Towards a Definition of Collocation

Laura Cowan

A Thesis

in

The TESL Centre

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Arts at
Concordia University
Montreal, Quebec, Canada

April 1989

ISBN 0-315-49083-7

ABSTRACT

In an effort to arrive at a model on which to base an operational definition of collocation, various theoretical issues raised in the literature were critically examined. These issues were treated under the following six questions: (1) what is the basic unit of analysis?; (2) are closed-class grammatical words potential elements in collocation?; (3) how is collocational span defined?; (4) how is the statistical significance of collocation determined?; (5) how are collocations and idioms distinguished?; and (6) is collocational meaning determined by the meaning of the words involved or by their tendency to co-occur?

An evaluation of how certain practical decisions concerning these issues were applied in three empirical studies has shown that the empirical studies have not been effective in revealing all the structural and semantic aspects of collocation. Suggestions have been made about what needs to be done so that the notion of collocation might be more rigorously defined, and thereby allow `collocation' to become a useful concept in the teaching of English as a second language.

Table of Contents

TOWARDS A DEFINITION OF COLLOCATION

1.0 Introduction

When J.R. Firth introduced the term `collocation', he said that "you shall know a word by the company it keeps" (1957:196). 'Collocation' is usually described as the tendency of certain linguistic items to habitually co-occur with certain others. However, neither Firth nor most subsequent researchers attempted to elaborate the concept or to define it rigorously. Therefore, most treatments of collocation are limited by their imprecision.

Nevertheless, there has been recognition that the term collocation, if rigorously defined, could have useful applications in fields such as language pedagogy (M. Korosadowicz-Struzynska, 1980) and lexicography (A.P. Cowie, 1978). Halliday (1966) has also noted the relevance of collocation in the study of register and literary style, of children's language, the language of aphasics, and in the field of information retrieval.

The purpose of this thesis is to attempt to work towards a model for a definition of collocation which will allow it to become a useful concept in the teaching of English as a second language. Halliday has noted that in foreign language teaching "many errors are best explained collocationally" (1966:160). Cowie emphasizes that including collocational information in a learner's dictionary would help foster language skills by making "...quite explicit the possibility of lexical choice in a given context" (1978:130), and perhaps helping the student avoid such errors as *light possibility, or *driving a bicycle. The COBUILD English Language Dictionary (Sinclair, 1986) addresses the problems students of English as a second language may have in finding appropriate English collocations by including examples which "...have been selected to show typical contexts, collocations and grammatical structures" (p.ix). The recently published BBI Combinatory Dictionary of English: A Guide to Word Combinations (Benson, Benson and Ilson, 1986) also addresses this problem. The authors state that "...knowledge of other languages is normally of no help in finding English collocations" (p.vii) since different languages exhibit different collocational ranges. Compare English at the height of summer, in the depths of winter, [hit someone] right in the stomach, plumb in the middle with French en plein ete/hiver/ventre/milieu (Mitchell, 1975:10). Cowie has also noted that "...limited collocability arising from (cultural) factors is of course ... baffling for the foreign learner" (1978:134). The ESL student does indeed require guidance in identifying and using English collocations.

There are many different definitions of collocation in the literature (Croft, 1967; Ridout and Clarke, 1970, cited in Seaton, 1982; Jones and Sinclair, 1974; Crystal, 1985, to name a few) none of which seem to take into account the semantic and structural complexities of collocation. Native-speaker intuition seems to have been the basis, in most theoretical discussions, for identifying collocations. Collocation is not grammar, yet includes it; it is not explainable in terms of lexical sets, yet these are also part of collocation.

Faced with inadequate definitions on the one hand, and a collection of characteristics noted in the literature on the other, this thesis will attempt to formulate a model that will lead to a rigorous definition of the term 'collocation'. The first step (Chapter 1) in achieving this aim will be a critical review of the literature, examining six fundamental questions, each of which must be resolved if a rigorous definition is to be arrived at. The second step (Chapter 2) will be a description and critical analysis of three empirical studies of collocation in terms of these six questions.

The first question to be asked of a study of collocation is: what is the basic unit of analysis?

The second question is: are closed-class, function, or structure words potential constituents of a collocation or does the study only consider open-class words like nouns and verbs? The answer to this will reveal whether the researcher considers grammar and lexis to be either two separable or two interpenetrating aspects

(Sinclair, 1966:411) of language.

Third, how is collocational span defined? The span considered for a collocational study and its justification - if any - will be examined to see, among other things, if the judgment of the strength of the mutual predictability of elements in collocation is based solely on span distance.

Fourth, how is the statistical significance of a collocation determined in the study? That is, how is probability of occurrence calculated, and how is this related to the length of text examined in the study?

Fifth, does the study distinguish between collocations and idioms? If so, how?

Sixth, is collocational meaning determined by the meaning of the words involved or by their tendency to co-occur? That is, are there any semantic factors which might usefully account for the collocation?

The conclusion (Chapter 3) will present suggestions as to how these six questions might be resolved so that future empirical studies of collocation might conform more closely to a more rigorous definition of collocation.

2.0 Critical Review of the Literature

2.1 What is the basic unit of analaysis?

The first question to be asked of a study of collocation is: what is the basic unit of analysis? That is, does the study deal with collocation at the 'lexeme' level or at the 'word' level? Are such words as drive, driving, drove, and driven analyzed as different words, or does each of these words designate a single lexeme and represent only inflected forms of a single stem?

Those authors who have proposed that the lexeme should be considered the basic unit in the analysis of collocation have felt that this would make it possible to account for the observed similarities in collocational patterning of derived and inflected forms of words. With the lexeme as a basic unit of analysis, the words found in different grammatical constructions could be related back to a 'basic collocation', and this would simplify the study of collocation.

Different authors have used different terms for what is referred to here as the lexeme. Sinclair uses the term 'lemma' (1985:84) while Mitchell, as well as most lexicographers, speak of the 'root'. Carter (1987), however, distinguishes 'root' and 'lexeme'. The lexeme, according to Carter, is the "abstract unit" (1987:6) underlying "word-forms" (p.6) including items consisting of more than one word form (as in phrasal verbs and idioms), while the 'root' is more closely related to 'free' morphemes which contrast with affixes, or "bound non-roots" (p.10). An example of a 'root' would be 'hope-' in 'hopeless', while '-less' would be a "bound morphemic non-root" (p.11). Since the lexeme is an abstraction, its form is not specifiable, and is indicated in this thesis as a word preceded by a slash (/).

Halliday, McIntosh and Strevens (1964) also suggest that the lexeme may be composite, that is, made up of more than just a single word. Thus:

> In grammar we distinguish four items: (1) a word 'took', (2) a word 'taking', (3) a morpheme 'take' and (4) a word 'take'; but these are the same items whether followed by 'off' or by 'over'. In lexis on the other hand we distinguish two items: (1) 'take off' and (2) 'take over'; but 'take off', 'taking off', 'takeoff', and 'took off' are all the same item. (1964:35)

A number of authors (Firth, 1957; Mitchell, 1971; Palmer, 1976; among others) have proposed that classifying inflected forms of words under a single `basic collocation' is justified because of their similarity of collocational patterning. Cowie writes that "a given collocation may of course recur in a number of distinct grammatical constructions, which can then be regarded as transformationally related to each other" (1978:132).

Using the same approach, Mitchell sees collocation as a recognizable regular association of 'roots'. He assumes that by adopting this view the analysis of collocation will be facilitated. He maintains that on the basis of such regular associations, one can write rules to generate particular collocations. For example, given the roots /heav- and /drink one can apply the rules of syntax to produce *he drinks heavily*, *he is a heavy drinker*, *he is putting in some heavy drinking*, and he *is drinking pretty heavily*.

Palmer feels that "it is with collocations between scatters, not words, that we should be concerned" (1976:101). He suggests that if what he calls collocational restrictions, the ways in which the derived and inflected forms of lexemes do or do not combine with others, are to be handled under lexis, as opposed to grammar, the basic unit of analysis should be the lexeme. It can be assumed that Palmer, like many researchers, felt that the study of collocations would be simplified if the importance of the grammatical differences between the different collocational realizations of two lexemes could be reduced. This would facilitate the indication, as in a learner's dictionary, of the restrictions applicable to the collocation. Several relatively recent learner's dictionaries have made efforts to indicate the usual syntactic patterns in which headwords are found (*Longman*

Dictionary of Contemporary English, Proctor, 1978; The Oxford Dictionary of Current Idiomatic English, Cowie, Mackin, and McCaig, 1983; and the COBUILD English Language Dictionary, Sinclair, 1986,to name just a few).

Sinclair, however, points out that deciding which form of a lexeme to use when ascribing words to various abstract lexemes ('lemmatization', as he puts it) is problematic (Sinclair, 1985:84). In order to speak of any association of lexemes, one must be able to identify them. Most researchers (primarily lexicographers) have used the uninflected form (ie. come, rather than coming or came), but an argument could be made that the most frequently occurring form should be used, especially when computer-readable corpora are involved (Sinclair, 1985:84).

Halliday and Hasan (1976:291) anticipated some difficulty in assigning words to particular lexemes. For example, if boy, boy's, boys and boys' all represent four forms of one lexeme, and go, goes, going, and went are all forms of one lexeme, as are good, better and best, do we consider tooth and dental, or oral and verbal, two forms of one lexeme or two distinct lexemes? The question here, Carter (1987:11) would argue, is whether grammatical paradigms, semantic criteria, and even the assignment of polysemous meanings are to be considered in deciding the extent of the paradigm ascribed to the lexeme.

The espousal of the lexeme as a basic unit of research, however, is hampered by what is traditionally termed collocational restriction or selectional restriction. Nagy uses the term 'transformational deficiency', by which he means "that not all possible syntactic permutations of the expression are equally acceptable" (1978:291). For example, while *reckless abandon* and *faint praise* are considered acceptable in English, *?the abandon was reckless* and *?he praised her faintly* are much less so. There are grammatical constraints on the combinatorial possibilities of the members of the paradigms. So, while one could certainly agree with Cowie and say that they are transformationally related to each other, it might still not be possible to maintain categorically that all co-occurrences of the lexemes /reckless and /abandon or /faint and /praise constitute a collocation.

So it would seem that however appealing it may be to attempt to establish the lexeme as the basic unit for analysis of collocations, this approach fails. Although all derived and inflected forms can be related back to an abstract lexeme, not all of those forms occur in all grammatical structures. It is thus not possible to say that a collocation is merely an association of lexemes. It may be that an approach that includes lexeme combination with syntactic restrictions will work better.

2.2 Are closed-class grammatical words potential elements in collocation?

In order to discuss this question, it is important to clarify what is meant by the term 'closed-class grammatical word'.

Jones and Sinclair have mentioned that the distinction between between 'grammatical' and 'lexical' words is not always clear-cut. This is so because semantic reference is often an important factor in the selection of one 'grammatical' word or another (compare 'a house in Toronto' to 'a house near Toronto'). The distinction, according to Halliday (1966:155), might depend on whether the distinction is important in terms of explaining the restrictions on their occurrence. For example, a grammatical explanation of the prepositions in the following two sentences would probably be less interesting than a lexical one:

He did it on purpose.
She met him by chance.

