NB. This is the author's post-refereed version of a paper to be published in the *International Journal of Corpus Linguistics 19/4* in November 2014. Copyright belongs for John Benjamins Publishing. This paper was first submitted for publication October 2012.

Corpus frequency and second language learners' knowledge of collocations

A meta-analysis

Philip Durrant
University of Exeter

Tests of second language learners' knowledge of collocation have lacked a principled strategy for item selection, making claims about learners' knowledge beyond the particular collocations tested difficult to evaluate. Corpus frequency may offer a good basis for item selection, if a reliable relationship can be demonstrated between frequency and learner knowledge. However, such a relationship is difficult to establish satisfactorily, given the small number of items and narrow range of test-takers involved in any individual study. In this study, a meta-analysis is used to determine the correlation between learner knowledge and frequency data across nineteen previously-reported tests. Frequency is shown to correlate moderately with knowledge, but the strength of this correlation varies widely across corpora. Strength of association measures (such as mutual information) do not to correlate with learner knowledge. These findings are discussed in terms of their implications for collocation testing and models of collocation learning.

Keywords: collocation, testing, frequency, formulaic language, vocabulary, SLA

1. Introduction

It has long been recognized that collocations are pervasive in language (Hoey 2005, Sinclair 2004), and that a healthy repertoire of collocations is essential to mastery of a

foreign language (e.g. Kjellmer 1990, Lewis 1993, Nattinger & DeCarrico 1992, Palmer 1933, Pawley & Syder 1983). Research on second language learners' knowledge and acquisition of collocations has gathered pace in recent years, with greater integration of corpus and psycholinguistic methods advancing our understanding and allowing ever more specific questions to be addressed (e.g. Durrant & Schmitt 2010; Webb & Kagimoto 2011; Wolter & Gyllstad 2011, 2013; Yamashita & Jiang 2010).

A major outstanding issue in this area is that of how learners' knowledge of collocations can be validly assessed While a number of recent studies have evaluated various test formats (Barfield 2003, Bonk 2001, Gyllstad 2007, Moreno Jaén 2009, Revier 2009), the key question of how collocations can be reliably sampled for inclusion as test items has not been addressed.

As with all 'selective' (Read 2000) vocabulary tests, collocation tests utilize samples of items which are small in proportion to the population from which they are drawn. Because we are usually interested not just in learners' knowledge of the particular items tested, but of collocations more generally, it is essential that items be selected in a principled way to allow inference beyond the sample.

In relation to single-word vocabulary, word frequency has been shown to correlate with likelihood of knowledge (Milton 2009) and it has become common practice to sample items according to this variable. Typically, words are grouped into frequency 'bands' and learners' performance on a sample of words is taken to reflect their knowledge of words in that band (Nation 1990). Since collocation can be seen as a type of vocabulary (in that collocations are linguistic items which need to be specifically learned, rather than being derivable from rules – see Section 2, below) and since models of collocation have claimed that L1 collocation learning is frequency-driven (e.g. Ellis 2001, Hoey 2005), it is tempting to extend this strategy to collocation tests. However, at least two considerations suggest that it would be unwise to do so without further evidence.

First, some researchers have doubted whether collocation learning is frequency-driven for second language learners. Wray (2002), for example, has claimed that adult L2 learners tend not to notice and remember the collocations they encounter. This view suggests that collocation learning is usually the result of explicit memorization of selected forms, rather than exposure, and so implies that collocation knowledge may not be sensitive to frequency.

A second reason to question the frequency-knowledge link for collocations lies in the nature of the corpus data on which frequency counts are based. The logic behind

using frequency to predict knowledge is that the more frequent an item is, the more likely learners are to have met it repeatedly. However, few corpora are likely to be representative of the language which any individual learner has encountered (Durrant & Doherty 2010). Corpora are generally designed to represent, not individuals' experiences, but rather particular types of discourse. A further problem is introduced by limitations in the ways that frequency counts are conducted. In particular, in the absence of fine-grained semantic tagging, counts do not distinguish different senses of polysemous words. Since it is likely that language learners do make such distinctions, this constitutes a further distortion of their experience of the language.

Studies have shown that corpus-based frequency counts are a reasonable guide to learners' knowledge of relatively frequent words (especially for the 4,000 most frequent words) (Milton 2009). However, the correlation weakens considerably at lower frequencies (Milton 2009). This is probably because, whereas frequent words tend to be frequent across a wide range of situations, lower frequency words are usually associated with particular contexts, and their frequency therefore tends to vary between corpora. For such words, the assumption of a correlation between frequency in a particular corpus and frequency in a given learner's experience is dubious. This raises problems for collocations because individual items tend to be relatively infrequent. Shin & Nation (2008), for example, find that only 891 collocation have frequencies similar to those of the 4,000 most frequent single words, where Milton (2009) finds frequency to be a reliable predictor.

