Words that go together: Measuring individual differences in native speakers' knowledge of collocations

Ewa Dąbrowska[1]

[1]Northumbria University

Abstract

Although formulaic language has been studied extensively from both a linguistic and psycholinguistic perspective, little is known about the relationship between individual speakers' knowledge of collocations and their linguistic experience, or between collocational knowledge and other aspects of linguistic knowledge. This is partly because work in these areas has been hampered by lack of an adequate instrument measuring speakers' knowledge of collocations. This paper describes the development of such an instrument, the "Words that go together" (WGT) test, and some preliminary research using it. The instrument is a multiple choice test consisting of 40 items of varying frequency and collocation strength. The test was validated with a sample of 80 adult native speakers of English. Test-retest reliability was 0.80 and split-half reliability was 0.88. Convergent validity was established by comparing participants' scores with measures expected to correlate with language experience (print exposure, education, and age) and other linguistic abilities (vocabulary size, grammatical comprehension); divergent validity was established by comparing test scores with nonverbal IQ. The results of the validation study are then used to compare speakers' performance on the WGT with corpus-based measures of collocation strength (mutual information, z-score, t-score and simple frequency); however, no statistically reliable relationships were found.

Keywords: collocation, language testing, formulaic language, age, individual differences, psychological reality, frequency

Corpus linguistic research (e.g. Ellis & Simpson-Vlach, 2009; Erman & Warren, 2000; Sinclair, 1991, 2004; Wray, 2002) has shown that virtually all texts contain a significant proportion of collocations, or recurrent clusters – combinations of words which are mostly regular (i.e. conform to the rules of grammar and have compositional meanings) but occur together much more frequently than one would predict given the frequencies of the words themselves. What is the mental status of such combinations? Some linguists (e.g. Bley-Vroman, 2002; Herbst, 1996) regard them as purely epiphenomenal: some words tend to co-occur relatively frequently because their referents often co-occur in the real world: for instance, *dark night* is a collocation because nights are dark. Others (e.g. Ellis and Simpson-Vlach, 2009; Sinclair, 1999) point out that many collocations are semi-idiomatic (cf. *break the law, keep a diary*) and not entirely predictable (as demonstrated by the fact that they are an area of particular difficulty for second language learners), and conclude that speakers store them as units.

However, although there is considerable evidence that recurrent word combinations are processed faster (see e.g. Arnon & Snider, 2010; Conklin & Schmidt, 2008; Tremblay, Derwing, Libben, & Westbury, 2011; Tremblay & Tucker, 2011; Siyanova-Chanturia, Conklin & van Heuven, 2011), attempts to test the psychological reality of the corpus-derived measures of association strength have not been particularly impressive: the correlations between corpus measures and speaker performance tend to be weak, and different studies have produced contradictory results. Hodgson (1991, cited in Durrant & Doherty, 2010) found priming for collocations in a lexical decision task but not in a naming task. Ellis and Simpson-Vlach (2009) examined the degree to which formulaic status facilitated processing. They found significant effects of mutual information (MI) – a widely used measure of association strength – but not of raw frequency in three different tasks. However, the observed effects were very small, with cluster length and MI together accounting for 8%-16% of the variance in different experiments. Schmitt, Grandage and Adolphs (2004) used an elicited imitation task in which participants were asked to repeat segments of a story which contained recurrent clusters of varying lengths, and examined how accurately speakers reproduced the clusters. They found vast differences between clusters: some (*go away, I don't know what to do*) were reproduced verbatim by

nearly all participants, while others (*in the same way as, aim of this study*) were nearly always either omitted or replaced by a different phrase. There was no relationship between corpus frequency and experimental performance. Finally, Durrant and Doherty (2010) used a lexical decision task to examine the effect of association strength on processing of words in collocations. Participants were primed with the first word of a collocation (e.g. *foreign*) and had to decide as quickly as possible whether or not the second word (e.g. *debt*) was a real word. There were four experimental conditions: rare combinations (MI<2), moderate collocations (MI between 4 and 5), frequent collocations (MI > 6), and associated frequent collocations (combinations with an MI > 5.5 which were also psychological associates, i.e., had a high ranking in lexical association norms). They found statistically significant effects only for frequent collocations and associated frequent collocations, suggesting that collocational priming may be restricted to words with very high MI scores (MI > 6, t > 7.5). In a second experiment, they used exactly the same stimuli and method, except that the prime was presented for only 60 ms, which is too short for participants to become aware of it. This time, they found significant facilitation only for the associated frequent collocations, and, as in the Ellis and Simpson-Vlach study, the effect sizes were very small. Moreover, rare combinations and moderate collocations had very similar effect sizes and approached significance. Nevertheless, Durrant and Doherty concluded that "priming between associates may be different in type from that between collocates, with the former controlled by automatic processes and the latter by strategic" (2010: 144).

