9-30-2018

The Error in Trial and Error: Exercises on Phrasal Verbs

Brian Strong
Victoria University of Wellington

Frank Boers
fboers@uwo.ca

The error in trial and error: Exercises on phrasal verbs

Brian Strong and **Frank Boers**

Abstract

An analysis of 44 commercially available EFL textbooks found that it is common for textbooks to present learners with exercises on phrasal verbs without first providing relevant input to help them. In these cases, the learners are likely to resort to trial-and-error and are then expected to learn from feedback. We report an experiment conducted with Japanese EFL students (N=140) in which we compare the effectiveness of such a trial-and-error method with a retrieval procedure in which students first study a set of phrasal verbs and then complete an exercise. Scores on both an immediate and a one-week delayed post-test suggest superiority of retrieval over the trial-and-error procedure, where, despite the provision of feedback, 25% of the wrong exercise responses were reproduced in the delayed post-test.

Introduction

English has a rich repertoire of phrasal verbs (e.g., *give up, go on*, *turn out* and *break down*). These are constructions made up of a verb and a preposition that acts as a particle. As a class, phrasal verbs are very common in English (e.g., Bolinger, 1971; Gardner & Davies; Garnier & Schmitt, 2015; Liu, 2011) and thus useful for language learners to master. Unfortunately, they also pose a considerable challenge for many language learners (e.g., Condon, 2008; Garnier & Schmitt, 2016; Liao & Fukuya, 2004; Side, 1990). It is not surprising, then, that virtually all mainstream EFL textbooks include sections on phrasal verbs, and some resources for language study are entirely devoted to them (e.g., McCarthy & O'Dell, 2004). These books often present learners with study material (e.g., a set of phrasal verbs paired with paraphrases and example sentences), followed by an exercise (e.g., a matching task). However, an analysis (reported below) of EFL textbooks reveals that it is also very common for such books to quiz learners with an exercise first, followed by feedback (from a teacher or in the form of an answer key). Put differently, exercises on phrasal verbs are implemented in EFL course books as study + retrieval procedures and also as trial-and-error + feedback procedures.[1]

Support for both approaches can be found in the memory research. On the one hand, when the exercise follows a study episode, learners' successful retrieval of the studied items is expected to entrench this new knowledge (Roediger & Karpicke, 2006). On the other hand, trying an exercise without a prior study episode may pique a learners' curiosity about the appropriate answers and hence stimulate intake of subsequent feedback (Hays, Kornell, & Bjork, 2013). To the best of our knowledge, however, no studies have as yet compared the learning outcomes of these two

approaches for learning phrasal verbs. Given the importance of phrasal verbs and given the space that EFL resources devote to exercises on phrasal verbs, such an evaluation seems overdue. We therefore designed an experiment in which we randomly assigned EFL learners to either a study + retrieval condition or a trial-and-error + feedback condition, and then compared the effectiveness of the two approaches. Before developing the two treatment conditions, we analysed a corpus of EFL course books to estimate how common these two general approaches are and to determine what exercise formats they typically use.

Literature review

The trouble with phrasal verbs

There are several reasons that, together, help to explain why mastering phrasal verbs can be challenging. One is the sheer number of phrasal verbs (Bolinger, 1971). Another is that a single verb-particle combination typically has multiple meanings (on average 5.6, according to Gardner & Davies, 2007). For example, the meaning of the combination *make + up* varies in *make up a story, make up after an argument, make up one's face, make up the difference, make up a bed*, *make up for something* and *make it up to someone*. Fortunately, the availability of large online corpora now makes it feasible to identify the highest-frequency verb-particle combinations (Gardner & Davies, 2007; Liu, 2011) and to identify the most frequent meanings of a given verb-particle combination (Garnier & Schmitt, 2015). This can usefully inform choices of prioritization in learning and teaching, but the resulting lists still represent a considerable learning burden. For example, Garnier and Schmitt's (2015) list includes 150 verb-particle combinations expressing 288 meanings altogether. In settings where

learners are not exposed to abundant amounts of English, unassisted acquisition of all of these is not very likely. EFL textbooks of course contain texts that include instances of phrasal verbs, but hardly in sufficient numbers to make acquisition likely (Alejo-González, Piquer-Píriz, & Reveriego-Sierra, 2010).

Understanding and remembering the meaning of phrasal verbs is often challenging also because many phrasal verbs are semantically non-compositional, that is, their meaning does not follow straightforwardly from the meanings of the constituent words. It is not obvious, for example, why *put up with something* should mean *to tolerate*. Semantic non-compositionality is a defining feature of idioms, and so phrasal verbs are sometimes considered a subset of idioms (e.g., Gairns & Redman, 2011; Kövecses & Szabó, 1996). Because of this particular learning challenge, it is not surprising that it is instances of non-compositional phrasal verbs—sometimes called *idiomatic* phrasal verbs—that are typically targeted in L2 learning experiments (e.g., Boers, 2000; Yasuda, 2010).

Because the meaning of many phrasal verbs is not transparent, it is also not surprising that many textbook exercises on phrasal verbs engage learners with the form-meaning mapping of the items, for example by requiring them to match intact phrasal verbs to their paraphrases or definitions (and thus testing meaning recognition). As we shall see further below, where we report an analysis of a corpus of EFL textbooks, other kinds of exercises aim in addition to engage learners with the composition of phrasal verbs, for example by requiring them to decide which particle from a set of options goes with which verb (i.e., testing form recognition) or to supply the missing particles in gap-fill exercises (i.e., testing form recall).

It is often the contribution made by the particle to the overall meaning of a phrasal verb that is far from obvious. It is not immediately obvious in what way *in* contributes to the meaning of *give in* or *on* to the meaning of *catch on*, for example. Instead, it is often the verb that appears the better clue for an approximate interpretation a phrasal verb (e.g., *drink* seems a stronger clue than *up* to interpret *drink up*). Consequently, learners may pay more attention to the verb than to the particle. In addition, particles are formally not salient. Particles are thus relatively unlikely to catch learners' attention in the absence of instruction that points them out.[2]

Yet another challenge for learners, but one which falls beyond the scope of this article, is to master the syntactic patterning of phrasal verbs, more specifically, determining when the verb and the particle of a phrasal verb can be separated. First, a distinction needs to be made between transitive and intransitive phrasal verbs, since the latter require the verb and the particle to be adjacent to one another (e.g., *My car broke down on the motorway / *My car broke on the motorway down*). Second, within the class of transitive phrasal verbs, some are separable (e.g., *I looked up this word in the dictionary / I looked this word up in the dictionary*), while others are not (e.g., *Look after the children / *Look the children after*). Third, in the case of separable transitive phrasal verbs, separation becomes mandatory when the object is a pronoun (e.g., *I looked it up / *I looked up it*). One more challenge that needs to be mentioned, but which is also outside the scope of the present article, is for learners to appreciate which phrasal verbs are typical of informal, conversational discourse and which ones are deemed appropriate in more formal genres, such as academic writing (Liu, 2011).

For many learners the above challenges are exacerbated by the absence of a structurally similar category of phrasal units in their L1 (e.g., Garnier & Schmitt, 2016;

see, Slobin, 1996, and Talmy, 1985, for broader typological accounts). This also helps to explain why such learners tend to avoid using them when single-word alternatives (e.g., *refuse* instead of *turn down*) are available (Dagut & Laufer, 1985; Liao & Fukuya, 2004; Siyanova & Schmitt, 2007). For learners whose L1 does have structural equivalents, the task of learning English phrasal verbs appears less daunting, in relative terms (Hulstijn & Marchena, 1989; Laufer & Eliasson, 1993). Learners whose L1 has not familiarized them with the phenomenon are thus the most likely to need assistance.

We mentioned above (and discuss further below) that EFL textbooks structure exercises on phrasal verbs in essentially two ways: Either learners are first presented with study material and are subsequently given an exercise where they are expected to retrieve the studied material from memory, or they are first given an exercise to reveal the gaps in their knowledge and feedback (study material) then follows to fill those knowledge gaps. It is therefore useful to examine what predictions regarding the merits of these two approaches might be distilled from previous experiments where retrieval or trial-and-error procedures were put to the test.

