

- MILROY, J. (1996) : *Gender factor as a factor in linguistic change*. Paper presented at a seminar at the University of London.
- MOUGEON, R. & BENIAK, E. (1991) : *Linguistic Consequences of Language Contact and Restriction. The case of French in Ontario, Canada*. Oxford : Clarendon Press.
- PHILIPS, S.U. & REYNOLDS, A. (1987) : "The interaction of variable syntax : and discourse structure in women's and men's speech", in : *Language, gender, and sex in comparative perspective*. S.U. Philips, S. Steele & C. Tanz (eds.), Cambridge : Cambridge University Press, 71-94.
- SANKOFF, G. (1974) : "A quantitative paradigm for the study of communicative competence", in : *Explorations in the ethnography of speaking*, R. Bauman & J. Scherzer (eds.), Cambridge : Cambridge University Press, 18-49.
- SHAYWITZ, B.A.; SHAYWITZ, S.E.; PUGH, K.R.; CONSTABLE, R.T.; SKUDLARSKI, P.; FULBRIGHT, R.K.; BRONEN, R.A.; FLETCHER, J.M.; SHANKWEILER, D.P.; KATZ, L. & GORE, J.C. (1995) : "Sex Differences in the Functional Organization of the Brain for Language", *Nature*, 373, 6515, 607-609.
- SILVERMAN, I. & EALS, M. (1992) : "Sex Differences in Spatial Abilities : Evolutionary Theory and Data", in : *The Adapted Mind, Evolutionary Psychology and the Generation of Culture*, J.H. Barkow; L. Cosmides & J. Tooby (eds.), New York-Oxford : Oxford University Press, 533-549.
- SMITH, P.M. (1979) : "Sex markers in speech", in : *Social markers in speech*, K.R. Scherer & H. Giles (eds.), Cambridge-Paris : Cambridge University Press & Editions de la Maison des sciences de l'Homme, 109-146.
- SWANN, J. (1992) : *Girls, boys and language*, Oxford : Blackwell.
- TANNEN, D. (1992) : *You just don't understand. Women and Men in Conversation*. London : Virago Press.
- TANNEN, D. (1993) : *Gender and Conversational Interaction*. Oxford : Oxford University Press.
- THORNE, B.; KRAMARAE, C. & HENLEY, N. (1983) : *Language, gender and society*. Rowley MA : Newbury House.
- TRUDGILL, P. (1974) : *The social differentiation of English in Norwich*. Cambridge : Cambridge University Press.
- WESTMORELAND, R.; STARR, D.P.; SHELTON, K. & PASADEOS, Y. (1977) "News Writing Styles of Male and Female students", *Journalism Quarterly*, 54 : 599-601.

Measuring collocational knowledge : key issues and an experimental assessment procedure

Norbert SCHMITT
University of Nottingham

Introduction

In the years since FIRTH (1957, p. 194) brought the idea of collocation into prominence, there has been an increasing awareness of its importance in using words naturally and appropriately in discourse. With researchers benefiting from the availability of larger corpora, we now have a much better understanding of the collocational relationships between words. For example, there is general agreement that there are two basic kinds of collocation : grammatical/syntactic collocations and semantic/lexical collocations (BENSON, 1985; BAHNS 1993), although NEWMARK (1973, cited in BROWN, 1974) distinguishes seven different types. We know that collocates do not have to be adjacent to a node (target) word, which has led to a debate about how far away they can be.⁽¹⁾ NATTINGER and DECARRICO (1992, p. 22), for instance, believe that we must look more than five words away to find every collocational relationship. We also know that the collocates of a word are not random and unrelated, rather "words may habitually collocate with other words from a definable semantic set", which may carry either positive or negative connotations (STUBBS, 1995).

(1) See SINCLAIR (1966) and SINCLAIR, DALEY, and JONES (1970) for early attempts to gain statistical answers to this kind of question.

There has been a limited amount of non-corpus-based collocation research as well. Psycholinguistic studies have found that collocational links are "powerful and long-lasting" links between words in the mind for native-speakers (AITCHISON, 1987, p. 79). In association tasks, collocations are the second most common response type after coordinates (JENKINS, 1970). Aphasics tend to retain collocational information, even though they have lost much of their other use of language (GOODGLASS and BAKER, 1976). In addition, collocational links do not seem to fade as people grow older (HOWARD, MCANDREWS, and LASAGA, 1981). So collocations are more than relationships derived from corpus computations; to some extent the mind does seem to organize words according to their collocational links.

