Words that go together well: developing test formats for measuring learner knowledge of English collocations

HENRIK GYLLSTAD (LUND UNIVERSITY)

Abstract
This paper describes the development of two test formats, COLLEX and COLLMATCH, measuring receptive recognition knowledge of English verb +NP collocations, and reports findings from pilots and initial test administrations involving Swedish upper-secondary-school and university level learners. Administrations of the test formats produced highly reliable scores and the performance of native speakers provided evidence of test validity. The tests discriminated significantly between upper secondary school learners, university learners, and native speakers. Significant differences were however not observed throughout between advanced learner groups only one term apart in terms of level of formal instruction. A vocabulary size measure was found to correlate highly with scores on both tests, which seems to suggest that learners with large vocabularies have a better receptive command of verb + NP collocations than learners with smaller vocabularies.

"Michelle, ma belle, sont les mots qui vont très bien ensemble"
(Lennon & McCartney 1965)[1]

1 Introduction

Vocabulary tests typically tap learners' knowledge of the meaning or form of single words. The task often involves translating foreign language words, occurring in or out of context, into L1, or vice versa. In other words, learners are either asked to supply the meaning of a given form, or to supply the form for a given meaning. This is often referred to as the passive – active distinction. Furthermore, a recall – recognition distinction can be introduced, where the recall type of knowledge involves the ability to retrieve from memory a form or a meaning, triggered by some sort of prompt, whereas the recognition type of knowledge involves the ability to recognize a form or a meaning, presented with a set of alternatives (see Laufer & Goldstein 2004).

[1] Lyric excerpt from the song *Michelle*, featured on the album *Rubber soul*, released by the Beatles in 1965.

Irrespective of what kind of task is used, there is a strong tendency in educational situations to focus on single words. For example, everyday vocabulary tests created by classroom teachers most often involve tasks that measure how many words are known from a defined pool of study (Schmitt 2000). Certainly, testing learners' knowledge of single words in the fashion described above is meaningful. The overall number of words for which a learner has at least some basic understanding is generally called 'vocabulary size' or 'vocabulary breadth'. Test formats aimed at measuring this construct[2] have been developed (see e.g. Meara & Buxton 1987, Meara & Milton 2003, Nation 1990, 2001; Schmitt 2000; Schmitt et al. 2001).

However, there is also a need for tests that tap word knowledge aspects beyond basic form-meaning mappings of single words, especially with post-beginner learners. In the literature, these word knowledge aspects are commonly captured by the umbrella term 'vocabulary depth'[3]. These aspects can be seen to involve associations that a word gives rise to, grammatical functions of the word, knowledge of register and style, and common collocates (see Nation 2001:347). Although such tests do exist (see e.g. Read 1993, 1998; Wesche & Paribakht 1996), they are few and far between.

Out of the aspects above, one seems to be particularly elusive for second language learners: words and their collocates. This paper describes the development of two test formats aimed at tapping learners' knowledge of collocations. The paper is organized as follows. In section 2, the reason why tests of collocation knowledge are warranted is discussed, and a review of previous studies that have bearing on the topic of this paper is provided. In section 3, the construction of two collocation test formats is described. Sections 4 and 5 report the results from pilots and two test administrations involving Swedish learners of English. Section 6, finally, sums up the results and points at further developments of the tests.

2 Collocation knowledge and previous research on learner performance

There are several co-existing definitions of what a collocation is, and it is not within the scope of this paper to give a thorough account of these. The sequence *take a decision* may be used to illustrate what the term typically denotes. In English, the concept of 'arriving at a mental state in which a step of action is

[2] Following Chapelle (1998:33), the term 'construct' is used here to denote "a meaningful interpretation of observed behaviour".

[3] Paul Meara and co-researchers (Meara 1996; Meara & Wolter 2004) advocate a view in which the depth construct is replaced by a dimension denoted 'organization'. In this view, organization refers more holistically to the way a lexicon is structured, and is a property of the system as a whole, not of the individual words that make up the system.

chosen', can be captured by this very string of words. Alternatively, one may *make a decision*, but one cannot **do a decision* or **set a decision*. Although the latter two sequences are grammatically well-formed, they are simply not used in an English speech community. Consequently, *take a decision* is a conventionalized recurring word string. Carter (1998: 51) gives the following definition:

> Collocation is a term used to describe a group of words which occur repeatedly in a language. These patterns of co-occurrence can be grammatical in that they result primarily from syntactic dependencies or they can be lexical in that, although syntactic relationships are involved, the patterns result from the fact that in a given linguistic environment certain lexical items will co-occur.

We may note here that the definition above captures the repeated co-occurrence of lexical items. Frequency is therefore an important aspect. However, not all repeatedly co-occurring items are of interest. The definite article *the*, for example, may co-occur frequently with basically any other word, whatever part-of-speech. Also, even though the verb *throw* may co-occur frequently with the noun *ball*, this may not be very interesting from a learner perspective. The reason for this is that *throw* + *ball* can be seen as a free combination. Knowing the argument structure and selectional restrictions of the verb *throw* and the semantic features of the noun *ball* will allow a learner to deduce and generate this sequence. In contrast, *throw* + *party* will not be a predictable combination, and therefore has to be learned. The last part of the quote above brings an important point to the fore: the often arbitrary link between two or more lexical items. We *say a prayer* and we *tell a joke,* but we do not **tell a prayer* or **say a joke.* The lexical items inherent in the latter two sequences simply do not co-occur in English. In order to sound native-like, a learner has to use the well-formed structures that native speakers use. Although the grammar of a language may allow several structures for expressing a single concept, very often the number of actually used structures in a speech community is limited. Certain structures become conventionalized. Thus, in addition to knowing the meanings of single words, knowing how to combine these words into bigger chunks—conventionalized phrases and clauses, collocations—is a very important skill, especially for more advanced learners.

That learners have problems with collocations is a well-established fact (see e.g. Biskup 1992, Bahns & Eldaw 1993, Howarth 1996, Granger 1998, Nesselhauf 2005). As has been pointed out by Wray (2002: 183), collocations can only be learned if they are present in the input learners are exposed to. Since there does not seem to be any reason to believe that the input directed to learners is simplified with regard to collocational content, Wray draws the conclusion that learners simply do not seem to pay attention to collocational relationships. She furthermore hypothesizes that there is a difference in the way native speakers and non-native speakers deal with language. First language learners,

she claims, start with large and complex strings, and do not break them down any more than necessary. In contrast, post-childhood second language learners are suggested to start with small units which they try to build up.

> Phrases and clauses may be what learners encounter in their input material, but what they notice and deal with are words and how they can be glued together. The result is that the classroom learner homes in on the individual words, and throws away all the really important information, namely, what they occurred with. (Wray 2002: 206)

Wray furthermore contends that native speakers' treatment of collocations, on the one hand, may be seen as fully formulaic pairings which have become loosened. That is, strings of words which can be separated under certain circumstances. Adult learners' treatment of collocations, on the other hand, can be seen as separate items, words, which have become paired (Wray 2002:211).

Thus, it seems like native speakers use a top – down strategy whereas learners adopt a bottom – up strategy. Adult learners are hypothesized to start with individual words and gradually build up bigger strings. Wray suggests that it is the pairing, and particularly the establishment of the strength of the association that causes difficulties for learners.

Warren (2005) claims that while native speakers construct generalized meanings of words by abstracting semantic commonalities from different contextual uses, the non-native speaker is likely to construct a generalized meaning of an L2 word by equating it with some core meaning in L1, i.e. a translation equivalent. As far as collocations are concerned, Warren's observation leads us to hypothesize that L1 influence will sometimes make it difficult for learners when producing L2 sequences, or when judging whether a certain sequence is idiomatic. For example, a German learner of English who want to capture the concept of taking a photograph might produce the infelicitous sequence **make a photo* since the German counterpart would be *ein Foto machen.* The learner is influenced by the verb *machen* and consequently uses the equivalent *make* when *take* would have been the idiomatic choice for this concept.

Although learners' problems with lexical collocations are widely attested, the overall number of studies investigating learners' command of collocation is on the whole scarce. A number of them involve analyses of learner essay corpora. Examples of these are Gitsaki (1996), Howarth (1996 and 1998), Granger (1998), Wiktorsson (2003), and Nesselhauf (2005). Although analyses of learners' production of collocations in essays can shed some light on the process involved in the associations learners build between words, the amount of data collectable from each learner is often limited. There are however also studies in which more test-like and experimental instruments are used to tap into learners' knowledge of collocations. These include: Biskup (1992), Bahns & Eldaw (1993), Farghal & Obiedat (1995), Bonk (2001), Mochizuki (2002), and Barfield (2003). Since this paper is concerned with the development of valid and reliable

test formats aimed at measuring learners' collocation knowledge, a review of these studies is warranted.

