meaning appears only 7 times in the same corpus. However, the core meaning of *turn* is recognised as *change* or *shift* by Korean students. Therefore, the first meaning is more difficult for them even though the first use is much more frequent in the corpus.

Consider another example involving the contrastive analysis between the two languages. In general, it seems that to some extent 'selection restrictions' of a pivot word correspond to semantic

associations as Newman (1988) shows in the contrast between Hebrew and English. Table 6.3 shows the contrast of the *dress* collocations between English and Korean.

Table 6.3
The contrast of the restricted selections
of some *dress* verbs between English and Korean

Restricted selections of some 'dress' verbs in English										
Object / Verb	jacket	trousers	bra	hat	glasses	shoes	ring	gloves	shawl	skirt
wear~	√	√	√	√	√	√	√	√	√	√
put~on	√	√	√	√	√	√	√	√	√	√
slip into~, on~	√	√	√	√	√	√	√	√	√	√
throw~on, ~over~	√	√	√	√		√		√	√	√
have~on (~)	√	√	√	√	√	√	√	√	√	√
be in~	√	√	√	√	√	√		√	√	√
Restricted selections of some 'dress' verbs in Korean										
Object / Verb	jacket	baji	bra	moja	ankyeong	sinbal	banji	janggab	shawl	chima
~ibda	√	√								√
~geolchida	√	√	√						√	√
~ggida					√		√	√		
~sseuda				√	√					
~sinda						√				
~dureuda									√	√
~chada			√							

- The same items are analysed in both languages (e.g. trousers=baji, hat=moja, skirt=chima, etc)

As shown in Table 6.3, some *dress* verbs in English, *wear, put~on,*

have~on and be in~ are broadly used with various kinds of clothes and accessories. *Slip into (on)~*, and *throw~on (over)* are more or less informal expressions. *Slip into (on)~* means "to put on clothes quickly and easily" or "to put some part of the body in clothes or accessories", for example, ...*slip one's clothes on*, and ...*slip a ring on one's finger*. *Throw~on (over)* means "to wear clothes in a quick and casual way", for example, ...*throw on a jacket*, and ...*throw a scarf over one's shoulders*. That is, English *dress* verbs are not restricted in their use, but they are just different in nuance where speakers imply their attitude or intention. Similarly, the *dress* verbs in Korean have their own specific functions. *Ibda* corresponds to *wear*, or *put~on* which has the most general meaning. Nevertheless, *ibda* is also under considerable selection restriction. *Ibda* can be used only with clothes, not with accessories such as *ankyeong* (glasses), *banji* (ring) and *janggab* (gloves). *Geolchida* is very close to *throw~on* which is one of the informal expressions, but its use is more restricted than *throw~on*. For example, *geolchida* cannot go with *sinbal* (shoes), *banji* (ring) and *janggab* (gloves) as collocates, though these may be possible in English. Similarly, *moja* (hat) collocates only with *sseuda, sinbal* (shoes) collocates with *sinda*, and *ankyeong* (glasses) and *janggab* (gloves) collocate with *ggida*. Therefore, we can see that Korean *dress* verbs are also differentiated in their use. If teachers are aware whether a L2 collocation is predictable or unpredictable from L1, they can help prevent the incorrect transfer between L1 and L2.

Finally, while practicing task-based problem-solving activities, students can pick up collocations. Consider Willis and Willis' (1988) model. In the pre-task stage, teachers select high frequency words and collocations and use them in dialogues. It is important that the dialogues be native-like. The modified material can help learners reduce their cognitive burden and focus on the target items by decreasing the redundancy of the material. Students engage in the task based learning after preparatory help by the teacher. The learners listen to the audio-taped dialogues based on a specific task such as opening a bank account, apologising, and giving directions. These examples help students understand the required tasks. Such tasks include collaboratively designing a dialogue of their own. The goal is to have the learners produce a dialogue using target items. Students' output is important, reinforcing long-term memory. The role-play as a dialogue form is one method used to support task based learning activities, and can effectively reinforce frequently used collocations.

6.2.2 Incidental learning through meaning-focused input

It is necessary for learners to meet and learn collocations incidentally through meaning-focused listening and through extensive reading of material at a suitable difficulty level. Interacting with peers at a level roughly similar to theirs, interacting with teachers who are sensitive to their level, simulating communicative activities which might

occur in real world, and working with graded listening and reading material at a suitable difficulty level could provide opportunities to obtain comprehensible input.

In this strand, the lists of collocations do not have much of a role to play. However, in listening activities, the teacher can note one or two useful collocations on the board. The lists could be good sources to guide the teacher to select some useful collocations to note.

Some studies have looked at the incidental learning of collocations. Kurnia (2003) examined the extent to which multi-word strings (collocations) were retained while reading for information in a L2 text and whether such retention was related to unknown vocabulary density, vocabulary knowledge, and text comprehension. 242 Indonesian learners showing mastery of at least the 2,000 word level participated in the study. A newspaper feature article in English about 1,000 words long was used at two different densities of unknown words - 2% and 5%. The target unknown or new words were replaced by pseudo-words to ensure that they were unknown to all participants and the other words in the text were kept within the first 2,000 words in English. The target words were the ones with the lowest frequencies in English and each one occurred only once in the text. The participants read the text for information, returned the text and were immediately tested on one of these three unexpected measures: retention of the meanings of new words, guessing from context, and retention of the target multi-word strings in the text. Afterwards, the participants were given back the

same text and they completed the comprehension test.

Around 10-11% of the multi-word strings containing familiar words were retained, while about 29-30% of the multi-word strings containing a new word were retained. New words might have caused some problem in decoding the meaning and therefore attracted attention not only to themselves but also to their surroundings during reading, which might result in more retention of the word combinations involving these new words. The retention of multi-word strings made up of familiar words was positively related to vocabulary knowledge, while the retention of multi-word strings containing a new word was positively related to both vocabulary knowledge and text comprehension. Text comprehension, in turn, was positively related to vocabulary knowledge. These findings suggest introducing and drawing attention to the word combinations that make up the multi-word strings when introducing a new word. This may encourage learners to give more attention to the new word and its surroundings. Kurnia's argument that the majority of multi-word units recur very infrequently supporting other similar studies (Moon, 1998a, 1998b; Grant, 2003) was proved wrong by the present study. There is a very large number of high frequency collocations.

6.2.3 Incidental learning through meaning-focused output

Receptive vocabulary knowledge is typically greater than productive vocabulary knowledge. However, it is difficult to be aware of

gaps between the two types of knowledge without having to produce language. Meaning-focused output involves speaking and writing. Having to speak and write makes learners aware of gaps in their knowledge. If reading material is used as a preparation for speaking and writing, then underlining useful collocations in the reading to use in the speaking or writing can have helpful effects. The teacher's feedback comments on speaking and writing can also be useful in learning collocations with the correct use of collocations being praised and errors corrected.

Native speakers as well as non-native speakers are likely to overlook collocations because even though the learners listen to or read a sentence that contains a collocation, this does not necessarily require them to consider it as a unit. Thus they may have a problem when they need to produce an expression including the collocation. However, an important factor affecting the productive use of collocations is the existence or non-existence of parallel constructions in the learners' L1. Learners may think of some collocational groups in the L1 and try to translate them into the Target Language (TL). In this case, the L1 can influence both linguistic forms and meanings of L2 collocations. However, Biskup's (1992) study shows that in the case of two languages that are perceived as distant, such as Polish and English, Polish students do not transfer forms from L1 to L2, instead their errors come from assumed semantic similarities that stimulate either loan translations or extension of the L2 item meaning on the basis of the L1 word. In other

words, in the case of two distant languages, learners are likely to transfer only core meanings to L2, and the transfer depends on the 'coreness' of a word. For instance, the verb *run* can have various meanings such as a) *I can run as fast as Magnus...*, b) *I run a company...*, c) *Next year he may run for president.* The first *run* is the core meaning, and when the core meaning is focused on, the other uses are unpredictable. On the other hand, in the case of closely related languages such as German and English, German students tend to make more errors resulting from assumed formal similarity, for example, German students produce *crunch* or *crunk* (*sic*) *nuts* for the collocation *crack nuts*, transferring the German word *knacken*.

In the case of Korean, similar results are expected as with Polish students. Because Korean and English are perceived as even more distant than Polish and English, Korean students are likely to transfer L1 semantic parallels to L2 collocations instead of similar linguistic forms. However, if there is no corresponding word to a target word in Korean or the meaning or use of the target word is even slightly different in Korean, it could cause Korean students to make errors. The data from the present research indicates that there is a one-in-three chance of making an error using word for word transfer. Table 6.4 shows some collocations that are unpredicatble for Korean students.

There are various reasons for the difficulties in learning the collocations marked in Table 6.4, but one common reason is L1 interference.

Table 6.4

Some examples of collocations that are difficult to be predicted in Korean

	Collocation	Part of speech	Frequency	Context including different uses of a collocation with its frequency
1	*at the moment*	PP	2,176	I cannot recall his name ***at the moment.***
2	*those sort (of sth)*	NP	108	We had ***those sort*** of jobs to do.
3	*come forward*	VP	100	a) (47)-…I will call out each name and If each one can come forward and receive the award…
				b) (29)-…have yet to ***come forward*** with some suggestions.
				c) (24)-…magistrates ***come forward*** to all these positions.
4	*go with it*	VP	53	a) (49)-…we should have the evidence to go with it.
				b) (4)-…your hairstyle to ***go with it***…
5	*boxing day*	NP	44	Although I enjoy ***boxing day*** cos I have all the family there.

- A bold phrase like ***grown up*** is a collocation which is not predictable from Korean. If a collocation has more than one different sense, it is subdivided into separate collocations.

The following list shows some of the reasons why difficulties occur in each item.

(1) **at the moment**: In Korean the sense of the present tense cannot be inferred from *at the moment*. When learners read the sentence *I cannot recall his name at the moment*, they could comprehend the present tense, however, they would use *could* instead of *can* when

producing that sentence. That is, for Korean learners *at the moment* would be confused with *at that moment.*

(2) **those sort (of sth)**: Korean students are likely to assume number agreement in a phrase or a sentence, so they expect *those sorts* instead of *those sort.* However, *those sorts of things* has 17 occurrences, while *those sort of things* has 25 occurrences in the corpus used in this study. A similar result is also identified with *those kinds* and *those kind. Those kinds of things* has 3 occurrences but *those kind of things* has 14 occurrences. As the plural forms *those sorts* or *those kinds* occur as well and are correct, we can justifiably teach the plural form and let students discover the singular form by exposure.

(3) **come forward**: The first meaning of the collocation is based on the literal meaning of the components, but the other two meanings are figurative expressions that are ambiguous in their meanings.

(4) **go with it**: The second use of the collocation means 'be a good match' or 'be suitable for it', so it is also a figurative expression that is difficult to infer from the literal meaning.

(5) **boxing day**: Korean students are unfamiliar with the special name for the 26th of December, the day after Christmas day. It cannot be inferred from its parts.

Most difficulties result from Korean students' attempt to match collocations with the L1 semantic system. It may be very difficult to

prevent all L1 interference on L2 but it is obviously possible to considerably reduce the interference. Therefore, to make sure of learners' proper comprehension and production of collocations, we need to provide as many different contexts as possible.