Nevertheless, collocations are quite complex; they are underutilized by advanced L2 learners (CHANNEL, 1981) and are often avoided by resorting to other lexical strategies like paraphrase (FARGHAL and OBIEDAT, 1995). BAHNS and ELDAW (1993) showed that the collocational knowledge of advanced German speakers of English lagged behind their general vocabulary knowledge.

Despite this psycholinguistic research, it must be said that our understanding of collocation is still almost entirely descriptive, stemming from the corpus style of research. As a result, even experts working primarily in this area have little idea of how to best test collocational knowledge. CHANNEL'S research (1981) required subjects to judge collocations given in a collocational grid, but this was a task of receptive knowledge only. Similarly, the translation and cloze tasks used by BAHNS and ELDAW (1993) also elicited receptive knowledge. While it is probable that receptive collocational knowledge aids reading, it seems intuitive that the most important role for collocation knowledge is in the production of language. Because the ability to use suitable collocates is a crucial skill in effective writing, a measurement of this skill would surely contribute to a more accurate evaluation of composition ability, as well as an improved measurement and description of lexis itself. Despite the apparent utility of such a measure, the author has been unable to uncover any previous attempt to measure productive collocational knowledge, either in the literature or from talks with colleagues. This paper attempts to stimulate interest in this area by discussing some of the issues

involved in collocation measurement, and by offering an experimental procedure for measuring productive collocational knowledge.

Preliminary Issues of Collocation

defining collocation

Before collocation can be measured, it must first be defined in such a way that the definition can be operationalized into some form of practical assessment procedure. Traditional definitions, typified by MCCARTHY (1990, p. 158), offer only a very broad notion of what collocation is :

collocation : the likelihood of co-occurrence between words. It is very likely that 'blonde' will occur with 'hair', but unlikely that it will occur with 'wallpaper'; 'blonde' and 'hair' are said to *collocate*.

Such definitions can be useful for students and others interested in a general statement about collocation, but they are not nearly rigorous enough to provide clear guidance in the development of a measurement procedure. They leave unanswered important questions such as what the necessary degree of likelihood is before words are said to collocate.⁽²⁾

A more precise definition needs to incorporate, at minimum, two major factors contributing to collocation. The first is the degree to which words *habitually* co-occur together. Almost any two words may occur together by chance, but this does not make them regular partners, as the following extract illustrates.

At a time when research in language acquisition is more versatile and widespread than ever before, it is hard to bring oneself to the brink of publication. (ENTWISLE, 1966, p. vii)

(2) There have been more detailed descriptions of collocation (eg. COWIE, 1978; SINCLAIR, 1991; MOON, 1997), but none have succeeded in overcoming this lack of explicitness.

Evidence from the Collins-COBUILD Bank of English Corpus confirms *language* and *acquisition* are frequent collocates. Although not listed among *language's* most frequent collocates, *research* must surely also be a collocate, at least in Applied Linguistics texts. But *brink* and *publication*, although they fit together and make perfect sense in this sentence, would not be expected to occur together very often at all. Indeed, there is only one instance of *brink of publication* in the entire 320 million word corpus. Thus co-occurrence is a continuum ranging from words which constantly occur together all the way down to those which may appear together only once in millions of words. On the assumption that these very infrequent partnerships do not constitute collocation, the problem lies in where to place the collocation cut-point on the frequency continuum.

Frequency alone is not sufficient to describe collocation however. The article *the* occurs very frequently with a huge number of nouns, but we wouldn't think of it as a collocate. Therefore, the second factor that needs to be taken into consideration is the *exclusiveness* of a word combination. *Hair* collocates with many words, such as *nice*, *long*, and *straight*, but its strength of collocation with *blonde* is said to be strong since *blonde* occurs with virtually no other word except *hair*. Thus the fewer the words a target word habitually occurs together with, the stronger the collocational relationship. So frequency and exclusiveness interact to determine which words collocate and how strongly.