Memory research on retrieval and trial-and-error procedures

The benefits of retrieval are well documented in the realm of memory research (Karpicke, Lehman, & Aue 2014; Roediger & Butler, 2011), where it is also known as the testing effect (Roediger & Karpicke, 2006). A common way in which the retrieval effect has been investigated in memory research is by presenting participants with word pairs (in their L1), followed by a test where one word is presented as cue to elicit the paired associate. This is compared to a condition where the same word pairs are presented twice, but without a test episode. Recall of the word pairs in post-tests has

almost invariably been found superior after the learning condition that included the retrieval component (Smith, Roediger, & Karpicke, 2011). One might argue that the learning involved in remembering paired words is different from learning vocabulary in a second language, because the word associations to be remembered concern familiar words in the participants' L1. However, the retrieval effect has also been attested for paired associates learning in L2s. In Barcroft (2007), participants were asked to study L2 word-picture pairs displayed on a screen. The second step in the retrieval condition presented the picture without its corresponding word and participants were asked to recall the missing word. The word-picture pairs were then displayed again so the participants could verify the accuracy of their recall. In the comparison condition, there was no retrieval task. Instead, the participants were asked to re-study the same word-picture pairs. Performance on post-tests was better in the retrieval than in the re-study condition.

A considerable body of memory research has investigated the effects of *multiple* study and test trials (e.g., Carpenter & DeLosh, 2005; Cull, 2000; Kang Lindsey, Mozer, & Pashler, 2014; Karpicke & Bauernschmidt, 2011; Karpicke & Roediger, 2007; Pyc & Rawson, 2007). Ways of implementing repeated retrieval practice have also attracted considerable attention in the domain of L2 vocabulary learning (e.g., Nakata, 2015, 2017; Schuetze, 2015). Multiple retrieval events concerning the same target items are not common in textbooks, however. According to our textbook analysis discussed below, a book will typically provide just one exercise on a set of target items.

As mentioned, textbooks also provide quizzes in a trial-and-error fashion, without a study episode preceding them. Learning in that case relies on students'

comparing their hunches against the feedback that follows. In other words, the exercise essentially serves as a pre-test and the feedback as the study episode. In memory research, pre-testing is associated with the so-called *generation effect* (Slamecka & Graf, 1978), where taking a test without any prior opportunity to study the target items is believed to enhance subsequent learning of these items. Slamecka and Graf (1978), for example, had native English speakers learn word pairs in a pre-test + study condition and in a study-only condition. The pre-test + study condition presented a word and the initial letter of a related word (e.g., rapid-f_) and asked participants to guess the missing word. The study-only condition presented both intact words together (e.g., *rapid-fast*). Final test performance was better under the test + study condition than the study-only condition. It is important to note, however, that the participants in the pre-test + study condition guessed 94% of the missing words in the pre-test, and so the feedback (or study episode) mostly served to confirm correct associations. In real educational contexts, students are probably less likely to generate so many correct responses when given a test on items they have not yet studied. It is therefore important to investigate the pre-testing effect also in the case of error-prone response generation.

Such a study was conducted by Kornell, Hays and Bjork (2009). In one condition, participants were presented with a cue and asked to guess what related word was missing (e.g., olive -?) within eight seconds. After a response was generated, the cue and the target word were displayed together for five seconds. In a comparison condition, participants simply studied the two words (e.g., olive – branch) together for 13 seconds. Although many wrong guesses were generated, the pre-test + study condition led to the better post-test scores. Superiority of a pre-test + study condition

over the study-only condition was also found in partial replications by Grimaldi and Karpicke (2012), Huelser and Metcalfe (2012) and Yang, Potts and Shanks (2017).

The targets for learning in these experiments were pairs of familiar words in the participants' L1, and so one may wonder if the findings are transferable to situations of L2 vocabulary learning. A study by Potts and Shanks (2014) suggests they are. Undergraduate English speakers were asked to learn words in Euskara, a language they had no prior knowledge of. There were three conditions. In one, participants were first asked to choose an L1 translation from a set of options for each of the L2 words, before the correct L1-L2 pairing was displayed. In the second, participants were asked to supply a translation themselves for the L2 words, before the correct pairing was displayed. The third learning condition involved no guessing or response generation. Instead, participants were given the L2-L1 word pairs to study. The participants were told a post test would follow. The best results on this near-immediate post-test, which was in a multiple-choice format requiring participants to choose the correct meaning of the L2 words, were obtained under the treatment condition where participants had supplied their own translations (which was inevitably the most error-prone pre-test), followed by the multiple-choice condition. Potts and Shanks' (2014) findings therefore suggest that generating errors in a trial-and-error event can positively influence subsequent learning of new form-meaning correspondences. This sounds promising if applied to phrasal verbs, since part of the challenge there is the same, namely, remembering their meaning.

It is important to mention, however, that Potts and Shanks' (2014) findings have not been confirmed in some other experiments. Warmington, Hitch and Gathercole (2013) examined children's learning of novel name-object pairs. In one condition, the

children were shown a new object and were told the first letter of its name (a pseudo-word). This was followed immediately by the full name (e.g., "This object's name begins with the letter P. It's a *prot*. Can you say *prot*?"). In the other condition, the children were asked to guess the name of the object, using the first letter as a cue. (e.g., "This object's name begins with the letter P. Can you guess its name?"). Only after the guess—which was almost invariably wrong—were the children told the name of the object. The post-test performance, where the children were asked to recall what the objects were called, was the poorest in the latter condition. Warmington and Hitch (2014) replicated this experiment with adults and again found that error-free encoding of new words led to better post-test performance than error-prone guessing followed by feedback. One might argue that these experiments by Warmington and colleagues are quite different from Potts and Shanks (2014) because they were conducted with L1 speakers rather than L2 learners. Still, the learning task was to establish a connection in memory between unfamiliar word forms (i.e., pseudo-words) and their meaning, and this is not dissimilar to the task of learning new L2 words.

As mentioned, when it comes to learning phrasal verbs, an additional task is to establish accurate memories of their composition, that is, knowledge of which particle goes with which verb to express a given meaning. Exercises intended to foster learning of syntagmatic word partnerships, or collocations, have attracted some scrutiny in recent years. For example, Boers, Demecheleer, Coxhead and Webb (2014) and Boers, Dang and Strong (2017) compared the effects of exercises on verb-noun collocations (e.g., *make an effort* vs. *do homework, tell the truth* vs. *talk nonsense; conduct research* vs *perform a task*). This included various exercise formats, such as ones where learners are asked to choose from a list of verbs the one that collocates with a given noun. Because

these exercises were not preceded by specific study materials on the collocations in the experiments, they can be characterized as trial-and-error events. When the exercises required students to assemble the verb-noun expressions, many wrong choices were made. A considerable number of these wrong responses re-occurred in the post-tests, despite the feedback that was given immediately after the exercise.

It is worth noting that none of the studies reviewed here directly compared the effectiveness of retrieval and trial-and-error procedures. The comparison conditions always consisted only of study (and re-study) episodes. In sum, although a plethora of studies in the realm of memory research have investigated the testing (or retrieval) effect and the pre-testing (or generation) effect, the available evidence looks insufficient to determine which is the most judicious choice in general, let alone in the specific case of exercises on phrasal verbs. What also complicates an evaluation of the available body of evidence is that conclusions have typically been drawn on the basis of (near-) immediate post-tests. The extent to which, for example, the attested benefits of learning from feedback after error generation are durable remains uncertain. The studies on collocation exercises by Boers et al. (2014) and Boers et al. (2017) are exceptions because they did involve delayed post-tests. On the other hand, they did not include immediate tests. Had immediate tests been administered, this could arguably have helped to consolidate the feedback on the exercises, and so the potential of learning from trial and error may have been underestimated in these studies.