One of the reasons why there is so much disagreement about the mental status of collocational knowledge is because we don't have an adequate instrument for assessing it. This paper describes the development of such an instrument and some preliminary research using it which examines the relationship between individual speakers' knowledge of collocations and their linguistic experience and between collocational knowledge and other aspects of linguistic knowledge. A second study examines the relationship between native speakers' performance on the collocations test and corpus-based measures of collocation strength.

Words that Go Together

"Words that go together" (WGT) tests receptive knowledge of collocations using a multiple choice format. Participants are presented with sets of five phrases and asked, for each set, to select one phrase that "sounds the most natural or familiar". Examples of test items are given in (1) below, with the target collocation marked with an X.

(1)	X blatant lie	X boost production	X odd remark
	clear lie	double production	peculiar remark
	conspicuous lie	enlarge production	queer remark
	distinct lie	extend production	unnatural remark
	recognizable lie	redouble production	weird remark

Admittedly, selecting a familiar phrase from a list is not a very natural task. However, since it involves simple recognition, it is less prone to task demands than other methods. Moreover, because the phrases are presented out of context, there are no context effects. This means that such a test will not tell us very much about the dynamics of normal language processing; but it will allow us to isolate and measure a particular aspect of linguistic knowledge.

An initial list of 80 items was extracted from a collocations dictionary (Douglas-Kozłowska & Dzierżanowska, 2004). Half of the items consisted of a verb followed by an abstract noun (e.g. *arouse suspicion, raise standards*), and the other half of an adjective followed by an abstract noun (*bitter dispute, full confession*). Abstract nouns were used in order to avoid combinations which could be epiphenomenal, since abstract nouns tend to be more idiosyncratic in their collocational preferences than concrete nouns. It is also much easier to construct good foils for abstract nouns. All the items chosen had fairly regular meanings: highly idiomatic combinations such as *white lie* were excluded. The candidate items' collocational status was verified using the British National Corpus. Only items

with an overall frequency of at least 5 AND an MI (mutual information) score of at least 4 were retained.

Next, 6-8 candidate foils were constructed for each item by replacing the adjective or verb with a synonym, or by taking another collocate of the noun and replacing it with a synonym. To ensure that none of the foil phrases were collocations, all the foils were tested against the collocations dictionary and the BNC; all the items which were either listed in the dictionary or whose MI was greater than 2 were deleted from the list. The remaining foils were checked by two native speakers of English who were asked to delete any items which were semantically or pragmatically implausible.

62 of the initial 80 items survived the selection process with at least four foils; these items were piloted with a group of 67 undergraduate students. The results of the pilot study revealed problems with a few of the items: either the majority of the participants chose one of the foils rather than the target collocation, or the item-whole test correlation was negative. All of these items were removed. The final test comprised 40 items (20 adjective-noun and 20 verb-noun collocations) selected so that mutual information and frequency did not correlate. The items were ordered from easiest to most difficult according to the pilot test results. A summary of the target item characteristics is provided in Table 1; further details can be found in Appendix 1 and the test itself in Appendix 2.

INSERT TABLE 1 HERE

There is some disagreement in the literature about what constitutes a collocation. One common criterion is a t-score of at least 2 or an MI score of 3 or more and a minimum frequency of 3-5 per 100 million (Hunston, 2002; Stubbs, 1995). As can be seen from the table, even the least frequent collocations used in the study exceed all three of these criteria. As explained earlier, they were taken from a dictionary of collocations and prefered over the foils by the majority of the pilot study participants, so their collocational status was confirmed by human judges. Thus, the target items meet virtually everyone’s definition of collocation.

Validation Study

The only way to learn that a particular combination of words constitutes a collocation is through repeated exposure to that particular combination. Thus, an adequate measure of collocation knowledge should correlate with language experience, and hence age (since, other things being equal, older speakers will have been exposed to more language than younger speakers), education (since more educated speakers will have experienced more varied language during their school/university years and are likely to have jobs which involve more and more varied linguistic interactions), and reading habits (since written language tends to be more complex than spoken language, and since skilled readers absorb more language per unit of time than skilled conversationalists).[1] Furthermore, a measure of collocation knowledge should correlate with other measures of linguistic proficiency, and especially vocabulary size, which to some extent depends on collocational knowledge (see Dąbrowska, 2009). Therefore, to establish convergent validity, the following study examines the correlation between participants' scores on WGT and various measures of language exposure and proficiency. Divergent validity is established by assessing non-verbal IQ.