The problem is that there is currently no research which would indicate numerical parameters for these factors. Thus we are still left with a vague description of collocation. This lack of a definition which can be objectively quantified makes it difficult to determine whether any word combination is a collocation or not, especially if it is a less frequent one. At some point, combinations must be considered chance occurrences rather than collocations.

elicitation options

To measure a subject's collocational knowledge, one of the first issues is how to elicit a demonstration of that knowledge. Following SCHOLFIELD'S (1995)

assertion that elicitation tasks should be as close as possible to 'real-life' activities, it seems desirable to elicit collocations embedded in discourse rather than in isolation. Creating sentences and larger discourse is an everyday language activity, while explicitly thinking of collocations as separate from discourse is not. This line of thought would eliminate elicitation formats such as the following :

What words frequently occur together with the word *massive*?
Write them on the line below.

If a productive measure of collocation is desired, a more natural and realistic task than that above is asking subjects to create complete sentences.⁽³⁾ Since the assessor will normally be looking at subjects' collocational knowledge for specific target words, having subjects produce sentences with those target words embedded can be a reasonably realistic elicitation method. The sentences can then be analyzed for how well the other words in the sentences collocate with the target words. The instructions given to subjects would be something like, "Please create a sentence which includes the word *massive* in it."

norming data for collocations

Once data has been elicited from the subjects, there must be some criteria as to what constitutes a collocation, even if the guiding definition is somewhat vague, as seen above. There seems to be two main ways to set this norming criteria. One is to consult large corpora and actually count the co-occurrences, and the other is to obtain native-speaker intuitions of how typical or natural collocations are. Native-speaker judgements may seem attractive from a naturalistic standpoint, but they suffer from a number of drawbacks. Chief among these is the logistical problem of arranging to have a number of

(3) L1 → L2 translation is another possibility, but translating from a number of different L1s makes it difficult to assume equivalence.

native-speakers judge the various collocational candidates that a subject may produce. Also, the judgements must come after the word combinations are given by the subjects and so it is difficult to have a principled *a priori* discussion of possible combinations. Finally, native-speakers may disagree on the typicality of different word combinations. Even if interrater reliability is high, there will always be some discrepancies which are difficult to reconcile.

If one derives norming criteria from corpus data, there are none of these disadvantages. First, as long as one has access to a large corpus⁽⁴⁾ and a standard computer, a single person can extract large amounts of norming information from a corpus. Second, computer programs can quickly provide figures which take account of both the frequency and exclusiveness of co-occurrence within a corpus. Third, the information obtained will have a degree of objectiveness. (It is important to note that information derived from corpora is not absolutely objective, since decisions must be made about issues such as where to place cut-off points in terms of the above figures, whether to work with individual word forms or lemmas [see below], and how balanced/representative any particular corpus is. Nevertheless, once these decisions have been made, there is a fixed standard to refer to.) Perhaps most importantly, a norming list of the most frequent collocates can be generated from a corpus, providing a convenient baseline from which to judge subject language samples. These norming lists can be used even before the first subject is tested, both to predict possible answers and to inform the design of a study/test (see the development of sentence prompts below).

While the above methods should have no problem accommodating common collocations, rare combinations, such as *brink of publication* above, are still likely to be problematic. Groups of native-speaking judges, even if they are trained, are unlikely to agree on these infrequent cases, whereas rare combinations may simply not occur in a corpus. The problem can be minimized with corpora by using very large ones, since this improves the chances of any particular combination being adequately represented. Corpus data may

(4) Larger corpora are becoming increasingly more accessible for general use. For example, Collins-COBUILD offers an online access facility to a 50 million word sample of their main corpus via CobuildDirect. Interested persons can get further information at : direct@cobuild.collins.co.uk

have the limitation of not being able to capture combinations at the very bottom end of the frequency continuum, but this should not matter, because rare combinations like *brink of publication* are clearly not typical and thus do not need to be catered for in testing learners.

From this discussion, it appears that corpus evidence is the most appropriate way to set norming criteria. We do have to be slightly cautious about what this evidence can tell us however. We have seen that due to an inadequate definition it is almost impossible at present to determine in absolute terms whether a word combination forms a collocation or not. Any measurement procedure is probably on much firmer ground if we interpret corpus evidence with the more restricted aim of determining whether a word combination is typical or not.

A Study into Collocation Measurement

Previous research gives less than adequate guidance when it comes to a number of other issues involved in collocation measurement. For example, how close must a word be to the target word in order to be a collocate? It has been assumed that words in closer proximity tend to collocate more strongly than those further apart. SINCLAIR (1991) suggests that the usual span employed in collocation research has been about $\pm 4 \sim \pm 5$ words, but this span seems to be based more on experience than any principled rationale. Another issue is whether there should be an exact match in word form between subject sample and norm criteria, or whether inflected and derivative forms should be accepted. In order to obtain at least initial answers to questions like these as a prerequisite to developing a productive collocation measurement procedure, the following study was undertaken.