While the factors discussed above suggest that the relationship between corpus frequency and L2 knowledge of collocations may not be entirely straightforward, a number of recent studies have suggested that some relationship does exist. Durrant & Schmitt (2010) find that – contrary to Wray's (2002) claims – adult second language learners do retain memories of which words appear together in the language they meet, and that greater repetition leads to greater retention. Similarly, both Siyanova-Chanturia et al. (2011) and Wolter & Gyllstad (2013) find that adult L2 users' speed of processing English collocations was affected by collocation frequency.

While these studies suggest that frequency is related to L2 knowledge, at least two caveats should be noted. Firstly, these studies were conducted with a relatively narrow range of high proficiency learners. Durrant & Schmitt (2010) and Siyanova-Chanturia et al.'s (2011) participants were all students at a single British university, while Wolter & Gyllstad's (2013) participants showed an impressive mean vocabulary size of 7,350 words, putting them above even the 3,750-5,000 words which are associated with the

highest levels (C1 and C2) of the Common European Framework (Wolter & Gyllstad 2013). Whether similar effects will be seen for learners below these high levels remains an open question.

Secondly, these studies did not aim to measure knowledge of the type which is tapped in typical test formats, but rather efficiency of processing (Siyanova-Chanturia, et al. 2011, Wolter & Gyllstad 2013) or priming relationships between words (Durrant & Schmitt 2010). Further work is needed to determine whether the frequency effects which these studies show through eye-fixation durations, response times to decision tasks, or priming is also found in students' responses on standard test tasks.

In response to these issues, the present paper will investigate the extent to which learners' knowledge of collocations, as measured by typical test formats, is related to collocations' frequency in a corpus. It aims both to establish whether corpus frequency is a valid strategy for sampling collocation test items and to give guidance on which types of frequency information are most relevant to collocation sampling.

One way of studying the frequency-knowledge relationship would be to create a test including a set of collocations of different frequencies and to determine whether the number of students knowing each collocation is correlated with collocation frequency. However, any individual test administration would be limited by the inevitably small sample of collocations used, the testing method employed and any peculiarities of the test-takers. To gain a more robust data set, therefore, existing literature was reviewed to identify studies which report collocation tests. Frequency data were then retrieved for the collocations in these tests, and correlations between frequency and the percentage of students answering correctly were determined. Meta-analytic techniques were then used to determine overall correlations across all studies. As with all meta-analyses, the logic is that, if the effect we are seeking is robust across multiple studies despite the different types of error inherent in each, we can have a high degree of confidence that the effect is real (Cooper 1998).

2. Defining collocation

The term 'collocation' has been used by researchers from a variety of traditions and has been defined in a several different ways (see e.g. Barfield & Gyllstad 2009, Nesselhauf 2004). It is therefore important for any study of collocation to define its scope clearly. Durrant & Mathews-Aydinli (2010) describe three main orientations:

(i) 'Phraseological approaches' (e.g. Cowie 1998, Nesselhauf 2004) define collocations as word combinations in which one element does not carry its usual meaning (e.g. *take a step*, *explode a myth*) or in which there are restrictions on which words can enter a combination (e.g. *commit* can only be followed by a small number of nouns, related to wrongdoing, *shrug* combines almost exclusively with *shoulders*).

(ii) 'Frequency-based approaches' (e.g. Biber 2009, Hoey 1991, Sinclair, 1991) define collocations as sets of words which have a statistical tendency to co-occur in texts. These are likely to include collocations as defined in the phraseological approach (e.g. *shrug* is statistically highly likely to co-occur with *shoulders*) but also include combinations which do not exhibit semantic specialization or restriction (e.g. *next week*, *drink tea*).

(iii)'Psycholinguistic approaches' (e.g. Hoey 2005, Wray 2002) define collocations as combinations of words which have psychological reality in that they are stored holistically or there is an associative link between their elements. This has clear overlaps with the previous categories in that both semantic specialization/restriction and high frequency of occurrence are likely to imply some form of mental representation.