Participants

80 adult native speakers of British English (37 males and 43 females) aged from 17 to 65 (mean age 38, median 32) were recruited through advertisements in the local press, church groups, schools and personal contacts. The participants came from a variety of educational backgrounds. 5 participants (i.e., 6% of the sample) had no formal qualifications; 54 (68%) held a secondary school and/or a vocational certificate (UK National Qualifications Framework levels 1-3); 11 (14%) held an undergraduate degree or were studying for one; and 10 (13%) held a postgraduate degree. The distribution of qualifications roughly reflects that of the general UK population, although the proportion of participants with postgraduate degrees was somewhat higher and the proportion of participants with no formal qualifications somewhat lower. The number of years spent in full-time

education ranged from 10 to 21 (mean 13.8, median 13). Thirteen of the participants were in full-time education, 3 were housewives, 13 retired and 5 unemployed; the rest were more or less evenly distributed between manual (e.g. cleaner, shop assistant, waitress, roofer, nail technician), clerical (office workers, IT support, etc.) and graduate-level jobs (e.g. teacher, web designer, quantity surveyor).

Procedure

Participants were presented with sets of five phrases and asked, for each set, to select one phrase that "sounds the most natural or familiar". They were given two examples (see (2) below) in which the target answer was marked with an X. Since collocational knowledge is largely implicit, people often claim that they do not know the answer even when objective measures demonstrate above chance performance; therefore, in order to ensure that the task offered as accurate a reflection of participants' knowledge as possible, they were asked to guess when they were not sure. Participants were tested individually in a quiet room. Each participant was given a written copy of the test and was asked to follow along as the experimenter read each item out loud. This dual presentation method was adopted to ensure that with low literacy participants were not disadvantaged; it also ensured that participants provided an answer to every question. The experimenter recorded the participants' responses; if a participant hesitated, the experimenter reminded them that they should follow their "gut feeling" and guess when they were not sure. Participants were given as much time as they needed, but typically completed the test within about 10 minutes.

(2)	delicate tea	X deliver a speech
	feeble tea	hold a speech
	frail tea	perform a speech
	powerless tea	present a speech
	X weak tea	utter a speech

In addition to the collocations test, participants also completed a grammatical comprehension test ("Pictures and Sentences", Dąbrowska, unpublished), a vocabulary test (a modified version of the Vocabulary Size Test, Nation & Beglar, 2007), two non-linguistic tests: a test of print exposure (the Author Recognition Test, Acheson, Wells, & MacDonald, 2008) and a nonverbal IQ test (Shipley-2 Block Design, Shipley, Gruber, Martin, & Klein, 2009), and a background questionnaire which included questions about their reading habits (see Dąbrowska, in preparation for further details).

To obtain measures of test-retest reliability, 30 of the participants were asked to return to the testing site three to six months later. They only took the three language tests during the second testing round.

Results and Discussion

Individual scores ranged from 11 (28%) to 39 (98%) correct (mean 29.5, median 20, *SD* 6.2). Note that chance performance would be 8 points (20%); hence, all participants performed above chance. Information about the percentile distribution of individual scores is presented in Appendix 3. Results for individual items, along with other item characteristics, are presented in Appendix 1. As can be seen from these figures, items range in difficulty, with the easiest ones (*blank expression*) being correctly identified by 98% of the participants, and the most difficult one (*striking example*) being correctly identified by only 29%.

Item discrimination was analysed using the standard procedure described by Weir (2005). Participants were ranked according to their scores on the test and the top third and the bottom third were assigned to the high group and low group respectively. The item discrimination index was computed using the following formula:

$$IDis = H_c/(H_c + L_c)$$

where H_c is the number of correct responses in the high group and L_c is the number of correct responses in the low group. The results for individual items are given in Appendix 1. Item

discrimination indices range from 0.51 to 0.90. Fifteen items are relatively easy, with 80% or more of the participants supplying the correct answer and correspondingly low item discrimination indices (between 0.51 and 0.61). This is because, as explained earlier, only items for which at least half of the pilot study participants chose the correct answer were included in the test. This was done mainly to ensure that all phrases were psychologically real collocations. A further advantage is that the test can be used with other populations: advanced L2 learners, older children, or individuals with a language impairment. Since the items on the test are arranged in order of difficultly, the larger number of easy items at the beginning helps the testees to grasp the nature of the task.

It should also be noted that for three items, *(memorable) phrase*, (*refuse an) application* and (*striking) example*, somewhat less than half of the validation study participants (48%, 45% and 29% respectively) chose the target answer. It was decided to retain these items, however, because of their very high discrimination index (see Appendix 1).