6.2.4 Fluency development

Another way to reinforce collocational knowledge is related to fluency development. Fluency means making the best use of what is already known. This involves two factors - recognition (listening and reading) speed, and production (speaking and writing) speed. Recognition speed can be developed by hearing and reading the same story several times. The listening and reading input should contain no unknown words and there should be some pressure to perform slightly faster in each exercise. The classic productive activity for developing speaking fluency is the 4/3/2 activity. In the 4/3/2 activity, a speaker gives the same talk three times changing partners (listeners) with a decreasing amount of time for each of the three deliveries (from the 4-minute talk to a 3-minute talk to a 2-minute talk). After this, all the listeners become speakers and the same procedure is followed. For writing fluency, the teacher needs to encourage learners to write on very easy topics, to write on closely related topics, and to write on the same topic several times. These productive fluency activities are also effective in decreasing errors and increasing grammatical complexity

(Arevart and Nation, 1991). As learners become more fluent in recognising and retrieving a target word, it makes it easier for the learners to move their focus onto the surroundings of the target word. This does not mean to simply combine discrete words, but means to restructure them so they are seen and stored as a whole unit. Fluency increases in two ways, by increasing speed of access, and by restructuring to work with larger units.

As shown above, each strand has different strengths and thus different contributions to make. Deliberate learning speeds up learning making efficient use of time. Incidental learning allows large quantities of small amounts of learning to occur. Meaning-focused output allows the noticing of gaps and fluency development automates the use of words and encourages the use of larger units like collocations. Learning high frequency collocations both deliberately and incidentally is important in second language acquisition. A balance across the strands is needed to achieve collocational competence.

One of the basic reasons motivating the present study is that non-native speakers are likely to overlook collocations if they are not meeting them often. This can be problematic when learners need to produce an expression using a collocation. Thus, there is value in giving some deliberate attention to collocations.

It is clear from this study that differences in collocations and their frequency of use, are important distinguishing features of spoken and written language. Even though there are clear differences between the

two language types, there are only a small number of studies on collocations based on a spoken corpus, and it is hoped that the present study will be a useful contribution to help fill this gap.

REFERENCES

Altenberg, B. (1994). On the functions of 'such' in spoken and written English. In N. Oostdijk & P. de Haan (Eds.), *Corpus-based research into language: In honour of Jan Aarts* (pp. 223-240). Amsterdam: Rodopi.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word combinations. In A. P. Cowie (Ed.), *Phraseology* (pp. 101-122). Oxford: Oxford University Press.

Arevart, S., & Nation, I. S. P. (1991). Fluency improvement in a second language. *RELC Journal, 22*, 84-94.

Bahns, J. (1993). Lexical collocations: A contrastive view. *English Language Teaching Journal, 47* (1), 56-63.

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System, 21* (1), 101-114.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6* (4), 253-279.

Becker, J. D. (1975). The phrasal lexicon. In B. Nash-Webber & R. Schank (Eds.), *Theoretical issues in natural language processing 1* (pp. 70-73). Cambridge, Mass: Bolt, Beranek and Newman.

Biber, D. (1989). A typology of English texts. *Linguistics, 27*, 3-43.

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

Biber, D., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English*. London: Longman.

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

Biskup, D. (1992). L1 influence on learners' renderings of English collocations: A Polish/German empirical study. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 85-93). London: Macmillan.

Bogaards, P. (2001). Lexical units and the learning of foreign language vocabulary. *Studies in Second Language Acquisition, 23*, 321-343.

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

Church, K., & Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics, 16* (1), 22-29.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisitions: Exploiting on-line resources to build a lexicon* (pp. 115-163). Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins English dictionary (1994). (3rd ed.). Glasgow: Harper Collins Publishers.

Collins COBUILD advanced learner's English dictionary. (2003). (4th ed.). Glasgow: HarperCollins Publishers.

Cortes, V. (2002). Lexical bundles in freshman composition. In R. Reppen, S. M. Fitzmaurice & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 131-145). Amsterdam: John Benjamins Publishing Company.

Coxhead, A. (1998). *The development and evaluation of an academic word list.* Unpublished master's thesis. Victoria University of Wellington, New

Zealand.

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

Cruse, D. A. (1986). *Lexical semantics.* Cambridge: Cambridge University Press.

Crystal, D. (1985). *A dictionary of linguistics and phonetics.* Oxford: Blackwell.

de Glopper, K. (2002). Lexical retrieval: An aspect of fluent second language production that can be enhanced. *Language Learning, 54* (4), 723-754.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking and points of order. *Studies in Second Language Acquisition, 18*, 91-126.

Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33-68). Cambridge: Cambridge University Press.

Essence Korean-English dictionary (2002). (7th ed.). Seoul: Minjungseorim.

Boers, F., Eyckmans, J., Kappel, J. Stengers, H., & Demecheleer, M. (2006). *Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. Language Teaching Research, 10* (3), 245-261.

Fano, R. (1961). *Transmission of information: A statistical theory of communications.* Cambridge, Mass: MIT Press.

Fernando, C. (1996). *Idioms and idiomaticity.* Oxford: Oxford University Press.

Fillmore, C. J. (1997). Lectures on construction grammar. Retrieved 22, November, 2005 from http://www.icsi.berkeley.edu/~kay/bcg/lec02.html.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in*

Linguistic Analysis (pp. 1-32), reprinted in F. Palmer (Ed.), *Selected Papers of J. R. Firth 1952-59* (pp. 168-205). London: Longman.

Fletcher, W. (2003/2004). PIE: Phrases in English. Retrieved 12 January, 2006 from http://pie.usna.edu.

Fries, C. C. (1952). The structure of English: An introduction to the construction of English sentences. New York: Harcourt, Brace & Company.

Gleason, H. A. (1955). *An introduction to descriptive linguistics.* New York: Holt, Rinehart and Winston.

Grant, L. (2003). *A corpus-based investigation of idiomatic multiword units.* Unpublished doctoral dissertation. Victoria University, Wellington.

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

Grant, L., & Nation, I. S. P. (2006). How many idioms are there in English? *ITL-International Journal of Applied Linguistics, 151,* 1-14.

Greenbaum, S., & Quirk, R. (1990). A student's grammar of the English language. London: Longman.

Gregg, K. R. (1995). [Review of the book *Linguistics and second language acquisition*]. *Second Language Research, 11* (1), 90-94.

Ha, L.Q., Sicilia-Garcia, E., Ming, J., & Smith, F. J. (2002). Extension of Zipf 's law to words and phrases. In *Proceedings of the International Conference on Computational Linguistics* (pp. 315-320). Taipei, Taiwan: COLING. Retrieved 20 September, 2005 from http://www.nslij-genetics.org/wli/zipf/ha02.pdf.

Halliday, M. A. K. (1985). *Spoken and written language*. Oxford: Oxford University Press.

Halliday, M. A. K., McIntosh, A., & Strevens, P. (1964). *The linguistic sciences and language teaching*. London: Longman.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). RANGE and FREQUENCY programs. Retrieved 13 July, 2004 from http://www.vuw.ac.nz/lals/staff /Paul_Nation.

Hockett, C. F. (1958). *A course in modern linguistics*. New York: Macmillan.

Kellerman, E. (1983). Now you see it, now you don't. In S. Gass & L. Selinker (Eds.), *Language transfer in language learning* (pp. 112-134). Rowley, Mass: Newbury House Publishers.

Kennedy, G. (2003). Amplifier collocations in the British National Corpus: Implications for English language teaching. *TESOL Quarterly, 37* (3), 467-487.

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 W. Meijs (Ed.), *Proceedings of the International Conference on English Language Research on Computerised Corpora* (pp. 133-140). Amsterdam: Rodopi.

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

Kurnia, N. S. (2003). *Retention of multiword strings and meaning derivation from L2 reading*. Unpublished doctoral dissertation, Victoria University, Wellington.

Laufer, B. (1997). What's in a word that makes it hard or easy: Some intralexical factors that affect the learning of words. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 140-155). Cambridge: Cambridge University Press.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Longman.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. Hove: Language Teaching Publications.

Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the Association for Computational Linguistics* (pp.317-324). University of Maryland, College Park, Maryland: ACL. Retrieved 9 October, 2004 from http://www.cs.ualberta.ca/~lindek/papers/noncomp.pdf.

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

Marton, W. (1977). Foreign vocabulary as problem No. 1 of language teaching at the advanced level. *Interlanguage Studies Bulletin, 2*, 33-57.

McCarthy, M. (1988). *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.

Moon, R. (1998a). Frequencies and forms of phrasal lexemes in English. In A. P. Cowie (Ed.), *Phraseology* (pp. 79-100). Oxford: Oxford University Press.

Moon, R. (1998b). *Fixed expressions and idioms in English: a corpus-based approach*. Oxford: Clarendon Press.

Nation, I. S. P. (2000). Learning vocabulary in lexical sets: Dangers and guidelines. *TESOL Journal, 9* (2), 6-10.

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

Nation, I. S. P. (2001b). How many high frequency words are there in English? In M. Gill, A. W. Johnson, L. Koski, R. Sell & B. Wårvik (Eds.), *Language, Learning and Literature: Studies presented to Håkan Ringbom* (pp. 167-181). Abo: Abo Akademi.

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/Philadelpha: John Benjamins Publishing Company.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24* (2), 223-242.

Newman, A. (1988). The contrastive analysis of Hebrew and English dress and cooking collocations: Some linguistic and pedagogic parameters. *Applied Linguistics, 9* (3), 293-305.

Ojeda, A. (2005). (2nd ed.). Discontinuous dependencies. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (pp. 624-630). Oxford: Elsevier. Retrieved 16 January, 2001 from http://linguistics.ucdavis.edu/FacultyPages/aeojeda/DiscCons.pdf.

Oxford advanced learner's English dictionary. (2000). (5th ed.). Oxford: Oxford University Press.

Palmer, H. E. (1933). *Second interim report on English collocations.* Tokyo: Kaitakusha.

Paul, D. B. & Baker, J. M. (1992).The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the International Conference on Spoken Language Processing* (pp 899-902), Banff, Alberta, Canada: University of Alberta Press.

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

Prator, C. H. (1967). *Hierarchy of difficulty.* Unpublished classroom lecture. University of California, Los Angeles.

Prime English-Korean dictionary. (2004). (4th ed.). Seoul: Doosan Dong-A Co.

Renouf, A., & Sinclair, J. (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 128-143), Harlow: Longman.

Richards, J. (1974). Word lists: Problems and prospects. *RELC Journal, 5* (2), 69-84.

Rommetveit, R. et al. (1971). Processing of utterances in context. In E. A. Carswell & R. Rommetveit (Eds.), *Social contexts of messages* (pp. 29-56). London: Academic Press.

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

Seaton, B. (1982). *A handbook of English language teaching terms and practice.* London: Macmillan.

Seymour, G. (2001). *The untouchable.* London: Bantam.

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

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

Sinclair, J. M. (Editor-in-chief). (1995). *COBUILD English collocations* version 1.1. London: Harper Collins Publishers.

Skehan, P. (1996). A framework for task-based approaches to instruction. *Applied Linguistics, 17*, 34-59.

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford: Oxford University Press.

Spöttl, C., & McCarthy, M. (2004). Comparing knowledge of formulaic sequences across L1, L2, L3, and L4. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 191-225). Amsterdam/Philadelphia: John Benjamins Publishing Company.

SPSS for windows version 10.0.5 (1999). Chicago: SPSS Inc.

Stubbs, M. (2000). Using very large text collections to study semantic schemas: A research note. Retrieved 8 June, 2006 from http://www.uni-trier.de/ uni/fb2/anglistik/Projekte/stubbs/largtext.htm.

