EVALUATING PHRASAL VERB EXERCISES: AN INVESTIGATION INTO THE EFFECTIVENESS OF ERROR-FREE AND TRIAL-AND-ERROR LEARNING

BY

BRIAN PHILIP STRONG

A thesis submitted to the Victoria University of Wellington
in fulfilment of the requirements for the degree of
Doctor of Philosophy

Victoria University of Wellington

2017

Table of Contents

List of Tables

Acknowledgements

To my research colleagues and mentors, thanks for offering me advice in the process of becoming a researcher in the field of applied linguistics. Frank Boers, thanks for providing guidance wrought with wisdom and for creating an enriching research space where I could pursue my interests. Stuart Webb, thanks for the guidance early on in my research, it was priceless.

To the examination committee, thank you for your thoughtful comments, and for the time you spent reading my dissertation and thinking about my work.

Cara, your patience is unending and support is priceless. Your help in all aspects of my research was invaluable, without it I might not have accomplished my goals. We journeyed together from one country to another, and you have been by my side. Now, it is my turn to give back to you as much as you have given to me. Thank you.

Amanda, sis, thanks for editing early drafts of this manuscript. Without your feedback, this dissertation would probably still be a jumble of letter and words. Thanks, too, for the positive support.

Mom and Dad, although you do not fully understand what my research is on, your love and encouragement were enough, thank you.

Fellow TESOL teachers in Japan, thank you so much for allowing me to carry out my data collection in your classrooms. Without the data your students provided, I would have nothing to report. Ariel Sorensen, you help is greatly appreciated.

Finally, to the Vocabulary Research Group in general and Mark Toomer in particular, our conversations on all things vocabulary related was extremely valuable in helping to shape my research. Thank you.

Abstract of Dissertation

Although phrasal verbs are perhaps the most challenging type of verb phrase for L2 students to learn, only a handful of studies have looked into the effects of methods to enhance their acquisition, but these studies focused exclusively on the use of exemplary study input materials. The present thesis investigates this topic by examining and comparing learning procedures that consist of a study trial followed by a test trial (retrieval conditions) with learning procedures that include a test trial followed by a study trial (generation conditions). In the field of cognitive psychology, these two procedures have received considerable attention for the learning of single words but less so in the field of applied linguistics for the learning of phrases. In essence, the use of retrieval and generation conditions for the learning of phrasal verbs by non-native speakers of English was examined in three separate studies in this thesis. Additionally, it investigated the extent to which versions of these conditions occur in general ESL/EFL course textbooks. The studies were carried out in L2 classrooms and data were collected electronically from a group of L2 learners whose L1 system lacks phrasal verbs. In the first study, 199 students from five parallel classes were assigned to either a retrieval condition or a generation condition. In the retrieval condition, the study trial presented a phrasal verb and a paraphrase of its meaning, and then the test trial displayed the initial letter of the verb followed by the paraphrase of the phrasal verb's meaning. The generation condition was comprised of the same two trials; however, the order of their presentation was reversed, so the test trial preceded the study trial. In the second study, 153 students from four parallel L2 classes were assigned to one of four conditions. All the conditions were comprised of the same study trial and test trial. The study trial presented a phrasal verb along with a paraphrase of its meaning. The test trial displayed the verb of the phrasal verb followed by the paraphrase of its definition. In the study-test condition, the study trial and the test trial occurred consecutively, while in the study-delay-test condition, they were separated by approximately 6.5 minutes. In the test-study condition, the study trial occurred immediately after the test trial, while in the test-delay-study condition, a 6.5-minute interval separated the two trials. The last study examined contextualized versions of the retrieval condition and the generation condition on 172 L2 students from six parallel L2 classes. This thesis produced the following main findings. First, a vast majority (72% to be exact) of phrase learning conditions in existing course textbooks are generation-oriented. Second, the experimental studies showed that retrieval learning conditions offer significantly better short-term learning of phrasal verbs than generation learning, although no such advantage was found in long-term learning retention. Third, contextualized learning (i.e., learning phrasal verbs with the contextual support of exemplars) yields more effective learning than decontextualized learning does. Fourth, errors produced in generation conditions are difficult to unlearn. Overall, these findings provide us with some new insights about the learning and teaching of phrasal verbs. These results also have direct, meaningful pedagogical implications for the teaching of phrasal verbs as they show specifically which teaching procedures are more effective and which ones are less effective or ineffective for the learning of these difficult English verb phrases.