Test/retest reliability was 0.80 and split-half reliability (Spearman-Brown corrected) was 0.88. Both of these figures are well above the standard reliability criterion for research instruments (0.70). They are, however, substantially lower than the corresponding figures for objectively scored receptive vocabulary tests developed for clinical and educational use. For instance, the split-half reliability of the Peabody Picture Vocabulary Test (Dunn & Dunn, 1997) is 0.94, and its test-retest reliability 0.92. This is to be expected, since in such tests the distractors are clearly incorrect, while on the WGT, all five options are possible phrases, and hence more difficult to rule out.

Information about correlations between scores on WGT and measures of language proficiency and exposure is presented in Table 2. Performance on WGT shows a robust correlation with vocabulary size and somewhat weaker, though still highly significant, relationship with performance on the grammatical comprehension test. (Grammar and vocabulary are also correlated: $r = 0.44$, $p < 0.001$). It also correlates with all four measures of language exposure, viz. the Author Recognition Test, self-reported reading, the number of years in full-time education, and age.

INSERT TABLE 2 HERE

The relationship between test scores and age shows an interesting pattern. This can be seen in Figure 1, which shows Loess (smoothed) curve of performance on WGT as a function of age. Performance on the vocabulary and grammar tests has also been provided for comparison. As we can see, collocation test scores increase sharply until about age 32, then level off and begin to decrease from about 50. This is confirmed if we compute separate correlations for the two age groups: for the under-32s, $r = 0.62$, $p < 0.001$, N=39; for over-32s: $r = -0.26$, $p = 0.10$, $N = 41$. Interestingly, the other two language measures show a different pattern: receptive vocabulary increases steadily until about 55 and then levels off, and performance on the grammatical comprehension test remains stable throughout the entire period.

INSERT FIGURE 1 HERE

Finally, divergent validity was established by computing the correlation between the collocation test score and non-verbal IQ (measured using Shipley's Block Design). The correlation was not significant ($r = 0.19$, $p = 0.099$), and the correlation coefficient approached 0 ($r = -0.009$) once vocabulary size was controlled for. This confirms that the test is not simply measuring general ability or cooperativeness.

The results discussed above confirm that the “Words that go together” test is a reliable measure of speakers’ knowledge of English collocations. Furthermore, the fact that its results correlate well with measures of language exposure and other measures of language proficiency but not with non-verbal IQ strongly suggests that it is a valid test, i.e., it measures what it is intended to measure, although of course further validation comparing performance on WGT and other measures of collocational knowledge would be desirable.

Relationship between WGT Scores and Corpus Measures of Collocation Strength

The second goal of this study was to examine the relationship between native speakers' knowledge of collocations and corpus-based measures of collocation strength. Four well-known corpus-based measures were used here: raw frequency, z-score (Dennis, 1965), t-score (Church, Hanks, & Hindle, 1991), and mutual information, or MI (Church & Hanks, 1990; for a discussion of all four measures, see Durrant, 2008; Evert, 2004, 2005). The results of the analysis are presented in Table 3. Since frequency measures are not normally distributed, the figures shown are Spearman correlations. As we can see, the corpus measures correlate to various degrees. There is no correlation between frequency and MI because of the way the test items were selected (see above). However, none of the measures correlates significantly with performance on the “Words that go together” test.

INSERT TABLE 3 HERE

Conclusions

As we have seen, “Words that go together” has high test-retest and split-half reliability, and individual scores on the test show robust correlations with measures of linguistic experience and with other measures of language proficiency, but do not correlate with non-verbal IQ. Thus we can conclude that it is a valid and reliable test of individual speakers' collocational knowledge. A word of caution is in order, however. WGT is an instrument intended for research purposes. It was developed as part of an ongoing project examining the relationship between speakers' knowledge of collocations and other linguistic and non-linguistic abilities (see Dąbrowska, in preparation). It can also be used to compare knowledge of collocations in different populations (e.g. younger v. older, native v. non-native, monolingual v. bilingual speakers). In conjunction with other tests, it can help identify speakers with unusual profiles (e.g. a low score on WGT in comparison with vocabulary would suggest problems with distributional learning and/or inadequate exposure). It was not intended as a diagnostic tool to measure language achievement in pedagogical settings. It could be used in such

contexts of course, but the results will provide only limited information about the learner, i.e., how well s/he performs relative to native speakers or to other leaners in the group. Without further research, the test score will not allow us to draw inferences about what the learner can or cannot do with collocations in real life, or what kind of instruction would be most appropriate for them.