The present study has two parts. The first is an analysis of phrasal verb exercises in a corpus of EFL textbooks. The second part is an experiment in which we compared the effectiveness of two implementations of phrasal verb exercises, retrieval versus trial-and-error. The findings from the textbook analysis served to ensure that the materials

and treatments designed for the experiment bore a reasonable resemblance to actual classroom, or, more precisely, textbook reality.

The nature and implementation of exercises on phrasal verbs in EFL textbooks

Research questions

The principal research question addressed by the textbook analysis is: what proportion of exercises on phrasal verbs are implemented in a way that invites trial and error rather than retrieval of previously presented information. Two secondary questions concerned (a) the proportion of exercises requiring learners to (re-)assemble phrasal verbs rather than perform operations with intact phrasal verbs and (b) the average number of phrasal verbs tackled in a single exercise.

Method

Selection of textbooks

The textbook analysis was conducted on a sample of 44 recent textbooks that had been purchased by the authors' university in 2014 for the purpose of a large-scale textbook analysis. They were the most recent editions available at that time of EFL textbooks published by three major publishing houses (Oxford University Press, MacMillan, and Pearson Education) and marketed globally. The books belonged to the following series: *English Result, Global, New English File, New Headway, New Inside Out, New Total English, Speak Out*, and *Straightforward* (see Appendix 1 for the full list). Each book was manually screened for sections containing exercises on phrasal verbs (although not all textbooks use the term *phrasal verb* to refer to the same structure).

Analysis

The manual screening of the textbooks produced a corpus of 140 exercises on phrasal verbs. For each of these exercises we reviewed all the input materials that preceded them in the book in which it occurred, in order to estimate whether the given exercise was intended as a retrieval activity. We examined the input material preceding each exercise for explicit information about the meaning and/or composition of the target phrasal verbs, and we also screened preceding texts in case these included instances of the target phrasal verbs (sometimes made salient through typographic enhancement). If such opportunities were found in the given textbook for studying (or at least encountering) the target items prior to doing the exercise, we categorized it as a retrieval exercise. If no such opportunities were found in the given book prior to the exercise, it was categorized as a trial-and-error exercise. The distinction refers to the way the exercises are incorporated in the textbooks and thus the ways students are *likely* to tackle them. We cannot rule out, of course, that a student will resort to guessing also in an exercise intended to induce retrieval. Neither can it be ruled out that a student acquired knowledge of a given target item from sources outside the given textbook, and so may recall it from longer-term memory.

We then also categorized the exercises according to whether they engaged learners with the form-meaning mapping of intact phrasal verbs (e.g., matching phrasal verbs to their paraphrases—meaning recognition) or with the composition of phrasal verbs (e.g., combining a verb and a particle to express a stipulated meaning—form recognition or form recall). As part of the analysis we also counted the number of phrasal verbs tackled per exercise.

Results and discussion

Of the 140 exercises on phrasal verbs identified in the textbook corpus, 54 qualified as retrieval exercises, while 86 (61%) qualified as trial-and-error exercises. In the latter case, students were sometimes directed to an answer key for feedback or it must be assumed that feedback would be provided by a teacher. Six of the eight textbook series were found to include a greater number of trial-and-error exercises than retrieval exercises (see Appendix 2). In sum, trial-and-error exercise implementations are quite common in the corpus of EFL textbooks we examined. We need to bear in mind, of course, that this was far from a comprehensive corpus of available textbooks, and so we need to be cautious not to generalize from the data extracted here. Still, it is a collection of textbooks that certainly includes internationally well-established ones and so it is reasonable to assume that many EFL learners around the globe experience not only retrieval exercises but also trial-and-error exercises for phrasal verb learning. This makes an empirical comparison of the two approaches all the more pertinent.

While the majority of the exercises were found to focus on the meaning of intact phrasal verbs, approximately 35% required learners to combine verbs and particles themselves to match or express a certain meaning (see Appendix 3). The latter is arguably the more challenging exercise and, if effective, also the closer to fostering productive knowledge. As to the number of phrasal verbs tackled per exercise, the average turned out to be 6.77, and so we decided to work with sets of seven phrasal verbs in our experiment.

Experiment: study episode + retrieval versus trial-and-error + feedback

Research questions

The experiment addressed the following principal research question: Do study + retrieval and trial-and-error + feedback procedures for learning idiomatic phrasal verbs produce similar amounts of learning, as gauged by an immediate and a one-week delayed post-test? A secondary research question concerns specifically the effect of feedback in the trial-and-error condition and asks how well this effect (if any) is sustained over time, i.e., between the immediate and the delayed post-test. This secondary research question is motivated by the observations in the literature review that experiments on learning through error-generation followed by feedback have tended to examine only short-term effects.

Method

Participants

The study involved the participation of 170 Japanese university EFL students. It is worth mentioning that Japanese does not have phrasal verbs and that these students could thus be expected to find phrasal verbs learning quite challenging. All the students were non-English majors from five parallel classes in their third year of study at the university. There were 107 females and 63 males, with an average age of 21.8 (*SD* = 1.2). Informed consent to participate was obtained from all participants. Prior to the experiment, they all took Version 2 of the Vocabulary Levels Test (VLT, Schmitt, Schmitt & Clapham, 2001) at the 2,000 frequency level. The mean score was 28/30 (SD = 2.2), suggesting that these students should have little trouble in understanding English

input materials made up of words from the 2,000 most frequent word families (see below).

The students from four of the classes were randomly divided into two treatment groups (each $n = 70$). The fifth class ($n = 30$) participated in a pilot test to help us select phrasal verbs to be targeted in the actual experiment. We compared the VLT scores of the two treatment groups and the pilot group, and no significant differences were found: $F(2, 167) = 0.263$, $p = .768$. We also obtained the participants' scores on a TOEIC Bridge Test, which they took approximately six months prior to the present study, and found no significant differences in these scores either: $F(2, 167) = 1.029$, $p = .359$. Overall, the two treatment groups and the pilot group thus appeared reasonably equivalent in terms of their English listening and reading comprehension as well as their vocabulary knowledge.

Selection of phrasal verbs

We first screened corpus-based lists of phrasal verbs (e.g., Gardner & Davies, 2007; Garnier & Schmitt, 2016; Liu, 2011) and recent course books with a focus on phrasal verbs (e.g., Gairns & Redman, 2011; McCarthy & O'Dell, 2004; 2007) for potential target items. This yielded an initial set of 35 phrasal verbs[3] which were (a) idiomatic (in the sense of non-compositional), (b) current (judging by their inclusion in several of the resources consulted), and (c) likely to be unfamiliar to the students participating in the study (as estimated by the first author, who has extensive experience teaching EFL at universities in Japan).

To ascertain that these phrasal verbs were indeed highly unlikely to be known by the 140 participants in the two treatment groups, they were piloted as test items with the

aforementioned group of 30 students drawn from the same population. The test format was the same as the post-test used in the actual experiment. Of the initial set of 35, 14 phrasal verbs were unknown by all 30 students according to this pilot test. These were thus the ones most likely to be unknown also to the students assigned to the treatment conditions, and so these were the ones chosen as the target items for the experiment. The 14 target phrasal verbs (and paraphrases) were the following: *Catch on* (become popular), *run out* (use all of something), *hang out* (spend time with friends), *pass away* (die), *brush up* (improve a skill), *head off* (go somewhere), *give in* (accept that you cannot win), *open up* (talk about your feelings), *drop out* (leave before finishing), *blow up* (become very angry), *chicken out* (not do something because you are scared), *hold on* (wait for a short time), *pop in* (visit for a short time), *back up* (repeat something you have said).

For the sake of ecological validity, we imitated a presentation format we encountered multiple times in the corpus of textbooks and in McCarthy and O'Dell's (2004; 2007) books for independent study. Each phrasal verb was embedded in a short dialogue and preceded by a paraphrase of its meaning. For example:

Hang out: spend time with friends.

Speaker A: Hey, Yuki, if you're not busy after work, do you want to hang out?

Speaker B: I'm sorry, Tomoko, but I'm not feeling well today. How about tomorrow?