Taylor, C. V. (1983). Vocabulary for education in English. *World Language English, 2* (2), 100-104.

Verstraten, L. (1992). Fixed phrases in monolingual learners' dictionaries. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 28-40). London: Macmillan.

Wells, R. S. (1947). Immediate constituents. *Language, 23*, 81-117.

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

Willis, J., & Willis, D. (1988). *Collins COBUILD English course*. London: Collins.

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

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

Zipf, G. K. (1935). *The psychobiology of language*. Boston: Houghton Mifflin.

Zipf, G. K. (1945). The repetition of words, time-perspective, and semantic balance. *The Journal of General Psychology, 32*, 127-148.

Zipf, G. K. (1949) *Human behaviour and the principle of least effort: An introduction to human ecology*. Cambridge, Mass: Addison-Wesley Publishing Co.

APPENDICES

Appendix 1

The pivot word-based collocation list

(The list ranked according to the frequency of the pivot word, for example, 'well' is the most frequent content pivot word of the 1,000 content words from the BNC spoken section)

1. well

as well	AVP	5754	Can I just say something else as well?
very well	AP/AVP	987	You can read very well can't you?
as well as	C/P	620	…other schools in the area used to use this facility as well as we did.
well done	VP	171	Top of the class, well done.
really well	AVP	118	I've done really well.
quite well	AP	94	I thought it was quite well organized considering.
so well	AP	80	…got the final ball wrong but a shame he'd done so well.
well known	VP	74	I mean it's more well known than it used to be.
very well	INT	68	Very well, thanks
here as well	AVP	47	Feels warm in here as well.
pretty well	AVP	43	Dave's got a job pretty well hasn't he?
very well	AP	12	he's not very well, he looks how pale he is, James is quite pale as well

2. know

you know	INT	27348	The big step is then getting the rest of the Council to take it on board, that's the big step, …the development budget there went on projects for young people, you know, so there are using there money.
know that (S V)	VP	889	…you know that this is the only room available.
know it	VP	469	It doesn't really worry me whether you know it or not.
know if {S	VP	321	I don't know if any of you are old enough

V}			to remember…
know whether {S V}	VP	247	I don't know whether anybody would disagree with that.
as I know	CA	95	...just as I know I must call you my lady Anne!
know this	VP	88	…issue of drawings to the client wished to know, wished to know this.
(not) know anything	VP	66	…I don't really know anything special about me.
know the one	VP	56	Ruth, you know the one I used to look after?
know the answer	VP	47	Well I supposed we'd all like to know the answer to that.
know one	VP	42	I know one who wouldn't stand for it.

3. so

so much	AP/AVP	1334	He loved the sea so much.
and so on	CA	872	…you know, brief films about different countries and, political systems and so on.
think so	VP	333	I think so because I don't know where we got…
so many	AVP	241	Especially now I've got so many.
and so forth	CA	194	…some things to write a few words in, some things to tick, some things to make a mark on one of the numbers and so forth…
so far	AVP	176	Now so far we've talked about things like recharging the batteries.
so good	AP	166	…I got stuck on a an article that was so good at the beginning
say so	VP	134	It's nice of you to say so, Mrs. Briant.
so often	AVP	116	…you're going along two at a time and it happens so often.
so well	AP	80	…got the final ball wrong but a shame

			he'd done so well.
(but [35]) even so	AVP	75	But even so, the government are still the people in charge…
so important	AP	67	…I'll explain that later cos it's so important.
so long	AP	65	…this has taken so long, I've completely forgot about it.
if so	CA	57	And if so, what were their findings?
so bad	AP	42	No it, it wasn't so bad really.
so to speak	AVP	40	…I have had to do in my work quite a lot of work coming up against that act, so to speak.

4. think

I think (that)	CA	25862	I think I've still got the piece about that.
think about it	VP	415	If you think about it, you slide it into where the joint would be along the line…
think so	VP	333	I think so because I don't know where we got…
think if {S V}	VP	177	Can you think if we have something meal facilities …
think of it	VP	164	What did you think of it when you first moved here?
think of one	VP	72	I know I have been but I can't think of one at the moment
think of something	VP	46	Think, think of something else.
think of anything	VP	41	…it was I really don't think of anything outstanding.

5. just

just one	NP	221	...I get a lower dragging sound than if I pick up just one.
just now	AVP	210	Well, I'm only nineteen just now!
just as	C/P	138	…it's bound to pick up just as I'm moving house…

just in case	PP	131	If they thought well I'd better go round just in case…
just before	C/P	116	…it was just before I went on holiday so my memory is kind of hazy…
just sit	VP	115	…you must just sit quiet and say nothing.
just because	C	99	…where millions of women were burnt as witches, but just because they were women!
just do it	VP	81	Just do it that way which is what you'd normally do…
just after	C/P	53	I spoke to today when I was over there just after lunch.
just here	AVP	39	Oh no I assumed it was just here...
just enough	AP	37	Then there's another long unit right at side of it, just enough to get one trailer in.

6. right

all right	AP	435	Yeah well, they were all right when we first come in but eventually…
right now	AVP	214	…What do you think they're doing all this firing on us right now.
right down	AVP	202	I mean I actually it was right down in there, right up, right round and it was painful.
right up	AVP	124	We used to walk through at back of that chapel right across Neddy 's fields, right up through Wood.
say right	VP	120	I'm just waiting until I see the granddaughters dear and then I shall say right!
{Det-the {65}} right one	NP	110	…her father's original view turned out to be the right one.
absolutely right	AP	98	You've just mentioned, you're absolutely right, of course.
quite right	AP	96	…despite the strains, and she's quite

			right !
{Det-the [85]} right way	NP	93	Fortunately it is the right way round.
{Det-the [61]} right time	NP	72	…they seem to have this ability to be able to be in the right place at the right time.
right here	AVP	70	We'll sit down right here and watch the boats on the river.
right away	AVP	63	So I went down there and well I suppose fell in love with the job right away.
my right honourable friend	NP	51	I thank my right honourable friend for that reply.
right there	AVP	49	Hold it right there!
in the right place	PP	39	I wasn't the best man, I was in the right place at the right time.
on the right hand side	NP	37	And on the right hand side, there's the goods shed.
right enough	AP	36	And then in the herring fishing time, oh it was very busy right enough…

7. up

go up	VP	676	…and in a time of rapid inflation, prices go up, salaries go up…
shut up	VP	590	I will answer if you shut up .
pick up {smo, sth}	VP	589	Now pick up the important ones, sort out your priorities.
up there	AVP	560	There he is up there, wanted for sabotage.
(be-verb) set up (sth)	VP	487	I want to set up the printer.
going up	VP	467	He said, Look up there, and you'll see a crane, he says, We're going up that ladder.
end up (with sth	VP	431	...because we could end up spending too much time on that

[152])			
coming up	VP	429	There are council elections in York coming up in May.
went up	VP	418	Ooh, I don't know what age I'd be when we went up there.
up here	AVP	323	…some of the tanks were away down there and some were up here…
get up	VP	316	Oh yeah often used to get up in the morning
up and down	AVP	311	And they're going up and down there, trying to catch hold of them.
came up (to swe, smo [60])	VP	294	...we used to go to see her, and she came up a couple of times.
(be-verb, have-verb, get-verb) picked up (sth)	VP	291	With this instruction, it will get picked up on the quality
go up	VP	288	…we used to go up and work in some of the mills.
pick {smo} up	VP	248	I'm gonna pick you up at quarter past four at the school.
up to you	AVP	242	Well that's entirely up to you, but we'll see how we go with that.
comes up	VP	229	...a man comes up last week and tells me that he's been paid off by the Daily Record he's been working for…
pick it up	VP	224	So you had to pick it up and carry it outside?
stand up	VP	217	Don't move just stand up.
gone up	VP	213	...if the rail fare's gone up as I expect it may well have done.
up to {No.}	AVP	212	Well we all put five initially, but we were told you could put it up to ten.
turn up	VP	210	…where maybe we'd sit here and nobody

			would turn up.
make up	VP	209	I had to make up an excuse why I went out!
(get-verb, be-verb) fed up (with [93])	VP	203	I don't know but I'm getting pretty fed up with this.
give up {sth}	VP	190	You either give up watching or you keep a score, don't you, yeah?
(be-verb) put up (sth)	VP	189	So another thing we said on our list or your list what you put up here was visuals.
(be-verb) made up ((of [53]) sth, sth)	VP	188	Companies are made up of hundred of individual citizens who may depend on the support services of the voluntary sector.
(be-verb) brought up (smo, sth)	VP	178	...ninety one of the women here were brought up by a mother and a father and only nine by one parent.
goes up	VP	175	Space craft goes up and when it gets to a certain height above the earth it just goes round and round and round.
build up (sth)	VP	154	The relationships you build up with our clients and the advertisers…
ended up (with [37])	VP	151	I ended up paying for him.
take up {sth}	VP	146	I'd like to take up some points as they've occurred this morning.
turned up	VP	144	…it's like nobody turned up.
picking up {sth, smo}	VP	143	Do you like picking up worms?
up to date	AVP	136	So don't start this work until everything else is up to date.
get up	VP	130	…people get up and speak for ten minutes
right up	AVP	124	We used to walk through at back of that

			chapel right across Neddy 's fields, right up through Wood.
got up	VP	121	You just automatically got up in the morning…
go up there	VP	112	…used to go up there, and I was very friendly with the old skipper there.
look up {sth, smo}	VP	111	So if they can do that, they can look up anything.
got up	VP	106	…after he got up the street near the church.
setting up {sth}	VP	105	You're setting up a place for people to live and conduct their business…
(be-verb, get-verb) taken up (by sth [14])	VP	103	Saturday and Sunday are taken up.
put up with {sth, smo}	VP	102	And they'll put up with any conditions.
phoned up {smo}	VP	99	I phoned up the bank this morning.
(be-verb) tied up (with sth, sth)	VP	97	It was all tied up together.
(be-verb) built up (sth)	VP	94	Craighead was built up high on the hill side.
woke up	VP	93	Now I woke up at five o'clock this morning.
(be-verb, get-verb) mixed up	VP	89	You'll get them all mixed up!
coming up to {swe, TIME, INF}	VP	89	She's she's a waitress in Chester so it can shows you how much cross section we've got and she said her and her

			boyfriend was coming up to Blaenau to be on the picket line and so on and we were quite friendly with her.
put it up	VP	79	Now he'll put it up and hold it.
rang up	VP	79	...she rang up and said that she wanted to go and would I go with her?
straight up	AVP	79	The smoke would just rise straight up…
building up {sth}	VP	78	Building up trust and friendship takes time …
picked it up	VP	71	Perhaps he just picked it up in the right place.
standing up	VP	71	They will be standing up for quite a long time.
came up with {sth}	VP	67	And she came up with a list which is pleas from the switchboard supervisor.
catch up (with[29]) {sth}	VP	66	We've got to catch up.
open up {sth}	VP	65	...let's not open up the debate about chairmanship of meetings …
walk up	VP	65	Did you walk up or did somebody take you up?
(be-verb) finished up (with sth [19], sth)	VP	62	That would be finished up?
finish up (with sth [38], sth)	VP	61	You still finish up with the same amount.
(be-verb) given up (smo, sth)	VP	61	I've given up my sewing.
(be-verb) grown up (sth)	VP	60	…at least until the children are grown up and probably well beyond?
(be-verb)	VP	59	Yes, he's caught up with two more cars

caught up (sth, with sth, in sth)			
got up	VP	59	…she got up and waved to Mrs…
picks up {sth}	VP	59	Chris picks up these things from school…
(be-verb) split up (with sth [8], into sth [7])	VP	59	What all split up, yeah?
keep up {sth}	VP	59	I am very concerned that in the present recession we shall not be able to keep up our work …
getting up	VP	56	Aha, that's what I do, it's like getting up in the morning.
(be-verb) opened up (sth)	VP	56	Do you know when a new branch is opened up?
bring up {sth}	VP	55	...perhaps if we could just bring up the item you wanted to mention.
taking up {sth}	VP	54	You know it was taking up so much space.
looked up (sth)	VP	51	And I looked up and said it's here!
picked {smo} up	VP	51	You should have phoned me at ten o'clock, I would have picked you up.
picking {smo} up	VP	51	They might be picking you up at ten o'clock.
clear up {sth}	VP	51	Yeah, still wanna clear up the spots on my back.
hands up	NP	50	Now have a look at the work now, hands up and see if you could tell me…
come up here	VP	48	When are you supposed to come up here again?
gave up	VP	47	she gave up her flat and went to live with