In contrast, a grammatical analysis of *he praised her faintly would help to explain why it is less acceptable than he was damned by faint praise.

Jones and Sinclair further suggest that an "intuitively satisfying" (1974:24) differentiation lies in Halliday's suggestion that the distinction may relate to word frequency. Thus, lexical words occur in a text less frequently than gramatical words. Nevertheless, "there is no reason to assume a correlation of 'most grammatical' with either 'least lexical' or 'most frequent'" (Halliday, 1966:155).

Some authors view collocation as a type of relationship that exists between lexical words only (Haskel, 1971), while others stress that grammatical words must also be taken into account (Sinclair, 1966). The former view is typical of researchers interested in collocation as a tool in stylistic analysis while the latter is typical of those whose focus is semantic analysis.

There are two distinct views about whether or not grammar plays any role in collocation. The majority of authors accept that grammatical analyses and lexical analyses both examine the same object, and that these two types of analyses are simply two ways of looking at language. Firth's views about the importance of context in the study of language calls for the recognition of the equal relevance of grammar and lexis, since any 'context' must include both these aspects of language form. Sinclair agrees, noting that lexis and grammar are "two interpenetrating ways of looking at language form" (1966:411), and that neither can be separated from the other. Muller states that a description of collocations "results in a certain redundancy, for collocations will...be doubly identified as grammatical constructions and as habitual constructions of the language" (1981:175).

Quirk points out that grammar alone cannot be used to identify collocations, since "...when grammar is a constant, ready comprehensibility may still vary sharply, according to the expectedness or unexpectedness in the selection or collocation of words" (Quirk, 1962:235). This can be illustrated with the following two sentences, which, though similar in grammatical structure, vary in 'comprehensibility' because of the choice of lexical items involved.

(1) The table was of polished mahogany and it gleamed in the bright light.

(2) The car was of corrugated plastic and it swayed in the ploughed sand. (Quirk, 1962:235)

The co-occurrences of <u>car</u>-<u>corrugated</u>-<u>plastic</u> and <u>swayed</u>-<u>ploughed</u>-<u>sand</u> is much less expected than those of <u>table</u>-<u>polished</u>-<u>mahogany</u> and <u>gleamed</u>-<u>bright</u>-<u>light</u>, and it is this lack of "expectedness" which interferes with the comprehensibility of the sentence.

Sinclair, Muller, Quirk and Gleason all agree that Firth's 'context' includes not only the surrounding lexical words, but also takes into account the relationships, grammatical as well as lexical, operating within the text. Kjellmer, also, recognizes the importance of both grammar and lexis in the study of collocation. He describes collocation as

> both lexically determined _and_ grammatically restricted sequences of words...'Lexical determination' here refers to the fact that only recurring sequences are accepted as potential collocations... The term 'grammatical restriction', on the other hand, is used to imply that only grammatically well-formed sequences are accepted as collocations. (1984:163)

Kjellmer's 'lexical determination' eliminates from what can be considered collocations those random combinations of words that occur simply by virtue of being language in use. 'Grammatical restriction' allows, for example, "the soldier fought with reckless abandon" to be considered a collocation between _reckless_ and _abandon_ (if it recurs) while disallowing "reckless soldiers abandon their posts", since in the latter _reckless_ and _abandon_ do not form a part of the same grammatical unit.

Consider another example:

(1) The gambler cursed his bad luck.
(2) The bad gambler cursed his luck.

In (1) _bad_ and _luck_ form part of a single grammatical unit, while in (2) they do not. 'Lexical determination' would eliminate _bad gambler_ and _his luck_ as collocations based on frequency of occurrence.

Kjellmer has refined his description of collocation to "a sequence of [from two to five] words that occurs more than once in identical form (in the Brown Corpus) and which is grammatically well-structured" (1987:137). This, unfortunately, requires Kjellmer to call such free constructions as _to be_, _one of_, _had been_, _would be_, etc.

(Kjellmer, 1987:133) collocations since they fulfill the two criteria proposed.

If we examine it closely, we find that the 'lexical determination' criterion cannot distinguish collocations from free constructions since grammatical words such as the examples above appear in texts more often than lexical words and so will, of course, be found in recurring combinations more frequently than lexical words. Most authors, in fact, have stipulated that a basic characteristic of collocations is that they are restricted in their distribution. Since Kjellmer's examples of collocations do not show even a slight restriction of distribution beyond those determined by the rules of syntax (one of/for/in/out/any noun; had been/any past participle, gerund, adjective or noun; would be/any gerund, adjective, noun) most authors would hesitate to call them collocations. Kjellmer's definition is valuable in that it accounts for some collocations, but ultimately fails because it accounts for too much: it does not distinguish between collocations and free combinations.

On the other side of this issue are those authors who do not consider the grammatical feature of collocations to be specifiable. Halliday (1966) notes that the rules governing grammatical patterning do not operate in the same ways as the rules for lexical patterning and therefore cannot be related to them. "Word class distinctions...are purely grammatical and irrelevant to lexis" (Halliday, 1964:36).

Both Halliday (1966) and Turner (1973) point out that there need not be any formal or grammatical considerations at all in the analysis of collocation, since the elements involved may be in different sentences or discontinuous within one sentence, as in "I wasn't altogether convinced by his argument. He had some strong points but they could all be met" (Halliday, 1966:150) which is intended to illustrate the collocation of *strong* and *argument*. Kjellmer would argue that this example illustrates the collocation between *strong* and *points* rather than between *strong* and *argue*. Halliday seems to have confused the concepts of collocation and lexical set. He had earlier defined a lexical set as "simply a grouping of items which have a similar range of collocation" (Halliday et al, 1964:33). Halliday does note that the discontinuity of lexical items in collocation has its limits, but believes that it is not possible to usefully define these limits grammatically (1966:151). Halliday does not, however, offer any alternative to a grammatically defined limit.

2.3 How is collocational span defined?

Collocational span refers to the distance, measured in individual words, between words collocating with each other. The word whose collocational patterning is being examined is called a 'node', while those words which enter into collocation with the node are called 'collocates'. In "he drew up his last will and testament", the collocate will occurs at a collocational span position of -2 from the node testament (where -2 indicates a position two words to the left of the node, and +2, two words to the right). For most authors, the span distance considered in a collocational study is usually not more, and often less, than five or six words on either side of the node.

The length of the span chosen in a study is a reflection of the author's opinion about how close together in the text words must be in order for the combination to remain a collocation. This issue, however, is problematic. Both Turner and Halliday have noted that collocating words need not be in the same sentences, though, as Turner says, "it is not quite clear how far apart words may be and still be said to collocate" (1973:129). Martin, Al, and van Sterkenburg (1983) have argued that, in theory, the further apart words are, the weaker their influence on each other (p.84) and so a certain span should be established beyond which collocations will not be studied. This position would assert that in "he was granted absolution", the collocation of granted with absolution (with a span of -1) is stronger than will with testament (with a span of -2) in the example above.

Sinclair, however, rejects the notion that the proximity of the words involved in collocation is directly related to the strength, or degree of mutual predictability, of the collocation, since some common collocations are habitually discontinuous without becoming any less mutually predictable (Sinclair, 1966:414).

Halliday and Hasan (1976) note the cohesive effects of collocations within a text of not only pairs of words, but also long "chains" of lexical relations "with word patterns... weaving in and out of successive sentences" (p.286). This relates to the question of how one determines span distances between lexical items which consist of more than one word and whose constituents may be discontinuous. If we consider, as Palmer (1974, 1976) does, look up to be a single lexical unit, calculating span distances in order to measure collocational strength becomes untenable. Given separable phrasal verbs such as look up, how can span distances be calculated in the following sentences: "She was obliged to look up the unknown word in the dictionary" as compared to "She was obliged to look the unknown word up in the dictionary"?

The problem this question raises for the study of collocation is substantial: where does one draw the line on limiting the length and placement of text in which a collocation can be said to occur? Martin et al adopt a statistical approach:

> Defining the optimal span for a collocational study is always a matter of dispute...Statistical tests lead us to the conclusion that more than 95% of all relevant information can be obtained by examining collocations within a span of +5 and -5 (disregarding punctuation). (Martin et al, 1983:84)

This seems to indicate that on a practical level the analyst must make a relatively arbitrary decision, based on statistical tests, about the number of consecutive words that will be examined in a study of collocation.

But Mitchell has also noted that "a collocation is not a mere juxtaposition or a co-occurrence" (1983:53) so it does not seem possible to say that it is meaningful to approach collocational span by simply counting words on either side of a node. However, if one adopts Kjellmer's idea that collocation occurs only within the same grammatical unit the question of span distance becomes much less complicated. The discontinuity of the lexical items, as in "he was <u>tried</u>, without the slightest doubt, <u>in</u> <u>absentia</u>", does not affect the recognition of the collocation between <u>tried</u> and <u>in</u> <u>absentia</u>, or its strength, since Kjellmer's criterion will mean that the interposed phrase is ignored in the analysis of the collocation.

The drawback to this approach is that it will not always account for collocations which occur across different sentences. Consider the following:

> He was tried in Ruritania. It all happened in absentia, of course, since he was here at the time.

It would require a very complete grammar, one that is text-based rather than sentence-based, to link _tried_ and _it_ _all_ to _in_ _absentia_, which would be necessary for this approach to be more usable.

Thus the question of how much text to take into account in the study of collocation has been approached from two different directions. One approach has been to simply count a certain number of words on either side of the node in order to examine collocations occurring within that limited span of text. Some authors have disagreed with this approach since collocatinq words may not be in a single sentence or they might be discontinuous. Another approach has been to limit collocating words to those which occur within the same grammatical unit, disregarding any intervening text, though this too fails to account for words which collocate but are found beyond the bounds of sentence level grammar. It would seem that this latter view, though flawed, would work better than the former in accounting for collocations in a text.

2.4 How is the statistical significance of collocation determined?

Collocation, it has been agreed, is a matter of tendencies, of degrees rather than absolutes. Muller (1981:175) states that "collocations are not functional units comparable to phonemes or morphemes" and compares the way in which words combining together may be judged more or less mutually predictable just as they may be judged more or less idiomatic or grammatical. This, he notes, excludes the possibility of "a simple dichotomous classification [between collocations and free-constructions] ...thus in principle precluding an exhaustive description" of collocation.

Because it is not possible to draw a definite line between collocations and free constructions, most researchers measure degrees of collocability by assigning a measure of statistical significance to different collocational patterns to denote their mutual predictive strength, that is, the probability that one particular word will occur within a specified environment of another word. "A 'significant' collocation is one in which the two items co-occur more often than could be predicted on the basis of their respective frequencies in the length of text under consideration" (Martin et al, 1983:84).

Halliday (1966) notes, however, that frequency of occurrence and collocability are not necessarily the same thing. Grammatical words, for example, occur in texts with much greater frequency than lexical words, but are not by that fact prone to greater collocability. Kjellmer (1984), Halliday and Hasan (1976), and Lehrer (1974) all agree that what is important is the ratio between overall frequency of occurrence and expected frequency of occurrence.