Six target words were extracted from a practice TOEFL test (1995) : *massive, peak, rare, surging, subtle, and trend*. These words were selected (and this measurement procedure developed) as part of a larger study into second language vocabulary acquisition (SCHMITT, 1997, in press). It was important to create the collocation norming lists for these target words from a large modern corpus, so the 319-million word (as of July 25, 1996) COBUILD

Bank of English Corpus was chosen.⁽⁵⁾

The computer program in use at COBUILD provides two measures of collocational strength. The first is a *Mutual Information* (MI) score. For an MI score, the computer calculates how often a combination of words in the corpus occurs compared to how often the words appear separately. This score is good for highlighting words which may be infrequent in the corpus overall, but which collocate often with a target word when they do occur. The *T*-score is similar, but factors in a standard deviation measure to gauge the strength of collocation. This results in a better indication of the collocates which most frequently occur in discourse. (For more on the technical details and formulas involved, see CHURCH, GALE, HANKS, & HINDLE, 1991; STUBBS, 1995). An example may clarify this. The MI method gives *downward* and *inegalitarian* as the most significant collocates immediately to the left of the target word *trend*. The *T*-score results give *downward* as the third most significant collocate in the same position, but does not show *inegalitarian* in the top 49 collocates. Thus, although *inegalitarian* collocates with *trend*, this combination does not occur frequently enough to be in common usage. The combination *downward trend*, on the other hand, is both collocationally exclusive and frequently used.

One MI list and one *T*-score list, both with a span of ± 4 words, were generated for each of the target words (4 words on either side of the target word). For each collocate position (-4,-3,-2,-1,+1,+2,+3, and +4), the 50 most frequent collocates were listed, once for either calculation method used (MI or *T*-score). The maximum length of the columns that would fit on the computer screen was 50 items, which accounts for the number of collocates in each column. Considering that the 50 items multiplied by eight columns gave 400 possibilities, this was considered sufficient to build a norming baseline. Also, if intuition is correct that collocates in closer proximity tend to be stronger than those with wider separations, limiting the span to ± 4 would capture the stronger collocates of the target words.

Having decided on a sentence elicitation task, it was thought desirable to find a way to constrain the range of collocations elicited, while at the

(5) My thanks to Gwyneth Fox and Rosamund Moon for access to the Bank of English Corpus.

same time allowing subjects a reasonable amount of freedom in creating sentences. If the subjects were allowed to produce any sentence, it might be on any subject and would be more likely to include the uncommon, but still acceptable, kind of combination this procedure is not designed to deal with. An additional problem with a completely unrestricted task is that many people find it difficult to compose sentences without some parameters to guide them. The tendency for collocates to cluster within certain semantic fields (CARTER, 1987, Chapter 3; STUBBS, 1995) allowed a principled way of restricting subjects' collocations. For example, for the word *massive*, collocates on the norming list included :

attack, damage, destruction, died, explosion, injuries, launched, military, refugees

amount, billion, budget, companies, debts, deficient, development, dollar, economic, expansion, financial, investment

cause, changes, increase, influx, reduce, rises, turned

These collocates refer to areas which can be roughly described as war, economics or finance, and change respectively. I decided to require subjects to give sentences using these semantic fields as sentence topics, hoping to elicit collocates from the semantic field clusters already identified. Prompts were given to indicate the semantic field the subjects were to create sentences about. The prompts for the word *massive* and the above three semantic fields were :

1. If you were talking about war
2. If you were talking about finance or the economy
3. If you were talking about statistics

The measurement procedure in this study required the subjects to produce three sentences per target word, but this actually raises an interesting issue : how many demonstrations of collocational ability are sufficient to prove the existence of underlying knowledge ? A word has different collocates depending on its meaning sense (*rare antique; rare steak*), so a separate demonstration seems necessary for each meaning sense. But even within a meaning

sense, any target word collocates with numerous other words, as seen with *massive*. It was decided to elicit three sentences, since every target word had collocates which clustered into at least three fields. It was felt that asking for three sentences would give some indication of a subject's depth of collocational knowledge, although there was no assumption that this is the optimum number of sentences.