{sth}			her mother.
keep up with {sth, smo}	VP	47	Do you ever keep up with people outside work?
set it up	VP	46	Well it could technically, it depends how you set it up.
(be-verb, get-verb) held up (by sth, smo [8], with sth [3], sth)	VP	45	There was a horse running round and we got held up with traffic.
take it up	VP	45	Could I have a copy of the letter, please, can I take it up?
making up (sth)	VP	41	I mean making up conversations all over the place.
speak up	VP	41	Chair, can you ask people to speak up?
up to now	AVP	41	That all sounds the best one up to now doesn't it?
use up {sth}	VP	41	And when did you use up your seventy thousand?
wake up	VP	41	…the baby wake up at seven so I come up and stayed up.
{Det-the [10]} back up	NP	40	...you can see the sort of backup from the brochure…
draw up {sth}	VP	40	I mean, in an ideal world what agenda would you draw up?
stood up	VP	40	She stood up straight and still.
add up (sth)	VP	39	The trick is simply to add up the numbers.
follow up	VP	39	…the Board has not failed to follow up its very good work.
giving up {sth}	VP	39	We've all heard of women giving up work for their children, but men?
grow up	VP	39	…if children grow up healthy, capable

			and ready to work for the good of their neighbours.
took up {sth}	VP	39	...the door took up so much room, you see.
putting up {sth}	VP	38	...they've only been able to balance their proposed production this year, by putting up the price by fifty pence.
started up	VP	38	I was as bad as him when I started up in the butchering.
(be-verb) cut up (sth)	VP	37	...worn blankets cut up to make sleeping bags for children.
all the way up	AVP	36	All the way up, the whole passageway was just bin bags.
(be-verb) brought up (sth)	VP	36	Yeah, this was brought up last night at the meeting at of Old Harlow and Potter Street forum.
get up	VP	36	…couldn't get up that steep hill.
(wear-verb, get-verb) make up	NP	36	I don't enjoy wearing make up, I feel dirty with it on, but, what ever you want to do, it's up to yourself.
running up	VP	35	He's over there, he's running up through there.
(be-verb) drawn up (sth)	VP	35	Unfortunately, whether the contract drawn up in the first place has been a correct one or not, I'm not sure.
break up {sth}	VP	35	…I didn't break up this marriage…
grew up	VP	35	…in the small Texas town where he grew up.
growing up	VP	35	They're growing up in the village.
add up to {No.}	VP	34	Add up to a hundred and eighty degrees.
looking up (sth)	VP	34	...see better, you was looking up it on the front rows, you see.
make up {smo's}	VP	34	...you have then got to make up your own mind what you're going to do for the

mind			casualty and your own safety···
run up (sth)	VP	34	I started to run up, straight up the hill, lucky the monster was running straight at me and sticking out of the crowd and he tripped over and after the end
bringing up {smo}	VP	34	If you're eighteen to twenty four years old and you're bringing up a child on your own.
come up with {sth}	VP	33	I don't know who comes up with these ideas.
comes up with {sth}	VP	33	If somebody comes up with a suggestion for a change to a procedure at this meeting···
(all) lined up	VP	32	He had us all lined up after we'd been riding round and he started enquiring how long we'd been out of hospital .
bring it up	VP	32	If you don't bring it up at the first part with plan to proceed, you can't do it here.
hurry up	VP	32	Hurry up otherwise you're out.
look it up	VP	31	...it'd be easier to look it up on a graph.
start up (sth)	VP	30	That was it, those particular shopkeepers didn't start up business again.
wind up	VP	28	Could you wind up now please?
tie up (with sth [11])	VP	26	...they don't tie up with you so you're alright.
{Det-the [7], a [6]} follow up	NP	24	···having done the follow up on it, it doesn't seem to me that···
going up (to swe)	VP	24	Take their time going up to Scotland.
getting up (swe)	VP	21	Well, it's getting up there.
bring up {smo}	VP	19	I think the best possible way to bring up children is where there's a father and mother provided the marriage is stable and balanced, cos the children have a

			role model from the father and the mother.
back up (sth)	VP	19	...if so, we have to look at the evidence then that was used to back up this claim.
break up	VP	19	…when they break up for Easter next week.
{Det-a [10]} build up	NP	17	You get a build up of black heads around the nose, and around the chin.
run up {to sth}	VP	16	We're gonna hear it time and time again in the run up to the Election.
built up (N)	AP	15	If you build a relief road which is fairly tightly in to the built up area, that relief road will cater for both the long distance bypassable traffic and the local traffic.
{Det-the [9], a [4]} set up	NP	13	This was all part of the set up.
shut up	VP	13	…so that you got the stables shut up before it got dark at four o'clock…
getting up	VP	12	…why am I getting up and speaking.
{Det-the [6]} break up	NP	10	The major impact was the break up of the family unit.
{Det-a [8]} grown up	NP	9	…you are a grown up now aren't you?
{Det-a [6]} turn up	NP	8	I was just lucky with a turn up though.
running up (to sth)	VP	7	You're running up to the thing, yeah?
bringing up {sth}	VP	7	I'm just bringing up a point that No problem about bringing up his previous convictions?
turn up	VP	7	Let's turn up the sound then…
follow up	AP	6	... we've got a follow up meeting about this training.

turned up {sth}	VP	6	Bishops then turned up the pressure in a search for an equaliser…
{Det-the [3]} tie up	NP	5	So what's the tie up between those?
keep up {sth}	VP	4	You do in a swimming pool because what happens to you, it makes you keep up the top, doesn't it?
put it up	VP	4	I mean he would of put it up for sale…
stood up for {sth}	VP	4	I mean lunch time we were hearing how, because they stood up for what was right…
wind up	VP	4	Can I suggest that we wind up the meeting?
running up {sth}	VP	2	I'm running up these debts…
{Det-the [1], a [1]} wind up	NP	2	...just look at it if we get a wind up, if we, a wind starts to blow it'll move a lot of this...
a stand up	NP	2	…has everybody got a stand up?
make up	VP	2	Where was the little girl, or younger I don't want to make up because, but now I feeling I'm aged it, I should be wearing some make up here…
a (lousy) turn up	NP	1	What a lousy turn up!
turned up {nose}	NP	1	I know Abbie's sort of like real sort of turned up nose.
wind up	AP	1	Mine's a good old fashioned proper mechanical wind up job yes.

8. go

go out	VP	711	…we'll go out and have a drink and have a meal like…
go up	VP	676	…and in a time of rapid inflation, prices go up, salaries go up…
got in (swe, sth)	VP	506	I'm not sure how they got in the shops.

go away	VP	345	And does this thing go away or does it stay for?
go home	VP	293	Or if you wanted to go home for your dinner…
go up	VP	288	…we used to go up and work in some of the mills.
go off	VP	205	…you get to the time when you're all going to get in the car and go off.
go in there	VP	187	No somebody could go in there who belonged to the church anyway…
go to bed	VP	184	…before you go to bed at night.
go there	VP	165	And the lads were glad to go there.
let's go	VP	161	You know I'm the one, Come on, let's go, man.
here we go	CA	148	Oh here we go, it's started, it's started.
go ahead	VP	131	…I can go ahead as soon as I've got some dates…
go up there	VP	112	…used to go up there, and I was very friendly with the old skipper there.
go to school	VP	102	My children do go to school and have been for some months now.
go wrong	VP	91	…there are many different reasons why things go wrong.
go along with {sth}	VP	89	…I could go along with that approach as well.
go down there	VP	88	So he says you ought to go down there, you can mention my name.
go again	VP	86	You want to go again?
go for it	VP	83	I'm sorry if you're interested in this then go for it.
go upstairs	VP	81	Right, I've got to go upstairs and put my face on, very quickly.
go along	VP	77	Right, we'll go along, we'll go the long way.
go around	VP	76	…when I go around seeing the conditions,

			some of our people work in
go through it	VP	74	…we could just go through it very quickly.
go forward	VP	66	…if those proposals go forward.
go outside	VP	63	Well you can go outside in a few minutes.
go over there	VP	59	How much does it cost to go over there?
go straight	VP	59	…the wall just going to go straight across the garden.
go with it	VP	49	…we should have the evidence to go with it.
let go	VP	48	She's not gonna let go of it!
go that way	VP	46	…we can't go that way when we're not staff
go now	VP	46	We'll go now okay.
go anywhere	VP	44	…I can't possibly go anywhere during the day on Wednesday.
in one go	PP	43	…it should be done all in one go.
go any further	VP	37	Before we even go any further!
go ahead	VP	36	Do we want them to go ahead without us.
go further	VP	36	Could we go further and suggest why they think that?
{have-verb} {Det-a [21]} go at it	VP	32	Don't be frightened to have a go at it.
go straight	VP	32	…we will go straight into matters…
go ahead	NP	17	…it still depends on the government to give the go ahead?
go away	VP	6	I mean you could always go away on holiday, if you'd like.
go with it	VP	4	…your hairstyle to go with it…
go around	NP	1	…whatever such programmes, there are go around these days!

9. now

right now	AVP	214	…What do you think they're doing all this firing on us right now.
just now	AVP	210	Well, I'm only nineteen just now!
here now	AVP	128	We've got lovely young doctors here now.
now and again	AVP	113	Every now and again I think about it and I'm deeply embarrassed.
{No.} years now	NP	93	I mean I've stayed at home for nearly four years now.
now now	INT	53	…now now here we go.
for now	PP	50	And you can use this little book for now.
go now	VP	46	We'll go now okay.
going now	VP	46	Are you going now?
by now	PP	43	It could be anywhere by now.
from now on	PP	43	From now on you'll see him taking a part.
up to now	AVP	41	That all sounds the best one up to now doesn't it?

10. said

said well (that, what) {S V}	VP	1135	…they said well it might take us two days.
said to {smo}	VP	1076	He said to me on the phone he didn't know a great deal…
said if {S V}	VP	254	Right, if you, he said if you don't take it you lose it, which is absolutely right.
said yes	VP	177	Fifty three people have said yes, they've been physically abused.
said yeah	VP	121	You asked me if I wanted salt out and I said yeah, so you don't get it out!
said something	VP	98	She said something about changing it.
said anything	VP	47	…nobody said anything at all.

- See the attached CD for more data.