The research conducted as part of the validation study revealed robust correlations between speakers' collocational, grammatical and vocabulary knowledge. Such correlations are predicted by usage-based theories of language (Barlow & Kemmer, 2000; Bybee, 2006, 2010, 2013; Langacker, 1988, 2000) and are problematic for modular theories (e.g. Chomsky, 1981; Pinker, 1997, 1999; Ullman, 2006), according to which these three types of knowledge, or at least grammatical knowledge and vocabulary size, are independent components of the language faculty. The particularly strong correlation between vocabulary knowledge and collocational knowledge is consistent with the hypothesis that the acquisition of non-basic vocabulary depends strongly on distributional learning mechanisms (see Dąbrowska, 2009).

As anticipated, collocational knowledge was also found to correlate with age, but a closer examination of the data revealed an interesting pattern: knowledge of collocations increases linearly with age until about 32, then levels off and begins to fall again after 50. At present, there is no explanation for this finding. Learning of collocational patterns is thought to rely on implicit tallying of frequencies of co-occurrence, and since implicit learning abilities do not change in adulthood (Verneau, van der Kamp, Savelsbergh, & de Loozem, 2014), there is no reason to expect collocational learning to stop as long as speakers are continually exposed to combinations they do not yet know.

As indicated in the introduction, one of the reasons for developing the “Words that go together” test was to assess the psychological reality of corpus-based measures of association strength. The traditional approach to this problem is to take a set of collocations attested in the corpus and try to determine to what extent measures of differences in collocation strength obtained from corpora predict human behaviour. The study described in this paper used the opposite approach: it started with collocations that we know to be psychologically real and measured speakers’ ability to recognise them

in order to assess how well corpus-derived measures predict differences in performance. Contrary to expectations, speakers' performance on the test did not correlate with standard corpus-based measures of association, i.e., mutual information, z-score, t-score, or raw frequency. Of course, the fact that such relationship was not found does not mean that none exists. There are many more measures of collocation strength, and it is possible that some of them would turn out to be more predictive; it is also possible that the relationship was masked by some other factor. Furthermore, collocations may be salient to speakers for reasons other than frequency. Words belonging to multi-word units often alliterate or rhyme (cf. *bite the bullet, gain some ground, publish or perish, fair share, the name of the game*; for discussion see Boers & Lindstromberg 2008, Gries 2011); such phonological properties presumably help speakers to remember them. They may also be more memorable because they are particularly "colourful" (*living death, spout venom, spread like wildfire*) or parasitic on another established expression (many respondents chose *instil* rather than *restore* as the verb that "goes with" *faith*, possibly on analogy with *instil confidence* or *instil belief*). Thus, patterns found in corpora need not necessarily reflect patterns in speakers' minds.

Most importantly, the research presented here provides striking evidence of our ability to memorize recurrent word combinations. Because the test was designed to be challenging for adult native speakers, the collocations included in it were rare, with a mean frequency of 0.09 per million words.[2] It is remarkable, therefore, that all participants performed above chance, and some were close to ceiling. This finding demonstrates that collocations, even when they are semantically and syntactically regular, are psychologically real, and that native speakers must know vast numbers of them.

References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods, 40,* 278-289.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language,* 62, 67-82.

Barlow, M., & Kemmer, S. (2000). *Usage-Based Models of Language*. Cambridge: Cambridge University Press.

Bley-Vroman, R. (2002). Frequency in production, comprehension and acquisition. *Studies in Second Language Acquisition, 24*, 209-213.

Boers, F., & Lindstromberg, S. (2008). Structural elaboration by the sound (and feel) of it. In F. Boers & S. Lindstromberg (Eds.), *Cognitive Linguistic Approaches to Teaching Vocabulary and Phraseology* (pp. 329-353). Berlin/New York: Mouton de Gruyter.

Bybee, J. (2006). From usage to grammar: the mind's response to repetition. *Language, 82*, 711-733.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Bybee, J. L. (2013). Usage-based theory and exemplar representation. In T. Hoffman & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 49-69). Oxford: Oxford University Press.

Chomsky, N. (1981/1993). *Lectures on Government and Binding: The Pisa Lectures*. Berlin: Walter de Gruyter.

Church, K. W., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical Acquisition: Using On-line Resources to Build a Lexicon* (pp. 115-164.). Hillsdale: Lawrence Erlbaum Associates.

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

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics,* 29, 72-89.

Dąbrowska, E. (2009). Words as constructions. In V. Evans & S. Pourcel (Eds.), *New Directions in Cognitive Linguistics* (pp. 201-223). Amsterdam: John Benjamins.

Dąbrowska, E. (in press). Individual differences in grammatical knowledge. In E. Dąbrowska & D. Divjak (Eds.), *Handbook of Cognitive Linguistics*. Berlin: De Gruyter Mouton.