The prompts were critical in this task, in that they needed to guide the subject without giving any information away. So far as was possible, the prompts were written so that they did not include any of the collocates on the norming lists. Thus in the third semantic set above, *change* would probably have been the best prompt, but it was thought too important a collocate, so *statistics* was used in an attempt to capture the same idea. The prompts were also written so as not to give away any meaning information about the target word. The subjects were told they did not have to use words from the prompts in their sentences, but if they did, they were not penalized. The final prompts, as they appeared after piloting and revision, are given in the Appendix.

Thirty international students, who were studying at summer pre-sessional courses before they entered universities in Britain, volunteered to participate. They were interviewed individually on the six target words, and were asked to produce three sentences if possible, one for each of the relevant prompts. This would have returned 540 sentences (6 words \times 3 sentences each \times 30 subjects) if every subject produced 18 sentences, but in many cases the subjects did not know all of the target words. This is largely because the target words extracted from TOEFL tests were not of particularly high frequency. In particular, many subjects did not know *subtle* and *surging*. The net result was that 414 sentences were produced, which was 77% of the possible cases (414/540).

Discussion

This large number of sentences provided a considerable data base from which to develop a formulation of the best criteria for matching a subject's sentence with the norming lists in order to most effectively measure colloca-

tional knowledge. Some conclusions about methodological issues are given below.

i) It became clear when checking the sentences that neither norming list had enough internal diversity of collocates to capture all of the collocates which my intuitions said were allowable. It seems necessary to use both norming lists (MI and T-score) in conjunction to provide the broadest and best norming standard. Since the collocates on the two lists are substantially different, this gives subjects the greatest opportunity for having words in their sentences match those on the list.

ii) In many of the sentences, the only matched collocates were function words. This means that the presence of a grammatical collocate did not imply the presence of a lexical (content word) collocate. Furthermore, some subjects guessed and produced sentences even though they had no idea what the target word meant (subjects were also interviewed for semantic knowledge of the target words). These guessed sentences contained a content word collocate 34% of the time (17/50) but almost always contained function word collocates which appeared on the T-score norming list. Since function words are so ubiquitous, a number appear on every T-score norming list, and subjects usually included at least one of these in the process of composing a grammatical sentence. Measuring grammatical collocations does not allow us to discriminate between subjects very well, therefore it seems best to concentrate on lexical collocations by excluding function words from the norming lists, leaving only content words as scoreable collocates. It is also probably advisable to disregard other words which do not carry a 'full' content load, such as delexicalized verbs (*take, do*), general nouns like *thing* and *stuff*, modal verbs (*can, will*), and content words which could appear in virtually any sentence, such as temporal words (*now, yesterday*).

iii) Typical collocations should correspond to exact lemma forms, unless corpus evidence indicates that several lemma forms are collocates. For example, *losses* appears on the collocate list for *massive*, but *loss* does not. Likewise, *start* and *started* appear for *trend*, but not *starting* or *starts*. On the other hand, the corpus shows that *breed, breeding, and breeds* all collocate with *rare*. The problem in applying this strictly lies in a weakness of non-

native performance — learners often do not have full control of inflectional suffixes (SCHMITT and MEARA, 1997) and frequently omit them. Considering the difficulty in determining whether an incorrectly-inflected lemma is a competence error or a performance mistake, it is probably necessary to accept any form of a lemma at this point. In the future, it may be possible to develop a grading system giving a higher score for exact lemma form matches and a lower score for inexact matches.

However, often the inflected form of words in English also represent multiple word classes. For example *used* is the past form of the verb *use*, but it is also an adjective. As an adjective, it may strongly collocate with words it would not as a verb, i.e. "He sold used cars" versus "He used cars to deliver pizza".⁽⁶⁾ But this problem can almost be eliminated by avoiding the use of such words as target words, i.e. select *use* or *useful* as a target word, but not *used*.

The various derivatives of a word may take quite different collocates. The most frequent T-score collocates of *massive* in the +1 position are *attack*, *scale*, *increase* and *amounts*. These seem intuitive, but it is not so easy to think of a sentence which would combine these words, in any position, with the nouns *mass* or *massiveness*. Translating these insights into scoring criteria, if subjects produce a word in any inflected form in their sentences which matches with a word on the norming lists (including irregular past forms like *find* → *found*), they should be given credit. However, derivative forms must exist separately on the norming lists to be counted. For example, if the norming lists include the collocate *produce*, subjects should gain credit for sentence words *produce*, *produced*, *producing*, *produces*, but not for the words *production* or *producer*, unless they are separately included on the norming lists.

iv) Words which are included in the prompt should not be scoreable as collocates, for the simple reason that a test of productive knowledge cannot allow answers which might have been assisted by information previously given on the test.