Appendix 2

The frequency-ranked collocation list

1	you know	INT	27348	The big step is then getting the rest of the Council to take it on board, that's the big step, …the development budget there went on projects for young people, you know, so there are using there money.
2	I think (that)	CA	25862	I think I've still got the piece about that.
3	a bit	NP	7766	Can you move round this side a bit?
4	(always [155], never [87]) used to {INF}	VP	7663	We used to look forward to them coming.
5	as well	AVP	5754	Can I just say something else as well?
6	a lot of {N}	NP	5750	I got a lot of letters from the children there and which was very gratifying.
7	No. pounds	NP	5598	I'd also ask you to consider costs of ten pounds.
8	thank you	VP	4789	Well thank you for that that's a very good start to the evening.
9	No. years	NP	4237	I've done it for seven years.
10	in fact	PP	3009	In fact, if the previous speaker has complained about waiting in patience, I have waited forty years to tell this story in the assembly…
11	very much	AVP	2818	…they enjoy it very much…
12	No. pound	NP	2719	I've got to take a taxi of one pound forty a day to shop...
13	talking about {sth}	VP	2489	It was a different from what you're talking about.
14	(about [91]) No. percent (of sth [580], in sth [54], on sth	NP	2312	As I said, we've already got forty one percent of them…

	[44], for sth [38])			
15	I suppose (that)	VP	2281	I suppose that was one way of nothing being done.
16	at the moment	PP	2176	Well I haven't done anything at the moment because I didn't think it was worth it actually.
17	a little bit	NP	1935	And the other times your concentration will drop a little bit.
18	looking at {sth}	VP	1849	I think there's another way of looking at that.
19	this morning	AVP/NP	1846	Oh she was screaming this morning.
20	(not) any more	AP	1793	…women shouldn't have off days any more.
21	come on	INT	1778	No, it's not a verb, come on, what is it?
22	number No.	NP	1661	Number six which is the thing that we have to look at.
23	come in (swe, sth)	VP	1571	We're about to finish, so please come in.
24	come back	VP	1547	We'll come back to you in a second.
25	have a look	VP	1471	You can go and have a look.
26	in terms of {sth}	PP	1463	I think it was one of the things which never really took off in terms of the accident.
27	last year	AVP/NP	1347	That was last year or was that two years ago?
28	so much	AP/AVP	1334	He loved the sea so much.
29	No. years ago	AVP	1314	That was last year or was that two years ago?
30	{Det-the [879], this [39], a [21]} county council	NP	1273	And we put in a report to the county council in the morning saying there was a fire.
31	this year	AVP/NP	1255	And these conditions apply from the first of April of this year.
32	go back	VP	1250	I must go back to the sea, she said.

33	last night	AVP/NP	1244	That's what I was doing last night.
34	rather than	C/P	1243	But I might do it by talking to them rather than by writing to them.
35	come out	VP	1163	I mean Henry won't come out
36	very good	AP	1160	The weather was very good, wasn't it?
37	I hope (that [455]) {N, S V}	CA	1155	Oh well I hope, I hope it goes right tomorrow.
38	No. times	NP	1147	Three times a week perhaps?
39	that way	NP	1145	…and I hope that I can train him to keep it that way
40	said well (that, what) {S V}	VP	1135	…they said well it might take us two days.
41	at the end (of sth [737])	PP	1122	Okay, I'll give you time to fill that in at the end.
42	{Det-that [425], this [146], the [142]} sort of thing	NP	1113	That's the sort of thing that the government encouraged.
43	for example (if S V [30])	PP	1107	For example, many people have four weeks' holiday a year, but not many people can take that four weeks all in one go…
44	as far as	C	1079	But of course as far as Britain was concerned, this could only be intervention against France.
45	said to {smo}	VP	1076	He said to me on the phone he didn't know a great deal…
46	mean (that) {S V}	VP	1066	I mean that that's another alternative is that we don't bother with pictures.
47	come on (to swe, smo [65])	VP	1059	Gary Mills who's missed half a dozen games through injury is going to come on now and he's going to take the place of Neil Lewis.
48	{FREQUENCY, QUANTITY} a week	NP	1056	I'm now sixty two and I go dancing twice a week!
49	all the time	NP	1044	…what's up all the time?

50	thank you very much	VP	1041	We just want to say thank you very much to all of you for coming and listening.
51	too much	AP/AVP	1034	Too long, waste too much time.
52	over there	AVP	1017	He's over there, he's running up through there…
53	that sort (of sth [953])	NP	1016	…and I were used to do that sort of work…
54	looking for {sth}	VP	990	What are you looking for?
55	make sure (that [394]) {S V}	VP	990	It is your responsibility to make sure that money is paid each and every week.
56	very well	AP/AVP	987	You can read very well can't you?
57	{Det-the [47]} last week	AVP/NP	956	The other thing that we mentioned last week as well is the moral message…
58	in the morning	PP	952	And I'd started at six in the morning.
59	it seems {N, A, to INF, that S V}	CA	945	It seems quite clear, I thought.
60	next week	AVP/NP	940	I want you to bring it in its completed state to next week's lesson please?
61	a number of {sth}	NP	929	Yeah, there are a number of courses which you do.
62	out there	AVP	929	Is there an audience for women's football out there?
63	what I mean	CA	929	Do you know what I mean?
64	get in (swe, sth)	VP	912	So you might want to remember that when you get in there…
65	find out {sth}	VP	908	Do you think she's got something to find out later on?
66	know that (S V)	VP	889	…you know that this is the only room available.
67	leave it	VP	886	Let's just leave it for the moment.
68	at home	PP	884	If you've got plenty of products at

				home, fine, use them, okay?
69	and so on	CA	872	…you know, brief films about different countries and, political systems and so on.
70	(about [226]) No. minutes	NP	867	This device should give us a single analysis in about five minutes rather than ninety as at present as we don't need to separate the mixtures, we can do the analysis directly.
71	(do) n't mind (sth)	VP	862	I'll have one a little later if you don't mind.
72	other people	NP	839	…you won't encourage other people to send them mail…
73	not really	AVP	837	Not really, can you?
74	talking to {smo}	VP	829	And you're talking to your mate and it's all just happening by magic.
75	mind you	INT	822	Mind you, you'd have to be quick, wouldn't you?
76	want it	VP	819	Mind, we wouldn't want it every time.
77	much more	AP/AVP	816	That's a much more accurate way of finding the gradient.
78	looked at {sth}	VP	805	Then I looked at it again and I realized that it wasn't too hard.
79	the other one	NP	805	What's the other one?
80	(at [207], about [110], till [50], by [24]} half past No.1~12	NP	798	...we had to start at half past eight and finish at six…
81	some people	NP	797	Some people use it as a perfume.
82	this week	AVP/NP	794	Something that occurred to me this week.
83	this time	AVP/NP	787	I shall write it down this time.
84	very nice	AP	784	And even in just the police station, in the small holding cells they have, it's not very nice.

85	I see	INT	756	Oh, I see, oh well in that case we will view it…
86	I bet (S V)	VP	746	I bet you'll be proud of her!
87	these things	NP	742	I'm just putting these things in your mind to put some doubt in your mind.
88	call it (A, N)	VP	737	Could you call it beverage?
89	(be-verb) not sure	AP	721	I'm not sure, I'm not sure they have been there.
90	at the time	PP	717	…what's going on in England at the time?
91	thought that {S V}	VP	714	So I thought that there was plenty of time as this is my first day off from work.
92	going out	VP	712	I'm going out now.
93	it comes	CA	712	If you add all that up it comes to about three million words.
94	go out	VP	711	…we'll go out and have a drink and have a meal like…
95	quite a lot	NP	711	...people probably need quite a lot of support.
96	even if	C	707	But even if no one else can help, I will take part in it.
97	last time	AVP/NP	704	Now what were we looking at last time?
98	hang on	VP	701	Actually, hang on!
99	believe that (S V, N)	VP	696	Oh I can't believe that !
100	(be-verb, become-verb) interested in {sth}	VP	689	I'm more interested in the, the kind of s so called philosophical ideas that have been raised…
101	I mean (if) {S V}	VP	689	I mean if we're gonna discuss that that needs to be, you need to go over that in more detail.
102	{Det-the [255]} only one	NP	682	I was the only one who stood.
103	anything else	NP	678	Anything else, yeah?

104	go up	VP	676	…and in a time of rapid inflation, prices go up, salaries go up…
105	listen to {smo, sth}	VP	674	Listen to these words.
106	(for [215]) a long time (ago [78])	NP	666	I haven't seen one of them for a long time
107	find it (A)	VP	656	I find it extraordinary.
108	in the middle (of sth [342])	PP	652	...he could have sunk in the middle of a gully or something.
109	(with [26], in [24]) all sorts (of sth [530])	NP	650	...it can be all sorts of funny things.
110	{Det-the [590]} other day	AVP/NP	650	I only asked him the other day…
111	see if {S V}	VP	643	We'll have to see if we can or not.
112	in front of {smo, sth}	PP	642	You have all the documents in front of you.
113	next year	AVP/NP	639	This should be brought up to date for next year.
114	looks like {N, S V}	VP	638	It looks like an allergic rash.
115	one thing	NP	638	Can I just ask one thing if we've finished?
116	agree with {smo, sth-that [154], it [21]}	VP	632	Anybody agree with that?
117	coming in (swe, sth)	VP	629	There might be a security camera there that catches them coming in.
118	as well as	C/P	620	…other schools in the area used to use this facility as well as we did.
119	{Det-the [272], a [225]} new settlement	NP	616	…the District Council considers it essential that there is a new settlement to take that amount of housing that cannot be accommodated within the Southern Ryedale area.
120	every time	AVP	615	Every time you've done the chip-pan fire I've always missed out.
121	use it	VP	615	Right, well you , how you going to use it?

122	(do) n't worry (about sth- it [158], that [73])	VP	606	Don't worry about that.
123	this afternoon	AVP/NP	600	I'm hoping he will ring me this afternoon.
124	(in [432]) this country	NP	599	And when did you come to live in this country?
125	a bit more	AP/AVP	596	Now we'll talk a bit more about signals because presently we're going to have a look at a a little bit of video which is showing commentary.
126	come here	VP	595	You can come here any time.
127	{Det-the [561]} other side	NP	594	I could reach the bottom when I got to the other side.
128	shut up	VP	590	I will answer if you shut up .
129	sit down	VP	590	…will you sit down for two more minutes.
130	like it	VP	589	I quite like it.
131	pick up {smo, sth}	VP	589	Now pick up the important ones, sort out your priorities.
132	in a minute	PP	585	I'm gonna to speak to you in a minute.
133	(for [46]) the week	NP	582	That was my achievements for the week.
134	at night	PP	578	On the other hand, at six o'clock at night I've gone home.
135	(a [435]) good idea	NP	577	That's a good idea.
136	(oh [426]) my god	INT	575	My God, will it really cost this much?
137	very difficult (N, P, to INF)	AP	569	They find it very difficult to manage a budget.
138	something else	NP	567	Can I just say something else as well?
139	very important	AP	563	And the names in the story are very important.