Biskup (1992) investigated how well a total of 62 Polish and German university students translated verb + noun and adjective + noun collocations from their respective mother tongues into English. She found that the two groups produced the same mean number of correct responses, but with more restricted collocations produced by the Polish learner group than German group. Also, the Polish learners more often refrained from answering, whereas German learners supplied more paraphrases. From a test perspective, it is not clear how many items were tested, or if the tested items were decontextualized items, sentences, or full texts with underlined items. Furthermore, no reliability measures of the test instrument are presented. By and large, the lack of clearly presented details about the items and the test instruments makes it difficult to fully evaluate Biskup's findings.

Bahns & Eldaw (1993) aimed at testing learners' productive knowledge of 15 verb + noun collocations. 58 German university students of English, in years 1-3, participated in the study. Of these 58 subjects, 34 were given a translation task in which 15 German sentences were to be translated into English, and 24 subjects were given a cloze format in which the target collocations were inserted into English sentences with the verb collocate of a noun missing. An example of the translation task item is given in (1), and an example of the cloze task item is given in (2) below:

(1) Als Teenager hatte sie damit begonnen, regelmäßig Tagebuch zu führen. [Translation task]
(2) When she was a teenager, she used to _____ a diary. [Cloze task]

The 15 verb + noun collocations were selected from various sources, such as learning materials and dictionaries, and were pre-tested on 2 native speakers as a validation measure. The subjects' answers were rated as acceptable or unacceptable by 3 native speakers. In terms of results, no significant differences were found between the two groups as to the mean number of correctly answered items, 7.2 for the cloze group and 8.1 for the translation group, respectively. Bahns & Eldaw also concluded that collocation knowledge does not develop alongside general lexical knowledge. This conclusion is interesting and it will be addressed in the discussion of my own results in sections 5 and 6. However, in Bahns & Eldaw's case, the conclusion was based on a rather odd analysis in which the measures of two assumed variables, general vocabulary knowledge and knowledge of collocations, were taken from the same data, and thus not independent. The analysis entailed taking the percentage of felicitously translated single lexical words in hypothetically ideal translations (83 lexical words x 34 students) and comparing this with the percentage of felicitously translated verbal collocates. No reliability measures of test instruments are

presented, and the number of items tested is fairly small. On the whole, since the measures of general vocabulary and collocation knowledge were confounded, together with the fact that very few items were tested, the conclusions drawn in this study cannot be seen as sufficiently robust.

In Farghal & Obiedat (1995), a total of 57 Arabic university students of English were tested for their knowledge of English collocations. Two groups were used, A and B. The aim of the study was to test knowledge of 22 common English collocations. The two groups were given separate tasks. Group A took an English fill-in-the-blank test with 11 items in which one member of a collocation pair was given, and one was missing, which was meant to be supplied. Group B took a test in which Arabic sentences were meant to be translated into English. This test was based on the same target collocation material as the fill-in-the-blank test. Two examples of the English test can be found in (3) and (4) below:

(3) I prefer _____ tea to strong tea
(4) Some people like salty soup, but others like _____ soup.

The collocate pairs targeted in the examples above are *strong tea*/*weak tea* and *salty soup*/*bland soup*. The English form was validated by two native speakers of English. Farghal & Obiedat found that 4 lexical simplification strategies were used. The use of synonymy was the most frequently used strategy by both groups when a correct collocation was not produced, followed by that of avoidance. The two other strategies identified were transfer and paraphrasing, used to varying extent by the two groups. The conclusion drawn in the study is that L2 learners cannot cope with collocations. This is because "they are not being made aware of collocations as a fundamental genre of multi-word units" (p.326). Farghal & Obiedat claim that vocabulary is taught as single lexical items, something that leads to lexical incompetence on the part of the L2 learners.

No reliability measures of test instruments are presented, the number of items tested is fairly small, and it is not clear how the test items were selected. Furthermore, the study seems to rest on the assumption that there is a self-evident relation of antonymy between the collocations used, an assumption that is scarcely tenable.

Bonk (2001) subjected 98 university students, a majority of whom were speakers of East-Asian languages, to a test battery consisting of 3 subtests of collocation knowledge and a general English proficiency measure. The overall aim of the study was to investigate the reliability and validity of the test instruments used, and to correlate collocation knowledge with general English proficiency. The subtests used were the following: a) a 17-item prompted recall verb+object collocations test of English sentences, each with a gap for a verb to be inserted, b) a 17-item prompted recall verb+preposition collocation test, also

with English sentences, but each with a gap for preposition to be inserted, and c) a 16-item receptive figurative use of verb phrases test, consisting of multiple-choice items with 4 sentences in each. The task for the testee was to judge which one of the four sentences does not contain a correct usage of the verb. Finally, d), a 49-item general language proficiency measure was administered in the form of a 49-item condensed TOEFL test. Examples of items in the three collocation subtests are given in (5), (6) and (7) below:

(5) Punk rockers dye their hair red and green because they want other people to _ _____ attention to them.

(6) Many of the birds in the area were killed _____ by local hunters. (to exterminate)

(7) a. Are the Johnsons throwing another party?
b. She threw him the advertising concept to see if he liked it.
c. The team from New Jersey was accused of throwing the game.
d. The new information from the Singapore office threw the meeting into confusion.

The test battery was validated by administration to 10 native speakers. 98 students participated in the main test administration. The students scored a mean of 25.3 (SD 7.3) out of 50 on the collocations test total, and their mean scores on the 3 subtests were close to 50% of the maximum score of the respective tests (8.7, 8.8 and 7.8) Their total mean score on the 49-item TOEFL test was 37.3 (SD 7.2). A Kuder-Richardson 20 analysis of internal consistency showed that the scores on the collocations test were reliably measured at .83. One of the subtests, however, the verb+preposition test, was found to display a rather low and unacceptable reliability value at .47. Bonk also carried out item analyses including item facility and item discrimination indices, and point biserial coefficients[4]. These analyses showed that a majority of the items functioned as good, well-discriminating items. The mean item facility for the three subtests was around .50, and the mean point-biserial correlation was .38, .27, and .34 respectively for the three collocation subtests. Based on Item Response Theory (IRT) Rasch analysis, and Generalizability analysis, Bonk concluded that the 50-item collocations test worked well on the whole for the population, but that subtest 2, the verb+preposition test, was a somewhat weak link and that it could practically be discarded in favour of extending subtests 1 and 3. Bonk found a moderately high level of correlation between general English proficiency and collocation proficiency (.73 after correction for attenuation). No instances of low proficiency and high collocation scores could be found, and no instances of high

[4] Point Biserial methods correlate binary item scores (0, 1) with continuous total scores on a test. As with Discrimination Indices, Point Biserial correlation coefficients indicate how well an item discriminates between testees with high total scores and testees with low total scores on a test (see Henning 1987).

proficiency and low collocation scores either, although the middle range of scores displayed some variation.

One of the advantages with Bonk's study is the attempt to include a larger number of items (k = 50). He also subjected his data to rigorous statistical analyses through which he attempted to support his conclusions. If several variables are to be compared and correlated with each other, it is important to show that these variables were reliably measured. On a more critical note, the task formats used by Bonk involve a fair bit of reading, and this raises the question of what is really measured. It could be the case that the subjects did not understand the sentence prompts and therefore did not answer an item correctly. If so, the test is more a measure of reading comprehension than collocation proficiency. Admittedly, Bonk tried to control for this by qualitatively examining 25% of the answer sheets, finding that the subjects seem to have understood the prompts "the great majority of the time" (p.134). A further remark on the minus side is the unsystematic selection of test items. The selection of items seems to have been made on the basis of intuition only.