> The more frequent a sequence is in relation to its *expected* frequency of occurrence, the more distinctive it is likely to be, other things being equal. If *to be* occurs a certain number of times in the corpus, this is less indicative of the distinctiveness of the collocation than if, say, *guerrilla warfare* should occur the same number of times, because the individual words *to* and *be* are much more frequent in the corpus than are *guerrilla* and *warfare*. (Kjellmer, 1984:166)

Most authors consider that collocations are properly analyzed in statistical terms, and the relationships between the items of a collocation are, in fact, usually expressed in terms of statistical probability (Berry, 1977:54). Halliday, McIntosh and Strevens (1964:33) state that "lexical sets...are bounded only by probabilities."

Martin et al (1983) see the results of statistical studies of collocation as a useful tool for further study of collocations. The use of various statistical tests, they point out, is useful in "establishing the most probable collocations. Evidently this will be of particular use to those interested in habitual and idiomatic collocations" (1983:87). In lexicography, this would be useful in selecting and limiting the collocational information included in dictionary entries to only those collocations which are most typical of the language.

The statistical tests used in calculating collocational significance are described by Martin et al:

> ...the probability that a particular collocate will follow or precede a particular node...can be calculated by dividing the number of collocations between the node and its collocate by the total number of node occurrences. The result will always be less than or equal to 1; the closer it is to 1, the greater is the probability of the collocation in question. (1983:86)

Statistical analysis of collocations, however, still leaves unidentifiable those collocations on the extreme ends of probability. Cowie (1981) raises the point that a statistical analysis of collocations may lead to an inability to distinguish completely restricted collocations (those having a statistical probability of occurrence of 100%) from idioms without bringing in additional distinguishing criteria. Conversely, the less restricted a collocation is, the more difficult it is to identify statistically, since the collocational behaviour of weak collocations may closely resemble that of free combinations.

Nor can statistical analyses tell the researcher whether it is the co-occurrence of two (or more) items which is unlikely, or merely their occurrence in one or more particular structures (Halliday, 1966:159). Muller has also pointed out that the bilateral dependencies between collocating words vary in strength. That is to say that the predictive power of one or the other item(s) will probably be greater in one direction than in the other. Statistical measurements, if not conducted carefully, might fail to show as significant a collocation in which the first lexical item is only weakly associated with the second while the second is much more strongly associated with the first.

Sinclair (1966) notes theoretical and methodological problems involved in the statistical study of texts to determine collocational sets. Pre-editing is necessary to rule out purposefully deviant sentences (such as poetic constructions, intentionally ungrammatical or unlexical sentences such as colourless green ideas sleep furiously) so that only collocations typical of the language remain. Also, texts representative of all kinds of discourse would have to be considered to rule out bias based on subject matter. Most of the computer-readable corpora available today do address this latter point by including a wide variety of 'genres' of texts.

It becomes obvious that although it is agreed that there are degrees of collocational strength which are quite properly measured in statistical terms, the use of statistics cannot reveal all the relevant information necessary for assessing whether or not a group of words is to be considered a collocation.

2.5 How are collocations and idioms distinguished?

All generally accepted definitions of idioms mention a lack of semantic transparency and at least imply that idioms consist of more than one orthographic word. Fraser's definition (as cited in Nagy, 1978:296) is an example:

> ...no expression can be considered a true idiom if it contains a lexical item whose literal meaning does contribute to the meaning of the overall expression.

Also generally accepted is that idioms are more or less syntactically frozen units whose elements resist lexical substitution. Mitchell (1971), and many other researchers, use these last two traits as criteria to distinguish between idioms and collocations. However, it is the observation that collocations, too, share three of these four characteristics (consisting of more than one orthographic word, a degree of syntactic frozenness, and resistance to lexical substitution) that has resulted in some confusion.

The muddle of existing terminology illustrates this point. Ridout and Clarke (as cited in Seaton, 1982:25) define collocation as "a group of words frequently found together and producing collectively a meaning not apparent from the meaning of each component part of the group." This definition seems to equate collocations with idioms. Collocations have been defined as "idioms of encoding" by

Makkai (1972:57) and "partial idioms" by Palmer (1976:98), and, conversely, idioms have been defined as collocations by Palmer (1938:iv, x-xi) and as frozen collocations by Cowan (cited in Makkai, 1972:26). Bolinger, on the other hand, seems to indicate that differentiating the two is not crucial:

> It is, of course, a matter of terminology whether collocations should be classed separately from idioms or as a major sub-class. (Bolinger, 1976:5)

Lehrer, however, takes a closer look at two of the three characteristics shared by idioms and collocations. In terms of the possibility of syntactic transformation, she notes that idioms may not be as frozen as current definitions might lead one to think:

> Idioms are not simply frozen phrases, however. Fraser (1970a) has shown that idioms differ with respect to the syntactic transformations they can undergo, and he establishes a hierarchy from the most frozen, which cannot undergo any transformation, to those that can undergo a wide variety. (Lehrer, 1974:184)

Muller (1981:185-6) contends that the greater syntactic freedom of collocations is what distinguishes them from idioms. Given Fraser's hierarchy (which is flawed, according to McCawley; see Makkai, 1972:54), however, it is difficult to see how syntactic frozenness alone can be used as a criterion to distinguish collocations from idioms.

Mitchell's view of collocation seems to be that it is the lack of 'fixity of association of roots' (that is, greater lexical substitutability) in collocations that distinguishes them from idioms. But Lehrer (1974:185) has found that not only do idioms differ in their potential for syntactic transformation, but also in their potential for lexical substitution. She offers the following two examples of idioms where substitution is possible: *keep/hold up one's end* and *hit/strike the high point*. One could also add *shoot the bull/breeze* as well as *give hell/shit to*. Since it is well noted that collocations too allow lexical substitution, and somewhat more freely than idioms, we can conclude that the possibility of lexical substitution alone cannot be used as a criterion to distinguish between idioms and collocations.

It seems that only on the point of semantic transparency do authors agree (Muller, 1981; Bolinger, 1976; Nagy, 1978; Mel'cuk, cited in Weinreich, 1980:228): idioms are semantically opaque and collocations are groupings of words peculiar to a language whose meanings might also, but not necessarily, be less than transparent. For example, the expression *red hair* does refer to hair colour, but not to red in strict colour terms (Palmer, 1976:98). Other examples of this might include *white man*, *white wine*, and *white coffee*, which are in reality respectively pinkish, yellow, and brown, and *blind alley* and *blind trust*, which are not blind in that word's most common sense.

Bolinger (1976:5) agrees, arguing that, in contrast with idioms, the meaning of a collocation can be predicted from the meanings of the individual elements involved, yet is nevertheless particularized. Cowie (1981) explains that a figurative meaning of one element in a collocation may be an important determinant of the limited collocability of the other(s).

Nagy (1978:296) objects to classifying as idioms such expressions as `bear witness to', `give chase to', `pluck up courage', etc., since they are semi-productive in that their literal meaning does contribute to the meaning of the overall expression. He argues that these cannot be considered `true' idioms because of their semantic transparency. It might be proposed, then, that if an expression contains at least one element which is semantically transparent, as in the above examples, it can be classified as a collocation rather than an idiom.

This classification, however, cannot be totally acceptable since, according to this criterion, raining cats and dogs (usually classed as an idiom) would necessarily be classed as a collocation because one element (raining) is given its literal meaning. However, considering that it is highly unlikely that anyone would think that raindrops resemble cats and dogs, it would seem that a highly metaphorical interpretation of cats and dogs must be made, and it is precisely this degree of abstraction that would have to be used to identify the construction as an idiom. Consider, on the other hand, that in the collocation <u>white coffee</u>, at least one of the elements (milk or cream) could be white, and thus there would be little or no abstraction.

It would seem that a subjective judgement about the levels of transparency and degree of metaphorical abstraction of an expression would thus be needed.

Thus, distinguishing collocations from idioms would seem to involve the consideration of four criteria: being composed of more than one orthographic word, a degree of syntactic frozenness, resistance to lexical substitution, and some degree of semantic opacity. While the first three of these criteria are necessary but insufficient for distinguishing between idioms and collocations, it is the fourth which ultimately distinguishes the two. It is also true that judging the degree of semantic opacity and metaphoric abstraction involves subjective judgements.

2.6 Is collocational meaning determined by the meaning of the words involved or by their tendency to co-occur?

Palmer (1976:96) has stated that although meaning by collocation is determined primarily by the meaning of the individual words within a collocation, in some instances collocational meaning is "idiosyncratic and cannot easily be predicted in terms of the associated words" (1976:76). For example, one could ask what it is about the word gaggle that attracts geese but not pigeons.

Firth's view on collocational meaning, as noted in his essay Modes of Meaning (1957), was that part of the meaning of a word is the set of other words with which it collocates. He understood `meaning by collocation' to be "an abstraction at the syntagmatic level [which] is not directly concerned with the conceptual or idea approach to the meaning of words" (1957:196). Quirk, for example, states that "...one of the meanings of PETTY is that it is frequently collocated with LARCENY" (1962:157).

Most authors agree with Firth, holding that, to a certain degree, the range of collocates of a node figures in establishing the meaning of the node. Carter has termed this "a structural semantic approach to word meaning...words do not exist in isolation: their meanings are defined through the sense relations they have with other words" (1987:18). However, some authors believe that it is rather the meaning of individual words that determines the range of words with which they will collocate. Cowie, a supporter of this latter position, states that the most important constraint on range or diversity of collocation appears to be the meaning of the word whose freedom of collocability one happens to be examining (1978).

Lehrer (1974:176-89) discusses the issues involved in these positions, which she terms the 'lexical position' and the 'semantic position', respectively, as well as a 'mixed position', which holds that both semantic and formal relationships play important roles in collocation. Firth's 'lexical position', however, already recognizes the importance of both semantics and structure in collocations, so the 'mixed position' will not be discussed here.

Within the 'semantic position', a number of semantic 'explanations' of the complex and diverse relations between elements in collocations have been put forth. Among these is Maher's 'salient feature copying' (1977), equivalent to what Backlund (1976) calls 'semantic redundancy'. More straightforwardly, Palmer calls it "having something in common semantically" (1976:97). Maher's 'salient feature copying' involves the pairing of a given lexical item with another that semantically copies a salient feature of the first. Examples of this include coal black (coal is by definition black, copying this feature in the adjective it precedes), as well as *snow white*, and *blood red*. Other examples such as *rose red*, *sea green*, and *lemon yellow* differ somewhat, but still reflect underlying comparisons. Other-than-colour term examples might include *ice cold* and *rock hard*. These examples, Cowie would claim, show how the semantic features of *coal*, *snow*, *blood*, etc. determine which word will collocate with the node in question. Not all collocations show salient feature copying, however: *face value* and *blind alley*, for example, do not.