Chapter 1

Introduction

As an English as a Foreign Language (EFL) Instructor in Japan for over 14 years, I was perplexed by how to effectively teach phrasal verbs (PVs) to Japanese EFL students. Based on my teaching experience, many of my students struggled with learning complex verbs, even the highly proficient ones.

As I learned more about these multi-word verbs, I came to realize that Japanese EFL learners are not the only group of students to find PVs challenging to learn. Like other students with different first languages (Dagut & Laufer, 1985; Laufer & Eliasson, 1993), they also appear to prefer to use single verb synonyms rather than PVs (Yasuda, 2010). If these verb phrases were of little importance in the English language, then they would infrequently be used by native English speakers, and this would make it somewhat acceptable for students to purposefully avoid their use. However, this is not the case. In fact, PVs are one of the most prolific types of verb phrases in the English language (Boulton, 2008; Gardner & Davies, 2007; Liu, 2010), particularly in informally spoken discourse (Garnier & Schmitt, 2016) and to a lesser extent in academic English. Given the significance of PVs, it is essential for learners to add them to their mental lexicon, even if these multiword verbs are extremely difficult to learn.

My pedagogical interest in PVs led me to examine how these multi-word verbs are handled in commercially available teaching and learning materials such as general ESL/EFL course textbooks. The textbooks were selected because they are used by countless students around the world to learn phrases, according to the publishers' promotional statements.

Additionally, they are used by many Japanese EFL teachers to teach phrases in L2 classrooms.

From my examination of these textbooks, I was able to discern broadly two classes of phrase-focused learning conditions. The first class I refer to as retrieval-oriented conditions. They are made up of two learning events: a study trial and a test trial. The study trial presents a phrase to be studied, while the test trial, which follows the study trial, elicits retrieval of the studied phrase. The second class I call generation-oriented conditions. They also consist of the two trials. The critical difference between the retrieval-oriented conditions and the generation-oriented conditions is that the trials are reversed, so in generation-oriented conditions, the test trial comes before the study trial.

Given the availability of these two classes of phrase-focused conditions, I wondered whether they were equally effective for learning PVs. In order to make predictions in this regard, I surveyed two bodies of literature: memory research in the field of cognitive psychology and L2 vocabulary studies in the field of applied linguistics. According to my readings in memory research, the effects of retrieval and generation conditions have often been found to be robust. However, in the L2 literature, I found that the effects of these two types of conditions are mixed. Therefore, the focus of this thesis is to shed new light on the effectiveness of these two types of conditions by comparing their effectiveness on the learning of PVs.

In this thesis, three studies were carried out to investigate and compare the effects of versions of retrieval and generation conditions on PV acquisition by Japanese EFL learners. In study 1, a retrieval condition was compared to a generation condition. In study two, the separation between the study trial and the test trial in the retrieval and generation conditions was manipulated to investigate massed and spaced learning. While Study 1 and 2 focused on decontextualized learning, Study 3 examined contextualized learning. The variants of the retrieval and generation conditions examined in these studies are especially relevant in light of the ongoing debate about whether retrieval conditions, which minimize the risk of erring in the test trial, are more effective than generation conditions, where learners are expected to learn from their mistakes.

Organization of the thesis

The following is a brief description of the chapters in this thesis. Following this introductory chapter is the literature review chapter (Chapter 2), which is divided into two parts. Chapter 3 presents a textbook analysis that sought to identify phrase learning conditions in general ESL/EFL course textbooks. Chapters 4, 5, and 6 present the experimental studies that were described above. Chapter 7 presents a discussion of the results and the limitations of the research. The conclusion chapter (Chapter 8) summarizes the thesis as a whole.