In Mochizuki (2002), 54 Japanese first-year university students, majors in German, Chinese, or Japanese, were tested on collocation knowledge, paradigmatic knowledge and overall vocabulary size. The aim of the study was to explore how Japanese learners of English develop two aspects of word knowledge, paradigmatic and collocational, and vocabulary size over one academic year. Over this period of time, the students received 75 hours of instruction (reading and conversation classes). The tests used were the following: a) a vocabulary size test, an adaptation of the Vocabulary Levels Test (Nation 1990, 2001), which in Mochizuki's version included 7 levels corresponding to 7 frequency bands, the 1st -7th thousand most frequent words, and the task involves matching English words with Japanese translation equivalents, b) a test of paradigmatic knowledge of 72 English words in a 4-choice format, and c) a collocation test of 72 words, the same words as in task b), also in a 4-choice format. Examples of subtests b) and c) are provided in (8) and (9), respectively, below:

(8) job (1) date (2) sort (3) star (4) work

(9) job (1) answer (2) find (3) lay (4) put

The task for the learner was to decide with which of the four alternatives there is a possible link – a paradigmatic one in the case of the paradigmatic knowledge test (8), and a syntagmatic one in the case of the collocation knowledge test (9). The target words in the tests were divided into four groups of 18, and each group consisted of six nouns, six verbs and six adjectives, all randomly selected, taken from one out of four word lists based on frequency counts.

When comparing the results obtained at the two administrations (April=T1 and January=T2), Mochizuki found that only in the case of the collocation test was a significant difference observable (41.7 (SD 5.4) at T1, and 42.8 (SD 6.4) at T2). In terms of internal reliability, the values calculated (Cronbach's alpha = α[5]) were α .71 and .75 for the two administrations of the paradigmatic knowledge test, and α .54 and .70 for the two administrations of the collocation knowledge test. Mochizuki deducts that the test administrations were moderately reliable. Mochizuki explains the very modest lack of increase over the two administrations by lack of motivation on the part of the learners. Following an argument advanced by Schmitt (1998), he furthermore explains the fact that there was a significant increase in collocation knowledge, and not in vocabulary size and paradigmatic word knowledge, by the inherent inertia of knowledge of meaning. It is assumed that a learner's knowledge of word meanings does not change radically over time, whereas knowledge of syntagmatic relationships does.

As with Bonk's study described above, Mochizuki's study attempted to test a larger number of items (k = 72), which is positive. Also, values of internal reliability were reported, even though no reliability values were given for the vocabulary size measure. One administration of the collocation knowledge test showed a relatively low value of α .54. The value might be partially explained by the rather homogeneous group of learners taking the test. Homogeneous group scores generally result in low internal reliability values, since the calculation relies on a certain variance (see Brown 1983:86)). In contrast to Bonk's study, decontextualized items were used. An analysis missing in the study, I think, is a correlation measure. It would be interesting to correlate the vocabulary size variable with the paradigmatic knowledge and collocation knowledge variables, respectively, to see whether and how these word knowledge aspects are interrelated. This will be discussed further in sections 5 and 6 below.

The final study under review here is that of Barfield (2003). In this study, a total of 93 Japanese university students, undergraduates and post-graduates, belonging to 4 different fields of study, participated. The overall aim of the study was to test a large number of decontextualized verb + noun collocations for recognition, and to compare recognition patterns with those of the single verbs and nouns. For this purpose, 40 lexical verbs from a previous study were used. These verbs were taken from the Academic Word List (AWL) (Coxhead 2000), and the General Service List (GSL) (West 1953). Furthermore, 3 noun collocates were chosen for each of the 40 verbs, based on data in the Cobuild Bank of English. Furthermore, 20 mis-collocations were created, intuitively, mainly based on other verbs' collocates. This was done as a means of checking

[5] See e.g. Bachman 2004 for an account of this reliability coefficient. Also, see section 3.1 below for a brief account of what test reliability is.

the realiability of the test instrument. The result was 120 items out of which 100 were real collocations and 20 mis-collocations. The learners were asked to rate each collocation on a 4-state scale:

1 I don't know this combination at all.
2 I think this is not a frequent combination.
3 I think this is a frequent combination.
4 This is definitely a frequent combination.

Figure 1. A 4-state scale of reported knowledge of verb+noun combinations, from Barfield (2003)

It is not clear how the tested items were presented to the learners, but examples of the tested items are *adopt* + *approach*, **adopt* + *child*, *adopt* + *profit*, *break* + *ground*, *break* + *record*, and *break* + *rules* (asterisk indicates mis-collocation)

Barfield first tested the learners' recognition knowledge of the 120 nouns, using a similar, but slightly differently worded rating scale than that above. He found that the noun recognition was very high, with a mean score of 3.87 (SD .079). Through comparing the reported recognition of the 4 groups on the noun test with data on general English proficiency, Barfield concluded that noun recognition scores harmonized well with general English proficiency. No reliability figure was calculated due to zero variance with 55 of the 120 nouns.

As for the verb+noun collocation test, the mean recognition for the total number of collocations was 2.56 (SD .39). For the real collocations, the mean score was 2.65 (SD .47), and for the mis-collocations, somewhat lower, 2.15 (SD .62). Looking at the recognition scores of the 100 real collocations, Barfield found that these scores showed high reliability as measured by Cronbach's alpha ($\alpha = .97$), and that there was a significant difference between two of the learner groups. Reliability was high also for the mis-collocations ($\alpha = .93$). No significant differences were found between the group mean scores.

Barfield concludes that the different groups shared a similarity in rejecting the mis-collocations, although many of them were rated as knowledge state 2, i.e. (I think this is not a frequent combination). Furthermore, he observes that one group (Medical Science), scored significantly higher than the other three groups on the real collocations. He suspects that the higher than expected recognition of the mis-collocations could point to a possible overestimation of recognition of the real collocations, on the part of the learners.

With one exception, all of the nouns and verbs of the top 20 most recognized collocations, e.g. *change mind*, *protect body*, *protect environment*, *explain reason* and *govern country*, were within the 3000 most common words of English according to frequencies in the British National Corpus (BNC), which leads Barfield to conclude that the relative frequency of the single words making up a collocation is a supporting factor in collocation recognition. Looking further at the 20 most recognized collocations, core sense in both the verb and the noun seemed to figure highly as the primary deciding factor (11 items).

Another factor seemed to be the combination of an abstract noun + a verb in its core sense (8 items). The remaining collocation residing in the top 20 was a verb in specialized sense + concrete noun. Based on these findings, a 4-way division of semantic transparency for collocational recognition is suggested (2003: 45, figure 2), in which field 1 is suggested to be the easiest and field 4 the most difficult for learners.

Taking results from previous studies into account, Barfield is able to juxtapose recognition scores for verbs, nouns, and verb+nouns. This juxtaposition shows that learners claimed a higher recognition of nouns (mean 3.87) than of verbs (mean 3.56), and a higher recognition of verbs than of verb+noun collocations (mean 2.65).

		NOUN	
		CORE	NON-CORE
VERB	CORE	1) semantic transparency in both components	2) semantic transparency driven by abstract noun
	NON-CORE	3) verb in specialized sense with core noun	4) semantic opacity in both components

Figure 2. A 4-way division of semantic transparency, taken from Barfield 2003, p.45.

Barfield's study is yet an example of efforts to use a large number of items. The selection of items is systematic. The 4-state scale of knowledge used is interesting, since word knowledge is not an all-or-nothing knowledge. Also interesting is the fact that recognition of the constituent parts of the collocations, the single verbs and nouns, is tested. This is good since learners claiming knowledge or not of a collocation may depend on their knowledge of the parts of the combination.

On the minus side can be noted the fact that some of the mis-collocations are possible in certain contexts, a shortcoming admitted by the author. Examples of these are *explain address*, *approve opportunity* and *create temperature*, all of which could be rather feasible combinations, conditioned by the insertion of one or more lexical items in-between and around the verb and the noun: *to explain an address to someone*, *to approve of a job opportunity*, and *to create a temperature at which certain solid elements become liquid*. A final observation concerns the fact that no delexical verbs were used. It is noted in the literature that delexical verbs, such as *make*, *take*, *do*, *give* and *have*, occur frequently in English and that they are difficult for learners (Källkvist 1999, Altenberg & Granger 2001, Nesselhauf 2004), even at advanced levels. For this reason,

investigating learners' knowledge of collocations in which delexical verbs appear seems to be warranted.