In spite of their differences, all three of these approaches share the idea that collocations are combinations whose behaviour cannot be fully explained in terms of features of their component words and therefore need to be handled as partially independent entities, with their own semantic/ distributional/psycholinguistic properties. For language learning purposes, this corresponds to the idea that collocations are combinations of words which need to be independently learned. This conception is captured well by Palmer's (1993) definition of collocations as:

> successions of words […] that (for various, different and overlapping reasons) […] must or should be learnt, or is best or most conveniently learnt as an integral whole or independent entity, rather than by the process of placing together their component parts. (Palmer 1933: 4)

To be of use, however, Palmer's (1993) definition needs to be developed in two ways. First, it is necessary to spell out the "various, different and overlapping reasons" (Palmer 1993: 4) why a succession of words might be best learnt as a whole. Three main reasons can be cited here:

(i) Semantic opacity: the collocation is semantically non-transparent. I.e. its meaning cannot be reliably predicted on the basis of a knowledge of the meaning its component parts have in other contexts. Examples include: *small talk* and *curry favour*. Without specific knowledge of the meanings of such collocations, a learner is unlikely to be able to understand or produce them accurately.

(ii) Received usage: particular collocations may become the conventional way of expressing a particular meaning, even though other phrasings are equally plausible. Examples include: *answer phone* and *slit throat*. Without specific knowledge of such pairings, a learner has a good chance of guessing the "wrong" combination and their language is likely to sound "inauthentic" (Pawley & Syder 1983).

(iii) Fluency: the combination occurs with such high frequency that learning it as an item is likely to promote fast and accurate (efficient) language processing. Possible examples include: *sunny day* and *salt and pepper*. Without knowledge of such collocations, a learner may not be able to achieve nativelike fluency (Pawley & Syder 1983).

Second, Palmer (1993) does not specify how many words collocations can have. His examples (e.g. *there is something the matter with you, to be difficult for someone to do something*) seem to indicate that he has no particular limit in mind. Some researchers in the frequency-based/psychological traditions have similarly called combinations of any number of words 'collocations' (Biber 2009, Hoey 2005, Kjellmer 1990, Sinclair 1991). However, corpus linguists often make a distinction between two-word combinations and longer sequences, with longer combinations commonly referred to by other terms, such as 'lexical bundle' (Biber et al. 1999, Ellis et al. 2008), 'n-gram' or 'concgram' (Cheng et al. 2006). The differences between these labels are important in corpus research because each involves a different search strategy. For example, whereas lexical

bundles are retrieved as fixed contiguous sequences of words, collocations are usually searched for as pairs of words frequently appearing within a certain distance of each other, so allowing greater flexibility regarding their relative positions. Similarly, the various measures which are used to quantify collocation frequency (reviewed below) can vary dramatically across combinations of different length, with frequency dropping and mutual information increasing sharply as the lengths of combinations increase.

For these reasons, frequency data about positionally-flexible two-word collocations are not strictly comparable with frequency data about other types of word combination. For studies, such as the present one, which make extensive use of such data, it is therefore important to maintain a distinction between combinations of different lengths. For this reason, the term 'collocation' will be used here to refer only to combinations of two words within a given span.

Taking these points into consideration, Palmer's (1933) formulation can be adapted to define collocations as:

> combinations of two words that are best learned as integral wholes or independent entities, rather than by the process of placing together their component parts, either because (i) they may not be understood or appropriately produced without specific knowledge, or (ii) they occur with sufficient frequency that their independent learning will facilitate fluency.

3. Material and methods

Meta-analysis is a technique for synthesizing existing research in order to clarify the relationships between the main variables and to understand the effects of moderating variables (e.g. Cooper 1998, Lipsey & Wilson 2000). Meta-analytic work in the field of language learning is usefully reviewed by Norris & Ortega (2006), who describe three main stages in a meta-analysis: sampling of relevant studies; coding of data relevant to each study; and analysis. The following sections will describe each of these stages in turn.

3.1 Sampling

The first step in the meta-analysis was a comprehensive search of the literature to identify relevant data. To ensure relevance to the research questions and comparability between studies, it is important at this stage to define clear criteria for inclusion in the review. In the present case, studies needed to include descriptions of selective tests of non-native speakers' knowledge of English collocations and provide information about the numbers of learners answering each test item correctly.

The first step in the review was to search five major databases using the search term: *collocation** AND (*test** OR *learn** OR *knowledge*). The databases searched were: (i) Web of Knowledge (topic search, refined to 'Arts Humanities' and 'Social Sciences'); (ii) ERIC (abstract search); (iii) Linguistics and Language Behavior Abstracts (abstract search); (iv) PsychInfo (abstract search); (v) PsycArticles (abstract search). Further, both Google and Google Scholar were searched using the search term *language collocation* test* learn* acquisition**. Because of the large number of (often irrelevant) hits returned by Google, only the first 250 results were used.

The abstracts of all retrieved items were checked to determine whether they included empirical studies which involved selective tests of non-native speakers' collocation knowledge in English. 35 such studies were identified.

The second step was to check the bibliographies of all relevant studies for further studies. Google Scholar was also checked for studies citing the retrieved works. Any publication whose title suggested it included some evaluation of learners' collocational knowledge was retrieved and checked to see if it met the inclusion criteria. This process was repeated recursively with all newly-identified studies.