Chapter 2 - Literature review

Part I: Phrasal Verbs

2.1 Introduction

This thesis investigates procedures to foster PV acquisition. This chapter is divided into two parts. Part I discusses issues concerning phrases in general and PVs in particular. Part II looks at procedures to enhance learning and retention. Part I begins with a discussion of phrases (Section 2.2). It then looks more closely at the characteristics of PVs (Section 2.3). After that, the importance of PVs in the English language is discussed (Section 2.4). Section 2.5 discusses factors that may hinder PV acquisition. Section 2.6 reviews empirical studies that have examined L2 learners' behaviour towards PVs. The next section of Part I reviews empirical studies that have investigated how to enhance L2 learners' knowledge of PVs (Section 2.7). The last section (Section 2.8) summarizes Part I. Part II of this chapter begins with an introduction (Section 2.9) followed by a discussion of the processes involved in learning with respect to encoding (Section 2.10), storage (Section 2.11), and retrieval (Section 2.12). Section 2.13 looks at prominent theories of retrieval, respectively. The Sections 2.14 and 2.15 discuss the testing effect and the pretesting effect, respectively. Comparisons between testing conditions and pretesting conditions are looked at in Section 2.16. The next section looks at factors that affect the effectiveness of retrieval and generation conditions (Section 2.17). The chapter ends by pointing out the gaps from the literature and shows how these gaps were addressed in this thesis (Section 2.18). Let us now look at each issue in turn.

2.2. Phrases

The term *phrase* is used as an umbrella term to refer to a wide range of expressions comprised of more than a single word. It also is used in the sense of multiword unit (or phraseology), not the way it is used in syntax. In the literature, various terms have been used to refer to the phraseological dimension of languages such as *prefabs, formulaic sequences, chunks, collocations,* and *idioms*. Over the past forty years, phrases of all sorts, including word combinations commonly referred to in the literature as formulaic sequences and multiword units (Cowie, 1992; Schmitt, 2004), have attracted the attention of linguists from various sub-disciplines. Corpus linguists have demonstrated that words tend to form partnerships with other words. Sinclair (1991) referred to this phenomenon as the idiom principle, which holds that the nature of language is “non-random” (p. 109). Psycholinguists have investigated whether native (and nonnative speakers) process phrases faster than non-phrases and, based on this information, speculate as to whether these word combinations are stored as single units or as separate words (e.g., Cacciari & Tabossi, 2014; Siaynova-Chantura & Martinez, 2014; Wray, 2002). Phraseologists have undertaken serious efforts to establish criteria to distinguish phrases consistent with particular structural and semantic characteristics. Nesselhauf (2003) distinguishes three groups (free combinations, collocations, and idioms) of word combinations, Cowie (1981, 1994), distinguishes five groups (free combinations, collocations, restricted collocations, idioms, and pure idioms) of word combination. However, after nearly half a century, linguists still have not agreed on a comprehensive definition of the phenomenon, and the likelihood of this occurring is slim to none due to the fact that phrases seem to exist in so many different forms (Schmitt & Carter, 2004)

Corpus studies have underscored the significance of phrases as vocabulary deserving of recognition by applied linguists, educational linguists, and L2 teachers. Linguists, using corpora (an extensive collection of written and spoken texts), have sought to determine whether phrases frequently occur in the English language. Although the estimates researchers provide vary widely depending on the size and the type of corpora they used as well as the criteria they employed to delimit phrases, their findings suggest that a large proportion of discourse is phrasal. Jackendoff (1995) analyzed the vocabulary used in a game show (Wheel of Fortune) and estimated that native speakers know approximately 80,000 phrases. Biber and Conrad (1999) calculated that around 30% of a conversation corpus consists of phrases. Erman and Warren (2000) found that phrases make up approximately 58.6% of a spoken corpus and 52.3% of a written corpus. Howarth (1998) and Foster (2001) suggested that phrases account for up to 30% of the corpora they examined. Overall, corpus studies propose that phrases constitute from one-third to one-half of any discourse (Conklin & Schmitt, 2012).