Dąbrowska, E. (in preparation) Grammar, vocabulary and collocations.

Dennis, S. F. (1965). The construction of a thesaurus automatically from a sample of text. In M. E. Stevens, V. E. Giuliano & L. B. Heilprin (Eds.), *Proceedings of the Symposium on Statistical Association Methods for Mechanized Documentation* (Vol. 269, pp. 61-148). Washington, DC: National Bureau of Standards Miscellaneous Pubcation.

Douglas-Kozłowska, C., & Dzierżanowska, H. (2004). *Selected English Collocations*. Warszawa: PWN.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory,, 6*, 125-155.

Durrant, P. L. (2008). *High frequency collocations and second language learning.* (PhD), University of Nottingham. Retrieved from http://etheses.nottingham.ac.uk/622/.

Ellis, N. C., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory, 5*, 61-78.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text, 20*, 29-62.

Evert, S. (2004). Computational approaches to collocations. Retrieved 8 February 2014, from www.collocations.de.

Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* (Ph.D.), University of Stuttgart. Retrieved from http://www.stefan-evert.de/PUB/Evert2004phd.pdf.

Gries, S. T. (2011). Phonological similarity in multi-word symbolic units. *Cognitive Linguistics, 22*, 491-510.

Herbst, T. (1996). What are collocations: sandy beaches or false teeth? *English Studies, 77*, 379-393.

Hodgson, J. M. (1991). Informational constraints on pre-lexical priming. *Language and Cognitive Processes, 6*, 169-205.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Langacker, R. W. (1988). A usage-based model. In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics* (pp. 127-161). Amsterdam: John Benjamins.

Langacker, R. W. (2000). A dynamic usage-based model. In M. Barlow & S. Kemmer (Eds.), *Usage-Based Models of Language* (pp. 1-63). Stanford, CA: CSLI Publications.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13.

Pinker, S. (1997). Words and rules in the human brain. *Nature, 387*, 547-548.

Pinker, S. (1999). *Words and Rules. The Ingredients of Language*. London: Weidenfeld and Nicolson.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing and Use* (pp. 127-151). Amsterdam: John Benjamins.

Shipley, W. C., Gruber, C. P., Martin, T. A., & Klein, A. M. (2009). *Shipley-2 Manual*. Los Angeles: Western Psychological Services.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (2004). *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. B. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multi-word sequences. *Experimental Psychology: Learning, Memory, and Cognition*, 37, 776-784.

Stubbs, M. (1995). Collocations and semantic profiles: on the cause of the trouble with quantitative methods. *Functions of Language, 2*, 1-33.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning,* 61, 569-613.

Tremblay, A., & Tucker, B. V. (2011). The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon,* 6, 302-324.

Ullman, M. T. (2006). The declarative/procedural model and the shallow structure hypothesis. *Applied Psycholinguistics, 27*, 97-105.

Verneau, M., van der Kamp, J., Savelsbergh, G. J., & P., de Loozem, M. (2014). Age and time effects on implicit and explicit learning. Experimental Aging Research, 40, 477-511.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-based Approach.* New York: Palgrave Mac.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Appendix 1: Item characteristics

Note: The frequency information was extracted from the BNC.

Collocation	Freq.	Node (Adj/V) freq.	Collocate (noun) freq.	Expected freq.	z-score	t-score	MI	% correct	IDis
absolute silence	27	3377	5162	0.18	63.3	5.2	7.2	80.0	0.61
achieve one's objectives	316	16553	7210	1.23	283.8	17.7	8.0	66.3	0.74
arouse suspicions	73	1330	2130	0.03	427.0	8.5	11.3	92.5	0.56
attract publicity	38	6229	2442	0.16	95.6	6.1	7.9	93.8	0.53
attractive proposition	48	5000	1987	0.10	149.7	6.9	8.9	75.0	0.68
bend rules	27	3259	18225	0.61	33.7	5.1	5.5	88.8	0.58
bitter dispute	54	2357	4435	0.11	164.2	7.3	9.0	70.0	0.69
blank expression	13	1393	8456	0.12	37.0	3.6	6.7	97.5	0.52
blatant lie	12	323	2150	0.01	141.7	3.5	10.7	92.5	0.55
boost production	93	1676	15791	0.27	177.5	9.6	8.4	78.8	0.60
close similarity	15	10326	1667	0.18	35.2	3.8	6.4	65.0	0.64
dim view	47	678	28193	0.20	105.4	6.8	7.9	71.3	0.64
divert attention	154	1156	13457	0.16	384.1	12.4	9.9	77.5	0.66
divert suspicion	6	1156	2130	0.03	37.5	2.4	7.9	70.0	0.68
fair share	272	7870	15830	1.28	238.9	16.4	7.7	91.3	0.53
full confession	10	27288	832	0.23	20.2	3.1	5.4	75.0	0.61
gain popularity	43	8601	1304	0.12	126.1	6.5	8.5	81.3	0.53
general direction	93	29308	10505	3.17	50.4	9.3	4.9	73.8	0.60
hazard a guess	44	110	799	0.00	1461.6	6.6	15.6	87.5	0.55
hear rumours	99	34199	1853	0.65	121.7	9.9	7.2	80.0	0.57
inflict punishment	18	1027	2423	0.03	112.2	4.2	9.5	60.0	0.66