In the retrieval condition, the example dialogue and clarification of the phrasal verb's meaning were presented first. Only after this study episode were the students given the exercise. In the exercise, the phrasal verb was embedded in a very similar dialogue (preserving the same meaning as in the study episode), but with the particle missing. For example:

Speaker A: Hi, Toko. What are you doing tomorrow? Do you want to hang __?
Speaker B: I would love to, but I have to stay home and clear my room. Sorry.
An empty box was provided under the blank for students to supply the missing particle.

In the trial-and-error condition, the above two stages were reversed. The students were first given the exercise and asked to supply the missing particle. After this, they were given feedback in the form of the same study material as provided as the first phase in the retrieval condition, i.e., the phrasal verb with its meaning clarification and an example dialogue including its intact form. In other words, the materials used in the two treatment conditions were identical; it was just the order of presentation that differed.

Also for the sake of ecological validity, we presented the 14 target items in two sets of seven, whereby in the retrieval condition seven phrasal verbs were first presented consecutively as study material before the same seven items were presented in exercise format, while in the trial-and-error condition the order of the two phases was reversed. The order of the two sets of items was counterbalanced between the conditions and the order of the items within each set was randomized. After the students completed their work on the first set of seven phrasal verbs, they were given a brief distractor task where they answered some trivia questions in their L1 and did some simple addition/subtraction tasks. The immediate post-test concerning the seven phrasal verbs followed. The post-test was identical in format to the exercise, that is, each of the phrasal verbs was embedded in a simple dialogue that was very similar to the ones previously met but now with a blank where the particle was missing (e.g., I'll be free later this evening. If you are free as well, shall we hang __?). A box was provided again for the students to supply the missing particle. After completing this post-test, the

students worked on the second set of seven phrasal verbs, following the same learning and testing procedures as before. A delayed post-test was administered seven days later. The delayed post-test was identical to the immediate post-test, except that it included all 14 items in one presentation. The order of the test items was randomized in both tests.

We ran the dialogues, the meaning definitions, and the target phrasal verbs through the lexical profiler on Tom Cobb's website (https://www.lextutor.ca), and ascertained that all the words (barring proper nouns) belonged to the 2,000 most frequent word families of English. Recall that the students' scores on the Vocabulary Levels Test suggested they had knowledge of most words at that frequency level. To minimize the risk that some of the words might nonetheless be unknown, we asked the students' teachers to indicate any of the words in the materials that they thought might still be unfamiliar to their students. A week prior to the actual experiment, the students were given electronic flashcards which paired these potentially difficult words with their L1 equivalents, and they were requested to learn these as part of a regular English class. This activity did not include any of the phrasal verbs, nor their components (verbs or particles). This class with electronic flashcards also served the purpose of familiarizing the students with the computer interface that would be used in the actual experiment.

For the sake of experimental control, all the study and test procedures were run in a classroom on a computer. When the students logged in, they were randomly assigned to one of the two treatment conditions (retrieval vs. trial-and-error). They were told explicitly that the aim of the session was to learn phrasal verbs, and they were asked to follow the instructions on the computer screen typing in their responses to exercise/test items when requested. The students were not forewarned that a post-test

would follow. The whole session (learning phases, distractor tasks, and immediate post-tests) took on average about 30 minutes. The one-week delayed post-test, which was also unannounced, was also administered on a computer, again under teacher supervision.

Analysis

Responses were scored dichotomously, with one point awarded for each correctly supplied particle.[4] We used a mixed effects logistic regression model (R package lme4, Bates, Maecher, Bolker, & Walter, 2014) to test whether performance on the post-tests was significantly mediated by the relevant predictors. The model is suitable because the post-tests were scored binomially with 0 for an incorrect response and 1 for a correct response. It also accounts for the general variability among participants and target items (Seltman, 2014). We preferred to analyze the observed data using mixed effects logistic regression because recent research has indicated that it is superior to repeated-measures ANOVA (for a discussion see Jaeger, 2008, and Linck & Cunnings, 2015).

The mixed effects logistic regression analysis was conducted with the following predictors entered as fixed effects: treatment condition (i.e., retrieval versus trial-and-error), test time (immediate versus delayed), performance on the exercises, VLT scores and TOEIC Bridge Test scores. Participants and target items were entered as random effects. Applying a backward stepwise selection, the VLT scores and the TOEIC Bridge scores were removed as they turned out not to contribute to the model. We then calculated the predicted probabilities of providing a correct response on the post-tests.

Results and discussion

Overall exercise and post-test scores compared

Table 1 shows the total number of correct responses on the exercises and on the post-tests under the two treatment conditions.[5] As expected, the retrieval group outperformed the trial-and-error group on the gap-fill exercise. Participants in the retrieval group filled in the answer correctly 95% of the time, whereas those in the trial-and-error group did so only 12.96% of the time. This confirms that the exercise treatments were indeed very different from one another, with the retrieval condition being almost error-free while the trial-and-error condition was error-prone. Regarding the 13% success rate in the trial-and-error condition, it is worth noting that particles make up a small class of words (and some members are more frequent and occur in more phrasal verbs than others), and so it is likely that a fair number of correct responses were lucky (or reasoned) guesses.

Table 1: Tallies of correct responses in the exercise, the immediate post-test and the delayed post-test. Max = 980 (K = 14 x n = 70).

Treatment group	Exercise	Immediate post-test	Delayed post-test
Retrieval	931 (95.00%)	699 (71.33%)	544 (55.51%)
Trial and error	127 (12.96%)	602 (61.43%)	405 (41.33%)

Turning now to the post-tests, the retrieval group provided the correct response (71.33%) on the immediate post-test more often than did the trial-and-error group (61.43%). Likewise, the number of correct responses on the delayed post-test was higher for the retrieval group (55.51%) than for the trial-and-error group (44.33%). Both

treatment groups showed attrition after a one-week period, with a 15.82 percentile score loss in knowledge for participants in the retrieval group and a 17.1 percentile score loss in knowledge for those in the trial-and-error group.

The output of the mixed effects logistic regression revealed that treatment condition ($z = -4.087$, $p < .0001$), test time ($z = -9.926$, $p < .001$) and performance on the exercise ($z = 5.750$, $p < .0001$) were significant predictors of post-test performance. The treatment effect detected by the mixed effects model thus confirms the impression from the descriptive statistics that the retrieval condition was significantly more effective than the trial-and-error condition. This answers our principal research question.

According to the model, test time (immediate vs. delayed) was also a significant predictor of test performance: There was significant loss in knowledge for participants in the retrieval group ($z = 7.888$, $p < .001$) and even more so for those in the trial-and-error group ($z = 9.237$, $p < .001$). Although at first glance the fall in test scores appears more pronounced in the trial-and-error condition, the mixed effects model indicates that the degrees of attrition were not significantly different.

Gauging the effect of feedback on trial and error

When estimating the effect of the feedback that was part of the trial-and-error procedure, it is worth distinguishing between the cases where students provided a correct exercise response (i.e., 13%) and those where the exercise response was wrong—which constituted the bulk (i.e., 87%) of the data. That close to 13% of the trial-and-error exercise responses were correct is somewhat surprising, since target phrasal verbs had been selected for the experiment that, according to the pilot study,

were unlikely to be known to these learners. It is impossible to tell with certainty whether these correct exercise responses were lucky guesses or if they reflected prior knowledge. Interestingly, however, only a very small fraction of these correct exercise responses was followed by correct post-test responses, which lends credibility to the assumption that most of the correct exercise responses were indeed guesses. Of the 127 correct exercise responses, only 15 (12%) were followed by correct responses in the immediate post-test and even fewer, just five (4%), were followed by correct responses in the delayed post-test. The probability of providing a correct response on the immediate post-test when a correct response was produced on the exercise is predicted to be only 0.10 (95% CI: 0.052, 0.181). By the delayed post-test, this probability diminishes to 0.03 (95% CI: 0.012, 0.083). If so little learning occurred when the feedback confirmed an exercise response that happened to be correct, then the overall gains observed for the trial-and-error condition must be due mostly to the instances of *corrective* feedback. Indeed, the probability of producing a correct test response after receiving feedback on a failed exercise response was 0.71 (95% CI: 0.632, 0.790) in the immediate post-test. It fell to 0.46 (95% CI: 0.396, 0.559) in the delayed post-test, however, which suggests that the effect of corrective feedback was often short-lived.