While there is a general consensus that phrases are prevalent in the English language, linguists continue to contemplate by virtue of what means to evaluate the semantic attributes of these word combinations. Phrases have meanings that can range on a semantic continuum from the literal to the idiomatic. The literal ones appear fairly straightforward to identify as the component words retain their literal meanings. *Read a newspaper* describes exactly what a person is doing to a newspaper: reading it. Previously, there was little debate as to what constitutes an idiom. Traditionally, they were viewed as phrases that admit meanings that are not the summation of the meanings of their individual words (e.g., Nayak & Cutting, 1989). A classic example is *kick the bucket.* Literally, kick specifies an action done to a bucket. The literal meanings of these words cease to represent their forms when they co-occur in a particular order. Instead, the words function in much the same way as a single orthographic unit and admit a single meaning. However, in recent years, linguists have ascertained the historical origins of many idioms and identified them as metaphorically motivated. For example, *bite the bullet,* meaning to accept something difficult or unpleasant, once regarded as semantically opaque, has its origin in days when anaesthesia was unavailable during a battle and doctors asked the injured soliders to bite down on a bullet to distract from the pain.

The bulk of phrases is assumed to lie somewhere between the extreme ends of the semantic continuum. Some have referred to the space between fully transparent and fully opaque as figurative (as well as semi-literal, semi-idiomatic, not fully compositional and so on). Therefore, figurative phrases are neither fully compositional nor idiomatic. A figurative phrase may consist of words where one retains its literal meaning while the other takes on a non-literal one. *Run a bath* is an example. The base meaning of *run* is to go faster than walking but is not used in its literal sense. However, *bath* continues to mean a large container of water. Why then is run used to express "prepare a bath?" One reason is that run elicits an image of fast movement which seems to capture the flow of water from the tap accurately. Linguists continue to document the semantic attributes of phrases.

The prevalence of phrases in the English language implies that they are a fundamental ingredient of fluency, where language production occurs without hesitation. Pawley & Syder (1983) argued that fluency is the natural outcome of phrases stored in memory as unanalyzed chunks available for use in much the same way as single words. As a result, native speakers tend to process idiomatic phrases faster than non-phrases (Swinney & Cutler, 1979; Gibbs, 1980). However, non-native speakers appear to process idiomatic phrases and novel phrases at similar rates of speed but process idiomatic phrases more slowly than literal phrases (Siyanova-Chanturia, Conklin, & Schmitt, 2011). Despite the fact that non-native speakers may not process idiomatic phrases like native speakers, evidence suggests that increasing their phrasal storehouse positively influences their fluency (Towell, Hawkins & Bazergui, 1996; Boers, Eyckmans, Kappel, Stengers & Demecheleer, 2006; Stengers, Boers, Housen, & Eyckmans, 2010; Wood, 2002, 2009, 2010). Towell, Hawkins, & Bazergui (1996), for example, found that after spending a year immersed in an L2 speaking environment, students' fluency significantly improved more than if they had stayed in their home country. They explained that fluency developed because procedural memory was constantly active while the students were listening to the phrases used by native speakers. Wood (2009), in a case study, also showed that the fluency of a non-native speaker greatly increased after six weeks of focused instruction on phrases.

Non-native speakers' problems with phrases become evident in their language output, which may sound unnatural and non-native-like. One reason for this is the development of a limited and unbalanced phrasal repertoire which leads to the overuse of familiar phrases (Granger, 1998; Foster, 2001) and underuse less familiar ones (De Cock, Granger, Leech, & McEnery, 1998; Hasselgren, 1994; Howarth, 1998). Schmitt (2013) pointed out that one type of phrase non-native speakers underuse is PVs, partly because many are composed of delexicalized verbs, which feature in hundreds of phrases, and partly because no L1 equivalent exists in many non-native speakers' first languages (Dagut & Laufer, 1985). Several studies have investigated which phrases non-native speakers underuse and overuse (Nesselhauf, 2003; Laufer & Waldman, 2011). For example, Laufer & Waldman (2011) compared the use of verb-noun phrases in a non-native corpus of argumentative and descriptive essays with the LOCNESS corpus of young adult native speakers of English. Their analysis revealed that non-native speakers' written texts contained far fewer verb-noun phrases than native speakers' and that the use of these phrases increased only at the advanced levels.