The review was restricted to papers written in English and either freely available online or accessible through my institution's library. In a small number of cases, references from other sources suggested that a source which was written in another language or which was not freely available contained information of the type required. These sources (i.e. Jaen 2009, Barfield 2003) were obtained through direct contact with the authors or through my institution's library.

This process returned a total of 85 studies. Of these, very few included the information required for the meta-analysis. Only 46 studies recorded which collocations were included in their tests. Of these, 14 provided data on the number of students answering each item correctly. Four of these were excluded because test items did not have unique correct answers and so did not allow show whether learners knew specific target collocations; one was excluded because it focused on collocations from a

narrowly-defined area of discourse (Maritime English) for which available corpora were unlikely to provide valid frequency data.

Some of the nine remaining publications included more than one test and so provided multiple data sources. As discussed above, tests were only included if items had a single correct answer. This meant that, for example, Gyllstad's (2007) COLLMATCH tests and Jaen's (2009) test 3 were not included in the analysis.

This process provided a total of 19 different tests, summarized in Table 1. The tests were conducted by nine different researchers in eight different countries. Participant numbers ranged from 18 to 340, with a total of 1,568 distinct test takers.

Table 1. Tests included in the meta-analysis

Source	Participants	Items	Test format
Abdul-Fattah 2001	340 10th grade students at 10 different schools in Jordan	12 V + N 2 Adj + N	4-option selected response sentence completion; node given, collocate selected
Barfield 2003 (Chp 3)	93 students at a university in Japan (various departments)	99 V + N	4-point self-report knowledge scale; no context given
Brashi 2009	20 senior undergraduate English Language students at a university in Saudi Arabia	20 V + N	4-option selected response completion of sentence context; noun given, verb selected
Farghal & Obeidat 1995 (Test 1)	34 junior/senior English majors at a university in Jordan	7 Adj + N	Sentence completion
Farghal & Obeidat 1995 (Test 2)	23 senior English majors at a university of Jordan	15 Adj + N 2 N + N	Whole sentence translation from L1 (Arabic)
Gyllstad 2007 (COLLEX 1)	18 2nd year undergraduate ELT students at a university in Sweden	59 V + N	2-option selected response; noun given, verb selected; no context given
Gyllstad 2007 (COLLEX 2)	84 1st year undergraduate English Language students at a university in Sweden	48 V + N 12 Adj + N 2 N + V 1 Adv + Adj	2-option selected response; noun given, collocate selected; no context given
Gyllstad 2007 (COLLEX 3)	116 1st-2nd year undergraduate English Language students at a university in Sweden	38 V + N 8 Adj + N 2 Adv + Adj	2-option selected response; noun given, collocate selected; no context given
Gyllstad 2007 (COLLEX 4)	188 students in Sweden (26 10th grade high school; 28 11th grade high school; 134 1st year English language undergraduates)	38 V + N 8 Adj + N 2 Adv + Adj	2-option selected response; noun given, collocate selected; no context given
Gyllstad 2007 (COLLEX 5)	24 students in Sweden (7 11th grade high school; 17 1st year undergraduate English Language students)	38 x V + N	3-option selected response; noun given, verb selected; no context given
Jaén 2009 (Test 1)	311 undergraduate English Philology/English Translation and Interpretation students at three universities in Spain	22 Adj + N 13 V + N 6 N + N 1 N + V	C-test with sentence context; noun and first letter of collocate given
Jaén 2009 (Test 2)	311 undergraduate English Philology/English Translation and Interpretation students at three universities in Spain	16 Adj + N 11 V + N 1 N + N	Translation: L1 phrase and English node given; test takers supply collocate
Jaén 2009 (Test 4)	311 undergraduate English Philology/English Translation and	23 Adj + N 15 V + N	4-option selected response sentence completion. Node

	Interpretation students at three universities in Spain	6 N + N 1 N + V	given; collocate selected
Koya 2005 (Test B)	130 students at a university in Japan (various departments)	68 V + N	3-option selected response completion of sentence context; noun given, verb selected
Kurosaki 2012 (selected response - French)	34 French undergraduate students studying English part-time in Paris	16 V + N 7 Adj + N 5 Adv + Adj	4-option selected response sentence completion; node given, collocate selected
Kurosaki 2012 (selected response - Japanese)	30 3rd/4th year non-English major undergraduate students in Japan	16 V + N 7 Adj + N 5 Adv + Adj	4-option selected response sentence completion; node given, collocate selected
Kurosaki 2012 (translation - French)	29 French undergraduate students studying English part-time in Paris	13 V + N 8 Adj + Adj 5 Adj + N	Translation from L1 - target sentence provided with whole collocation removed
Kurosaki 2012 (translation - Japanese)	38 3rd/4th year non-English major undergraduate students in Japan	13 V + N 9 Adj + Adj 5 Adj + N	Translation from L1 - target sentence provided with whole collocation removed
Revier 2009	56 students in Denmark (20 10th grade high school; 17 11th grade high school; 19 1st year undergraduate)	19 V + N	3-option selected response completion of sentence contexts; each component of collocation selected separately