While it is undeniable that the corrective feedback as part of the trial-and-error procedure stimulated learning, its benefits were clearly outweighed by the benefits of the study + retrieval procedure, especially if we exclude the small number (5%) of instances where a student failed at the exercise stage to successfully retrieve the studied material. When a correct response was successfully retrieved from memory, the probability of a correct test response is predicted to be as high as 0.78 (95% CI: 0.705, 0.839) for the immediate post-test and 0.59 (95% CI: 0.498, 0.68) for the delayed post-

test. It is these comparatively high probabilities that account for the overall superiority of the retrieval condition over the trial-and-error condition indicated by the mixed effects model reported above.

For completeness' sake, it is perhaps worth mentioning that, on those rare occasions when learners in the retrieval group failed in the exercise to accurately retrieve a studied item (5%, i.e., 49 out of 980 responses), they were extremely unlikely to produce the correct response in the post-tests. The probability of a correct test response when these participants had produced an incorrect exercise response is predicted to be only 0.04 (95% CI: 0.013, 0.175) for the immediate post-test and nil for the delayed post-test. This is not surprising, as learners in the retrieval condition received no feedback on their exercise responses. They may therefore not have realized when one of their exercise responses was incorrect, and consequently this was not repaired in the post-tests. This demonstrates that feedback is of course also useful after a retrieval effort, an issue we return to further below.

In summary, whereas in the retrieval condition it was the correct exercise responses (i.e., successful retrievals) that were associated with post-test success, in the trial-and-error condition it was the *in*correct exercise responses (followed by corrective feedback) that were more often associated with post-test success. It is therefore not surprising that the mixed effects model indicates an interaction between treatment condition and exercise performance ($z = -9.128$, $p < .0001$). Table 2 presents the pair-wise comparisons of the probabilities of post-test successes under the different 'scenarios', that is, whether exercise responses were correct or incorrect in a given treatment condition.

Table 2: Pairwise comparisons of probabilities of post-test successes.

Retrieval	vs	Trial and error	Immediate post-test	Delayed post-test
Exercise correct	>	Exercise correct	$z = 14.21$ ($p < .01$)	$z = 11.42$ ($p < .01$)
Exercise wrong	<	Exercise wrong	$z = -8.82$ ($p < .01$)	$z = -6.04$ ($p < .01$)
Exercise correct	>	Exercise wrong	$z = 3.27$ ($p < .01$)	$z = 4.65$ ($p < .01$)
Exercise wrong	=	Exercise correct	$z = -1.17$ ($p = 0.13$)	$z = -0.65$ ($p = .26$)

As an additional analysis, we examined where students made an error in the trial-and-error exercises, and we compared this response to the one supplied for the equivalent item in the post-tests. This was done with a view to calculating the proportion of *duplicated errors*, because such duplications may reflect the failure of corrective feedback to override memories left by initial responses (see also Warmington & Hitch, 2014). Of the 853 exercise errors made under the trial-and-error condition, 120 (14%) were duplicated in the immediate post-test and 213 (25%) were duplicated in the delayed post-test. This suggests that the corrective feedback often prevented the initial wrong answer from re-emerging in the short-term (while not necessarily supplanting it by the correct answer). However, the fact that no fewer than 25% of the wrong exercise responses resurfaced in the one-week delayed post-test indicates that this effect of the corrective feedback was often short-lived. Interestingly, only 52 (i.e., 6%) of the 853 incorrect responses on the exercise were reiterated across *both* post-tests. Instead, learners sometimes (19%) reverted to their initial, incorrect hunch when they took the delayed post-test, despite having temporarily discarded it when they took the immediate post-test shortly after receiving the corrective feedback. This helps to explain why the gap between the two groups' scores was wider in the delayed than the immediate test.

It is worth putting the above figures into perspective, though, lest they leave too bleak an impression about the potential of corrective feedback. We also tallied the duplicated errors in the retrieval condition, that is, where no feedback was given. Of the 49 exercise errors made under this condition, 28 (57%) were duplicated in the immediate post-test, 30 (61%) were duplicated in the delayed post-test, and 18 (36.5%) of the exercise errors were reiterated across the two post-tests. Of course, the duplication of errors is unsurprising here, as the students had not been alerted to them in the first place. What nevertheless matters is that these are greater proportions than those observed above for the trial-and-error procedure, and this suggests that the feedback received in the latter condition did undeniably serve its purpose in a number of cases. The bottom line remains, however, that it was not effective enough to trump the learning gains obtained under the retrieval procedure.

General discussion

While trial-and-error implementations of exercises appear to be quite common in mainstream EFL textbooks, the results of our experiment indicate that, overall, a trial-and-error + feedback exercise procedure for learning phrasal verbs is less effective than a study + retrieval procedure. Our tallies of duplicated errors also demonstrate that, although corrective feedback on responses generated through trial and error can relatively often prevent learners from supplying the same erroneous responses in an immediate post-test, the initial erroneous responses are quite likely (25% in this study) to resurface in a delayed test.

At first sight, this contradicts the study by Potts and Shanks (2014) and other studies in favour of error generation followed by feedback. It needs to be borne in mind,

though, that those experiments involved a comparison with a study condition (where to-be-learned associations were simply presented to participants), not a retrieval condition. It is also worth pointing out that Potts and Shanks's (2014) experiment concerned the learning of L2 content words with discrete meanings, a task which may be different from the challenge of establishing syntagmatic partnerships among familiar L2 words (including function words such as particles) with vague or multiple meanings. For one thing, when asked to take a wild guess about the meaning of an unknown word form (as was done in Potts and Shanks, 2014), participants are not likely to give the wild guess much thought, will expect the guess to be wrong anyway, and may therefore not be committed to it. For another, the meaning proposed by the participant is unlikely to be similar to the correct answer that is given as feedback, and so the risk of confusion due to a pre-existing semantic relatedness between the two meanings is probably small. By contrast, when trying to guess, for example, if the appropriate word combination to express a given meaning is *catch on* or *catch up, give in* or *give up, brush up* or *brush off,* and *hold on* or *hold up*, then the guess one settles on will seem to stand a reasonable chance of being correct (if only because particles constitute a closed class, and so the number of choices is limited), and it may not look very distinct from alternative combinations (see, e.g., Hunt & Worthen, 2006, for a collection of studies on the role of distinctiveness in memory). This may then make it harder for the learner to prevent the erroneous hunch from resurfacing when re-presented with the same matching task (see, Boers et al., 2014, and Boers et al., 2017, for a similar account in relation to exercises on verb-noun collocations). In sum, it seems likely that the effectiveness of trial-and-error + feedback procedures will depend on the nature of the learning challenge. If the learning challenge is to establish durable memories of word partnerships, especially

ones comprising semantically vague or confusable constituents, then generating errors is probably not the most judicious first step in the learning process.

The finding that the study + retrieval procedure was more effective than the trial-and-error + feedback procedure in our study does align with those of Warmington et al. (2013) and Warmington and Hitch (2014). However, although participants in these studies were L1 speakers, they—like Potts and Shanks (2014)—also focused on the learning of new words with discrete meanings. Another difference with our experimental design regards the timing of retrieval and feedback. Warmington and colleagues presented their participants with the items to be learned one at a time, thus ensuring that the retrieval was effectively errorless and, in fact, virtually effortless. By the same token, the erroneous guesses in the trial-and-error condition were immediately followed by feedback. We shall return below to the potential relevance of these design choices for approximate replication studies.