Having reviewed a number of studies relevant to the topic of this paper, a couple of trends emerge. Firstly, on the whole, few studies have been carried out investigating learners' knowledge of collocations through more test-like, experimental measures. Secondly, in the few studies that do exist, a rather small number of items are tested, usually 10-20, with the exception of the last three reviewed above (Bonk 2001, Mochizuki 2002, and Barfield 2003). The drawback of using few test items is that it is very difficult to draw any well-founded conclusions, especially so when the item selection is made in an unsystematic way, or not described at all. Thirdly, far too often, no reliability values of the test instruments per se are reported. Again, the three studies reviewed above are exceptions to this trend. Especially when different variables are compared, it is essential that the operationalized measures of the variables, the scores, show a decent degree of reliability. If too high a percentage of a score is marred by unsystematic variance, inconsistencies, not attributable to the underlying language ability of the test-taker, then less trust can be put into any conclusions drawn from the score. As pointed out by Bachman (1990: 160), "in order for a test score to be valid, it must be reliable". Reliability is thus a necessary condition for validity. Fourthly, and finally, none of the studies reviewed here, or any other studies to the best of my knowledge, compare learners at different levels of formal instruction when it comes to collocation knowledge measured in a more test-like manner.

In the remainder of this paper the development of two test formats aimed at measuring L2 learners' command of English collocations will be described.

3 Developing two collocation knowledge test formats: COLLEX and COLLMATCH

3.1 Assumption and criteria guiding the construction of the formats

Four main criteria guided the initial construction of my two formats, called COLLEX and COLLMATCH: a) measuring receptive recognition knowledge of primarily verb + object combinations, b) constructing tests that display evidence of high reliability, c) focusing on collocations which are combinations of relatively high frequency words, and d) using decontextualized items. The four criteria will be elaborated on, one by one, below.

As for criterion a), verb + object combinations have primarily been chosen because they are frequent, notoriously difficult for learners, and since they "tend to form the communicative core of utterances where the most important information is placed" (Altenberg 1993:227; cited in Nesselhauf 2005:9). Although production of collocations is the type of knowledge one would ideally

like to measure, receptive knowledge is more readily measurable, especially from a language testing perspective, and can certainly be a useful indication of the learners' phraseological knowledge, since presumably the collocations which a learner understands or uses with any depth of meaning will also be recognized in the test formats.

Criterion b) concerns test reliability. High reliability is not an end in itself but rather a step on a way to a goal. Unless test scores are consistent, they cannot be related to other variables with any degree of confidence. Thus reliability places limits on validity, and the crucial question is whether a test's reliability is high enough to allow satisfactory validity. Analyses of internal consistency seek to determine the degree to which the test items are interrelated. If the scores on the various items comprising a test inter-correlate positively, the test is homogeneous. Internal consistency estimates are obtained from one administration of a test, as opposed to measures based on several administrations. Because the estimates are derived from item inter-correlations, the primary source of error lies in the items comprising the test (Brown 1983).

A reliability coefficient, such as Cronbach's Alpha (see, e.g. Bachman 2004), indicates the degree of inconsistency, but not the causes of the lack of consistency. However, the literature describes a number of probable reasons for low or high reliability values. A longer test is generally more reliable than a shorter one (Brown 1983:85), since, as the number of items increases, random measurement errors tend to cancel each other out. Very easy items and very hard items have low variances and thus decrease internal consistency. The nature of the groups tested also influences reliability coefficients. In general, as groups become more homogeneous, variability decreases.

Criterion c) has to do with word frequency. Even though the constituent words of a collocation are very frequent, it does not follow that the collocation itself is "easy" for learners. Knowing what combinations frequent words enter into is an important skill for any learner, especially intermediate and advanced ones.

As for the final criterion, d), there is always a trade-off between the number of items one can test and the depth with which one can measure. In order to be able to test a large number of items, I have chosen to use decontextualized items. Certainly, providing some sort of linguistic context around targeted test items makes any task more natural in that it is the way language appears to us as language users. However, as pointed out by Cameron (2002), it is reasonable to assume that learners presented with decontextualized test items do not make sense of the tested items in a decontextualized mental void. Rather, she claims, the recognition process may activate recall of previous encounters and their contexts. Also, it is arguable that the more context one adds to a test item, the more relevant is the question of what one is really measuring. More context means that reading comprehension and inferencing skills come into play, and this may in a way muddle the measure of the intended construct.

3.2 The COLLEX format

In the format called COLLEX, the testee is presented with a large number of V+N collocations (~50). The format is a kind of forced-choice preference test, and it is inspired by a single word test of vocabulary size, suggested by Eyckmans (2004). In each item, two lexical combinations are juxtaposed, one real collocation and one pseudo-collocation. The testee's task is to decide which one of the two is a real collocation. In the test instruction, the testee is told to select the one which is deemed more frequent and used by native speakers of English. In the initial version of the format, the testee is also instructed to indicate if he or she does not know, but is guessing, by ticking the box to the right of the item pair.

Here are some examples of COLLEX items:

1)	tell a prayer	say a prayer	☐
2)	pay a visit	do a visit	☐
3)	run a diary	keep a diary	☐
4)	do a mistake	make a mistake	☐

Figure 3. Example of items in the COLLEX format.

As far as the selection of items is concerned, a large majority of the test items in COLLEX are high frequency words. For example, almost 85% of the verbs come from the 3000 most frequent words of English, based on calculations in the BNC (Kilgarriff 1996), and so do almost 80% of the nouns. Thus, it is assumed that university students of English will know the great majority of the single words that make up the collocations. Examples of lower frequency verbs in the test are *dial*, *fell*, *undo*, *polish* and *shed*. Furthermore, examples of lower frequency nouns are *sacrifice*, *visibility*, *apology*, *errand* and *fuse*. In addition to the verb + noun collocations, a small number of adjective + noun collocations are used as items in the test. Table 1 below summarizes the distribution of the test items in COLLEX version 3 with regard to what frequency bands they were taken from.

A guiding criterion when constructing the test item pairs was the need for the distracting pseudo-collocation to be tempting as an alternative in relation to the real collocation. Intuition about possible L1 influence, transfer, from Swedish constructions was used in this respect. Initially, in the pilot versions of the test, care was taken to try to sample the verbs in each test item, that of the real collocation and the pseudo-collocation, from the same frequency bands. Furthermore, the nouns were taken from 3 frequencies: 1K, 2K and 3K+. This, however, proved to be difficult in some cases since it very often resulted in

pseudo-collocations which were not deemed plausible. This was seen in the item analysis of the piloted items (see section 4). Consequently, this criterion had to be dropped. In practice, however, a great majority of the two verbs in an item belong to the same or adjoining frequency bands (the same thousand word band or, for example, one verb from band 1K and one from 2K). On the whole, even though care was taken to concentrate on high to moderately high frequency words for the items, sometimes obtaining distractor "credibility" took priority.

	1K	2K	3K	4K	5K	6K	7K	7K+	Total
Verbs	23	10	4	2	1	1	-	3	44
Nouns	15	17	7	1	1	2	-	7	50
Adjectives	9	3	2	-	2	-	-	3	19
total	47	30	13	3	4	3	0	13	

Table 1. Number of test item words (types) from different frequency bands in COLLEX 3.

The real collocations and the pseudo-collocations were checked in the BNC, using a span of +3 words to the right of the verb, which was used as the keyword. A z-score (Barnbrook 1996:95) of >3 was the minimum level of acceptance for the real collocations. For most collocations, though, very high z-scores were obtained, indicating that they are indeed frequent combinations in English. The pseudo-collocations were also checked in the BNC to make sure that they were not possible combinations in English. If an intended pseudo-collocation was found to occur in the corpus, concordance lines were retrieved and investigated in order to see if the use of the combination could be seen as conventionalized in any way. Furthermore, the items were checked with a native speaker of English in order to avoid poor and unclear items.

The test taps receptive knowledge, recognition knowledge to be more precise. Since the learner is presented with only two L2 sequences, and no L1 equivalent, the format tests whether a certain sequence form is salient compared to another sequence. It seems plausible to suggest that the learner may employ the following types of cognitive processes when answering a test item. Either a) the learner has been exposed through prior input to one of the two sequences to the extent that it triggers the retrieval of a stored mental representation: a chunk. According to Newell (1990, cited in Ellis 2001: 39), a chunk is "a unit of memory organisation, formed by bringing together a set of already formed chunks in memory and welding them together into a larger unit". This stored chunk will prove stronger than a competing sequence, if the competing sequence is not stored as a chunk. If exemplified, this means that the collocation pair **make suicide – commit suicide*, will trigger, in the best case, with the learner, a mental representation in long-term storage of the verb *commit* together with the object *suicide*. The sequence **make suicide* will however not be expected to trigger a stored representation since it is unlikely that the learner has been exposed to this string of input. The stored mental representations may include sequences where the verb appears with tense and/or number inflections, e.g.

commits suicide, committed suicide, committing suicide. What sequences are stored depends on the input the learner has been subjected to.