For various reasons, not all items on all tests were included in the present analysis. Specifically, items were omitted if they did not test collocations as defined in this study (e.g. if they included more than one word or included a non-lexical word) or if more than one answer was accepted by the researchers as correct. Table 1 shows the number and grammatical type of collocations remaining in each test. After adjustments, the tests comprised between 7 and 100 items each, with a total of 724 items across the 19 tests. There was some overlap between tests in the items used. For this reason, the total number of unique (lemmatized) collocations was lower, at 476. The majority of items were verb + noun combinations (349), followed by adjective + noun (99), noun + noun (15) and adverb + adjective (13).

A common problem with meta-analyses is that of publication bias – i.e. that studies tend only to get published if they achieve significant results. This means that meta-analyses which incorporate only published studies may inadvertently exclude contrary evidence. However, the present study is unusual amongst meta-analyses in that the main effect it studies (the relationship between frequency and knowledge) was not a focus on the original studies reviewed. There is therefore no reason to believe that the studies included will demonstrate a greater or lesser relationship between frequency and knowledge than would unpublished studies.

3.2 Coding

The second stage of the meta-analysis was that of coding studies for variables of interest. In this study, the main variables are the percentage of participants correctly answering each item and the frequency of each collocation. The former was provided by the original studies. The latter was retrieved directly from corpora. Because quantification of collocation frequency is a complex issue, involving a number of decisions, this will be described in detail below (Section 3.3).

As well as the main variables, studies need to be coded for any potential moderator variables that might be relevant to the analysis. Four such variables were identified in the current set of studies:

(i) Students' experience of studying English. Tests can be broadly divided into those in which the text-takers were full-time students on university programmes directly related to English language and those which were not (Gyllstad's (2007) COLLEX 4 and 5 drew on a mix of university and pre-university students and so will not be included in this analysis);

(ii) Students' L1. These can be divided into European languages (Danish, French, Spanish and Swedish), Arabic and Japanese;

(iii) Test task type. The main types used are selected response and translation. Three other task types (self-report, sentence-completion, and C-test) are combined under the category 'other';

(iv) Whether test-takers are asked to provide the whole collocation or only the collocate.

Table 2 shows how the 19 tests are categorized on each of these variables.

Table 2. Categorization of tests according to possible moderators

Source	English majors	L1	Task type	Whole collocation required
Abdul-Fattah	No	Arabic	Selected response	No
Barfield	No	Japanese	Other	Yes
Brashi	Yes	Arabic	Selected response	No
Farghal & Obeidat (Test 1)	Yes	Arabic	Other	No
Farghal & Obeidat (Test 2)	Yes	Arabic	Translation	Yes
Gyllstad (COLLEX 1)	Yes	European	Selected response	No

Gyllstad (COLLEX 2)	Yes	European	Selected response	No
Gyllstad (COLLEX 3)	Yes	European	Selected response	No
Gyllstad (COLLEX 4)	Mixed	European	Selected response	No
Gyllstad (COLLEX 5)	Mixed	European	Selected response	No
Jaén (Test 1)	Yes	European	Other	No
Jaén (Test 2)	Yes	European	Translation	No
Jaén (Test 4)	Yes	European	Selected response	No
Koya	No	Japanese	Selected response	No
Kurosaki (MC Fr)	No	European	Selected response	No
Kurosaki (MC Jp)	No	Japanese	Selected response	No
Kurosaki (trans. Fr)	No	European	Translation	Yes
Kurosaki (trans. JP)	No	Japanese	Translation	Yes
Revier	No	European	Selected response	Yes

3.3 Frequency data

Collocation frequency can be quantified in a number of different ways (see Schmitt 2010 for a review). Since it is unclear which of these is most likely to be related to learner knowledge, several different methods were used.

The first variable which needs to be considered in counting collocations is the 'span' of text within which two words need to occur to be counted as an example of the collocation. Collocates can occur at quite some distance from each other, as the following Example (1) of the collocation *realize dream*, taken from the Corpus of Contemporary American (COCA) (Davies 2008-), illustrates:

(1) The old *dream* of wireless communication through space has now been *realized*

Thus if the span used in our search of collocations is too narrow, many genuine examples will be missed. However, as the span is widened, the chances of counting word pairs which are not in a collocational relationship increases. Consider Example (2), again taken from COCA:

(2) she *realizes* that the buzzing sound from her *dream* is present in her bedroom.