It is worth mentioning that the retrieval condition in our experiment (at least as implemented here as a one-off retrieval event) did not bring about spectacular learning gains either—delayed post-test success rates were well below 60%. Discrete-point, minimally contextualized exercises such as the ones tried here (and mimicking common practice in EFL course materials) arguably fail to invite sufficient cognitive investment from learners. It would be interesting to compare their effectiveness to, for example, communicative tasks where learners purposefully incorporate studied phrasal verbs in their own L2 production, which would be more in keeping with constructs such as Laufer and Hulstijn's (2001) task-induced involvement. Another potential alternative approach that merits further exploration is to help learners appreciate the underpinnings of particle selection in those cases where a plausible explanation can be offered. Rather

obvious cases include the associations of *up* and *down* with 'more/less' (e.g., *turn up the volume*), with 'happy/unhappy' (e.g., *cheer up*), and with 'active/inactive' (e.g., *break down*). Others, such as the use of *out* in *find out (the truth)*, may necessitate slightly more elaborate reasoning (e.g., we associate 'knowing' with 'seeing'; if something is inside a container it may not be visible—it needs to be 'out' to become 'knowable') (see, e.g., Lindstromberg, 2010, for a helpful resource of explanations of this kind). Although studies have shown the benefits of this approach (Boers, 2013, for a review), it is not the approach taken in the mainstream EFL textbooks that we have analysed.

Limitations

Several limitations to this study need to be acknowledged. One is that we cannot tell on the basis of the exercise and test responses alone how the participants actually experienced the procedures and what thought processes drove their response behaviour. In future investigations, it would be useful to complement the offline measures with online processing data, so as to gain insight into learners' thought processes as they tackle the exercises, read the study materials, and tackle the post-tests. Think-aloud procedures may be suitable for this purpose.

Other limitations concern the way we operationalized the two learning conditions. For the sake of a fair comparison, we kept the quantity and content of the study and exercise materials identical in the two conditions, the order of presentation being the only difference. As a consequence of this methodological decision, the retrieval exercise was not followed by any feedback. Feedback could nevertheless have been useful (recall that there was a 5% error rate when the students did the retrieval exercise) and would in any case be realistic in actual practice—where students will

probably verify their exercise responses by re-visiting the study material, by checking an answer key where available, or by seeking confirmation from their teacher. Another consequence of using a balanced design with the same input for both treatment groups was that the feedback given on the trial-and-error exercise consisted of the intact form of the phrasal verb, a clarification of its meaning, and an example of usage. This is probably more than what is typically comprised in an answer key at the back of a textbook. We can only speculate about what the outcome would have been if we had imitated textbook content more closely, but it seems likely that (a) adding a feedback stage to the retrieval condition and (b) reducing the feedback material for the trial-and-error condition to a mere answer key would have widened the gap between the two groups' post-test scores.

While our experimental design was informed by the textbook analysis, we decided to use a computer interface that prevented students from navigating back and forth between the study materials and the exercises. In the case of print textbooks, however, students and teachers may not consistently adhere to the order in which materials are presented nor follow all the instructions to the letter. For example, there is little to prevent students from consulting study materials while tackling a given exercise, especially if both are on the same page or on opposite pages (e.g., the layout used in McCarthy & O'Dell's books for independent study). In a similar vein, students may not wish to make wild guesses when given a trial-and-error exercise, and peek at the answer key instead if they know one is available. It is also conceivable that students will review exercises they have already done, as they prepare for tests or exams. Although a given textbook may include just one exercise on a set of items, this does not exclude the possibility that the textbook user will revisit that exercise. Nor does it

exclude the possibility that the learner will seek additional exercise materials. This would in fact be highly advisable as a way of adding retrieval opportunities to the learning process.

Five more choices we made as to the experimental design should be mentioned, because they all seem worth reconsidering in approximate replications. First, only items that were deemed to be new to the learners were selected as targets. In real practice, it is likely that at least a few of the items targeted in a given section of a textbook are already familiar to (some of) the students. Second, we selected target phrasal verbs through piloting with a group of students who were not the participants in the actual experiment. This was an alternative to pre-testing the actual participants (because this would have constituted a trial-and-error event in its own right). On the downside, we cannot rule out the possibility that some items that were unknown to the pilot group might have been known to some of the students in the treatment groups. Additional measures are conceivable, such as a pre-test that is administered long enough in advance of the actual study, querying students after the procedure which items, if any, they already knew, and administering the pre- and post-tests also to a control group with a view to obtaining a baseline estimate of the (pre-)testing effect. Third, the students were required to tackle seven phrasal verbs in one go. The retrieval exercise must therefore have involved a certain amount of effort and was not error-free. For the trial-and-error condition, this also meant that feedback was given after a certain delay (i.e., after seven items were completed). In a different design, retrieval could be rendered fully errorless by prompting retrieval immediately after the presentation of each study item. Alternatively, a feedback stage could be added to the retrieval exercise to alert learners to unsuccessfully retrieved items. Also, feedback could be provided immediately after

each trial-and-error response (e.g., Hays, Kornell, & Bjork, 2013; Warmington, et al., 2013). Fourth, we required participants to work alone on the exercises. In practice, students may be invited to work collaboratively (e.g., Nassaji & Tian, 2010), so they can pool knowledge or at least compare the plausibility of their guesses. Recording the students' interactions during collaborative work could then be another way of revealing the thought processes triggered by exercise formats and learning procedures. Fifth, we included only one exercise/test format, i.e., eliciting the particle of phrasal verbs. Other popular formats elicit the verb instead of the particle, or elicit the complete phrasal verb, or present learners with options to select from.

Finally, we should bear in mind that it is possible to create learning conditions which combine trial-and-error and retrieval procedures. For example, one could begin with a trial-and-error exercise followed by feedback and later on present a similar exercise again to induce retrieval, which could again be followed by feedback. However, as already mentioned, such repeated revisiting of the same target items is not a common feature of the textbooks we have analysed. It will therefore be up to the teacher and/or learner to create or seek such repeated learning opportunities.

Conclusions and Implications

Because we discerned two rather different implementations of exercises on phrasal verbs in mainstream EFL textbooks, where they either promote retrieval of previously studied material or promote learning of material after a trial-and-error episode, we designed a controlled experiment to compare the effectiveness of these two types of implementation. The results of the experiment suggest that a procedure where the exercise follows study materials and thus promotes retrieval trumps a procedure where

the exercise functions as a trial-and-error event followed by feedback. We have argued that constructions such as phrasal verbs are perhaps non-optimal candidates for a trial-and-error + feedback procedure, owing to the limited semantic distinctiveness and high degree of confusability of their constituents, making it hard for corrective feedback to eradicate memories left by error-prone response generation.

Although much more research is necessary for a conclusive picture to emerge, these findings may begin to inform pedagogy in the following ways. One concerns the design of textbooks. If textbook authors wish to include exercises on phrasal verbs of the kind examined here—and we are by no means implying that they should—, then our findings suggest that precedence should be given to implementations that promote retrieval over ones where students are invited to make guesses under the assumption they will remember the correct answers presented as feedback. Another regards teachers' and learners' use of their textbooks. Even if a given textbook appears to rely on trial-and-error + feedback for the learning of phrasal verbs, it is possible to adapt the order in which materials are dealt with and to supplement what is missing as pre-exercise study material. A textbook is no straightjacket.

While the impetus for the present study came from patterns we discerned in print textbooks, the results can also be informative for the design of online study materials and for advice that could be given to users of such materials. ESL/EFL websites (e.g., https://www.ego4u.com/en/cram-up/grammar/phrasal-verbs) will typically present users with navigational options where they can click links to exercises or links to information about the target items. Users may feel tempted to try the exercises straightaway (if only because testing oneself may be more appealing than passively reading explanations

beforehand). In the case of phrasal verbs, however, learners may need to be aware that the aphorism *we learn from our mistakes* may not necessarily hold true.

Notes

1. In the realm of grammar instruction, learning through trial-and-error would qualify as inductive learning, where learners gradually work out a general pattern, system, or 'rule'. However, because phrasal verbs learning in this study is a matter of item learning, we have refrained from the use of the terms inductive versus deductive learning procedures.

2. It has been argued (e.g., Wray, 2002) that adult second language learners will often resort to a word-by-word analysis of multiword expressions. This is different from how most multiword expressions are acquired in the native tongue, where they tend to be processed holistically. However, even if L2 learners are generally inclined to take recourse to a word-by-word analysis, this will require that they discern and identify the words first—which in fluent aural input may not be self-evident in the case of short, non-salient words such as particles.