Backlund's semantic redundancy (a feature he also terms 'bidirectional semantic overflow') is similar to salient feature copying in that it deals with words which are attracted to collocationally restricted nodes with which they share a common semantic element (ie. in the phrase *a good mixer*, *mixer* has a positive meaning, as does *good*, and in *brazen hussy* there is also a common semantic element between the two words). These attractions, Muller has pointed out, vary in strength, i.e., the attraction of one or the other item(s) will probably be greater in one direction than in the other.

Lyons (1977:v.2:612) and others have also noted semantic factors operating within collocations, especially in terms of the cohesive effects they produce. Firth himself mentioned "the association of synonyms, contraries, and complementary couples in one collocation"

(1957:199). Examples of these types of relationships are:

synonyms: "They named me executrix of their last will and testament." (or any example of doublets, where two words, of Latin and English origin respectively, occur together as in goods and chattels).

contraries: "It was a matter of life and death."

complementary couples: "All she needs me for is to fetch and carry."

It is also worth noting that if we accept that grammatical words and grammatical relations are essential to collocations, the 'semantic' position would perforce hold that syntactic relations, too, would have to play a role predicting in which words would enter into collocation. However, since the definitions of grammatical words state that they have minimal semantic content, it would be difficult to determine their exact role under the 'semantic' position.

However, the main problem with this approach to collocational meaning is that the list of different semantic features - semantic redundancy, salient feature copying, etc. - could turn out to be virtually endless, and by that fact, of little use in developing generalizations about collocations.

Most authors who adhere to the lexical position see collocation as a formal, rather than semantic, statement of co-occurrence (Halliday et al, 1964; Crystal, 1985; Lyons, 1977; Porzig, cited in Lyons, 1977; among others).

This co-occurrence has been called "powerful mutual predictability" (Crystal and Davy, 1962:56) or "the mutual expectancy of words" (Firth, 1957:196). Lyons (1977, vol.2:613) and others have stated that there is frequently so high a degree of interdependence between lexical items which tend to occur in texts in collocation with one another that their potentiality for collocation is reasonably described as part of their meaning.

Porzig (cited in Lyons, 1977, vol.1:261) has also mentioned the impossibility of describing the meaning of collocationally restricted lexical items without taking into account the set of lexemes with which they are syntagmatically connected. It would, for example, be difficult to explain the meaning of blond without mentioning hair (or tobacco or wood).

Halliday et al (1964) favor the lexical position because it is easier to handle than the semantic position:

> The formal criterion of collocation is taken as crucial because it is more objective and susceptible to observation than the contextual criterion of referential or conceptual similarity...the fact that shop and emporium are conceptually similar should not be used as a criterion to say that their collocational distributions should also be similar because, clearly, they are not. (p.34)

Katz (1972, cited in Fodor, 1977:98), Halliday and Hasan (1976) and Crystal and Davy (1969:56) are all of the opinion that there is no systematic semantic basis for some collocational restrictions. Leech (1974) thinks that collocational meaning is inexplicable, that it is "simply an idiosyncratic property of individual words" (p.20). This, Palmer points out, can be seen in the collective words such as _flock of sheep_, _herd of cows_, _school of whales_, _pride of lions_, _chattering of magpies_, and _exaltation of larks_ (p.77). Bolinger illustrates this when he states that

> If a child defines a hole as 'a hole in the ground', he is giving an example of the only kind of concrete existence a word has, which is in its remembered associations. (1976: 1)

Lyons (1977:vol2:265), however, cautions against defining the meaning of a word to be no more than the set of its collocations.

Palmer (1976:77) has pointed out that the meanings of words may vary according to the collocations into which they enter. For example, a _blind_ beggar and a _blind_ alley are not both blind in the same way. In each case the meaning of _blind_ is conditional on its association with either _alley_ or _beggar_. In order to devise a semantic model which would account for this, greater and greater degrees of delicacy would be required. The model would have to be extended to the point where it would be necessary to include overwhelmingly detailed information in the definition of the words. For example, _rancid_ certainly cannot be said to mean rotten in a butter-like or tuna-like way, for there is nothing inherent in the rottenness of butter or tuna that distinguishes it from the rottenness of anything else.

Mitchell shows himself to be a supporter of the lexical approach when he states that the dependencies between words in collocation lies in their association more than in how they might modify each other. He illustrates this with the collocation between /green and /grass (a 'simulative intensifier', in Mitchell's terminology) in (as) green as grass

> The association of /green with /grass does not 'modify' the meaning of /green any more than collocation with /green 'modifies' /envy in green with envy. (Mitchell, 1978:55)

Mitchell's statement can be seen as an illustration of collocation as a formal, rather than semantic, statement of co-occurrence. In the example above, /grass, the simulative intensifier, should be seen as comparable to very green, extremely green, or greener than, etc. The term 'simulative intensifier' itself, in fact, echos Carter's (1987:18) description of the lexical approach as a 'structural semantic' approach, one in which part of the meaning of a word, though by no means all of it, is determined by its collocations.

Other examples of simulative intensifiers could be the collocations between /smoke and /chimney, /swear and /trooper, /blue and /cold, and /easy and /pie.

It seems, however, that it might be difficult to set limits on what can or cannot be termed a "simulative intensifier" even though it would be useful to be able to define them formally. As it stands, they seem to occur only within a limited range of structures. For example, the bidirectional relationship between *easy* and *pie* is very evident in the construction *as easy as pie*, but is not so in ?*the ease with which this pie was made*, or ?*it was an easy pie to make*. There seems to be a gradation in this feature, from associations which accept transformations, to ones that are questionable, to many which are not acceptable in other structures. Compare the following:

as black as coal: coal black
as white as snow: snow white
as colo as ice: ice cold

But: ?as red as blood: blood red

and these ones which require an article:

as high as the sky: sky high
as hard as a rock: rock hard

But:

as green as grass: *grass green
as easy as pie: *pie easy
as strong as an ox: *ox strong
as black as the night: *night black
*as black as jet: jet black

The acceptability of the following might be doubtful:

as fast as lightning: ?lightning fast (or: ?lightning quick)

In conclusion, although the major authors agree that there exists an interdependence between lexical items and the text which surrounds them, there seem to be two distinct opinions as to the degree to which the potential meanings of a lexical unit are determined by its context.

Both the semantic position and the lexical position have strengths and weaknesses. It is hard to see how the semantic position could be used in practical applications because it requires the specification of a great many different semantic features, each of which can account for only a limited number of collocational patterns, but none of which can be useful in the formulation of a model to account for collocational meaning.

Although the lexical position maintains that there is no single explanation for collocational meaning, it does uphold the importance of the context in which lexical items are found.

3.0 Chapter II: Empirical Studies of Collocation

3.1 Description of three empirical studies of collocation

3.1.1 Berry-Rogghe, G.L.M. 1973. The computation of collocations and their relevance in lexical studies

The primary purpose of this pilot study was to "make explicit the notion of 'collocation' in statistical and computational terms" (Berry -Rogghe, 1973:103) in order to establish a methodology for the study of collocation. The author was interested not so much in developing collocational profiles, or lists of collocations (although that too was a stated secondary aim), but in attempting to answer two questions: 'what is the optimal span size?', and 'should grammatical words be ignored?' (1973:105).

Berry-Rogghe adopts the following definition of collocation from Halliday:

> The syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at n removes (a distance of n lexical items) from an item x, the items a, b, c... (Halliday, 1961)

In order to answer the two methodological questions raised, the author first compiled a list of collocates which significantly co-occurred with a given lexical item (defined as the graphic word) within a specified span, using the z-score to determine statistical significance. A z-score is a score expressed in standard deviation units which tells where one score lies in relation to other scores. It is used to compare a score's relative position in two or more score sets (Agnew and Pyke, 1982:179). The corpus used for this study consisted of three texts: A Christmas Carol by Charles Dickens, Each his own Wilderness by Doris Lessing (1959) and Everything in the Garden by Gils Cooper (1963), chosen because they were available in machine-readable form, not for any "stylistic considerations" (p.105). The number of running words was 71,595, which "proved sufficient for an initial methodological investigation" (1973:104).

Having chosen the word 'house' as the node, all items occurring within a span of three words on either side of the node (disregarding sentence boundaries) were conflated into an alphabetical list. For each item, the number of co-occurrences with the node was counted and 'tested' against a previously compiled 'dictionary' consisting of an alphabetic frequency table for the entire corpus. The total number of occurrences of each collocate were recorded and their z-scores computed. The 'significance limit' was set at a z-score of 2.576. Those items co-occurring with the node 'house' only once were rejected, as were those items with a negative z-score. It was found that the most frequent collocates of 'house' were grammatical words.

A second study was conducted in order to examine the question of span distance more closely. Having tested other (unreported) nodes for statistical significance, it was found that most significant collocates occurred within a span of four words on either side of the node. The author concludes that

> it seemed appropriate to provisionally adopt a span of four for both types of data [Dickens and the two modern writers] and for all nodes which were non-grammatical items, except in the case of adjectives where a span of only two seemed indicated. (1977:108)

Furthermore, the author felt that a number of grammatical words, "such as _and_, _but_, _it_, _nor_, _or_, _that_, _what_, _whether_, _which_, _who_, and all forms of the auxiliaries _be_, _do_, and _have_" (1973:110) should be excluded as possible collocates. This, the author points out, would permit the desired flexibility of span size because these words would not be counted.

3.1.2 **Haskel, P.I. 1971. Collocations as a measure of stylistic variety**

The purpose of this study was to establish a preliminary list for a proposed dictionary of collocations. This list would include 100 keywords along with a complete list of collocating words, with the percentage that each collocating word appears with the keyword in question relative to its overall frequency in the text.

Starting with Carl Darling Buck's A Dictionary of Selected Synonyms in the Principal Indo-European Languages (Chicago, 1949, no publisher's name provided in Haskel's report), a list was made of keywords from the dictionary which also occurred in the Brown Corpus (Francis and Kucera,1964). Eliminated were words which occurred both "very infrequently" (p.163) and with "the highest frequency" (p.164) in the Brown Corpus. Also excluded were grammatical words, both as keywords and as collocates. The keywords were then marked with their frequency of occurrence in the Brown Corpus, and those which appeared 200 or more times were selected. With the elimination of the grammatical words, 157 words remained on the list.

It was intended that the final list would be chosen by random methods, but since the list was to be limited to 28 words (for unspecified reasons), those on the list were checked for their "formal and semantic characteristics" (p.164), i.e. syllable length and part of speech. Words with "multiple diverse meanings" (p.165) were eliminated, and then others were added to the list so that all of Buck's twenty-two word categories (such as Parts of the Body, Food and Drink, etc.) would be represented. Words chosen were those falling nearest to the 200 frequency mark, for a total of 28 words.

In the first stage of the study, the computer scanned Buck's word categories separately, ignoring function words and focusing on keywords. When a keyword occurred the entire sentence, including function words, was printed. A span of four was used, which served about 75% of the time to collect the "words that have a direct structural or semantic relationship" (p.166) to the keyword. All collocating words within a span of four were placed into an array with the appropriate keyword and the words in each array were counted and keypunched.

In the second stage of the study, a computer program collected all of the collocating words from the Brown Corpus into one alphabetized dictionary complete with counts for keywords and counts for collocates. A raw count was kept rather than a percentage for this preliminary study since the percentage kept changing as more material was scanned. Only 7% of the million-word corpus was sampled.