140	(on [171]) the other side	NP	561	And then I forgot to wipe the other side
141	up there	AVP	560	There he is up there, wanted for sabotage.
142	the number of {sth}	NP	559	You only have to look at the number of visitors going to places such as Nepal to see the increase there…
143	somebody else	NP	556	Let's ask somebody else, shall we?
144	many people	NP	555	…how many people have walked out…
145	my lord (Mayor [85])	INT	551	My Lord Mayor, I started by saying how angry and dismayed I was on the twenty fourth of July.
146	as soon as	C	546	I want you to do it as soon as possible!
147	no way	NP	542	There's no way of removing the tape.
148	get out	VP	540	...that was the only way that these children were able to get out!
149	came in (swe, sth)	VP	537	I was alright when I first came in because I'm under the doctor under the hospital at the moment.
150	at school	PP	533	If you got in trouble at school, you got in trouble at home.
151	get rid of {sth}	VP	531	...she'd tried to get rid of it.
152	on top (of sth [359])	PP	528	I mean you're never going to control, you're never going to be on top of it…
153	every day	AVP/NP	527	I mean if you bank every day it costs you a fortune.
154	of the day	PP	522	...anybody can come into this building at virtually any time of the day or evening…
155	excuse me	VP	521	Excuse me for a moment.
156	remember that (S V, N)	VP	520	I think, we must remember that cigarettes are a drug.

157	(at [459]) the same time	NP	519	And if we all talk at the same time, he can't hear anything.
158	go round	VP	519	Go round and ask everybody in the area.
159	live in {swe}	VP	519	Do your children still live in Harlow?
160	other things	NP	515	What other things did you do at this school?
161	keep it (AV, A)	VP	507	How did they keep it cool on the boat?
162	one day	AVP	507	One day she said, I'll become the queen…
163	got in (swe, sth)	VP	506	I'm not sure how they got in the shops.
164	in other words	PP	502	In other words it could be correct me if I'm wrong
165	read it	VP	500	You can read it if you've got time.
166	in the past	PP	497	We have had that in the past, haven't we?
167	(in [442]) those days	NP	496	Oh yes, it was marvellous, that's how they used to make it, in those days.
168	(just [32]) carry on (with sth [45], sth)	VP	496	We cannot carry on increasing water service charges by ten to fifteen percent, year in and year out.
169	{Det-the [476]} only thing	NP	496	...the only thing to bear in mind is that we should be a little bit careful with the changing environment...
170	that much	AP/AVP	496	But I don't think they change that much.
171	at that time	PP	493	Tell me something of the hours you used to work at that time.
172	say well (that) {S V}	VP	491	Of course, they'll say well there's no demand.
173	going back	VP	490	So we'll have to, I think, keep going back to these throughout the year…
174	work out {sth}	VP	489	I'm not trying to make it hard for you, I'm trying to work out what is

				useful for you…
175	of time	PP	488	…I think that would be a better use of time.
176	these people	NP	488	…how many of these people ever travel by bus…
177	(be-verb) set up (sth)	VP	487	I want to set up the printer.
178	at No. o'clock	PP	483	I 'm able to run about banging doors at four o'clock in the morning.
179	some time	NP	483	I think it would have to be some time this week.
180	this sort (of sth [459])	NP	483	In black wool, navy or this sort of colour.
181	at work	PP	481	He was always at work, always Sunday.
182	{be-verb} aware of {sth}	VP	478	Were you aware of that?
183	no problem	NP	478	There's no problem!
184	went out (to swe [59])	VP	478	There were animals who lived in the forest who went out hunting only at night.
185	{S V} and yet {S V}	C	475	You're not so relaxed that you want to do this but you're relaxed and yet you're happy with life as well.
186	get back	VP	470	...the men mainly wanted to get back to work, because they saw time running out.
187	know it	VP	469	It doesn't really worry me whether you know it or not.
188	speak to {smo}	VP	469	Shall I speak to Paula about that then?
189	not necessarily (V, A, N, P)	AVP	468	...actually that's not necessarily true…
190	seen it	VP	468	You haven't seen it?
191	going up	VP	467	He said, Look up there, and you'll see a crane, he says, We're going up that ladder.

192	{Det-the [372]} other thing	NP	466	The other thing we're looking at, of course, is your technique as a driver.
193	No. days	NP	466	We arranged it two days or three days...
194	goes on	VP	462	...it shows that we're receptive to people's needs and we care about what goes on.
195	this way	NP	460	We shouldn't be allowed to present animals in this way.
196	later on	AVP	457	Oh I see so you are going out with him later on?
197	more or less	AVP	457	I think last year, more or less, shows that we really have to concentrate on one
198	listening to {smo, sth}	VP	456	I'm listening to somebody else now.
199	no good	NP	455	It's no good
200	went down	VP	454	Well, I went down and had a look at it as a kid, you know.

- See the attached CD for more data.

Appendix 3-1

The top 50 collocational groups from the top 150
content pivot words based on the 10 million word spoken section of the BNC

1	*you know*	INT	27348	The big step is then getting the rest of the Council to take it on board, that's the big step, …the development budget there went on projects for young people, you know, so there are using there money.
2	*I think (that)*	INT/CA	25862	And I think that the last year has actually improved the attendance at the theatre…
3	a bit	AVP/NP	7766	Can you move round this side a bit?
4	*as well*	AVP	5754	Can I just say something else as well?
5	a lot of {N}	NP	5750	I got a lot of letters from the children there and which was very gratifying.
6	No. pounds	NP	5598	I'd also ask you to consider costs of ten pounds.
7	thank you	VP	4789	Well thank you for that that's a very good start to the evening.
8	*(for) No. years*	NP/PP	4237	I've done it for seven years.
9	*in fact*	AVP/PP	3009	In fact, if the previous speaker has complained about waiting in patience, I have waited forty years to tell this story in the assembly…
10	*very much*	AP/AVP	2818	...the water temperatures didn't change very much.
11	No. pound	NP	2719	I've got to take a taxi of one pound forty a day to shop...
12	talking about {sth}	VP	2489	It was a different from what you're talking about.

13	at the moment	PP	2176	Well I haven't done anything at the moment because I didn't think it was worth it actually.
14	a little bit	AVP/NP	1935	And the other times your concentration will drop a little bit.
15	looking at {sth}	VP	1849	I think there's another way of looking at that.
16	this morning	AVP/NP	1846	Oh she was screaming this morning.
17	(not) any more	AP/AVP	1793	…women shouldn't have off days any more.
18	come on	INT	1778	No, it's not a verb, come on, what is it?
19	number No.	NP	1661	Number six which is the thing that we have to look at.
20	come in (swe)	VP	1571	We're about to finish, so please come in.
21	come back	VP	1547	…you can come back in a minute and see if it's what you think.
22	have a look	VP	1471	You can go and have a look.
23	*(Det-the [86]) last year*	AVP/NP	1347	That was last year or was that two years ago?
24	*so much*	AP/AVP	1334	He loved the sea so much.
25	*No. years ago*	AVP	1314	That was last year or was that two years ago?
26	{Det-the [879], this [39], a [21]} county council	NP	1273	And we put in a report to the county council in the morning saying there was a fire.
27	*this year*	AVP/NP	1255	And these conditions apply from the first of April of this year.
28	go back	VP	1250	I must go back to the sea, she said.
29	*last night*	AVP/NP	1244	That's what I was doing last night.
30	*{Det-the [1178], a [30]} fact that {S V}*	NP	1208	And the fact that the license is not in puts you in difficulty!
31	come out	VP	1163	I mean Henry won't come out
32	very good	AP	1160	The weather was very good, wasn't

				it?
33	No. times	NP	1147	Three times a week perhaps?
34	(in) that way	AVP/NP	1145	…and I hope that I can train him to keep it that way
35	said well (that, what) {S V}	VP	1135	…they said well it might take us two days.
36	***at the end (of sth [737])***	PP	1122	Okay, I'll give you time to fill that in at the end.
37	{Det-that [425], this [146], the [142]} sort of thing	NP	1113	That's the sort of thing that the government encouraged.
38	mean (that) {S V}	VP	1066	I mean that that's another alternative is that we don't bother with pictures.
39	come on	VP	1059	Yeah now I'll come on to that in a minute.
40	{FREQUENCY, QUANTITY} a week	AVP/NP	1056	I'm now sixty two and I go dancing twice a week!
41	all the time	AVP/NP	1044	…what's up all the time?
42	thank you very much	VP	1041	We just want to say thank you very much to all of you for coming and listening.
43	***too much (N)***	AVP/AP	1034	Too long, waste too much time.
44	over there	AVP	1017	He's over there, he's running up through there…
45	that sort (of sth)	NP	1016	…and I were used to do that sort of work…
46	looking for {sth}	VP	990	What are you looking for?
47	make sure (that) {S V}	VP	990	It is your responsibility to make sure that money is paid each and every week.
48	very well	AP/AVP	987	You can read very well can't you?
49	(be-verb) said to {smo}	VP	985	He said to me on the phone he didn't know a great deal…
50	***(Det-the [47]) last***	AVP/NP	956	The other thing that we mentioned

	week			last week as well is the moral message…

- The 15 italicised items *you know, as well, very much*, etc are included in both lists.

Appendix 3-2

The top 50 collocational groups from the top 150 content pivot words based on the 10 million word written corpus (including ACE, Brown, LOB, FROWN, FLOB, Kolhapur, WW corpora and 3 million token texts from the BNC written section)

1	of course	INT	2698	During the hottest part of the day, of course, the sun comes straight down...
2	as well as	C/P	1979	...it is time to start making decisions on a political as well as a financial basis.
3	*in fact*	AVP/PP	1679	In fact, however, both principles have always been nebulous and loosely defined.
4	*I think (that)*	INT/CA	1565	I think that it reminded us, we used to be friends.
5	*{Det-the [1351], a [28]} fact that {S V}*	NP	1482	It was even more annoying to have to face the fact that he didn't stop thinking about her.
6	{the [209], smo's, A} way to {INF}	NP	1467	Politics means the way to rule a country. But a country is made up of people.
7	said that (N, S V)	VP	1345	The spokesman said that the interview had been conducted six months ago…
8	a number of {sth}	NP	1313	I only found a number of heads turned in my direction.
9	*(for) No. years*	NP/PP	1287	The house they have lived in for four years is spacious.
10	*so much*	AVP/AP	1238	There is so much talk about planning but so little visible evidence of it.
11	*(No.) years ago*	AVP	1233	I met Perrig in Switzerland more than two years ago.
12	away from {swe, sth}	AVP	1220	Harry had a dreadfully mangled

				finger, which would have kept most men away from work.
13	*this year*	AVP/NP	1156	One study this year indicated that the nutritional value of refugee rations is less than…
14	know that (N, S V)	VP	1104	He didn't even know that she was there.
15	*you know*	INT	1074	"This sounds like nonsense. I am a comic artist, you know."
16	each other	AVP	1042	Twelve years ago Mary and Daniel fell in love with each other.
17	*(Det-the (58)) last year*	NP	1027	The baseball team had their awards banquet here last year.
18	look at {sth}	VP	981	"Look at me. I'm interesting"
19	the number of {sth}	NP	949	The number of books issued has dropped by 10,000 compared with last year.
20	say that (N, S V)	VP	932	It is too much to say that now a new sense of beauty has overtaken our congregations…
21	just as	C/P	923	...but only glad to know that she kissed in the Western fashion and not just as Moslems do.
22	*as well*	AVP	917	...they found they were being helped and were helping others as well.
23	*at the end (of sth [797])*	PP	877	There was something else, a grey shadow at the end of his bed.
24	even if	C	862	Even if there are no livestock, the farmer cannot leave the farm for long periods…
25	at home	PP	839	This had influenced her decision to stay at home.
26	this time	AVP/NP	839	This time justice has been done

27	in the world	PP	828	My child and my wife are the most important things in the world to me.
28	*very much*	AP/AVP	823	It sounds very much like Hindi spoken with a bad accent.
29	(in [370], for [231]) No. days	NP/PP	799	He expects to be out in two or three days if all goes well.
30	(the) only one (N)	NP	796	The only one I liked was my mate, Rob.
31	at the same time	PP	778	The two events are taking place at the same time.
32	much more (A, N)	AP/AVP	776	We shall have to face the issue of Maori sovereignty much more openly if we are to cope with it.
33	at the time (of sth)	PP	765	I intend to make clear that the Administration followed a prudent policy toward Iraq at the time...
34	*too much (N)*	AVP/AP	764	She meant too much to me.
35	*last night*	AVP/NP	755	Anna said. "Where were you last night?"
36	so many (N)	AVP/AP	721	At best these plays may be said to be successful attempts.
37	(be-verb) set up (sth)	VP	711	A Parliamentary Committee had been set up in 1945 to consider the question…
38	in No. years	PP	707	And in ten years the average height of a ten - year - old has increased by half an inch…
39	{Det-the [214], a [158], smo's} (A) state of {sth}	NP	705	The answer depends upon the state of our consciousness.
40	any other {N}	NP	672	This is not an investigation conducted on mice, monkey, guinea pig, or rabbit or any other small animal.