3. It is worth mentioning that all the target items used here were 'true' phrasal verbs, that is, with the particle functioning as an adverbial. This distinguishes phrasal verbs from prepositional verbs (e.g., *look into* [something]), where the 'spatial' word functions as a preposition instead.

4. It was decided in advance that the responses would be were dichotomously and that no partial credit would be given in the case of misspelled answers. Particles are very short words, after all, so that changing one letter can either make the particle unrecognizable (e.g., ?*ot* could be intended as *out* or as *on*) or change it into another one

(e.g., *in* and *on*). This would have made it difficult to determine a participant's intended choice of particle when it was misspelled, and hence also difficult to determine if that intended choice might have been correct.

5. For readers who are unfamiliar with mixed logits modelling, we here add an alternative table with the mean scores (and standard deviations) obtained on the exercises and the two post-tests.

Table 3: Exercise and post-test scores by treatment group (max score = 14)

	Exercise		Immediate post-test		Delayed post-test	
	Mean	SD	Mean	SD	Mean	SD
Retrieval	13.3	1.23	9.9	2.2	7.7	2.4
Trial and error	1.81	1.89	8.6	2.8	5.8	1.9

Acknowledgements

We are grateful for the financial support from the Faculty of Humanities and Social Sciences (grant # 207988) of Victoria University of Wellington which enabled the first author to conduct the experiment in Japan. We would like to thank the students for their participation in the experiment, Ariel Sorensen for assisting with the data collection, and Lisa Woods for her advice on statistical analyses.

This article is based on a part of the first author's PhD work, which received constructive feedback from examiners Dilin Liu, Shaofeng Li and Anya Siyanova-Chanturia. The quality of the article itself benefited substantially from three anonymous reviewers' insightful comments and suggestions. We also extend our gratitude to the journal editor, Charlene Polio, for her guidance and detailed feedback on the pre-final version of our text.

References

Boers, F. (2000). Metaphor awareness and vocabulary retention. *Applied Linguistics*, 21, 553–571. doi: 10.1093/applin/21.4.553

Boers, F. (2013). Cognitive Linguistic approaches to second language vocabulary: Assessment and integration. *Language Teaching*, 46, 208–224. doi: 10.1017/S0261444811000450

Boers, F., Dang, C.T., & Strong, B. (2017). Comparing the effectiveness of phrase-focused exercises. A partial replication of Boers, Demecheleer, Coxhead, and Webb (2014). *Language Teaching Research*, 21, 362–280. doi: 10.1177/1362168816651464

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effects of exercises on verb–noun collocations. *Language Teaching Research*, 18, 54–74. doi: 10.1177/1362168813505389

Alejo-Gonzáles, R., Piquer-Píriz, A.M., & Reveriego-Sierra, G. (2010). Phrasal verbs in EFL course books. In S. De Knop, F. Boers & A. De Rycker (Eds.), *Fostering language teaching efficiency through Cognitive Linguistics* (pp. 59-78). Berlin: Mouton de Gruyter.

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, 57, 35-56. doi: 10.1111/j.1467-9922.2007.00398.x

Bates, D., Maecher, M., Bolker, B.M., & Walker, S. (2014). *Linear mixed-effects models using Eigen and S4*. http://CRANR-projectorg/package5lme4.

Bolinger, D.L.M. (1971). *The phrasal verb in English*. Cambridge, Mass: Harvard University Press.

Carpenter, S.K., & DeLosh, E. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition,* 34, 268-276. doi: 10.3758/BF03193405

Cull, W.L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology,* 14, 215-235. doi: 10.1002/(SICI)1099-0720(200005/06)14:3<215::AID-ACP640>3.0.CO;2-1

Dagut, M., & Laufer, B. (1985). Avoidance of phrasal verbs: A case for contrastive analysis. *Studies in Second Language Acquisition*, 7, 73-80. doi: 10.1017/S0272263100005167

Gairns, R., & Redman, S. (2011). *Idioms and phrasal verbs: Intermediate*. Oxford: Oxford University Press.

Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly,* 41, 339-360. doi: 10.1002/j.1545 7249.2007.tb00062.x

Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research,* 19, 645-666. doi: 10.1177/1362168814559798

Garnier, M., & Schmitt, N. (2016). Picking up polysemous phrasal verbs: How many do learners know and what facilitates this knowledge? *System,* 59, 29-44. doi: 10.1016/j.system.2016.04.004

Grimaldi, P.J., & Karpicke, J.D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, 40, 505-513. doi: 10.3758/s13421-011-0174-0

Hays, M.J., Kornell, N, & Bjork, R.A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 39, 290-296. doi: 10.1037/a0028468

Huelser, B.J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition*, 40, 514-527. doi: 10.3758/s13421-011-0167-z

Hulstijn, J., & Marchena, E. (1989). Avoidance: Grammatical or semantic causes? *Studies in Second Language Acquisition*, 11, 241-255. doi: 10.1017/S0272263100008123

Hunt, R.R., & Worthen, J.B. (Eds.) (2006). *Distinctiveness and memory*. New York: Oxford University Press.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434-446. doi: 10.1016/j.jml.2007.11.007

Kang, S.H., Lindsey, R.V., Mozer, M.C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review,* 21, 1544-1550. doi: 10.3758/s13423-014-0636-z

Karpicke, J.D. & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1250-1257. doi: 10.1037/a0023436

Karpicke, J.D., & Roediger III, H.L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language,* 57, 151-162. doi: 10.1016/j.jml.2006.09.004

Karpicke, J.D., & Grimaldi, P.J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review,* 24, 401-418. doi: 10.1007/s10648-012-9202-2

Karpicke, J.D., Lehman, M., & Aue, W.R. (2014). Retrieval-based learning: An episodic context account. *Psychology of Learning and Motivation*, 61, 237-284. doi: 10.1016/B978-0-12-800283-4.00007-1

Kornell, N., Hays, M.J. & Bjork, R.A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 35, 989-998. doi: 10.1037/a0015729

Kövecses, Z., & P. Szabó, P. (1996). Idioms: A view from cognitive semantics. *Applied Linguistics, 17*, 326-355. doi: 10.1093/applin/17.3.326

Laufer, B., & Eliasson, S. (1993). What causes avoidance in L2 learning: L1-L2 differences, L1-L2 similarity, or L2 complexity? *Studies in Second Language Acquisition*, 15, 35-48. doi: 10.1017/S0272263100011657

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics,* 22, 1-26. doi: 10.1093/applin/22.1.1

Liao, Y., & Fukuya, Y. (2002). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning,* 54, 193-226. doi: 10.1111/j.1467-9922.2004.00254.x

Lindstromberg, S. (2010). *English prepositions explained* (2nd ed.). Amsterdam: John Benjamins.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, 65, 185-207. doi: 10.1111/lang.12117

Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly*, 45, 661-688. doi: 10.5054/tq.2011.247707

McCarthy, M., & O'Dell, F. (2004). *English phrasal verbs in use*. Cambridge: Cambridge University Press.

McCarthy, M., & O'Dell, F. (2007). *English phrasal verbs in use: Advanced*. Cambridge: Cambridge University Press.

Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, 37, 677-711. doi: 10.1017/S0272263114000825

Nakata, T. (2017). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition, 39*, 653–679. doi: 10.1017/S0272263116000280

Nassaji, H., & Tian, J. (2010). Collaborative and individual output tasks and their effects on learning English phrasal verbs. *Language Teaching Research*, 14, 397-419. doi: 10.1177/1362168810375364

Potts, R., & Shanks, D.R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, 143, 644-667. doi: 10.1037/a0033194

Pyc, M.A., & Rawson, K.A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition,* 35, 1917-1927 doi: 10.3758/BF03192925

Roediger III, H. L., & Butler, A.C. **(**2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences,* 15, 20-27. doi: 10.1016/j.tics.2010.09.003

Roediger III, H. L., & Karpicke, J. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science,* 1, 181-210. doi: 10.1111/j.1745-6916.2006.00012.x

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing,* 18, 55-88. doi: 10.1177/026553220101800103

Schuetze, U. (2015). Spacing techniques in second language vocabulary acquisition: Short-term gains vs. long-term memory. *Language Teaching Research*, 19, 28-42. doi: 10.1177/1362168814541726

Seltman (2014). *Experimental design and analysis*. Pittsburgh, PA: Carnegie Mellon University.