(6) I am grateful to Jeremy Clear for bringing this to my attention.

v) The different sentences should include different collocates to earn a score. If this were not the case, a single collocate could be counted as many as three times. The purpose of using three sentences is to capture in some part the range of a subject's collocational knowledge; allowing multiple scores for a single collocate would defeat this purpose.

vi) Since this procedure is not measuring spelling ability, spelling mistakes should be ignored as long as the intended word is clear. In contrast, it is not so obvious how to handle grammatical mistakes. They come into play mainly when collocates are placed in odd positions relative to the target word. For instance, is *massive money* acceptable (because *money* is on the norming lists), even though 1) a more natural and grammatical structuring would be *massive amount of money*, 2) it occurs only nine times in the Bank of English corpus? The expedient of disregarding collocational position has obvious advantages and disadvantages. The major advantage is that having to make fine judgements about grammaticality or naturalness can be avoided; all that is required is checking to see if any words on the norming lists exist in the sentence in question. The disadvantage is that it limits the inference one can make about the subject's collocational knowledge. One can only say that the subject knows a common collocate for a target word; one cannot say that the subject can use that collocate in a sentence in a natural and appropriate way. A possible solution is to develop separate collocation norming lists for each position in relation to the target word (i.e. have a +1 list, a +2 list etc.), but this would be far more complex and time consuming. In this study, collocational position was disregarded.

vii) If we are to use elicited sentences as evidence of collocation knowledge, we cannot control for the number of collocations produced per sentence. No previous research gives guidance as to how many collocates we should expect in a sentence, but the thirty subjects in this study produced the numbers indicated in Table 1.

Table 1 : Number of Collocates Produced Per Sentence N=540

	Collocates Produced						Sentence Not Produced
	12	3	4	5	0		
Number of Sentences	153	68	19	12	1	161	126
% of produced sentences (n/414)	37	16	5	3	0	39	

We can see that requiring two or more collocates per sentence would mean that only about 24% of the sentences would be scored positively. This excludes a large number of sentences which seem perfectly appropriate in a collocational sense, but have only one collocate match, such as the following examples from the data :

- a) The **peak** season for sale of fruit is in April.
- b) The **rare** animals are protected by the government in many countries.
- c) The investment in this country is **massive**.
- d) The view from the **peak** was beautiful.

***Target words** Collocates

These sentences are the kind that native-speakers produce in everyday communication, albeit with slightly better syntax. So it seems that a required minimum of one collocation per sentence is the best criteria we can set from the present evidence.

viii) Another key issue is how broad a span to use in collocation measurement. A native-speaker judgement task was used to find a workable solution.

First, all words in sentences which appeared on the norming lists, according to the above criteria, were marked. Then the author judged whether the marked words seemed appropriate and natural collocates of the target word. Finally, the words were tallied according to their position in the sentences in respect to the target word. This was obviously a very subjective procedure, but it proved useful in supplying at least some indication of an appropriate span length. The results are given in Table 2.

Table 2 : Confirmatory Judgement of Collocates Indicated by the Norming Lists (Number of Words) N=385

	Proximity to Target Word									
	±1	±2	±3	±4	±5	±6	±7	±8	±9	±10 >±10
Judged as Collocates	101	65	55	28	25	15	11	6	5	1
Not Judged as Collocates	7	9	16	14	9	7	5	3	1	2

Although the collocates on the norming lists came from a span of ±4, the table shows that many of these collocates occur outside this span in the subjects' sentences. As NATTINGER and DECARRICO (1992, p. 22) believed, collocates can appear a considerable distance from the node word. There seems no reason to score the collocates on the norming list which occur further afield in the sentences as uniformly unacceptable. So although the vast majority of matching collocates (80%) occurred within a proximity of ±4 within the subjects' sentences, it may well be that the best course is to accept a word matching the norming list, regardless of where it occurs in the sentence, in order to capture the 20% which occur outside the ±4 span.