41	so far	AVP	660	Mrs. Gandhi has not, so far, made such a gesture.
42	on {Det-the [269], smo's} way (AV, to swe)	PP	642	The next morning, Gary dropped by her place on his way to work.
43	went on {N, verb-ing}	VP	635	Through a mist of tears she went on smiling - the most wonderful smile I'd ever seen.
44	think of {sth}	VP	629	"Well, what did you think of it?"
45	even though	C	628	This is true even though we may be exposed to it for only a few hours every now and then.
46	for the first time	PP	620	He turned the key softly in the lock for the first time since he'd slept in this house.
47	one day	AVP/NP	612	One day, while teaching, he suddenly felt dizzy.
48	thought that (N, S V)	VP	585	They all thought that he was rich and lazy.
49	all right	AP	579	Everything will be all right.
50	*(Det-the [18]) last week*	AVP/NP	564	Lucy and I went out last week and bought it, secretly.

Appendix 4

The results of the analysis of *predictability in L1* of the first 100 collocations

Rank	Collocations	The primary meaning according to the two English dictionaries		The primary meaning according to the two Korean dictionaries			Korean translation of the first constituent	Korean translation of the second constituent	Korean translation of the third constituent	Korean translation of the whole meaning of the collocation	Do all the Korean translations match with the whole translation?	Predictable/Unpredictable	Sample sentences
1	you know	O/C	O/C	E/P	E/P		너/당신- deferential	생각/아시죠 - deferential		(너/당신) 생각/아시죠	grammatical	*	The big step is then getting the rest of the Council to take it on board, that's the big step, ..the development budget there went on projects for young people, you know, so there are using there money.
2	I think (that)	O/C	O/C	E/P	E/P		나는	생각하다/생각 하다고		나는 생각하/생각 하다	match	P	I think I've still got the piece about that.
3	a bit		O/_		-/_			조금		조금	different	U	Can you move round this side a bit?
4	(always (153), never (87)) used to (UNP)	O/C		E/P			~하곤 했다			~하곤 했다	match	P	We used to look forward to them coming.

5	as well	?	?		?	?		?	?		또한	different	U	Can I just say something else as well?
6	a lot of (N)		O/C	O/_		E/P	E/P		제미/추억	~의/~에서	많은	different	U	I got a lot of letters from the children there and which was very gratifying.
7	(No.) pounds		O/C			E/_			파운드		파운드	match	P	I'd also ask you to consider costs of ten pounds.
8	thank you	O/C	O/C		E/P	E/P		감사하다	너/당신		감사합니다	grammatical	*	Well thank you for that that's a very good start to the evening.
9	(No.) years		O/C			E/P			해/년		해/년	grammatical	*	I've done it for seven years.
10	in fact	O/_	O/C		E/P	E/P		~에/속에	사실		사실	grammatical	*	In fact, if the previous speaker has complained about waiting in patience, I have waited forty years to tell this story in the assembly…
11	very much	O/C	O/C		E/P	E/P		매우	많이/매우		매우 많이/매우	match	P	…they enjoy it very much…
12	(No.) pound		O/C			E/_			파운드		파운드	match	P	I've got to take a taxi of one pound forty a day to

														shop...
13	talking about (sth)	O/C	O/C		E/P	E/P		말하는	~관하여		~관하여 말하는	match	P	It was a different from what you're talking about.
14	(about [91]) (No.) percent (of sth [580], in sth [54], on sth [44], for sth [38])		O/C			E/P			퍼센트		퍼센트	match	P	As I said, we've already got forty one percent of the main funding required for the project...
15	I suppose (that)	O/C	O/C		E/P	E/P		나는	가정하다/생각하다		나는 ~라고 가정한다/생각한다	match	P	I suppose that was one way of nothing being done.
16	at the moment	O/C		O/C	E/P		E/P	~에(place)		순간	바로 지금(에(time))	different	U	Well I haven't done anything at the moment because I didn't think it was worth it actually.
17	a little bit		O/C	O/_		_/P	_/_		조금	조각	조금	grammatical	*	And the other times your concentration will drop a little bit.
18	looking at (sth)	O/C	O/C		E/P	E/P		보다	~에		~을 보다	grammatical	*	I think there's another way of looking at that.
19	this morning	O/C	O/C		E/P	E/P		이/이것	아침		오늘 아침	1 matches	U	Oh she was screaming this morning.

20	(not) any more	O/C	O/C		E/P	E/P		조금이라도/더	더 많은		더이상은	grammatical	*	…women shouldn't have off days any more.
21	come on	O/C	O/C		E/P	E/P		오다	~에/위에		자이/언마/제발	different	U	No, it's not a verb, come on, what is it?
22	number (No.)	O/C			E/P			수/숫자			~번	different	U	Number six which is the thing that we have to look at.
23	come in (swe. sth)	O/C	O/C		E/P	E/P		오다	~안에/안으로		안으로 오다/들어오다	match	P	We're about to finish, so please come in.
24	come back	O/C	O/C		E/P	E/P		오다	되로		돌아오다/되돌아오다	1 matches	U	…you can come back in a minute and see if it's what you think.
25	have a look	_ /C		O/C	_ /_		E/P	하다		봄/일견	보다	different	U	You can go and have a look.
26	in terms of (sth)	O/C	O/C	O/C	E/P	E/P	E/P	~안에/~에	조건	~의/~에서	~의 관점/측면에서	2 match	U	I think it was one of the things which never really took off in terms of the accident.
27	last year	_ /C	O/C		_ /_	E/P		마지막/최후의	해/년		지난해/작년	1 matches	U	That was last year or was that two years ago?
28	so much	O/C	O/C		E/P	E/P		매우/아주	많이/많은		매우 많이/매우 많은	match	P	He loved the sea so much.
29	(No.) years		O/C	O/C		E/P	E/P		해/년	전에	~해/년 전에	grammatic	*	That was last year or was

	ago											al		that two years ago?
30	(Det~the [879], this [39], a [21]) county council	O/C	O/C		E/P	E/P		주	의회		주의회	match	P	And we put in a report to the county council in the morning saying there was a fire.
31	this year	O/C	O/C		E/P	E/P		이번/올	해/년		이번해/올해	match	P	And these conditions apply from the first of April of this year.
32	go back	O/C	O/C		E/P	E/P		가다	뒤로		되돌아가다	1 matches	U	I must go back to the sea, she said.
33	last night	O/C	O/C		E/P	E/P		마지막/최후의	밤		지난밤	1 matches	U	That's what I was doing last night.
34	rather than	O/C	O/C		E/P	E/P		오히려/차라리	보다는		~보다는 차라리	match	P	But I might do it by talking to them rather than by writing to them.
35	come out	O/C	O/C		E/P	E/P		오다	밖으로		나오다/밖으로 나오다	match	P	I mean Henry won't come out
36	very good	O/C	O_C		E/P	E/P		매우/아주	좋은		매우 좋은	match	P	The weather was very good, wasn't it?
37	I hope (that [455]) (N, S V)	O/C	O/C		E/P	E/P		나는/저는	바라다/희망하다		나는 바란다/저는 바랍니다	match	P	Oh well I hope. I hope it goes right tomorrow.

38	(No.) times		O/C			E/P			시대/시간		번/차례	different	U	Three times a week perhaps?
39	that way	O/C	_ /C		E/P	_ /_		그/그것	길		그 방식/방법	1 matches	U	…and I hope that I can train him to keep it that way
40	said well (that, what) (S V)	O/C	O/C		E/P	E/_		말했다	글쎄/어쩌 참		글쎄,~라고 말했다	match	P	…they said well it might take us two days.
41	at the end (of sth [737])	O/C		O/C	E/P		E/P	~에		끝/마지막	끝내/마지막에/ 결국에	match	P	Okay. I'll give you time to fill that in at the end.
42	(Det-that [425], this [146], the [142]) sort of thing	O/C	O/C	O/C	E/P	E/P	E/P	종류/류	~의/~에서	종건/~것	~류의 것	match	P	That's the sort of thing that the government encouraged.
43	for example (if S V [30])	O/C	O/C		E/P	E/P		~위해	예		예를 들어/예로	1 matches	U	For example, many people have four weeks' holiday a year, but not many people can take that four weeks all in one go…
44	as far as	O/C	O/C	O/C	E/P	E/P	E/P	~와 같이	길/선	~와 같이	~에 한해서	different	U	But of course as far as Britain was concerned, this could only be intervention

														against France.
45	said to (smo)	O/C	O/C		E/P	E/_		말했다	~쪽으로/~로(direction)		~에게(dative) 말했다	grammatical	*	He said to me on the phone he didn't know a great deal…
46	mean (that) (S V)	O/C			E/P			의미하다/뜻하다			의미하다/뜻하다	match	P	I mean that that's another alternative is that we don't bother with pictures.
47	come on (to swe, smo (65))	O/C	O/C		E/P	E/P		오다	~에/쪽로		~에/쪽로 오다	match	P	Gary Mills who's missed half a dozen games through injury is going to come on now and he's going to take the place of Neil Lewis.
48	(FREQUENCY, QUANTITY) a week		O/C			E/P			주		주	match	P	I'm now sixty two and I go dancing twice a week!
49	all the time	O/C		O/C	E/P		E/P	모든		시간	모든 시간/항상/내내	match	P	…what's up all the time?
50	thank you very much	O/C	O/C	O/C (*)	E/P	E/P	E/P(*)	감사하다	너/당신	매우 많이(*)	매우 (많이) 감사합니다	grammatical	*	Thank you very much indeed.
51	too much	_ /C	O/C		_ /_	E/P		또한	많은/많이		너무 많은/많이	1 matches	U	Too long, waste too much time.

52	over there	_ /C	O/C		_ /_	E/P		~위에	저기/저기		저기/저쪽으로	1 matches	U	He's over there, he's running up through there…
53	that sort (of sth [963])	O/C	O/C		E/P	E/P		그/그것/그런	종류/류		그런 류	match	P	…and I were used to do that sort of work…
54	looking for (sth)	O/C	O/C		E/P	E/P		보다	~위해		~을 찾다	different	U	What are you looking for?
55	make sure (that [394]) (S V)	O/C	O/C		_ /_	E/P		만들다(create)	확실히		확실히 하다/만들다 (make sth to be (become) in a particular state)	*	P	It is your responsibility to make sure that money is paid each and every week.
56	very well	O/C	O/C		E/P	E/P		매우	잘		아주 잘	match	P	You can read very well can't you?
57	(Det~the [47]) last week	O/C	O/C		E/P	E/P		마지막/최후의	주		지난주	1 matches	U	The other thing that we mentioned last week as well is the moral message…
58	in the morning	O/_		O/C	_ /_		E/P	~에(place)		아침	아침에(time)	grammatical	*	And I'd started at six in the morning.
59	it seems (N, A, to INF, that S V)	O/C	O/C		R/P	E/P		그것	~으로 보이다		그것은 ~으로 보인다	match	P	It seems quite clear, I thought.