Alternatively, b), the learner may have developed an analytical knowledge of the verb frame, thus knowing or having a feeling for the semantic preference of the verb (see Stubbs 2001). For example, trough input, a learner may have induced that the verb *commit* may exclusively be combined with objects of more negative connotations, e.g. *crime, suicide, murder.* This may help the learner in deciding that *commit suicide* is possible as a combination. However, it will not help the learner to rule out the possibility of **make suicide* being an equally acceptable combination.

3.3 The COLLMATCH format

In the format COLLMATCH, the testee is presented with a number of grids, each consisting of 3 verbs and 6 noun phrase (NP) objects. The testee is asked to indicate which of the 6 objects each verb felicitously combines with. The number of possible combinations is not known to the testee and in theory all or none, and every possible number in-between is possible. The same object may be combined with more than one of the three verbs. The instruction asks the subjects to tick the combinations they think exist in use in English.

An example of a COLLMATCH grid (version 1), can be seen in Figure 4 below:

	charges	patience	weight	hints	anchor	blood
drop						
lose						
shed						

Figure 4. Example of a COLLMATCH grid, version 1.

Just as in the COLLEX format, this format is a measure of receptive recognition knowledge. However, the cognitive effort involved is somewhat more demanding than the COLLEX format, since the number of alternatives is large. In each grid, there are 18 items. To a great extent, the format can be seen to elicit answers to the question: '*What can be* V-ed?' Thus, based on the items in the grid above, the questions would be: *what can be dropped*?; *what can be lost*?; *what can be shed*? This should give us a picture of learners' knowledge of the lexical restrictions, motivated or arbitrary, that must be abided by, if native-like sequences are the norm (cf Howarth 1998; Stubbs 2001). In some grids, the combinations are overlapping in the sense that two or even all three verbs may share the same object. An example of this can be seen in the grid above, where both *shed* + *weight* and *lose* + *weight* are possible combinations.

Since some of the verbs may enter into combinations in which the verb does not display its most common core meaning, the format can also been seen as measuring knowledge of word polysemy to some extent.

As with the COLLEX format, the items used in COLLMATCH are predominately words of high frequency. In version 1 of the format, verbs like *break*, *hold*, *keep*, *drop*, *lose*, *shed*, *say*, *tell*, *speak*, *beat*, *strike*, *perform*, *throw*, *draw*, *take*, *make* and *pay* are used together with their collocates and pseudo-collocates. The collocates of these verbs were retrieved from the BNC, in the same fashion as for the COLLEX items. Z-scores were checked both for the intended real collocations as well as for the intended pseudo-collocations.

4 Pilot findings and results from test session 1

Pilot versions of COLLEX and COLLMATCH were administered to groups of university learners of English with 9-10 years of classroom exposure to English prior to entering university. Initially a 60-item COLLEX (version 1) was taken by a group (n =19) of teacher students, who studied English in their second year. The item content of the test was improved on the basis of the results of this pilot. A second pilot was taken by a group (n = 84) of 1st-term students of English. This test session contained COLLEX (version 2) and a small-scale trial of the COLLMATCH format. Based on the results, the two formats were further improved. This was done through analyses of reliability (internal consistency) and by item analysis. Poor items were either discarded and replaced, or changed in a way which was thought to make them function better. The dominating problem was items with zero variance, they were simply too easy for the tested population. Also, an unacceptably large number of items got low item-total correlation values.

In the first main test administration, a 50-item COLLEX (version 3), a 144-item COLLMATCH (version 1), and a 40-item test of single words were administered as a test battery to a total of 118 university students. The students were 1st, 2nd, 3rd, and 4th term learners of English at Lund University, and they voluntarily took the test battery in connection with lectures in the courses they were following. The three tests were administered in a pencil-and-paper form, and it was completed by most subjects after 20 minutes and by all after 30 minutes. A number of students reported on the test form that they were not native speakers of Swedish. The answer sheets of these learners were scored but not used in the further analysis of the results. The reason for why subjects who were not native speakers of Swedish were excluded was to allow for specific comparisons of groups of Swedish learners of English. Subjects who reported English as their native tongue were used as a small validation control group in the analysis. There were only three 4th term students, and these were also excluded from inclusion in the group comparison. All in all, this left 102 students. The 40-item single word test was used to control for the fact that some single words might not be known to the students, which may in turn affect their recognition of the collocations. It was assumed that words of very high

frequency would be known by the students, who where deemed to be advanced students of English. Therefore, only words that were intuitively expected to cause problems to these learners were selected for inclusion in the single word test. The words were selected by the researcher in collaboration with two senior university teachers of English. The single word test format was a 3-choice test, with the targeted English word together with three Swedish options.

In order to check whether the scores on COLLEX would be different if a Swedish prompt was inserted to the left of the collocation pairs, half of the Swedish learners received a bilingual version of the test, whereas the other half received the monolingual version. Since we cannot be sure that the testee has the same concept in mind as the test constructor, for the intended real collocation in each item, this move was worth trying. A bilingual COLLEX item thus looked like in figure 5.

1) (be en bön) say a prayer tell a prayer

Figure 5. A test item in the bilingual version of COLLEX 3.

The scoring of COLLEX was done in the following way. 1 point was awarded for each correctly chosen collocation in each item. If the pseudo-collocation was circled, if no collocation was circled, or if both collocations were circled, 0 was given. No correction for guessing was applied.

The results on the test battery turned out as follows. Firstly, the results on the 40-item single word test showed that the single words featured in the collocations in COLLEX and COLLMATCH were known to a great extent. The total mean score was 37.2 (n = 96) with the following submeans for the different groups: 1st-term students (35.5), 2nd-term students (38.2), and 3rd-term students (37.8). From this we conclude that insufficient core meaning knowledge of the constituent words in the subsequently tested collocations would not be a decisive factor determining our results.

The results obtained from the first main test administration are given in Tables 2, 3 and 4 below. As can be seen in Table 2, the six native speakers scored a mean of 48.5, in turn followed by the third term students (45.6), the second term students (43.5), and the first term students (40.4). Overall reliability of the test scores, as measured by Cronbach's alpha (see Bachman 2004) for internal consistency, was .83, which is satisfactory. However, as can be seen in Table 2, the reliability coefficients for the various groups were lower. The low reliability for the native speaker group is partly due to the fact that as many as 43 out of 50 items had zero variance.

COLLEX 3					
Group	**n**	**k**	**M**	**S.D.**	**reliability α**
1st term university students	39	50	40.4	5.8	.82
2nd term university students	37	50	43.5	4.4	.76
3rd term university students	20	50	45.6	3.1	.66
Native speakers	6	50	48.5	1.6	.50
Total	102	50	42.8	5.2	.83

Table 2. Results from administration of COLLEX 3, February 2005.

The unbalanced design (different group sizes) and unequal variance between the groups violated the assumptions of a regular ANOVA. For this reason, appropriate alternative tests were conducted. The analysis was done in the SPSS 11.5 statistical software. A Welch and a Brown Forsythe test signalled a highly significant effect of learner group affiliation on scores on the test. A subsequent Games-Howell test showed that there was a significant difference between 1st-term learners on the one hand and 3rd-term learners and native speakers on the other hand. There were no significant differences between 2nd-term and 3rd-term learners. The difference between 1st-term and 2nd-term learners was not significant, but it was very close to being so ($p = .056$), and is therefore interesting. The native speakers' scores were significantly different from all three Swedish learner groups.

There were no to very small differences between the means of the group of students who took the monolingual version of the test and the means of the group who took the bilingual version. This can be seen in Table 3, below:

COLLEX 3		
Group	**Monolingual version COLLEX Mean score**	**Bilingual version COLLEX Mean score**
1st term university students	40.6 (n = 19)	40.2 (n = 20)
2nd term university students	43.5 (n = 19)	43.4 (n = 18)
3rd term university students	45.1 (n = 10)	46.2 (n = 10)

Table 3. Comparison of group means on two different versions of COLLEX: monolingual and bilingual.