Unfortunately, there is always the danger that these 'distant' collocations are random or accidental. But that can also happen in closer proximity. If fact, the whole notion of using proximity as a criterion may be flawed. The following example illustrates this :

The investment in this country is **massive**.

Whereas the norm list collocate *investment* is clearly connected to *massive*, it lies five words away. However, the closer norm list collocate *country*, as part of a prepositional phrase, does not seem to be directly related, because the investment could be in virtually anything. This suggests that syntactic structure may also have a part to play. It might be worth exploring whether syntactic concepts are useful in more precisely defining the boundaries of collocation.

It is interesting to note that the collocates identified by the norming list procedure outlined in this paper were largely confirmed by the judgement task. The author's intuitions indicated that 81% of the collocates identified were indeed natural and appropriate; only 19% were considered misidentified as collocates. Admittedly, these results are only from one rater (myself), but they suggest that this procedure (using norm lists alone as criteria) is able to achieve a reasonable identification of collocates in the subject sentences.

ix) Using this procedure, scores can range from 0 to 3. A subject with a 0 score may know the target word's meaning(s), but has not shown the ability to use the word with its common collocates in sentences. On the other hand, a score of 3 indicates the subject probably has a good sense of which words collocate with the target word, even though there may still be problems in using the collocates appropriately. This 0-3 scale is admittedly a rather crude measure, but at least it is a first step in giving some indication of collocational knowledge.

Limitations of the Procedure

Since this is one of the first attempts to measure productive collocational knowledge, the procedure proposed inevitably suffers from weaknesses. It gives some evidence of a subject's knowledge of the most common collocates for a target word, but is not designed to make statements about less significant collocates. There is also some element of chance; subjects creating sentences may accidentally include a norm list collocate even if they have no

collocational knowledge of the target word. Future measures of collocation will need to develop a principled way to control for this.

Another weakness is indicated by the fact that subjects were able to include a content word collocate in 34% of the cases in which they did not know the meaning of the target word. There is a positive way of looking at this however. It is possible that learners can remember word combinations without knowing the meaning of the constituents, that is, perhaps collocation can be learned without semantics. If so, this procedure can tap that collocational knowledge, even if semantic knowledge is deficient.

Conclusion

This experimental procedure obviously suffers from the lack of a precise definition of what collocation is. Nevertheless, discussion is always easier with something tangible to evaluate and constructively criticise, and this procedure at least provides that. It has already proved informative in one vocabulary acquisition study (SCHMITT, 1997, in press), although there are indications that the criteria as stated are too generous and need to be made stricter. It may be worth exploring whether this could be done by either shortening the span or requiring more exact positioning of the collocates.

It is hoped that this paper will inspire readers to try their hand at developing better measurement procedures in the future. As long as vocabulary and composition skills are considered important, a valid way to measure collocation can only be a valuable addition to the field.

Acknowledgement

Ramesh Krishnamurthy kindly provided insightful comments on an earlier draft of this paper, and I gratefully acknowledge that in several places the final version owes a debt to his thinking on the subject.

References

- AITCHISON, J. (1987) : *Words in the Mind*. Oxford : Blackwell.
- BAHNS, J. (1993) : *Lexical Collocations : A Contrastive View*. *ELT Journal*, 47, 56-63.
- BAHNS, J. and ELDAW, M. (1993) : *Should We Teach EFL Students Collocations? System*, 21, 101-114.
- BENSON, M. (1985) : *Collocations and Idioms*. In : ILSON, R. (ed.), *Dictionaries, Lexicography and Language Learning*. Oxford : Pergamon.
- BROWN, D. (1974) : *Advanced Vocabulary Teaching : The Problem of Collocation*. *RELC Journal*, 5, 1-11.
- CARTER, R. (1987) : *Vocabulary : Applied Linguistic Perspectives*. London : Unwin Hyman.
- CHANNELL, J. (1981) : *Applying Semantic Theory to Vocabulary Teaching*. *ELT Journal*, 35, 115-122.
- CHURCH, K. GALE W. HANKS P. and HINDLE, D. (1991) : *Using Statistics in Lexical Analysis*. In ZERNIK, U. (ed.), *Using On-line Resources to Build a Lexicon*. New Jersey : Lawrence Erlbaum.
- COWIE, A.P. (1978) : *Vocabulary Exercises Within an Individualized Study Programme*. *ELT Documents*, 103, 37-44.
- ENTWISLE, D.R. (1966) : *Word Associations of Young Children*. Maryland : Johns Hopkins.
- FARGHAL, M. and OBIEDAT, H. (1995) : *Collocations : A Neglected Variable in EFL*. *IRAL*, 33, 315-331.
- FIRTH, J.R. (1957) : *Papers in Linguistics 1934-1951*. London : Oxford University Press.
- GOODGLASS, H. and BAKER, E. (1976). *Semantic Field, Naming and Auditory Comprehension in Aphasia*. *Brain and Language*, 3, 359-374.
- HOWARD, D.V. MCANDREWS M.P. and LASAGA, M.I. (1981) : *Semantic Priming of Lexical Decisions in Young and Old Adults*. *Journal of*