The balance we need to achieve in setting a search span, therefore, is to maximize the number of genuine collocations while minimizing the number of false hits. The former pushes us to widen our search span, while the latter pushes us to keep it narrow. Jones & Sinclair's (1974) claim that most collocates are found within four words to the left or

right of their node has led to the widespread adoption of a 4:4 span. However, there has been little direct validation of this claim. The present research will therefore adopt two spans: a conservative 4:4 and a more liberal 9:9. Results from both types of search will be compared with student scores to see which is the better predictor of knowledge.

A second variable that must be considered is that of whether counts for separate forms of a word should be combined – such that, for example, *argue strongly* and *argued strongly* would count as two occurrences of a single collocation – or whether separate counts should be made for each form. While Halliday (1966) argues for the former on the grounds that treating different forms separately would add complexity without a gain in descriptive power, many corpus linguists have noted that conflating forms risks disguising important differences between the collocations of different forms of a word (Clear 1993, Hoey 2005, Sinclair 1991, Stubbs 1996). Both of these arguments, it should be noted, are based on the priorities of descriptive linguists. For our purposes, the important question is which approach produces counts which are relevant to students' likelihood of knowing a collocation. While there is some evidence that learners do not always transfer their knowledge of one form of a word to another (Schmitt & Zimmerman 2002), I would argue that the default assumption should be that learning will usually take place at least at the lemma level – for example, encountering *argue strongly* will increase a learners' chances of recognizing *argued strongly* as an appropriate collocation. Most of the frequency counts used in this study therefore combined counts of differently inflected forms of the component words. However, since the assumption that lemmatised counts provides a better estimate of knowledge is yet to be substantiated, one frequency count based on unlemmatised word forms was also provided for comparison.

A third factor that needs to be considered is the measure used to quantify collocation frequency. The simplest approach is to record the number of times a combination appears. However, such counts tend to give undue prominence to combinations of very high-frequency words (*of the*, *and a*, etc.), which co-occur very frequently by chance alone, while sidelining genuine collocations which consist of low-frequency words (*abject poverty, battering ram,* etc.). A number of methods have been suggested to overcome these problems. Perhaps the most widely used are the 't-score' and 'mutual information' (MI) statistics. The rationale for and calculation of these statistics are discussed in detail elsewhere (Manning & Schütze 1999) so will be described only briefly here.

Both statistics work by comparing the actual frequency of co-occurrence of a pair of words with the frequency we would expect them to co-occur by chance alone, given the individual frequency of each word. Expected frequency *E* is calculated using the formula

$$E = C * \frac{w1 * w2}{C^2}$$

where *C* is the total number of word tokens in the corpus, and *w*1 and *w*2 are the frequencies of each of the component words.

T-score and MI are then calculated with the formulas

$$t = \frac{O - E}{\sqrt{O}}$$

$$MI = \log_2 \frac{O}{E}$$

where *O* is the observed frequency of a combination.

The logic behind these two statistics is rather different, and this results in characteristically different types of collocations being highlighted by each. MI is a measure of the extent to which the probability of meeting one word increases once we encounter the other. T-score, on the other hand, is a hypothesis testing technique, which evaluates how much evidence there is that a particular combination occurs more frequently than we would expect by chance alone, given the frequencies of its component parts. As Clear (1993) puts it, whereas "MI is a measure of *the strength of association between two words*", t-score indicates "*the confidence with which we can claim there is some association*" (Clear 1993: 279-282, original emphases). Clear (1993) gives the example of *taste arbiters* as a combination with a high MI. Though the pairing is not particularly frequent, a high proportion of occurrences of each of its component words are found as part of this collocation, with, according to Clear's (1993) data, one quarter of all occurrences of *arbiters* being found within a two word span of an occurrence of *taste*. The two words are therefore strongly associated in that, where we find *arbiters*, we are also likely to find *taste*. However, the relatively low frequency of the collocation means that we cannot be confident that the association is generalisable – i.e. that we would encounter it in other samples of language. The pairing

taste for, on the other hand, is an example of a collocation with a high t-score. Though the association between these words is weaker than that between *arbiter* and *taste*, in that neither word is a strong predictor of the other, the pair occurs much more frequently, so we can be more confident in the generalisability of the association.