As can be seen in Table 3, only with the 3rd-term students could a small tangible difference be found (1.1). This suggests that the insertion of a Swedish prompt had very little effect on the scores. Thus, to a great extent, learners are able to

choose the more frequent form and reject the pseudo-collocation without getting the help of a Swedish prompt[6].

There is a tendency towards a ceiling effect in the test which might explain the absence of difference between the groups. The COLLEX 3 version might not be sensitive enough to pick up any actual differences in the construct measured. It could also be the case that testing learners who are only one term apart in terms of formal instruction will not yield any marked differences in the ability measured.

Moving on to the COLLMATCH 1 test, the format was scored in the following way. 1 point was given for each correctly chosen real collocation, as well as for each correctly rejected pseudo-collocation. Thus each of the 144-items was scored dichotomously, either 1 or 0. The test version consisted of 51 real collocations and consequently 93 pseudo-collocations.

The results of the COLLMATCH test, presented in Table 4 below, mirrored those of COLLEX 3 in that the highest scores were obtained by the six native speakers, followed in turn by 3rd-term, 2nd-term and 1st-term students.

COLLMATCH 1					
Group	**n**	**k**	**M**	**S.D.**	**reliability α**
1st term university students	39	144	116.2	8.6	.78
2nd term university students	37	144	122.2	6.4	.68
3rd term university students	20	144	124.4	4.6	.46
Native speakers	6	144	132.8	6.4	.83
Total	102	144	120.4	7.8	.80

Table 4. Results from administration of COLLMATCH 1, February 2005.

All groups scored relatively high, considering the many items of the test (k = 144).

The native speaker group scored 92% on the test which is high, and supports the validity of the test to an acceptable degree, but an even higher percentage might have been expected. The reason for the group not scoring even higher is believed to be the existence of poorly constructed items in the test, and possibly the fact that not even native speakers always know frequent collocations in their language. As pointed out by Bachman, "[t]he language use of native speakers has frequently been suggested as a criterion of absolute language ability, but this is inadequate because native speakers show considerable variation in ability" (1990:39). Furthermore, the inclusion of items in the test was based on findings in the 100 million word corpus BNC, which is a predominately written corpus (90% written texts, 10% spoken language transcriptions) of British English

[6] Britt Erman (personal communication) has suggested that the insertion of a Swedish prompt could actually be detrimental to the learners, since negative transfer from the L1 could affect their ability to select the correct collocation in the items negatively.

(Meyer 2002: 30). Native speakers of other varieties than British English have slightly other experiences of certain words and their collocates. For example, one of the native speaker respondents was of Canadian origin.

Overall reliability of the test scores was satisfactory at α .80. However, as with the COLLEX data, a low reliability coefficient was obtained for the 3rd-term students (α .46).

In terms of differences between the groups, an ANOVA test pointed to significant differences between them and a post-hoc Tukey test revealed the following: 1st-term students performed significantly differently from all the other groups. No significant difference could be established between 2nd-term students and 3rd-term students. Neither could any significant difference be established between 3rd-term students and native speakers, even though the difference was very close to being significant.

Summing up, we may conclude that scores on COLLEX 3 and COLLMATCH 1 seem to increase with length of exposure, and native speakers score very close to maximum. However, tendencies for ceiling effects were observed, and differences between groups were not in all cases significant. Overall reliability coefficients were acceptable at α .80-.83. However, low reliabilities for the more advanced Swedish learners were obtained. Reasons for this were believed to be homogeneously high scores and possibly that a large number of test items discriminated poorly between students with higher and lower total scores, respectively, for this group.

5 Results from test session 2

In this study, in addition to university students, two whole classes of upper-secondary-school students, 10th graders and 11th graders, judged to be low intermediate to high intermediate learners of English, were subjected to further developed and modified versions of COLLEX and COLLMATCH. The modifications were based on analyses of the results from test session 1 reported above. Item analyses pointed at a number of weak items and these were either discarded and replaced by new ones, or improved. The total number of students in the study was 188. As part of the test battery, all subjects also took the Vocabulary Levels Test, a test of English vocabulary size (see Nation 1990 and 2001, and Schmitt et al. 2001). This was done in order to investigate the degree of relationship between a vocabulary size measure and the two collocation knowledge measures. For the university students, it was possible to administer the whole test battery as the obligatory departmental vocabulary exam, given at the end of term. In this exam, primarily 2nd and 3rd-term university students participated, and only a small number of 1st-term students took the test (n = 7).

The 188 students were subjected to the following test battery:

a) Version 1 of the Vocabulary Levels Test (150 items).
b) COLLEX 4 (50 items)
c) COLLMATCH 2 (100 items; new format design)

The use of the new COLLMATCH 2 format needs to be described further. There were some obvious drawbacks with the COLLMATCH 1 format. One was that although as many as 144 items were tested, only 51 were real collocations. This meant the test primarily measured learners' ability to reject pseudo-collocations (65%), rather than their ability to recognize real collocations (35%). The large number of pseudo-collocations was a result of the format per se, i.e. the grid with three verbs and six shared potential objects. It was difficult to find objects that fit with two or all three of the verbs, and this meant that a majority of the points of intersection in the grid were not intended to be ticked as real collocations. The format also invited the potential inclusion of combinations on which not even native speakers agreed as to their acceptability.

For the above reasons, a modified format was constructed. Twenty high-frequency verbs, all taken from the first thousand most common words of English according to frequency counts based on the BNC (Kilgarriff 1996), were checked for frequent collocates. The 20 verbs were *have, do, make, take, give, keep, hold, run, set, lose, draw, say, break, raise, bear, serve, catch, pull, throw*, and *drop*. For each of the 20 verbs, five test items, consisting of the verb + a NP, were constructed. The NP was either a bare noun or an article plus a noun. A varying number of the 5 items was made up by a verb plus a pseudo-object. In the new version, the 100-item COLLMATCH 2 consisted of 65 real collocations and 35 pseudo-collocations. As a result, the new format measures learners' recognition knowledge of real collocations to a greater extent than the old format. A verb row of five items is illustrated below in Figure 6:

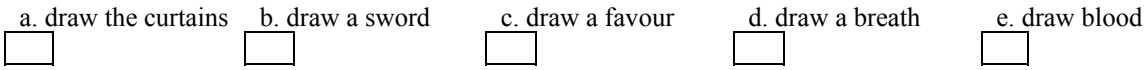

Figure 6. A modified test item format in COLLMATCH 2.

The task for the learner taking the test is to tick the collocations they think exist in the English language, and leave the boxes of the non-existing collocations blank.

The university students taking the test (n =134) had a maximum of 3 hours to complete the test battery, which for an overwhelming majority of the students was ample time. A majority of the students handed in after 60 to 90 minutes. Out of the 134 students, 5 students used the full 3 hours to complete the test form. A majority of the upper-secondary-school students who took the test (n = 54) completed the form in 40 minutes. A few students finished and handed in after 60 minutes. The big difference in time spent on the test was primarily due to the fact that the test battery constituted an end of term exam, a high-stakes

event, for the university students, a fact that meant that many students took their time, and double-checked their answers several times before handing in. For the upper secondary school students, the test session had no impact on their grades. The test was run in class at the end of term, after the final grades had been presented to the students.

Starting with the scores on the vocabulary size measure, the following results were obtained in the test administration:

The Vocabulary Levels Test (VLT), version 1					
Group	**n**	**k**	**M**	**S.D.**	**reliability α**
10th graders	26	150	95.3	17.1	.93
11th graders	28	150	80.4	20.2	.95
1st term university students	7	150	129.0	10.6	.90
2nd term university students	91	150	140.5	7.6	.89
3rd term university students	36	150	140.8	5.5	.81
Total	188	150	125.2	26.6	.98

Table 5. Results on Vocabulary Levels Test, test session 2.

As can be seen in Table 5, scores on the Vocabulary Levels Test increase with higher level of study/instruction, with the exception of 10th graders who scored better than 11th graders throughout. Also, only a minuscule difference could be observed between mean scores of 2nd and 3rd-term university students. Significant differences at $p < .05$ were observed between the 10th graders and the 11th graders, and between these two and all three university student groups. No significant differences were found between the three university students groups.

The administration of the vocabulary size measure provided excellent total reliability coefficients. Cronbach's alpha was estimated at α .98. The subgroups varied between α .81 and α .95. These coefficients are in line with earlier reported reliability values obtained for learner scores on the test (see Schmitt et al. 2001).