issue a statement	390	7833	13648	1.10	370.4	19.7	8.5	68.8	0.64
join the ranks	97	16701	3433	0.59	125.4	9.8	7.4	82.5	0.57
lodge a complaint	33	1066	4425	0.05	149.4	5.7	9.4	70.0	0.70
memorable phrase	13	832	4143	0.04	68.8	3.6	8.5	47.5	0.76
obvious conclusion	36	8234	7320	0.62	44.9	5.9	5.9	63.8	0.69
odd remark	8	4255	3049	0.13	21.5	2.8	5.9	66.3	0.69
outspoken critic	41	293	3690	0.01	388.2	6.4	11.8	56.3	0.66
overall responsibility	93	5897	11809	0.72	108.9	9.6	7.0	57.5	0.76
precise details	67	2834	17294	0.51	93.5	8.1	7.1	63.8	0.63
raise prices	109	18786	27440	5.32	45.0	9.9	4.4	93.8	0.51
raise standards	173	18786	14878	2.88	100.2	12.9	5.9	87.5	0.59
refuse an application	82	10172	15869	1.66	62.3	8.9	5.6	45.0	0.68
regular employment	31	7387	10600	0.81	33.6	5.4	5.3	83.8	0.60
restore faith	25	3839	5160	0.20	54.9	5.0	6.9	61.3	0.55
serious problem	619	11903	54555	6.70	236.6	24.6	6.5	73.8	0.57
striking example	92	1667	19265	0.33	159.3	9.6	8.1	28.8	0.90
thorough search	25	1081	5378	0.06	101.9	5.0	8.7	72.5	0.59
urgent matters	36	2066	23720	0.51	49.9	5.9	6.2	70.0	0.63
witness an incident	20	2015	5033	0.10	61.5	4.4	7.6	81.3	0.60

Appendix 2: Words that Go Together

This questionnaire consists of sets of five phrases. From each set, choose **one** phrase that sounds the most natural or familiar. If you are not sure, guess. Here are two examples:

delicate tea	X deliver a speech
feeble tea	hold a speech
frail tea	perform a speech
powerless tea	present a speech
X weak tea	utter a speech

The words *delicate, feeble, frail, powerless* and *weak* are similar in meaning; but with *tea,* we would normally use *weak.* In the second example, *deliver a speech* sounds more natural than the other choices.

1	blatant lie	7	chance a guess	13	bitter dispute
	clear lie		dare a guess		cruel dispute
	conspicuous lie		gamble a guess		hard dispute
	distinct lie		hazard a guess		harsh dispute
	recognizable lie		risk a guess		savage dispute
2	blank expression	8	bend rules	14	absolute silence
	frightful expression		honour rules		pure silence
	plain expression		institute rules		sheer silence
	sinister expression		reject rules		stark silence
	terrible expression		validate rules		supreme silence
3	attain publicity	9	believe a statement	15	complete confession
	attract publicity		change a statement		exhaustive confession
	bring publicity		issue a statement		extensive confession
	make publicity		offer a statement		full confession
	win publicity		revise a statement		thorough confession
4	fair share	10	advance standards	16	acquire popularity
	honest share		boost standards		attract popularity
	just share		elevate standards		earn popularity
	legitimate share		lift standards		gain popularity
	reasonable share		raise standards		get popularity
5	arouse suspicions	11	boost production	17	constant employment
	incite suspicions		double production		normal employment
	kindle suspicions		enlarge production		ordinary employment
	revive suspicions		extend production		regular employment
	stimulate suspicions		redouble production		unbroken employment
6	elevate prices	12	combine the ranks	18	glimpse an incident
	grow prices		conjoin the ranks		notice an incident
	lift prices		join the ranks		observe an incident
	raise prices		merge the ranks		see an incident
	stimulate prices		unify the ranks		witness an incident

19 achieve one’s objectives
complete one’s objectives
finish one’s objectives
follow one’s objectives
tackle one’s objectives