Both of these measures of association are non-directional, in that it makes no difference which word is taken as node and which as collocate. Clearly, however, the relationship between two parts of a collocation is often not symmetrical. The association from *arbiters* to *taste*, for example, is likely to be much stronger than that from *taste* to *arbiters* since, while a very high proportion of occurrences of *arbiters* is found in co-occurrence with *taste*, the reverse is not true. Since many of the task types included in the present analysis ask test-takers to identify a collocate when a particular node is given, this directionality may be important. For this reason, the analysis will also include the 'conditional probability' measure described by Durrant (2008: 84-85). This shows the probability of a particular word appearing, given that another particular word has appeared. It is calculated as:

$$P(w1 \mid w2) = \frac{w1 * w2}{w1}$$

A further point that needs to be taken into account when quantifying collocation frequency is the nature of the corpus consulted. To determine the extent to which learners' knowledge of collocation is frequency-driven, the best corpus would be one representative of each students' lifetime exposure to the language. Since such corpora do not exist, we need to work instead with more general corpora which may approximate to the types of exposure a variety of learners, on average, experience. With this aim, two widely used corpora were used: the British National Corpus (accessed through Davies's BYU-BNC interface (Davies 2004-)) and the Corpus of Contemporary American (Davies 2008-). Both of these corpora are intended to be representative of a national variety of English. The BNC is a corpus of approximately 100 million words of British English from the late 20th century. It includes around 10 million words of transcribed spoken language and 90 million words of written language, sampled from across five genres (academic, fiction, magazine, newspapers, non-academic non-fiction) plus one "miscellaneous" category. At the time of writing, the COCA includes around 450 million words of American English from the years 1990 to 2012. It is sampled in roughly equal amounts from spoken, academic, fiction, newspaper and magazine genres. Since it is possible that certain genres within each corpus will be more

representative of learners' experience than others, frequency information was rerieved both for the corpora as wholes and separately for each genre within them.

A related issue is that of 'dispersion' – i.e. the extent to which a collocation's occurrences are evenly spread throughout a corpus. Items which are frequent only because they are used intensively in a narrow range of texts represent a different learning prospect from items which occur regularly throughout the language. In general, it seems likely that more learners will have more exposure to a collocation that is widely dispersed than one which is restricted to a small range of texts. It is therefore worth asking whether learners have a better chance of knowing more widely dispersed collocations than those which are more restricted in their use. Several measures of dispersion have been proposed in the literature (Gries 2008). The measure adopted here was Gries's (2008) DP. This is calculated by (i) dividing the corpus into sections (in the present analysis, the sections will be the separate genres within each corpus); (ii) determining the size of each section and normalizing this against the overall size of the corpus to determine what percentage of occurrences of a collocation can be expected to appear in that section, if the collocation is equally distributed across sections; (iii) determining the actual percentage of occurrences of the collocation which is found in each section; (iv) computing the differences between expected and actual occurrences of the collocation in each section, summing these differences and dividing them by two. This provides a number, ranging between 0 and 1, where values close to 0 show an even distribution of the collocation across sections and values close to 1 show a strong bias towards particular sections.

As the discussion so far shows, collocation frequency can be quantified in many ways. The present research aims to determine both whether frequency in general is related to learners' likelihood of knowing a collocation and which of the methods of quantifying frequency are the best predictors of knowledge. With this aim, several different frequency statistics were employed. The first analyses employed frequency data from BNC and COCA as wholes. Collocation frequency was calculated in a number of ways. As the 4:4 span appears to be the most commonly-used in the literature (Hoey 2005) and as lemmatized frequencies have been argued to be the more relevant, the main analysis used lemmatized frequency with a span of 4:4 words. To determine whether different results are obtained when span and lemmatization change, additional counts were made based on lemmatized frequency with a span of 9:9 words and non-lemmatized frequency with span of 4:4 words.

In addition, the three measures of association (t-score; MI; conditional probability) and the measure of dispersion (DP) discussed above were calculated. To avoid an unmanageable multiplication of analyses, these measures were not calculated separately for all of the three collocation counts. For the reasons described in the previous paragraph, counts of lemmatized frequency with a span of 4:4 were used for this purpose. As a second step, separate frequency data were provided for each genre within the two corpora, i.e. in COCA: Academic; Fiction; Magazine; Newspapers; Spoken. In BNC: Academic; Fiction; Magazine; Newspapers; Non-academic; Spoken. Again to avoid an unsustainable multiplication of analyses, only lemmatized collocation frequency with a span of 4:4 were used for each genre.

3.4 Analysis

Data analysis took part in two stages. First, for each test, the percentage of learners correctly answering each item was correlated with each of the frequency measures described above. Second, a meta-analysis was conducted to find the average correlations across all 19 tests. While the first stage is straightforward, the second is more complicated and will be described here in detail. The procedures described here draw on the guidance provided by Lipsey & Wilson (2000).