Side, R. (1990). Phrasal verbs: Sorting them out. *ELT Journal*, 44, 144-152. doi: 10.1093/elt/44.2.144

Siyanova, A., & Schmitt, N. 2007. Native and non-native use of multi-word vs. one-word verbs. *International Review of Applied Linguistics in Language Teaching,* 45, 119-139.

Slamecka, N., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory,* 4, 592-604. doi: 10.1037/0278-7393.4.6.592

Slobin, D. (1996). From thought and language to thinking for speaking. In J. Gumperz & S. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70-96). Cambridge: Cambridge University Press.

Smith, M.A., Roediger III, H.L, & Karpicke, J.D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 1712. doi: 10.1037/a0033569

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description, Vol. III: Grammatical categories and the lexicon* (pp. 56-149). Cambridge: Cambridge University Press.

Warmington, M., & Hitch, G.J. (2014). Enhancing the learning of new words using an errorless learning procedure: Evidence from typical adults. *Memory, 22*, 582-594. doi: 10.1080/09658211.2013.807841

Warmington, M., Hitch, G.J., & Gathercole, S.E. (2013). Improving word learning in children using an errorless technique. *Journal of Experimental Child Psychology, 114*, 456-465. doi: 10.1016/j.jecp.2012.10.007

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

Yang, C., Potts, R., & Shanks, D.R. (2017). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1073-1092. doi: 10.1037/xlm0000363

Yasuda, S. (2010). Learning phrasal verbs through conceptual metaphor. *TESOL Quarterly*, *44*, 250-273. doi: 10.5054/tq.2010.219945

Appendix 1: **Textbooks analysed**

Bygrave, J. (2012). *New total English: Starter student's book*. Harlow: Pearson Education.

Bygrave, J. (2014). *New total English: Elementary student's book*. Harlow: Pearson Education.

Clandfield, L. (2006). *Straightforward: Elementary student's book*. Oxford: MacMillan.

Clandfield, L. (2007). *Straightforward: Beginner student's book*. Oxford: MacMillan.

Clandfield, L. (2011). *Global: Intermediate student's book*. Oxford: MacMillan.

Clandfield, L., & Jeffries, A. (2010). *Global: Pre-intermediate student's book.* Oxford: MacMillan.

Clandfield, L., Benne, R., & Jeffries, A. (2011). *Global: Upper intermediate student's book*. Oxford: MacMillan.

Clandfield, L., McAvoy, J., & Pickering, K. (2010a). *Global: Beginner student's book*. Oxford: Macmillan.

Clandfield, L., McAvoy, J., & Pickering, K. (2010b). *Global: Elementary student's book*. Oxford: MacMillan.

Clandfield, L., Jeffries, A., Benne, R., & Vince, M. (2011). *Global: Advanced student's book.* Oxford: MacMillan.

Clare, A., & Wilson, J. (2011a). *Speakout: Pre-intermediate student's book*. Harlow: Pearson Education.

Clare, A., & Wilson, J. (2011b). *Speakout: Intermediate students' book*. Harlow: Pearson Education.

Clare, A., & Wilson, J. (2012). *Speakout: Advanced student's book*. Harlow: Pearson Education.

Crace, A., & Arcklam, R. (2011). *New total English: Pre-intermediate student's book.* Harlow: Pearson Education.

Crace, A., & Arcklam, R. (2012). *New total English: Advanced student's book*. Harlow: Pearson Education.

Eales, F., & Oakes, S. (2011a). *Speakout: Elementary student's book*. Harlow: Pearson Education.

Eales, F., & Oakes, S. (2011b). *Speakout: Upper intermediate student's book*. Harlow: Pearson Education.

Eales, F., & Oakes, S. (2012). *Speakout: Starter student's book*. Harlow: Pearson Education.

Hancock, M., & McDonald, A. (2008). *English result: Elementary student's book*. New York: OUP.

Hancock, M., & McDonald, A. (2009). *English result: Intermediate student's book*. New York: OUP.

Hancock, M., & McDonald, A. (2010a). *English result: Pre-intermediate student's book*. New York: OUP.

Hancock, M., & McDonald, A. (2010b). *English result: Upper-intermediate student's book*. New York: OUP.

Jones, C., Bastow, T., & Jeffries, A. (2010). *New inside out: Advanced student's book*. Oxford: MacMillan.

Kay, S., & Jones, V. (2007). *New inside out: Beginner student's book*. Oxford: MacMillan.

Kay, S., & Jones, V. (2008a). *New inside out: Elementary student's book*. Oxford: MacMillan.

Kay, S., & Jones, V. (2008b). *New inside out: Pre-intermediate student's book*. Oxford: Macmillan

Kay, S., & Jones, V. (2009a). *New inside out: Intermediate student's book*. Oxford: MacMillan.

Kay, S., & Jones, V. (2009b). *New inside out: Upper intermediate student's book*. Oxford: MacMillan.

Kerr, P. (2007). *Straightforward: Pre-intermediate student's book*. Oxford: MacMillan.

Kerr, P., & Jones, C. (2005). *Straightforward: Intermediate student's book.* Oxford: MacMillan.

Kerr, P., & Jones, C. (2007). *Straightforward: Upper intermediate student's book*. Oxford: MacMillan.

Norris, R. (2008). *Straightforward: Advanced student's book*. Oxford: MacMillan.

Oxenden, C., & Latham-Koenig, C. (2006). *New English file: Intermediate. Student's book.* New York: OUP.

Oxenden, C., & Latham-Koenig, C. (2008). *New English file: Upper-intermediate. Student's book.* New York: OUP.

Oxenden, C., & Latham-Koenig, C. (2009). *New English file: Beginner. Student's book*. New York: OUP.

Oxenden, C., & Latham-Koenig, C. (2010). *New English file: Advanced. Student's book.* New York: OUP.

Oxenden, C., Latham-Koenig, C., & Seligson, P. (2006a). *New English file: Elementary. Student's book*. New York: OUP.

Oxenden, C., Latham-Koenig, C., & Seligson, P. (2006b). *New English file: Pre-intermediate. Student's book*. New York: OUP.

Roberts, R., A. Clare, A., & Wilson, J. (2011). *New total English: Intermediate student's book.* Harlow: Pearson Education.

Soars, L., & Soars, J. **(**2003). *New headway: Intermediate student's book*. New York: OUP.

Soars, L., & Soars, J. (2008). *New headway: Advanced. Student's book*. New York: OUP.

Soars, L., & Soars, J. (2011). *New headway: Elementary student's book*. New York: OUP.

Soars, L., & Soars, J. (2012). *New headway: Pre-intermediate student's book*. New York: OUP.

Soars, L., & Soars, J. (2014). *New headway: Upper-intermediate student's book*. New York: OUP.

Appendix 2: Number of exercises in the textbooks focusing on the meaning of intact phrasal verbs and on the composition of phrasal verbs.

Textbook series	**Exercises on intact phrasal verbs and their meaning**	**Exercises on the composition (verb – particle) of phrasal verbs**	**Total**
English Result	8	8	16
Global	8	1	9
New English File	10	8	18
New Headway	8	8	16
New Inside Out	11	5	16
New Total English	19	7	26
Speak Out	16	5	21
Straightforward	12	6	18
Total	**92**	**48**	**140**

Appendix 3: Number of retrieval and trial-and-error exercises in the textbooks.

Textbook series	**Retrieval exercises**	**Trial-and-error exercises**	**Total**
English Result	2	14	16
Global	3	6	9
New English File	6	12	18
New Headway	10	6	16
New Inside Out	6	10	16
New Total English	9	17	26
Speak Out	12	9	21
Straight Forward	6	12	18
Total	**54**	**86**	**140**