COLLEX 4					
Group	**n**	**k**	**M**	**S.D.**	**reliability α**
10th graders	26	50	29.9	5.1	.64
11th graders	28	50	28.6	4.1	.45
1st term university students	7	50	34.5	6.7	.81
2nd term university students	91	50	43.8	4.7	.81
3rd term university students	36	50	44.2	3.3	.64
Total	188	50	39.3	8.0	.91

Table 6. Results on COLLEX 4, test session 2.

Table 6 shows that there was a clear difference in performance on COLLEX 4 between upper secondary school students and university learners, as was tentatively predicted in subsection 3.4 above.

The 3rd-term students scored the highest mean (44.2), followed by the slightly lower mean score for 2nd-term students (43.8). The small group of 1st-term learners scored considerably lower, with a mean score of 34.5. As in the vocabulary size measure, the 10th graders scored slightly higher means than the 11th graders. When analyzed through a Games-Howell post hoc test, the observed differences were significant between 10th graders and 2nd and 3rd-term students, respectively. A significant difference was also observed between 11th graders and 2nd and 3rd-term students, respectively. Finally, a significant difference was also found between the scores of the 3rd-term and the 1st-term university students. All differences were reached at $p < .05$.

The overall scores were highly reliable with a internal consistency of α .91. As can be seen in the reliability column in the table, the coefficients for the 10th and 11th graders' scores, together with the university 3rd-term students' scores, were low (.64, .45 and .64). The reason for the high proportion of measurement error in the scores of the upper secondary school students is believed to come from a great deal of guessing. If there is much guessing, then this results in a lot of variance that is unsystematic. The measure will not reflect their true ability. Looking closer at the item-total correlation values for the 50 tested items in the scores of the 11th graders group (n = 28), we see that as many as 17 out of the 50 items, almost 40%, have negative values. This means that on these items, many learners with low total scores on the test gave correct answers, whereas learners with high total scores gave wrong answers. Clearly the test does not discriminate well between learners of different abilities in this group. All of these observations points to guessing as a highly probable cause. In the scores of the 10th graders (n = 26), this negative trend is not so strong but we find 8 items with negative values. As for the scores of the 3rd-term university students, we find 5 items with negative item-total correlations. In their case, the low overall reliability is also believed to stem from high and homogeneous group scores.

COLLMATCH 2					
Group	**n**	**k**	**M**	**S.D.**	**reliability α**
10th graders	26	100	62.1	8.6	.79
11th graders	28	100	60.5	7.5	.71
1st term university students	7	100	71.5	11.4	.90
2nd term university students	91	100	84.3	7.3	.83
3rd term university students	36	100	84.5	5.7	.73
Total	188	100	77.2	12.7	.92

Table 7. Results on COLLMATCH 2, test session 2.

The scores on the COLLMATCH 2 test, displayed in Table 7 above, mirrored those both on the Vocabulary Levels Test and COLLEX 4. Again, the 10th-graders scored better than the 11th graders (62.1 compared to 60.5). The small group of 1st-term university learners scored a mean of 71.5, and almost no difference was observed between the means of the 2nd term and 3rd-term university students.

The observed differences between the means of the groups were significant at $p < .05$ except for 10th-graders and 11th-graders, and 2nd-term and 3rd-term university students.

The overall reliability of the new version of the test was found to be very high at α .92. The coefficient values for the different groups were lower, ranging between α .71 and α .90. These values are all acceptable, but they might still be a bit low considering the large number of items in the test (k = 100).

If analysing only the reliability of the scores on the 65 real collocations, the data is highly reliable at α .92. An analysis of the 35 pseudo-collocations yields a reliability coefficient of α .76. Thus, the students' ability to recognize real collocations was more reliably measured than their ability to reject pseudo-collocations.

The score distributions of the three tests were all negatively skewed. Tests of normality of distribution confirmed that all three sets of test scores were significantly non-normal. In order to find out the relation between the scores on the Vocabulary Levels Test and the scores on COLLEX 4 and COLLMATCH 2 a correlation analysis was carried out. Since the data violated the parametric assumption of normally distributed data, a Spearman's correlation coefficient test was used. The students' scores on the three variables were first rank-ordered and then correlated in SPSS. A positive correlation was expected and therefore a one-tailed test was used. The following coefficients were found:

Inter-test correlations			
	Vocabulary Levels Test	COLLEX 4	COLLMATCH 2
Vocabulary Levels Test		.87	.90
COLLEX 4	.87		.89
COLLMATCH 2	.90	.89	

Table 8. Spearman correlation coefficient test (Spearman's Rho) between the different test parts. N = 188. All values significant at p < .01, one-tailed test.

As can be seen in Table 8, the three sets of scores showed very high positive correlations. The correlation between the vocabulary size measure, The Vocabulary Levels Test, and the COLLEX collocation measure, was 0.87, and for the COLLMATCH collocation measure it was 0.90. The correlation between the two collocation measures was observed at 0.89.

That the three variables would correlate positively was to some extent expected considering that a large vocabulary will have a positive influence on

practically any measured language ability. The very high positive correlation observed was however somewhat surprising. In section 2, in one of the reviewed studies, Bahns & Eldaw (1993) concluded that collocation knowledge does not develop alongside general lexical knowledge. Firstly, the review identified the method employed to reach this conclusion as an invalid one. Secondly, the current results point in the opposite direction. A correlation coefficient does not however tell us anything about causation. Thus, we cannot say that the results show that a large vocabulary causes high scores on the collocation tests. However, as Meara (1996: 37) puts it: "All other things being equal, learners with big vocabularies are more proficient in a wide range of language skills than learners with smaller vocabularies, and there is some evidence to support the view that vocabulary skills make a significant contribution to almost all aspects of L2 proficiency". The current results seem to support this view. Learners with large vocabularies seem to perform well on the receptive recognition tests of collocations used in this study. This could be interpreted to mean that having a large vocabulary also entails knowing a large number of, in this case, verb + NP collocations. Inversely, the results seem to suggest that learners with smaller vocabularies are not proficient in recognizing real collocations and rejecting pseudo-collocations.

The fact that the two collocation tests, COLLEX and COLLMATCH, showed a high level of positive correlation most likely stems from the fact that they basically test the same type of items: frequent word verb + noun collocation. The high correlation between the two tests (.89) tells us that they can be seen to measure the same underlying ability.

6 Concluding discussion

The results presented above from two administrations of the COLLEX and COLLMATCH formats have given rise to several observations.

Firstly, measuring receptive recognition knowledge of verb + NP collocations can be done seemingly reliably. High to very high overall reliability coefficients were observed in the two formats. However, slightly lower values were obtained for certain learner subgroups.

Secondly, differences could be observed between different learner groups taking the two test formats. These differences were however not always statistically significant. With one exception, scores on the two tests seemed to increase with length of exposure to English. In effect, the more advanced level the learners were studying at, the better they performed on the tests. This meant that native speakers of English scored higher results than Swedish university and upper-secondary-school learners; that 3rd-term university learners scored higher mean results than 2nd-term learners, who in turn scored higher mean results than 1st-term learners. The 1st-term learners performed better than upper-secondary-

school students. The anomaly in the scale was found when comparing 11th-graders with 10th-graders, where the latter performed better than the former. That the 10th-graders were more proficient overall in English than the 11th graders was confirmed by the teacher of the two groups, supported by their respective results on a nation-wide proficiency test of English. On the whole, the lack of significant differences observed between some of the learner groups is believed to stem from the fact that only 1 term, 5 months, separate the groups. It may be that this small difference seldom gives rise to statistically significant differences in performance on tests like those described here. This could mean that this type of knowledge doesn't develop in such a way that a difference is measurable. Another interpretation is that the two test formats used are not sensitive enough to pick up any existing differences.

The fact that the most advanced Swedish university learners performed very close to the native speakers' scores on the tests suggests that the former, in terms of receptive recognition knowledge of verb + NP collocations, have developed near-native speaker skills. It is probable that a productive measure would show bigger differences between the two groups.