The aims of a meta-analysis are to provide a single mean effect size which summarizes results from different studies and to determine the variation between different studies. While the former gives an overall indication of the influence of the main predictor variable, the latter allows examination of what other variables moderate this effect. Because studies which are conducted with a large number of participants are, other things being equal, more likely to provide a reliable effect size than studies based on smaller samples, the mean effect size is weighted to give more importance to studies with larger subject samples. Weighting is achieved by multiplying each effect size by the inverse of the standard error for the sample. Because correlations have problematic standard error formulations, they are usually transformed using Fisher's Z-transform before the weighting takes place. Z-transformed correlations are calculated using the formula:

$$ES_{z_r} = .5\log_e\left[\frac{1+r}{1-r}\right]$$

Once Fisher's Z transformation has been made, the mean weighted effect size is found by:

(i) Calculating a weighting for each effect size. This is the inverse of the variance for the sample. In the present case

$$SE_{z_r} = \frac{1}{\sqrt{n-3}}$$

$$w_{z_r} = \frac{1}{SE_{z_r}{}^2} = n - 3$$

where n is the sample size;

(ii) Calculating weighted effect sizes by multiplying each effect size by its weighting;

(iii) Calculating mean weighted effect size by dividing the sum of weighted effect sizes by the sum of weightings;

(iv) Calculating the standard error of the weighted mean effect size. This is calculated as:

$$SE_{\overline{ES}} = \sqrt{\frac{1}{\sum w_i}}$$

(v) Calculating the 95% confidence intervals for the mean using the standard error. This is calculated by adding/subtracting the product of the standard error and the critical value for the z-distribution (1.96) to/from the mean weighted effect size:

$$\overline{ES_L} = \overline{ES} - 1.96(SE_{\overline{ES}})$$
$$\overline{ES_U} = \overline{ES} + 1.96(SE_{\overline{ES}})$$

(vi) Converting the mean correlation and confidence interval from Z-transformed figures back to the original correlation type using the inverse transformation:

$$r = \frac{e^{2ES_{z_r}} - 1}{e^{2ES_{z_r}} + 1}$$

As discussed above, the aim of this meta-analysis is to allow generalization both to a wider body of L2 learners and to a broader population of collocations. For this reason, there are two sample sizes of relevance: the number of participants taking a test, and the number of collocations included on that test. For this reason, two meta-analyses were performed, one for each sample size.

Meta-analyses rely on the assumption that results from the different effect sizes they combine are independent of each other. This assumption is usually thought to be met if no more than one effect size in the analysis is taken from a single subject sample, though some researchers have argued that results conducted by the same team should also be considered dependent (Lipsey & Wilson 2000: 112). In the present meta-analysis, three types of violation of independence are relevant. Firstly, as Table 2 showed, there is some overlap between the collocations sampled in each test. In most cases, the overlaps are small. However, the two versions of Farghal & Obiedat's (1995) test, the five versions of Gyllstad's (2007) COLLEX test and the four versions of Kurosaki's (2012) test have substantial overlaps. It is therefore not likely that the effect sizes from these four tests will be independent of each other. Secondly, the three tests conducted by Jaén (2009) were all carried out with the same group of participants. Again, therefore, the assumption of independence is likely not to have been met. Thirdly, the four sets of studies just mentioned were each conducted by the same researchers. In addition to the overlaps in their samples, therefore, they also fail to meet the stricter criterion that effect sizes from studies conducted by the same researchers not be considered independent. For this reason, the correlations from each of these four sets of tests were combined into four single values by taking weighted averages of the correlations from each test. These average correlations were then used in the meta-analysis, rather than separate correlations for each test.

4. Results

Results from the first stage in the analysis are shown in Table 3 (for COCA data) and 4 (for BNC data). As collocation frequencies are not normally distributed, spearman's *r* was used to quantify correlation. All three counts of collocation frequency showed positive correlations with learner knowledge for the majority of tests, though the size of the correlation varied widely across tests and between COCA and BNC counts (with the former producing the higher correlations). The same pattern holds for correlations with

t-scores and conditional probability. DP shows the expected negative correlation in a majority of cases. The results for MI show a high degree of variability, with positive correlations in 13 tests using COCA data and in 9 tests using BNC data.

There are not sufficient data here to enable a reliable analysis of factors that might affect variation in scores between tests. However, it is worth looking at how these data vary across potential moderators. This is important both to provide clues as to potential effects that future research might investigate and to support interpretation of the meta-analysis, which relies on the assumption that effect sizes come from a single population and that differences between effect sizes are due to random errors, rather than systematic moderating factors. Section 3.2 described four variables that might moderate the current findings: learners' experience of studying English (English majors vs. non-English majors); learners' L1; task type (selected-response vs. translation); and whether test-takers are asked to provide the whole collocation or only the collocate.