In the third, projected, stage a list would be compiled showing the percentage of time that collocating words appear with each of the several keywords.

The author noted only several tentative conclusions, specifically that word frequency differed in the different genres of the Brown Corpus. For ten of the test words there were "noticeable patterns...which could be indicative of final results" (p.166), and the most frequent words "fulfilled the expectation that the greater the frequency the less would be the occurrence of _hapax legomena_" (p.166). Also, differences in meanings of individual words were noted in the different genres in the Corpus, and the author claims that this justifies looking at the genres separately before combining them for a standard dictionary.

The hypothesis of this study was that:

> ...originality is represented by numerous different collocates for a given word, or better, by unusual collocates: lack of originality is represented by stereotyped word groups or cliches. (p.167-8)

The study, however, fails to support this hypothesis.

3.1.3 **S. Jones and J.McH. Sinclair. 1974. English lexical collocations: a study in computational linguistics.**

This study was conducted in order to address two questions: (a) how can collocation be objectively described?, and (b) what is the relationship between collocation and meaning? The purpose was not to come up with usable collocational information, but rather to "establish some basic methods and principles" (p.18).

Two texts were examined in this study; 135,000 words of spontaneous conversation, recorded at Edinburgh University and University College London (referred to as the 'spoken text'), and about 12,000 words of written scientific English (the 'scientific text'). This amount of text was admitted to be an arbitrary choice. However, only words with a frequency of ten or more, a total of 1,155, were considered, and these were tested for significance using contingency tables and Fisher's Exact Probability Test. A span of 4 was adopted because it was found that "a very high proportion of relevant information could be obtained" (p.21) using this span. All of the collocational output in the study is taken from the spoken text, while the scientific text was used to test whether grammatical words occur in significant collocations.

The first part of the study, using a shortened version of the spoken text plus the scientific text, was conducted to examine the assumption that grammatical words would be characterized by a low ability to predict their environment. The words used as nodes were the following: <u>the</u>, <u>a</u>, <u>and</u>, <u>of</u>, <u>in</u>, <u>to</u>, <u>I</u>, <u>you</u>, and <u>it</u>. Using statistical methods to identify significant collocations, it was found that grammatical words are not, in fact, collocationally neutral. Even though they are not good predictors of specific words, they do show an ability to predict word classes at specific span positions. It was concluded that:

> Where the class is an open one [ie 'lexical words'] there will be little true lexical selection; the collocates appearing will be determined by the subject matter of the text. With closed classes [ie 'grammatical words'] the individual items may be predicted fairly accurately, due to the high frequency of certain grammatical structures or fixed word sequences in the language. (p.31)

A second part of the study was conducted in order to describe lexical behaviour through some specific examples. Twenty words with frequencies of between 90 and 290 occurrences in the spoken text were chosen as nodes. Only those collocates which occurred ten times or more within the text and at least three times in collocation with the nodes were considered. From these combinations, detailed tables of collocational patterns were produced showing frequencies, distributions, and measures of significance of the collocates. This information was then studied under four different headings: (a) collocation with grammatical words, (b) collocation with other lexical words, (c) position-dependent and position-free collocation, and (d) self-collocation (where a node collocates with itself, as in "from time to time").

In terms of collocation with grammatical words, it was found that collocations are governed by the rules of syntax. Thus, four specific verbs attracted subject pronouns on the left, object pronouns on the right, and were preceded by modal auxiliary verbs.

Collocations between words were determined partly by word class and partly by frequency. Frequent nodes displayed collocational patterns "with a considerable amount of lexical organization" (p.38). For example, adjectives are consistently preceded by adverbs and followed by nouns. The authors were looking for evidence of lexical sets in collocational patterning but found only "potential lexical sets" (p.42).

Position-free collocations are those in which the word class of the items is unimportant in terms of the significance of the collocation. For example, in the collocation between work and hard, where work is a verb, hard occurs as an adverb, and where work is a noun, hard occurs as an adjective. This type of collocation, the authors found, is relatively rare. Most significant collocations show a great deal of position dependence. The authors note that "freedom from fixed position is characteristic of lexical rather than grammatical associations between items" (p.43). The authors then used a Principal Components Analysis of seventeen nodes to statistically confirm that grammar and lexis are distinct forms of organization.

Self-collocation was found to be a "minor phenomenon" (p.45), and proved to be examples of `idiomatic phrases' (e.g. from time to time), repetition for rhetorical effect, and repetition arising from the conversational situation itself (question and answer, conversational repairs, hesitations, etc.). It was thus deemed to be unconnected with the theory of lexis and, in the context of the study, not worth pursuing.

No final comment was made concerning whether or not collocation can be objectively described or on the relationship between collocation and meaning.

3.2 Critical analysis of the three empirical studies

3.2.1 Unit of analysis

It is perhaps not surprising that none of the researchers considered using the lexeme as the unit of analysis in the three collocational studies being examined. Earlier it was noted that, theoretically, using the lexeme would be an ideal approach to studying collocations, but that syntactic restrictions prohibited it. However, what is surprising is that none of the studies gave any indication of having explored the collocational patterns of inflected or derived forms for similarities of distribution in order to see to what degree or in what ways syntactic restrictions might limit the use of the lexeme in collocational analyses. Given that much of the theoretical literature on the topic has argued that a large number of collocations would be amenable to this type of analysis, one would have expected specific mention of derived and inflected forms.

All three studies examined used the orthographic word as the unit of analysis. Jones and Sinclair (1974) provide a very specific definition of the unit their study is concerned with. The authors propose to examine the collocational behaviour of "lexical units", which are defined as:

> a) a morpheme
> b) a homograph - one 'meaning' of an orthographic word which may have several meanings
> c) a pair or group of words associated paradigmatically
> d) a pair or group of words associated syntagmatically, to form an 'idiom'. (1973:16)

As this study represents only an initial examination of collocational behaviour of lexical items, the authors have limited it to those lexical units which adhere to the form of the orthographic word and which can be distinguished by 'meaning' from other identical words (definition b). They do mention a proposed follow-up study which will examine the collocational behaviour of lexical units items falling into one of the other three categories.

The two appendices to the study are frequency tables of words in the text which occur ten or more times. These tables contain various forms of different lexemes. For example, at least seven forms of the lexeme /have are listed (<u>had</u>, <u>hadn't</u>, <u>has</u>, <u>hasn't</u>, <u>have</u>, <u>haven't</u>, and <u>having</u>), though there is no attempt to link these different forms. What makes a more thorough analysis of this study difficult is that since no accompanying list of collocates is given, it is unclear how the authors intended to distinguish between homographs. That is, if the node <u>quack</u>, for example, were being examined, a list of collocates would distinguish <u>quack</u> (a bad doctor) from <u>quack</u> (a sound made by a duck), and presumably would warrant a separate entry in the word frequency table. A list of collocates would then help distinguish which of the seven forms of /have might be functioning simply as auxiliaries, and which might be part of distinct collocations.

Despite this lack of information on collocational range, the authors do mention associations between words operating without regard to the form of the lexeme, but conclude that when an association of this nature occurs, it is unlikely that the combination is a collocation. That is, restriction on the freedom of association between different forms of lexemes is a strong indication of collocation.

It would have been interesting had the researchers been able to find, for example, instances of associations between different inflected or derived forms of work and hard, such as working, worked, worker, harder, and hardest and compared them to distributions of associations of forms of work with other collocates to see if they operate collocationally as a single item.

So although Jones and Sinclair note that the grammatical patterning of these words does not necessarily alter their association, they neither comment on nor explore the possibilities that this issue might raise in terms of the methodology of the study.

Berry-Rogghe (1973) mentions that, for her study, whether or not to consider the lexeme as a unit of analysis was not at issue, since the aim of the study was to compile a list of significant collocates for a number of nodes within a certain span, and not to explore the possible relationships these words had with each other. She used the same conventions as were used in the statistical analysis of American English at Brown University, that is: "For computational purposes, the lexical unit was defined as the `graphic word'..." (p.105). This meant that no lexical units consisting of more than one word were considered, and no distinction was made, for example, between come and come into [some money].

It is perhaps unfortunate that the word house was chosen as the node for this study, since the use of a more restricted item might have shed more light on how derivated and inflected forms could be treated. As it is, the only derived or inflected form which appeared on the list of collocates of house were the verbs do, be, and have, none of which can be considered verbs with restricted distributions and which probably functioned as auxiliaries in the corpus rather than as elements in a collocation. The seven most significant collocates of the node were, in order of significance, sold, commons, decorate, this, empty, buying, and painting (but not sell, decorating, emptied, bought, or painter).

Berry-Rogghe does, however, recognize that the approach in this study is only an initial venture into collocational analysis and that a more inclusive unit of analysis, perhaps the lexeme, might be useful.

> the eventual aim of collocational analysis is not just to establish sets of syntagmatically related items but to extend these to include paradigmatically related items so that eventually a 'semantic field' might be established. (p.111)

Haskel's (1971) study was also conducted at the word level. As in the other studies, collocation was considered at the level of the association of only two words, even though collocations may be longer, as in the combination of a phrasal verb and a noun. An interesting point noted in the study was that homographs were revealed as such through differences in the range or patterning of collocations in different genres of text. It is uncertain how the author was able to note these differences, though, since she specifically states that "...most words of multiple diverse meanings [i.e. homographs] were removed from the list" (p.165).

3.2.2 Closed-class grammatical words in collocations

We Have seen that there are two opposing views on whether or not grammatical words should be taken into account in the study of collocations and that there can sometimes exist abmiguity as to a word's status as either a grammatical word or a lexical word.

For the purposes of their study, Jones and Sinclair have defined the term 'grammatical item' as

> a unit of language whose presence in the text is due to its grammatical function rather than to any`meaning' it may represent. Possible examples are the words the, a, and, and the morphemes ing, ed, s. Co-occurrence with these items is, in theory, due to the influence of grammar alone and not to lexis. (1974:16)

They further include prepositions, personal pronouns and deictics. All of these classes of words, these 'grammatical items', according to Halliday (1966:155), are characterized by membership in 'closed' or finite word classes (i.e. not open-ended), high frequency in texts, and low ability to predict their environments.

Halliday hypothesized that grammatical items would be collocationally neutral. Jones and Sinclair, however, claim that the findings of the study show evidence contrary to Halliday's hypothesis. They found that grammatical words attracted a large number of significant collocates but that their collocational patterns "showed a distinctly grammatical influence" (1974:27). In fact, grammatical words were found to be good predictors of word classes, but not of individual words, and they tended to achieve significance over long stretches of text, primarily as a function of their relatively great frequency as well as their tendency to be more evenly distributed in the text than lexical words. The authors also mention that the associations between grammatical words and lexical words could not always be explained in grammatical terms. They noted, however, that certain verbs attracted certain grammatical words and not others.