60	next week	O/C	O/C		E/P	E/P		다음	주		다음주	match	P	I want you to bring it in its completed state to next week's lesson please?
61	a number of (sth)		O/C	O/C		E/P	E/P		수/숫자	~의	많은	different	U	Yeah, there are a number of courses which you do.
62	out there	O/C	O/C		E/P	E/P		밖에	저기/저기		저기 밖에	match	P	Is there an audience for women's football out there?
63	what I mean	_ /C	O/C	O/C	E/P	E/P	E/P	~하는 것	내가	의미하다	내가 의미하는 것은	match	P	Do you know what I mean?
64	get in (swe. sth)	_ /C	O/C		_ /_	E/P		~이 되다	~안에/~에		~에 이르다/안으로 들어가다	1 matches	U	So you might want to remember that when you get in there…
65	find out (sth)	O/C	O/C		E/P	E/P		발견하다	밖에/밖으로		발견하다/깨닫다	grammatical	*	Do you think she's got something to find out later on?
66	know that (S V)	O/C			E/P			~을 알다			~을 알다	match	P	…you know that this is the only room available.
67	leave it	O/C	O/C		_ /P	E/P		떠나다	그것		그것을 두고가다	1 matches	U	Let's just leave it for the moment.
68	at home	O/C	O/C		E/P	_ /P		~에	(자기)집		(자기)집에	match	P	If you've got plenty of products at home, fine, use them, okay?

69	and so on	O/C	_ /C	O/C	E/P	E/P	E/P	그리고/및	그와같이	~와에	등등 (그리고 그와같은 여러 예)	2 match	U	…you know, brief films about different countries and, political systems and so on.
70	(about {226}) (No.) minutes		O/C			E/P			분		분	grammatical	*	This device should give us a single analysis in about five minutes rather than ninety as at present as we don't need to separate the mixtures, we can do the analysis directly.
71	(do) n't mind (sth)	O/C	O/C		E/P	_ /_		아니다/않다	주의를 기울이다/조심하다		괘정하지 않다/신경쓰지 않다	1 matches	U	I'll have one a little later if you don't mind.
72	other people	O/C	O/C		E/P	E/P		다른/그밖의	사람들		다른 사람들	match	P	…you won't encourage other people to send them mail…
73	not really	O/C	O/C		E/P	E/P		않다/아니다	정말로/실제로		실다/쭉 그런건 아니다 (not necessarily)	1 matches	U	Not really, can you?
74	talking to {smo}	O/C	O/C		E/P	E/P		말하는	~쪽으로/~로(direction)		~에게(dative) 말하는	grammatical	*	And you're talking to your mate and it's all just happening by magic.

75	mind you	O/C	O/C		_/_	E/P		신경쓰다/걱정하다	너/당신		주의를 기울이다/상등이/안았니	different	U	Mind you, you'd have to be quick, wouldn't you?
76	want it	O/C	O/C		E/P	E/P		원한다	그것		그것을 원하다	match	P	Mind, we wouldn't want it every time.
77	much more	O/C	O/C		E/P	E/P		매우	더/더 많이		매우 더	match	P	That's a much more accurate way of finding the gradient.
78	looked at (sth)	O/C	O/C		E/P	E/P		보았다	~에		~을 보았다	grammatic al	*	Then I looked at it again and I realized that it wasn't too hard.
79	the other one		O/C	O/C	E/P	E/P			다른/그밖의	하나/한개/한 사람	다른/그밖의 하나	match	P	What's the other one?
80	(at [207], about [110], till [50], by [24]) half past (No.1~12)	O/C	O/C		E/P	E/P		반	지나서		반 시간이 지나서	1 matches	U	...we had to start at half past eight and finish at six…
81	some people	O/C	O/C		E/P	E/P		일부(의)	사람들		어떤 사람들	1 matches	U	Some people use it as a perfume.
82	this week	O/C	O/C		E/P	E/P		이번	주		이번주	match	P	Something that occurred to me this week.

83	this time	O/C	O/C		E/P	E/P		이/요/이것	시간		이번/요번/이 기회	1 matches	U	I shall write it down this time.
84	very nice	O/C	O/C		E/P	E/P		매우/아주	좋은		매우 좋은	match	P	And even in just the police station, in the small holding cells they have, it's not very nice.
85	I see	O/C	O/C		E/P	E/P		나는	보다		알았다/그렇군	different	U	Oh, I see, oh well in that case we will view it…
86	I bet (S V)	O/C	O/C		E/P	E/P		나는	걸다		확실히	different	U	I bet you'll be proud of her!
87	these things	O/C	O/C		E/P	E/P		이러한	~것들		이런한 것들	match	P	I'm just putting these things in your mind to put some doubt in your mind.
88	call it (A N)	O/C	O/C		E/P	E/P		~라고 부르다	그것		그것을 ~라고 부르다	match	P	Could you call it beverage?
89	(be-verb) not sure	O/C	O/C		E/P	E/P		아니다/않다	확실한		확실하지 않은	match	P	I'm not sure, I'm not sure they have been there.
90	at the time	O/C		O/C	E/P		E/P	~에(place)		시간/때	그 시각/시간/때에 (time)	grammatical	*	…what's going on in England at the time?
91	thought that (S V)	O/C			E/P			생각했다			생각했다	match	P	So I thought that there was plenty of time as this is my first day off from work.

92	going out	O/C	O/C		가다	밖으로	밖으로 가는/나가는	match	P	I'm going out now.
93	it comes	O/C	O/C	E/P	그것	오다	그것은 ~이 된다	I matches	U	If you add all that up it comes to about three million words.
94	go out	O/C	O/C	E/P	가다	밖으로	밖으로 가다/나가다	match	P	...we'll go out and have a drink and have a meal like...
95	quite a lot	O/C		O/C	상당히/많이		제법/꽤	I matches	U	...people probably need quite a lot of support.
96	even if	O/C	O_	E/_	심지어/조차	만약~한다면		I matches	U	But even if no one else can help, I will take part in it.
97	last time	O/C	O/C	E/P	마지막/최후의	시간		different	U	Now what were we looking at last time?
98	hang on	O/C	O/C	E/P	매달다	~한에		계속하다/기다리다	U	Actually, hang on!
99	believe that (S V, N)	O/C		E/P	믿다/생각하다			match	P	Oh I can't believe that !
100	(be-verb, become-verb) interested in (sth)	O/C	O/C	E/P	흥미를 가진	~에		~에 흥미를 가진	P	I'm more interested in the, the kind of s so called philosophical ideas that have been raised...

- See the attached CD for more data.

Appendix 5

Predictability of English Collocations in Korean

The following 50 items are English collocations which are all included in the most frequent 500 collocations of English. The purpose of this survey is to examine if native or near-native speakers of English recognise the difficulty of English collocations for learners of English. You are not expected to know Korean. We want to find out if people can guess which collocations are predictable or not. If you think the collocation in column 2 could be translated word for word into Korean, enter the letter P for *Predictable* in column 4. For example, *next year* in Korean is translated by the Korean word for *next* and the Korean word for *year*. It is not necessary for the English and Korean expression to have exactly the same word order. The verb is located at the end of the sentence in Korean. For example, *go home* has the Korean equivalent *집에 가다* (home go), so it is predictable in Korean. On the other hand, if you think that Koreans might express this another way, enter the letter U for *Unpredictable* in the fourth column. For example, *strong coffee* is the same as *thick coffee* in Korean, so it is unpredictable in Korean. One more thing we should note is that only the bold italicised phrases in the second column are collocations to focus on. For example, the word *the* of "the ***county council***" only shows a grammatical feature and it could be replaced by other words such as *a, this*, or *that*, so you should focus only on "***county council***".

	Collocations	Sample sentences	P (-redictables) / U (-npredictables)
1	*I think*	***I think*** I've still got the piece about that.	
2	*a bit*	Can you move round this side ***a bit***?	
3	*come back*	…you can ***come back*** in a minute…	
4	the *county council*	We put in a report to the ***county council*** …	
5	*three times*	***Three times*** a week perhaps?	
6	*come on*	Gary Mills who's missed half a dozen games through injury is going to ***come on*** now…	
7	*looking for*	What are you ***looking for***?	
8	*this year*	These conditions apply from the first of	

		April of *this year*.	
9	*don't mind*	I'll have one a little later if you *don't mind*.	
10	*I see*	*I see*, oh well, in that case we will view it···	
11	*thought that*	I *thought that* there was plenty of time···	
12	*hang on*	Actually, *hang on*!	
13	*find it*	I *find it* extraordinary.	
14	*this afternoon*	···he will ring me *this afternoon*.	
15	*like it*	I quite *like it*.	
16	*many people*	···how *many people* have walked out···	
17	*my lord*	*My Lord* Mayor, I started by saying how angry and dismayed I was···	
18	*one day*	*One day* she said, I'll become the queen.	
19	the *only thing*	...the *only thing* to bear in mind is that we should be a little bit careful with the changing environment...	
20	*set up*	I want to *set up* the printer.	
21	*going up*	We're *going up* that ladder.	
22	*this way*	We shouldn't be allowed to present animals in *this way*.	
23	*thinking of*	I'm just *thinking of* myself really.	
24	*for people*	···some of the treatments that were available *for people*!	
25	*over here*	Why don't you wander *over here*?	
26	can*not afford*	You can*not afford* to make mistakes like that on aircraft!	
27	*next door*	That was *next door* to the tannery.	
28	*at the top*	...we lived *at the top* of the second hill···	
29	*any way*	Would that in *any way* affect what's happening now?	
30	*it depends* on	...*it depends* on your attitude towards food.	
31	*in particular*	We were thinking about coffee at the time *in particular*··	

32	*what's happening*	He's going to phone them up Monday and see *what's happening.*	
33	*next time*	*Next time* I go to see a doctor.	
34	*comes out*	I'll see Dougie when he *comes out.*	
35	*get up*	…often used to *get up* in the morning…	
36	*both* teachers *and* pupils	…in order to provide maximum stimulation for *both* teachers *and* pupils.	
37	*really nice*	It is *really nice*…	
38	at *quarter to four*	I got back at *quarter to four.*	
39	*quite happy*	She's *quite happy.*	
40	*hold on*	*Hold on* a second!	
41	*eat it*	Don't you want to *eat it*?	
42	*going round*	...I've not seen this wagon *going round* so frequently as it used to.	
43	at *this stage*	So we don't really need to consider that at *this stage.*	
44	*the whole lot* of	I had to look through *the whole lot* of them.	
45	*on page eight*	Those are actually listed *on page eight.*	
46	*said if* S V	He *said if* you don't take it you lose it…	
47	*on the other hand*	*On the other hand*, we realize how difficult it is to use the law.	
48	*up to you*	That's entirely *up to you*…	
49	*city council*	I can tell you that Birmingham *City Council* told me this…	
50	have *no idea*	I have *no idea.*	

1) What languages do you know?

2) Have you studied Korean at all?

3) Are you a native speaker of English?

- You can comment on why you chose P or U for some specific items if you want.