Thirdly, in addition to the two collocation measures, a test of vocabulary size was included in the test battery of the second administration. The measure used was Nation's (1990) Vocabulary Levels Test. The scores on the size test were highly reliable in terms of internal consistency. As to the results, Swedish university learners scored higher than upper secondary school learners. This difference in vocabulary size was expected. Small to no differences were observed between the two more advanced subgroups of university learners. The reason for this lack of difference could be that vocabulary size does not develop to the extent that a difference emerges when the tested learners are only one term apart. As pointed out by Meara & Wolter, the lexicons of different learners may be in different growth and consolidation phases: "…we might find learners with similar vocabulary sizes, but very different degrees of organisations in their lexicons" (2004:95). In the test administrations reported here, however, no significant differences in receptive collocation knowledge were discernible between groups who scored similar results on the vocabulary size measure. Another interpretation of the size measure scores could be that the Vocabulary Levels Test is not sensitive enough to pick up any existing differences among very advanced learners.

Fourtly, a correlation analysis was carried out in order to investigate the relationship between the Vocabulary Levels Test and COLLEX and COLLMATCH, respectively. That the variables would correlate was expected considering that a large vocabulary will have a positive influence on practically any measured language ability. The very high positive correlation observed was however somewhat surprising. The result can be interpreted in a number of ways. Learners with large vocabularies seem to perform well on the receptive recognition tests of collocations used in this study. This could be interpreted to

mean that having a large vocabulary also entails knowing a large number of, in this case, verb + NP collocations. Inversely, the results seem to suggest that learners with smaller vocabularies are not proficient in recognizing real collocations and rejecting pseudo-collocations.

In terms of further research on the COLLEX and COLMATCH formats, more qualitative analyses looking at what type of collocations the different groups answered correctly or incorrectly may be carried out. These analyses could show whether support verb constructions are easier or more difficult than non-support verb constructions (see Nesselhauf 2004); whether collocation frequency is a factor, in terms of absolute frequency of words and/or significance of co-occurrence, z-score; whether type of collocation is a factor: free, restricted, figurative (cf. Barfield 2003, reviewed in section 2 of this paper, and Howarth 1998). Analyses will also be carried out looking at the possibility of an existing threshold score in a vocabulary size measure determining the score on either of the collocation knowledge tests.

As a means to investigate test validity, think-aloud protocol analyses of learners taking the tests will be carried out. Such analyses can provide important insight into what strategies learners use as they answer the test items in the test. For example, do learners mostly know what to answer, do they think they know, do they employ some sort of test-wiseness or do they simply resort to guessing to a high extent? Further research on COLLEX and COLLMATCH should shed light on these issues.

7 References

Altenberg, B. 1993. Recurrent Verb-complement Constructions in the London-Lund Corpus. In *English Language Corpora: Design, Analysis and Exploitation.* eds. J. Aarts, P. de Haan, and N. Oostdijk, 227-245. Amsterdam: Rodopi.

Altenberg, B and Granger, S. 2001. The Grammatical and Lexical Patterning of MAKE in Native and Non-native Student Writing. *Applied Linguistics* 22/2: 173-195.

Bachman, L.F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L.F. 2004. *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Bahns, J. and Eldaw, M. 1993. Should we teach EFL students collocations? *System,* 21: 101-114.

Barfield, A. 2003. *Collocation Recognition and Production: Research Insights*. Chuo University, Japan.

Barnbrook, G. 1996. *Language and Computers*. Edinburgh: Edinburgh University Press.

Biskup, D. 1992. L1 influence on learners' renderings of English collocations. A Polish/German empirical study. In *Vocabulary and Applied Linguistics*, eds. P.J.L. Arnaud and H. Béjoint, 85-93, London: Macmillan.

Bonk, W.J. 2001. Testing ESL Learners' Knowledge of Collocations. In *A Focus on Language Test Development: Expanding the Language Proficiency Construct Across a Variety of Tests. (Technical Report #21)*, eds. T. Hudson and J.D. Brown, 113-142, Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Brown, F.G. 1983. *Principles of Educational and Psychological Testing.* New York: Holt, Rinehart and Winston.

Cameron, L. 2002. Measuring Vocabulary Size in English as an Additional Language. *Language Teaching Research* 6/2: 145-173.

Carter, R. 1998. *Vocabulary*. London: Routledge.

Chapelle, C. 1998. Construct definition and validity inquiry in SLA research. In *Interfaces Between Second Language Acquisition and Language Testing Research*, eds. L. Bachman & A. Cohen, 32-70, Cambridge: Cambridge University Press.

Coxhead, A. 2000. A new academic word list. *TESOL Quarterly* 34: 213-239.

Ellis, N. 2001. Memory for Language. In *Cognition and Second Language Instruction*, ed. P. Robinson, 33-68, Cambridge: Cambridge University Press.

Eyckmans, J. 2004. *Measuring Receptive Vocabulary Size*. Utrecht: LOT.

Farghal, M. & Obiedat, H. 1995. Collocations: A neglected variable in EFL. *International Journal of Applied Linguistics* 28(4): 313-331.

Gitsaki, C. 1996. The development of ESL collocational knowledge. PhD thesis. University of Queensland.

Granger, S. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In *Phraseology: Theory, analysis, and applications*, ed. A.P. Cowie, 145-160, Oxford: Oxford University Press.

Henning, G. 1987. *A Guide to Language Testing*. Boston: Heinle & Heinle.

Howarth, P. 1996. *Phraseology and Second Language Proficicncy*. Tübingen: Max Nietnayer Verlag.

Howarth, P. 1998. The Phraseology of Learners' Academic Writing. In *Phraseology: Theory, analysis, and applications*, ed. A.P. Cowie, 161-186, Oxford: Oxford University Press.

Kilgarriff, A. 1996. BNC database and word frequency list. http://www.itri.brighton.ac.uk/ ~Adam.Kilgarriff/bnc-readme.html.

Källkvist, M. 1999. *Form-Class and Task Type Effects in Learner English.* Lund: Lund University Press.

Laufer, B. and Goldstein, Z. 2004. Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness. *Language Learning* 54: 399-436

Meara, P. 1996. The dimensions of lexical competence. In *Performance and Competence in Second Language Acquisition*, ed. G. Brown, K. Malmkjaer and J. Williams, 35-53. Cambridge: Cambridge University Press.

Meara, P. and Buxton, B. 1987. An alternative to multiple choice vocabulary tests. *Language Testing* 4: 142-154.

Meara, P. and Milton, J. 2003. *X_Lex: the Swansea Vocabulary Levels Test*. Swansea: Lognostics.

Meara, P. and Wolter, B. 2004. V_Links: beyond vocabulary depth. In *Angles on the English-speaking world 4*, eds. D. Albrechtsen, K. Haastrup and B. Henriksen, 85-96, Copenhagen: Museum Tusculanum Press.

Meyer, C.F. 2002. *English Corpus Linguistics*. Cambridge: Cambridge University Press.

Mochizuki, M. 2002. Exploration of two aspects of vocabulary knowledge: Paradigmatic and collocational. *Annual Review of English Language Education in Japan* 13:121-129.

Nation, I.S.P. 1990. *Teaching and Learning Vocabulary*. New York: Heinle and Heinle.

Nation, I.S.P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nesselhauf, N. 2004. How learner corpus analysis can contribute to language teaching: A study of support verb constructions. In *Corpora and Language Learners,* eds. G. Aston, S. Bernardini, and D. Stewart, 109-124, Amsterdam: John Benjamins.

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

Newell, A. 1990. *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Read, J. 1993. The development of a new measure of L2 vocabulary knowledge. *Language Testing* 10: 355-371

Read, J. 1998. Validating a Test to Measure Depth of Vocabulary Knowledge. In *Validation in Language Assessment*, ed. A. Kunnan, 41-60, Mahwah, NJ: Lawrence Erlbaum.

Schmitt, N. 1998. Tracking the Incremental Acquisition of Second Language Vocabulary: A Longitudinal Study. *Language Learning* 48: 281-317.

Schmitt, N. 2000. *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

Schmitt, N., Schmitt, D. and Clapham, C. 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing* 18/1: 55-88.

Stubbs, M. 2001. *Words and Phrases*. Oxford: Blackwell.

Warren, B. 2005. A Model of Idiomaticity. *Nordic Journal of English Studies* 4, 1.

Wesche, M and Paribakht, T.S. 1996. Assessing vocabulary knowledge: depth vs. breadth. *Canadian Modern Language Review* 53/1: 13-40.
West, M. 1953. A General Service List of English Words. London: Longman.
Wiktorsson, M. 2003. *Learning Idiomaticity: A Corpus-Based Study of Idiomatic Expressions in Learners' Written Production*. Stockholm: Almqvist&Wiksell International
Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Henrik Gyllstad
henrik.gyllstad@englund.lu.se