> This selection may be due to some lexical influence...It is (more) plausible to attribute choice of preposition after the verb to lexis: find + out, for instance, forms a phrasal verb whereas find + in does not. (p.38)

Perhaps if the concept of the lexical item were to be redefined so as to include phrasal verbs as separate units, this problem might be clarified. Phrasal verbs might be seen as quasi-idiomatic units which act in much the same way as idioms proper. Although this issue was mentioned briefly, no real investigation was conducted and no conclusions reported.

Jones and Sinclair are aware that using grammatical items as nodes is of little interest because of their inability to predict specific lexical items. For this reason they chose to exclude them from the choice of words to be tested for significant collocation. Since words with more than 300 occurrences were usually grammatical words, frequency ranges were limited to 90-290 (p.33). An analysis of grammatical words as collocates would be valuable as it would allow the inclusion of co-occurrences like by chance, on purpose, and come into, while precluding examination of on + any noun, by + any gerund, or any verb + into.

One of the questions Berry-Rogghe was concerned with was whether or not to ignore grammatical words in the study of collocations. She found that the most frequent collocates of the node chosen for the study (house) were indeed grammatical, and that this tended to hamper the gathering of "relevant items" (p.108) near the node because of a fixed span size. It was decided that the span size should be made flexible by excluding certain grammatical items (or "functors', which have only minimal semantic content" p.110) as collocates.

> A proposed 'exclusion list' would include such items as and, but, it, its, nor, or, that, what, whether, which, who, and all forms of the auxiliaries be, do, and have. (p.110)

It is unclear, however, why Berry-Rogghe excludes conjunctions, personal and relative pronouns, as well as auxiliaries, yet does not exclude articles and prepositions. Nonetheless, she concludes that while some grammatical items should be ignored in the study of collocations, grammatical relationships should not be:

> A case might be put forward for considering as potential collocates only those items which stand in

> a grammatical relation to the node. Generally speaking, this would have the effect of including only relevant items. However, unless a very sophisticated grammatical analysis which takes into account anaphoric reference is applied, a great many important collocates might be lost. (p.108)

In addition to the fact that such a sophisticated grammatical analysis is not yet feasible, Berry-Rogghe's suggestion is open to the same criticism as Kjellmer's (see page 13 above). A grammatical restriction criterion alone fails to distinguish between collocations and free constructions.

Haskel (1971) very straightforwardly eliminates not only grammatical words from consideration in her study, but also any word, like will and can, which can function as a grammatical word:

> It was determined that function words would be excluded from the list of collocating words as well as from the list of keywords. Function words can, of course, reveal a great deal about the structure of a language. Here, however, the emphasis is on lexical word selection as an independent parameter of style. (p.164)

A problem with this approach, as Jones and Sinclair (1974) and Halliday (1966) have pointed out, is that the selection of a grammatical word is often due to lexical influence, especially in phrasal verbs. It must be assumed, then, that Haskel's study is not concerned with those collocations which include grammatical words, but only those which involve associations among nouns, adjectives, verbs and adverbs.

Although among the three studies there seems to be no agreed upon list of what is or is not considered a grammatical word, there is overlap on some points. Both Jones and Sinclair and Haskel agree that grammatical words can be identified if they fulfill a grammatical function in the text, and Berry-Rogghe agrees with Jones and Sinclair that grammatical words have only minimal semantic content. Both Berry-Rogghe and Haskel believe that grammatical words should be excluded from the study of collocations. Jones and Sinclair think they should be included in collocational analyses as collocates rather than nodes.

Given the limitations of Haskel's and Berry-Rogghe's studies, as well as the fact that most of the literature on the topic recognizes grammatical restriction as a major characteristic of collocations, it would seem to make sense to consider using grammatical words as elements in collocations.

3.2.3 Collocational Span

In all three of the empirical studies the rationale for selection of span size was discussed, and it was agreed that the choice is ultimately arbitrary. Different types of collocational information were gathered at different span sizes, but all three studies concluded that the most `useful' span size was three to four words on either side of the node. It was also agreed that grammatical structures and sentence boundaries should not be taken into account when calculating the span.

Jones and Sinclair examined collocates up to span positions of +/-10, but found that the influence of the node did not extend beyond span position +/-4. Sentence boundaries did not enter into limiting the span size since it was recognized that words which enter into collocations may be in different sentences. And since sentence boundaries are often impossible to determine accurately in spoken English, it was expected that significant collocations would occur across, as well as within, individual utterances (p.17).

Another reason for limiting the span size to +/-4 was that elements in collocations were found to adhere to the rules of syntax, where words of a specific class are predictibly found in certain positions. In the example provided of the sort of collocational profile produced in the study (in this case, using the adjective 'good' as the node), it was noted, predictably, that 19 of the collocates occured at the N-4 to N-1 positions (examples being as good, jolly good, extremely good, pretty good, quite good; that's [] good, was [] good, is [] good; it's [] [] good; thought [] [] [] good). Fourteen collocates occurred at the N+1 position, (good for, good at, good indeed, good fun, good thing, good memory, good example, good God, good heavens, good subject, among others). Four collocates occurred in the N+2 to N+4 positions. This seems to indicate a strong focus on the N+1 position, and thereafter, a stronger focus to the left of the node than to the right. This may merely be a function of syntax, since adjectives usually precede the nouns they modify.

Berry-Rogghe (1977) suggested that span size should be flexible. Within a single study, it would vary according to the word class of the node in question.

Haskel (1971) also makes the same suggestion, holding that, in many cases, collocation of semantically and structurally related words occurs within a span of two, though the span size may increase to as many as eight words.

Flexibility of span size, according to Berry-Rogghe (1977), would also be obtained through the exclusion of certain grammatical items from consideration. As we have already seen, though, this would effectively eliminate a large number of collocations from a study since grammatical items often play important roles in collocations.

Berry-Rogghe seems to feel that flexibility of span size is important because of an observed connection between language variety and distances between nodes and their collocates. She states that "the stylistic nature of the corpus is highly relevant in defining the optimal span size" (p.108). It was found that there were noticeable differences in mean sentence length between the modern texts and the Dickens story, and that this was a factor in whether or not certain collocates fell within the span being used in the study.

It would seem that both Haskel and Berry-Rogghe have approached the question of span size by first finding the collocations in their texts and then establishing span sizes to include them. This belies Berry-Rogghe's stated aim of establishing an "optimal" (p.105) span size for the study of collocations. It may suggest, however, that given this approach to the study of collocation, it is not possible to establish an 'optimal' span size.

3.2.4 Statistical Significance

In two of the studies, Jones and Sinclair and Berry-Rogghe, significant collocations were defined as those associations of words which co-occur more often than their respective frequencies and the length of text in which they appear would predict. Both of these studies used a significance level of 0.1%, to which Berry-Rogghe added a z-score limit of 2.576 since this "yielded a gradation among collocates which largely corresponded with our semantic intuitions" (p.104). No comments or conclusions can be made from Haskel's study, however, since only about 7% of the corpus was sampled before the writing of the report. It seems that significance was, however, to be judged in terms of the ratio of each collocate to its total occurrence in the text.

Jones and Sinclair state that the reason for performing a significance test at all was to filter out "casual collocations", or ones which co-occur for reasons not strongly connected with the node. This, they assume, would leave "only those words whose association with the node is part of a regularly recurring pattern in the language" (p.32). Haskel (1971) noted that low-frequency words were the ones which accounted for many of the `casual collocations' in her study.

The choice of the 0.1% of significance was made for the two studies because it was found to be acceptably reliable. For every 1,000 times a test is performed at this level, it judges 1 collocation to be significant when it is not. That is, the test is almost always right in saying something is significant. Berry-Rogghe qualifies this by saying that the significance level chosen for any study, as well as the formula used to calculate it, is subject to the judgement of the individual investigator and to the purpose of the study.

The words chosen to be tested for significance in the Jones and Sinclair study also needed to meet certain criteria. Because the authors wanted primarily lexical nodes, they established a frequency range which would effectively exclude grammatical words as potential nodes. For collocates, a lower limit of 10 occurrences was chosen because the significance test would otherwise be unreliable. An upper limit for collocates was not set.

In asking how many times two words should co-occur before their collocation was submitted to the significance test, it was decided that no collocation consisting of only one co-occurrence would be tested since statistical tests in this case would give unsatisfactory results. Collocates with only two occurrences tended to bring out `casual collocations', and therefore the minimum number of co-occurrences was established at 3.

Two problems were noted, however. First, with the criterion of three or more co-occurrences, a great deal of information on the lexical behaviour of lower frequency nodes was lost. It was suggested that in a longer text this would not be a problem, though. Second, there was the question of certain kinds of collocations that arose because of the significance limit. It was found that calculations of significance produced such things as `accidental' collocations which arose because of repetition and other characteristics of spoken language which differ noticeably from written or scripted language. Berry-Rogghe (1973) also noted that the use of the z-score failed to account for self collocation, and that limits on significance might exclude from consideration "'unusual' but 'creative' collocations...along with obviously irrelevant ones" (p.107). It was unclear exactly what an 'irrelevant' collocation could be, though.

3.2.5 Distinguishing collocations and idioms

It has previously been noted that confusion in distinguishing between idioms and collocations might be a result of the fact that they share a number of characteristics. Both consist of more than one word, are syntactically frozen to some degree, and resist lexical substitution. It has been agreed, in fact, that any real distinction between the two is made on the basis of semantic transparency.

It is precisely a focus on a systematic semantic analysis that is lacking in all three studies, since all of them used a computational approach to analyze collocation within texts. Berry-Rogghe, despite referring directly to Firth's statement that collocation is more than simple co-occurrence of words in a text (p.103), still adopted a methodology based on statistical measurements of simple co-occurrence for analyzing collocation.

Following Halliday (1961), Berry-Rogghe defined collocations strictly on the basis of probabilities. This, of course, is in accordance with her stated aim of "attempting to make explicit the notion of 'collocation' in statistical and computational terms" (p.103). The particular formulas she used in the study, however, were adopted because the results they gave "largely corresponded with our semantic intuitions" (p.104). It would seem, then, that some sort of informal semantic analysis was done.

Without the inclusion of a semantic analysis in the study, no distinction between idioms and collocations could be made. However, Berry-Rogghe does mention the presence of an `idiom', "House of Commons" (p.106), in the results of her study, and includes the word full in a list of significant collocates of the word house, without commenting on whether or not it formed part of the idiomatic poker term "full house".

Jones and Sinclair proposed to test a number of predictions about the nature of lexis made in Sinclair (1966), using `statistical techniques', that is, using a computational approach to the study of collocations. However, although there seems to be very little semantic analysis in the report, Jones and Sinclair do attempt to distinguish idioms from other `lexical items'.

A 'lexical item', in this study, is a "unit of language representing a particular area of meaning which has a unique pattern of co-occurrence with other lexical items" (p.16). It can take one of four forms: 1) a morpheme, 2) a homograph, 3) a pair or group of words associated paradigmatically, or 4) a pair or group of words associated syntagmatically, to form an `idiom' (p.16). A further description of the lexical item states that:

> If the collocational behaviour of the unit of language under discussion turns out to be significantly different, within a large sample of natural language, from the behaviour of any other unit, it is a lexical item, whatever its form. (p.16)

On a purely syntactic level, this classification of idioms as lexical units cannot be upheld. Idioms cannot be considered single lexical units here because their identity as lexical units has to be revealed through their collocational patterning, and it is not likely that an idiom would collocate with another idiom.