- Gerontology*, 36, 707-714.
- JENKINS, J.J. (1970) : *The 1952 Minnesota Word Association Norms*. In : POSTMAN, L. and KEPPEL, G. *Norms of Word Associations*. New York : Academic Press.
- MCCARTHY, M. (1990) : *Vocabulary*. Oxford : Oxford University Press.
- MOON, R. (1997) : *Vocabulary Connections : Multi-word Items in English*. In : Schmitt, N. and McCarthy, M. (eds.) *Vocabulary : Description, Acquisition, and Pedagogy*. Cambridge : Cambridge University Press.
- NATTINGER, J.R. and DECARRICO, J.S. (1992) : *Lexical Phrases and Language Teaching*. Oxford : Oxford University Press.
- NEWMARK, P. (1973) : *Twenty-three Restricted Rules of Translation*. *The Incorporated Linguist*, January, 1993.
- SCHMITT, N. (in press) : *Tracking the Incremental Acquisition of Second Language Vocabulary : A Longitudinal Study*. *Language Learning*.
- SCHMITT, N. (1997) : *Using a Word Knowledge Framework to Research Vocabulary*. Unpublished doctoral dissertation, University of Nottingham.
- SCHMITT, N. and MEARA, P. (1997) : *Researching Vocabulary through a Word Knowledge Framework : Word Associations and Verbal Suffixes*. *Studies in Second Language Acquisition*, 19, 17-35.
- SCHOLFIELD, P. (1995) : *Quantifying Language*. Clevedon, England : Multilingual Matters.
- SINCLAIR, J.M. (1966) : *Beginning the Study of Lexis*. In : BAZELL, C. CATFORD J.C., HALLIDAY M.A.K. and ROBINS, R.H. (eds.), *In Memory of J.R. Firth*. London : Longman.
- SINCLAIR, J.M. (1991) : *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.
- SINCLAIR, J.M., DALEY R. and JONES, S. (1970) : *English Lexical Studies*, Report no. 5060. London : OSTI.
- STUBBS, M. (1995) : *Collocations and Semantic Profiles : On the Cause of the Trouble with Quantitative Studies*. *Functions of Language*, 2, 1-33.
- TOEFL Practice Tests. (1995) : New Jersey : Educational Testing Service.

Appendix

massive

1. If you were talking about war
(attack, destruction, explosion, launched, military)
2. If you were talking about finance
(amount, budget, debts, deficit, economic, investment)
3. If you were talking about statistics
(changes, increase, influx, reduce, rises)

peak

1. If you were talking about a business
(career, demand, levels, season)
2. If you were talking about a house
(roof)
3. If you were talking about geography
(Himalayan, mountain, snow, top)

rare

1. If you were talking about living things
(animals, beasts, breeds, butterflies, species)
2. If you were talking about cooking
(beef, steak)
3. If you were talking about a special person/entertainer
(appearance, gift, talent)

subtle

1. If you were talking about food
(aroma, flavour, sweet, taste)
2. If you were talking about communication between people
(approach, convey, hint, message, nuances, perceive)
3. If you were talking about a painting, like a Monet
(colors, hues, lighting, shades, technique, variation)

surging

1. If you were talking about the natural world
(floodwater, water, tide, waves)
2. If you were talking about business, finance, or economics
(costs, development, exports, growth, inflation)
3. If you were talking about people at a big sports or entertainment event
(crowd, emotions, mob)

trend

1. If you were talking about economics
(business, figures, inflation, market, prices, rates)
2. If you were talking about the clothing industry
(fad, fashion, setter)
3. Use any topic you like using trend as a noun. But you must include an adjective which describes the noun trend. (a(n) ____ trend)
(current, dangerous, disturbing, growing, increasing)