20 accurate direction
appropriate direction
convenient direction
general direction
specific direction

21 apply attention
dedicate attention
divert attention
grasp attention
sidetrack attention

22 extensive problem
extreme problem
serious problem
significant problem
vital problem

23 compelling matters
critical matters
desperate matters
major matters
urgent matters

24 close similarity
doubtful similarity
evident similarity
extreme similarity
near similarity

25 contradict rumours
discover rumours
hear rumours
know rumours
tell rumours

26 effective phrase
helpful phrase
memorable phrase
noteworthy phrase
significant phrase

27 distract suspicion
divert suspicion
mislead suspicion
redirect suspicion
sidetrack suspicion

28 bring faith
instil faith
offer faith
refresh faith
restore faith

29 complete search
full search
scrupulous search
thorough search
total search

30 abundant details
complete details
definite details
precise details
small details

31 apply punishment
deliver punishment
inflict punishment
perform punishment
provide punishment

32 appealing proposition
attractive proposition
charming proposition
inviting proposition
seductive proposition

33 dark view
dim view
murky view
shadowy view
shady view

34 aggressive critic
forthright critic
frank critic
open critic
outspoken critic

35 odd remark
peculiar remark
queer remark
unnatural remark
weird remark

36 distinct example
gross example
recognizable example
shocking example
striking example

37 formulate a complaint
lodge a complaint
place a complaint
record a complaint
write a complaint

38 confident conclusion
evident conclusion
obvious conclusion
solid conclusion
sure conclusion

39 general responsibility
large responsibility
overall responsibility
single responsibility
unique responsibility

40 decline an application
deny an application
ignore an application
refuse an application
scrap an application

Appendix 3: Percentile distribution of WGT scores

Score	Percentile
20	5
21	6
22	8
23	9
24	18
25	20
26	25
27	30
28	35
29	39
30	44
31	50
32	55
33	60
34	68

35	71
36	80
37	85
38	95
39	99

Author note

I would like to thank Sarah Duffy, Robyn Gibson and Steven McCarthy for help with the data collection, and Stefanie Wulff, Debra Titone, and two anonymous reviewers for their helpful comments on the paper. Correpondence should be addressed to Ewa Dąbrowska, Department of Humanities, Faculty of Arts, Design and Social Sciences, Northumbria University, Newcastle upon Tyne, NE1 8ST, United Kingdom, e-mail: ewa.dabrowska@northumbria.ac.uk.

Footnotes

[1] See Dąbrowska (in press) for a more detailed discussion of the relationship between language experience, education and reading habits.

[2] Note that the frequencies given in Table 1 are for the entire BNC corpus (i.e., approximately 100 million words).

Table 1

Target collocations in the British National Corpus

	Frequency	t	MI
Minimum	6	2.4	4.4
Maximum	619	24.6	15.6
Median	44	6.6	7.7
Mean	87	8.0	7.8

Table 2

Correlations (Pearson) between performance on the WTGT and measures of language exposure, language proficiency, and non-verbal IQ

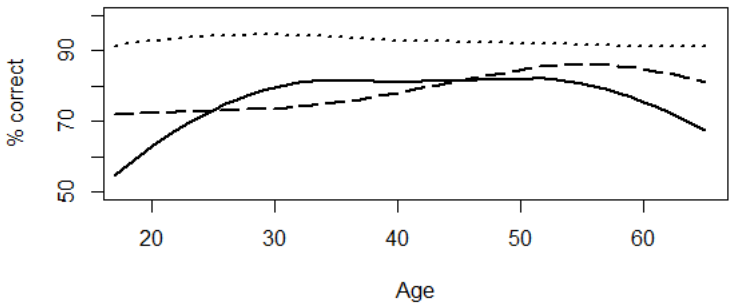
Language exposure measure	Correlation coefficient	*p* value
Author Recognition Test	0.51	<0.001
Self-reported reading	0.34	0.002
Education	0.35	0.001
Age	0.24	0.033
Grammar	0.43	<0.001
Vocabulary	0.53	<0.001
Blocks	0.19	0.099

Table 3

Correlations (Spearman) between performance on WTGT and corpus-based measures of collocation strength

Measure	freq	z	t	MI	WTGT
freq	1.00	0.60	1.00	0.01	0.08
z	0.60	1.00	0.61	0.78	0.01
t	1.00	0.61	1.00	0.03	0.07
MI	0.01	0.78	0.03	1.00	-0.05
WTGT	0.08	0.01	0.07	-0.05	1.00

Figure 1. Relationship between age and collocations (solid line), vocabulary (dashed line) and grammar (dotted line)


% correct
50
70
90
20
30
40
50
60
Age