This point is moot, in any case, since this study is limited to collocations between "certain orthographic words" (p.16), that is, between pairs of words rather than among more than two, and since idioms are often more than two words long, they fall outside the scope of this study. The authors do hint at another study with a wider range, which would examine the problems of identifying lexical items which do not take the form of a single orthographic word.

Given this exclusion of idioms from the scope of this study, it is noteworthy that Jones and Sinclair do include several two-word idioms in one of their tables (Table 3, p.35). The table illustrates the collocational patterning of the node good. There we find good God, good lord and good heavens, though no comment on their idiomaticity is found in the report of the study. It would seem probable that two-word idioms were not identified as idioms proper since a semantic analysis would have been necessary to distinguish them from collocations.

Although Haskel does not address the question of distinguishing idioms and collocations, she does note two of the characteristics shared by idioms and collocations; syntactic frozenness and resistance to lexical substitution. She describes collocations as sometimes being "...little more than stereotyped word groups or cliches..." and "...ready-made expressions..." (p.160), i.e., syntactically frozen. However, they are not as resistant to lexical substitution as idioms are, since she speaks of investigating `unusual collocations' as a measure of originality. These two criteria, as we have noted, are not sufficient for distinguishing collocations and idioms.

It is strange that although Firth's original motivation for introducing the idea of collocation stemmed from an observation of how meaning was conveyed by words in collocation, none of the three studies examined here have even mentioned the semantics of collocation. Distinguishing between collocations and idioms would be a major step in clarifying the concept of collocation itself. Even though the statistical approach used in the th studies is less complicated than a semantic approach, it cannot by itself contribute to a better understanding of the nature of collocation.

3.2.6 Collocational meaning

Because the primary focus of the three studies was on analyzing collocations on a computational basis, there is little overt mention of whether collocational meaning is determined by the words in question or by their tendency to co-occur. Nonetheless, both the Jones and Sinclair and Haskel studies seem to view collocational meaning as a result of interaction between the text and the word whose collocational behaviour is being examined. That is, they both support the lexical position as outlined by Lehrer (1974). The report on Berry-Rogghe's study is more difficult to evaluate because it is presented in such a way that no judgement can be made as to her position on this issue.

Jones and Sinclair start by qualifying any discussion on this issue by saying that because of the limited size of the corpus examined in their study, their results may not be representative of tendencies in the language as a whole. Nonetheless, they did find that some meaning results from the context within which a node is found.

They examined collocational patterning for both grammatical and lexical words, and found that while the latter predict specific words, and therefore specific meanings, the former do not. Collocations with grammatical words as nodes were more arbitrary and more text dependent than collocations with lexical nodes. That is, when grammatical nodes were studied, the lexical items they attracted depended on the nature and topic of the text.

Unlike grammatical words, lexical words in collocations were found to be "much more a part of the language structure" (p.39). Especially when examining lexical words which attracted grammatical ones, as in phrasal verbs, the authors found little reason to think that semantic factors are responsible for the choice of collocate.

> ...there seems no particular reason why *take* should attract *away* and *put* attract *down* [in the corpus] and not the other way round. (p.38)

Jones and Sinclair (1974:40) also noted that lexical nodes which share some semantic element - time, for example - also share a number of collocates. They found, though, that this tendency is specific to some lexical items and not others. Words with general reference, such as *thing* and *put*, do not share collocational ranges with other words. But, as Halliday has mentioned, the fact that two words are synonomous cannot be used as a basis for establishing similarity in collocational ranges (Halliday et al, 1964). Nor, as noted earlier, is it feasible to establish semantic explanations for collocational ranges that are not generalizable.

Haskel states her position quite directly. She cites Firth's position that the meaning of a word can be determined by collocation and she states that "collocations can...define the words of a language and reveal aspects of its structure" (p.160).

Most interestingly, in her study, homographs were revealed as such through their patterns of collocation. The meanings of words, either literal or figurative, became clear when the collocates of those words were studied, especially in different genres of texts. Examples she provides were how cut in its literal meaning collocated in fiction with open, belly, concussion, and boy; whereas in press reportage it showed a more figurative meaning ('decrease') in its collocations with inflate, modest, expenses and estimates (p.167).

Berry-Rogghe's study offers no discussion of this issue. This, perhaps, is partly due to the approach she used for the study, and partly due to the choice of texts used as her corpus. She combined a 19th century story with two modern stories, so although the genre is the same, the 'language' is not. Since language changes over time, Dickens' usual collocations would probably not be the same as Lessing's or Cooper's. Such a significant difference in language varieties would make it difficult to draw any conclusions or describe any tendencies in the results.

4.0 Suggestions and Conclusion

Three empirical studies of collocation have been examined in terms of how they dealt with the six questions raised in the review of the literature. In general, none of these studies analyzed collocations in such a manner that all of the structural and semantic complexities of this linguistic 'tendency' were reflected in the results. There were two main problems with the methodologies of the studies: 1) the basic approach to collocations was statistical in nature, including little or no semantic or structural analysis, and 2) there was no integration of the six aspects of collocation examined in this thesis.

A computational approach to the study of collocation can be invaluable if it is based on an analysis of corpora which are large enough to establish reliable collocational profiles. However, two of the three studies examined in this thesis drew conclusions about the nature of collocations from corpora of about 70,000 running words. The third used a corpus of only 147,000 words. In contrast, Halliday (1966) has asserted that any empirical study of collocation should use a corpus of at least 20 million words, or twenty times the size of the Brown Corpus.

Collocation has been described by Carter as a "structural semantic" (1987:18) aspect of language. This description underlines the fact that a semantic analysis is essential in any empirical study of collocation, for it is often only along semantic lines that collocations can be distinguished from other types of linguistic constructions. For example, a purely computational analysis would not account for the essentially semantic difference between the idiom *red herring* and the collocation *red hair*, and a structural analysis would be necessary to help explain why *reckless abandon* could be classified as a collocation but *to abandon recklessly* could not.

It becomes clear, therefore, that on this question the studies of Jones and Sinclair, Haskel, and Berry-Rogghe all suffered from two distinct problems; limitations in terms of the size of the corpora examined (which, if a computational approach is used in a study, must be substantial), and limitations on the comprehensiveness of the computational approach itself.

While a remedy to the first of these problems might be found in the use of larger corpora, such as the 20-million-word corpus used in the compilation of the COBUILD dictionary (Sinclair, 1986), the second problem is more complicated. It would seem that if a model for the study of collocation could be found which would resolve the problems associated with finding satisfactory answers to the six questions addressed in this thesis, an operational definition of the term `collocation' could be achieved.

It is not within the scope of this thesis to propose such a model. However, the analysis of the particular difficulties which the three empirical studies faced in applying the theories outlined in the literature has suggested certain essential features of a model for studying collocations, and hence certain requirements for an operational definition of the term.

What is required in the study of collocation is a means of analyzing corpora so that collocational information, and only collocational information, can be abstracted. This would entail the adoption of a method of analysis which would incorporate the six aspects of collocation discussed in this thesis, and would preclude the use of a purely computational approach. Before undertaking a study of collocation, a researcher would have to decide, in this order, a) what it is that is being examined, including where grammatical items fit into the study, b) where the meaning of the whole comes from, c) how much text to look at, and d) how typical the structure is of the language as a whole.

This thesis would suggest that the following answers to these questions, considered in the following order, would lead to a more definitive view of collocation.

1. What unit of analysis will best reveal the paradigmatic nature of collocation?

While it would seem that the unit of analysis in an empirical study of collocation must remain the orthographic word for the time being, there are at least three possible refinements available to the researcher. Keeping track of the distribution of the different forms of a lexeme (once they are established) might lead to a method of classifying collocating lexemes according to the type of syntactic restrictions under which they operate. Another approach a researcher might adopt, as suggested by Jones and Sinclair (1974), is to extend the unit of analysis to lexical units which consist of more than one word so that collocations between, for example, phrasal verbs and other items, whether lexical or grammatical, might be studied. This would have the great advantage of allowing the inclusion of grammatical words in the study of collocations. A third approach would be to do as the authors of the three studies have done, which is to examine the collocational patterning of selected words or groups of words in a corpus. This approach could be further refined to a comparative analysis of the collocational patterning of synonyms (including those which are phrasal verbs), which would effectively reveal the lexical restrictions, as well as some of the structural restrictions, which characterize collocation.

2. How can grammatical items best be included in the study of collocations?

Because the context of any linguistic item includes both lexical and grammatical relationships, grammatical items must be considered integral parts of collocations. It is suggested that if grammatical items were considered in the context of being parts of phrasal verbs or compounds, such as in take off and takeoff, their `lexicalness' could increase, as would their value in terms of determining collocational patterns. However, grammatical items should not be considered as nodes, since they have been found to be good predictors of word classes rather than of specific words.

Kjellmer's (1984) grammatical well-formedness criterion is necessary because it effectively puts limits on the string of text considered and helps establish collocations as distinct structures. However, this approach is not sufficient since, by itself, it does not distinguish collocations from free constructions

Thus, if any new research into collocation is to be conducted, it would seem essential that a grammatical analysis be incorporated into that research, so that a more thorough and complete picture of collocation might result. Using a text-based grammar would be required in order to account for anaphoric reference, for example, in combination with a re-evaluation of the unit of analysis, whether it be a single orthographic word, or a sequence, sometimes discontinuous, as with some phrasal verbs.

3. Does the meaning of the words in collocation depend on their association or is it independent of the association?

The context in which a lexical item is found, according to Firth, plays a large role in determining the meaning of that lexical item.

Collocational meaning can best be described as idiosyncratic in nature and attributable to a lexical item's presence (though not a grammatical item's presence) in surrounding text. Therefore, future empirical studies of collocation must examine collocations within a span of text large enough to establish a context, either oral or written, in which collocational meaning would become apparent.

4. How are idioms and collocations distinguished?

It has been noted that only on the basis of semantic transparency can collocations and idioms be distinguished, since the other two commonly accepted characteristics of idioms, lexical substitutability and syntactic frozenness, are also shared by collocations. Semantic transparency, moreover, must also be judged as to what degree of abstractness is acceptable before a string of language is judged to be an idiom. It is obvious that a computational approach to collocations cannot incorporate these types of analyses, and would therefore be of limited practical use. Future empirical studies should, therefore, include not only a semantic analysis, but also a means of judging the degree of semantic abstractness of a string of language in order to eliminate from a study of collocation any word combination that is semantically opaque or which is highly metaphorical in nature.

5. What span should be considered for a study of collocation?

The approach to collocational span used by most authors has been shown to be untenable since the strength of a collocation cannot be equated with the proximity of collocating words, especially for collocations between words which are habitually discontinuous.

A span size might be usefully established for empirical studies of collocation if it were to take into account Kjellmer's criterion that collocating words be within a single grammatical unit. In this way, much intervening text could be eliminated from the string of text considered in collocations. It seems, though, that a method of accounting for collocations which occur across sentence boundaries, or are made up of more than two words, as well as those which involve anaphoric reference must supplement the grammatical criterion. Therefore, an ideal span size would be one only as large as necessary to include both lexical and grammatical information.

6. How can collocational significance be measured?

Establishing the statistical probability that one lexical item will collocate with another has been used as a means of determining how typical a certain combination is in the language. Statistical measurements of the strength of a collocation, however, can only provide a partial description of this tendency. They cannot distinguish collocations from idioms, nor can they account for differences in syntactic patternings of derived or inflected word forms. Frequency of a particular combination in a corpus can only be suggestive of collocational status, and must be supplemented by other types of analysis.

Future empirical studies of collocation can, and should, use statistical measurements of significance. However, semantic and structural analyses should precede them so that rules for eliminating grammatical structures like *to be* as potential collocates can be properly formulated. Also, using measurements of significance could be useful for establishing a cutoff point of mutual predictability so that lexicographers, for example, might be able to include in learner's dictionaries only those collocations most typical of the language.

The value of collocational studies to the teaching of English as a second language cannot be overestimated; the recent appearance of learner's dictionaries containing pertinent collocational information bears witness to this. A rigorous definition of collocation will make it possible for lexicographers to provide reliable information on many of the semantic and structural complexities of the language, thereby assisting second language learners in their task.

References

Agnew, N. McK., and S.W. Pyke. (1982). *The Science Game: An Introduction to Research in the Behavioral Sciences*. (3rd ed.) Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Aisenstadt, E. (1979). Collocability restrictions in dictionaries. In R.R.K. Hartmann (Ed.), *Dictionaries and Their Users* (pp.71-74) Exeter: University of Exeter Press.

Anthony, Edward M. (1973). *Towards a Theory of Lexical Meaning: An Essay*. Singapore: Singapore University Press for SEAMEO Regional Language Centre

Backlund, U. (1976). Frozen adjective-noun collocations in English. *Cahiers de Lexicologie*, *28*(1), 74-88.

Bazell, C.E., J.C. Catford, M.A.K. Halliday, & R.H. Robins. (Eds.). (1966). *In Memory of J.R. Firth*. London: Longmans, Green and Co. Ltd.

Benson, M. (1985). Collocations and idioms. In R. Ilson (Ed.), *Dictionaries, lexicography and language learning* (ELT documents 120). (pp.61-68). Oxford: Pergamon Press.

Benson, M., E. Benson, & R. Ilson (comps.). (1986). *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. Philadelphia: John Benjamins.

Berry, M. (1977). *An Introduction to Systematic Linguistics*. London: B.T. Batsford Ltd.

Berry-Rogghe, G. (1973). The computation of collocations and their relevance in lexical studies. In A.J. Aitken et al. (Eds.),*The Computer and Literary Studies* (pp.103-112). Edinburgh: Edinburgh University press.

Bolinger, D.L. (1968). *Aspects of Language*. New York: Harcourt Brace and World Inc.

Bolinger, D.L. (1976). Meaning and memory. *Forum Linguisticum*, *1*, 1-14.

Bolinger, D.L., & D.A. Sears. (1981). *Aspects of Language*. (3rd ed.) New York: Harcourt Brace and Jovanovich.

Carter, R. (1987). *Vocabulary: Applied Linguistic Perspectives*. London: Allen & Unwin.

Cowan, G.M. (1965). *Some aspects of the lexical structure of a Mazatec historical text*. Publications in Linguistics and Related Fields, No. 11, Norman, Oklahoma, Summer Institute of Linguistics.

Cowie, A.P. (1978). The place of illustrative material and collocations in the design of a learner's dictionary. In P. Strevens (Ed.), *In Honour of A.S. Hornby* (pp.127-130). Oxford: Oxford University Press

Cowie, A.P. (1981). The treatment of collocations and idioms in learner's dictionaries. *Applied Linguistics*, *2*(3), 223-235.

Cowie, A.P. (1983). The pedagogical/learner's dictionary: I. English dictionaries for the foreign learner. in R.R.K. Hartmann (ed.), *Lexicography: Principles and Practice* (pp.135-152). London: Academic Press.

Cowie, A.P., R. Mackin, & I.R. McCaig. (1983). *The Oxford Dictionary of Current Idiomatic English*. Oxford: Oxford University Press.

Croft, K. (1967). Some co-occurences in American cliches. *TESOL Quarterly*, *1*(2), 47-49.

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

Crystal, D. (1985). *Dictionary of Linguistics and Phonetics*. (2nd ed). Oxford: Basil Blackwell Ltd.

Crystal, D., & D. Davy. (1969). *Investigating English Style*. London: Longmans, Green and Co., Ltd.

Firth, J.R. (1951). Modes of meaning. reprinted in *Papers in Linguistics 1934-1951*.

Firth, J.R. (1957). *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Fodor, J. (1974). *Semantics: Theories of Meaning in Generative Grammar*. Hassocks, Sussex: Harvester Press, Ltd.

Francis, W.N., & H. Kucera. (1964). *Manual of Information to Accompany A Standard Sample of Present-Day Edited American English, For Use With Digital Computers*. Providence, RI: Brown University Press.

Gleason, H.A. (1962). The relation of lexicon and grammar. In F.W. Householder and S. Saporta (Eds.), *Problems in Lexicography* (pp.85-102). Bloomington: Indiana University Press.

Gruber, J. (1976). *Lexical Structures in Syntax and Semantics*. Amsterdam: North-Holland Publishing Co.

Halliday, M.A.K. (1966). Lexis as a linguistic level. In C.E. Bazell, J.C. Catford, M.A.K. Halliday, & R.H. Robins (Eds.), *In Memory of J.R. Firth* (pp.148-162). London: Longman.

Halliday, M.A.K., A. McIntosh, & P. Strevens. (1964). *The Linguistic Sciences and Language Teaching*. London: Longmans, Green and Co., Inc.

Halliday, M.A.K., & R. Hasan. (1976). *Cohesion in English*. London: Longman.

Hartmann, R.R.K. (Ed.). (1983). *Lexicography: Principles and Practice*. London: Academic Press.

Haskel, P.I. (1971). Collocations as a measure of stylistic variety. In R.A. Wisbey (Ed.), *The Computer in Literary and Linguistic Research* (pp.159-168). Cambridge: Cambridge University Press.

Jones, S., & J. McH. Sinclair. (1974). English lexical collocations: a study in computational linguistics. *Cahiers de Lexicologie*, *24*(1), 15-61.

Kjellmer, G. (1984). Some thoughts on collocational distinctiveness. In J. Aarts and W. Meijs (Eds.), *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research* Costerus, N.S., *45*, 163-171.

Kjellmer, G. (1987). Aspects of English collocations. In W. Meijs (Ed.), *Corpus Linguistics and Beyond: Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora* Costerus, N.S., *59*, 133-140.

Korosadowicz-Struzynska, M. (1980). Word collocations in FL vocabulary instruction. *Studia Anglica Posnaniensa*, *12*, 109-120.

Langendoen, D.T. (1968). The London School of Linguistics: A Study of the Linguistic Theories of B. Malinowski and J.R. Firth (pp.37-75). Cambridge, Mass.: M.I.T. Press.

Leech, G. (1974). Semantics. Hamondsworth: Penguin.

Lehrer, A. (1974). Semantic Fields and Lexical Structure. (North Holland Linguistic Series 11) Amsterdam: North Holland Publishing Co.

Lyons, J. (1977). Semantics (Vols. 1-2). Cambridge: Cambridge University Press.

Mackin, R. (1978). On collocations: words shall be known by the company they keep. In P. Strevens (Ed.), In Honour of A.S. Hornby (pp.149-165). Oxford: Oxford University Press.

Maher, J.P. (1977). Papers on language theory and history: creation and tradition in language (Vol. 3). Amsterdam: Amsterdam Studies in the Theory and History of Linguistic Sciences.

Makkai, A. (1972). Idiom Structure in English. The Hague: Mouton.

Martin, W.J.R., B.F.P. Al, and P.J.G. van Sterkenburg. (1983). On the processing of a text corpus: from textual data to lexicographic information. In R.R.K. Hartmann (Ed.), Lexicography: Principles and Practice (pp.77-87). London: Academic Press.

Mel'cuk, I.A. (1960). [On the Terms `Stability' and `Idiomaticity']. Voprosy jazykoznanija, 4:73-80.

Mitchell, T.F. (1971). Linguistic `goings on': collocations and other lexical matters arising on the syntagmatic record. Archivum Linguisticum, 2 (N.S.): 35-69.

Mitchell, T.F. (1975). Principles of Firthian Linguistics. London: Longman.

Muller, E.A. (1981). Towards a stratificational description of collocations. In J.E. Copeland and P.W. Davis (Eds.), The Seventh LACUS Forum, 1980 (pp.175-187). Columbia, S.C.: Hornbeam Press, Inc.

Nagy, W. (1978). Some non-idiom larger-than-word units in the lexicon. In D. Farkas, W.W. Jacobsen, & K.W.

Todrys (Eds.), Papers from the Parasession on the Lexicon (April 14-15, 1978) (pp.289-300). Chicago Linguistic Society, Chicago: University of Chicago Press.

Palmer, F.R. (1965/74). The English Verb. London: Longman.

Palmer, F.R. (1976). Semantics: A New Outlook. Cambridge: Cambridge University Press.

Porzig, U. (1934). Westenhafte Bedeutungsbeziehungen. Beitrage zur deutschen Sprache une Literatur, 58:70-97.

Proctor, P. (1978). Longman Dictionary of Contemporary English. Burnt Mill, Harlow, Essex: Longman.

Quirk, R. (1962). The Use of English. London: Longman, Green and Co., Ltd.

Richards, J., J. Platt, & H.Weber. (1985). Longman Dictionary of Applied Linguistics. Harlow: Longman Group.

Ridout, R., & D.W. Clarke. (1970). A Reference Book of English. London: Macmillan.

Seaton, B. (1982). A Handbook of English Language Teaching Terms and Practice. London: Macmillan Press.

Sinclair, J. McH. (1966). Beginning the study of lexis. In C.E. Bazell, J.C. Catford, M.A.K. Halliday, & R.H. Robins (Eds.), In Memory of J.R.Firth (pp.410-430). London: Longman, Green and Co., Ltd.

Sinclair, J. McH. (1985). Lexicographic evidence. In R. Ilson (Ed.), Dictionaries, lexicography and language learning (ELT documents 120). (pp.81-94). Oxford: Pergamon Press.

Sinclair, J. McH. (Ed. in chief). (1986). Collins Birmingham University International Language Database English Language Dictionary. London: Collins.

Turner, G.W. (1973). Stylistics. Harmondsworth: Penguin.

Urdang, L. (1979). Meaning: denotative, connotative, allusive. In R.R.K. Hartmann (Ed.), Dictonaries and Their Users (pp.47-52). Exeter: University of Exeter.

Weinreich, U. (1980). On Semantics. (Philadelphia): University of Pennsylvania Press.