Another problem that non-native speakers face is maintaining the integrity of phrases. They may have trouble deploying phrases accurately due to (1) the component words' lack of salience and (2) L1 interference. Boers, Dang & Strong (2017) point out that the partnership of the component words comprising phrases can be difficult for learners to remember. The ones most likely to be erroneously produced are consisting of high-frequency verbs or have high-frequency synonym competitors. Thus, it can be difficult for non-native speakers to remember which partnerships are restricted and which ones are less restricted (Nesselhauf, 2005, pp. 14-15). Several studies have observed that the infelicitous phrases used by non-native speakers show evidence of L1 interference. Nesselhauf's (2003, 2005) study showed non-native speakers' failure to use phrases correctly. She extracted 1072 verb-object-noun phrases from 32 essays written by German learners of English in the ICLE corpus. These phrases were judged according to an acceptability rating, which showed that nearly one-quarter (24%) were erroneous and that the L1 was partly responsible for 45% of the errors (Nesselhauf, 2003, p. 235). It was also observed that errors were mainly due to the wrong choice of verb (p. 231). For Nesselhauf (2003), this finding is not surprising considering that verbs in phrases are restricted, in the sense that some verbs do not collocate with every noun, which makes their use difficult (e.g., "do one's homework" rather than "make one's homework").

Vocabulary acquisition is an incremental process, and there is little reason to think that learning phrases is anything different. Estimates suggest that anywhere from six encounters to 50 encounters with single words may be necessary for them to be picked up incidentally (Brown et al., 2008; Rott, 1999). Therefore, this number of exposures may also be necessary for incidental learning of phrases. However, phrases naturally occur less frequently than single words in the English language. While exposure to massive amounts of language input will increase the probability of multiple encounters with phrases, this is unlikely to occur in foreign language learning environments where the input is insufficient and controlled. One way to increase exposure to phrases then is for learners to be immersed in an English-speaking environment. Another option is to increase the number of occurrences of phrases in the texts that they read (Pellicer-Sanchez, 2015; Webb, Newton & Chang, 2013). However, it is still too early to tell whether this approach makes any notable contribution to increasing learners' phrasal repertoire.

Most recently, two studies have investigated how many exposures might facilitate knowledge of phrases (Pellicer-Sanchez, 2015; Webb, Newton & Chang, 2013). Webb, Newton & Chang (2013), examined the extent to which infusing four graded readers with different numbers of encounters of phrases (1, 5, 10 and 15 encounters) would enhance knowledge of the phrases. Participants read and simultaneously listened to the graded readers. Before this learning phase, participants completed a pre-test of receptive knowledge of written form, whereby broken up phrases were rejoined by matching the verb of a phrase to its noun located in a list of distractor nouns. Following the learning phase, a series of post-tests were administered which included the same pre-test. Pre-post-test comparisons revealed that 15 encounters might allow for some knowledge of the phrases to be picked up incidentally. The rate of incidental learning of phrases was also examined by Pellicer-Sanchez (2015), who manipulated the frequency of pseudo-phrases (four and eight encounters each) occurring in a text. Participants read the text and then answered comprehension questions. The results of a one-week delayed post-test showed no significant differences in receptive and productive knowledge between the phrases that occurred at different frequency rates. One possible reason frequency had little effect on retention is that inferring the meaning of the pseudo-phrases was too difficult.

Another issue that affects the learnability of English phrases is their translatability to the learner's first language (Irujo, 1986; Laufer, 2000). English phrases with L1 translation equivalents are easier to learn than English phrases without any clear L1 translation equivalents. Laufer (2000) outlined four degrees of similarity between English phrases and L1 translation equivalents, which can be used as a guide to predict which phrases students will find harder to learn. First, learners will experience the most trouble when the English phrase has no counterpart in the learners' L1. Second, an English phrase with a meaning similar to an L1 phrase with a different form can cause learning difficulty. Third, an English phrase with a meaning similar to an L1 phrase with a similar form is less difficult to learn. Fourth, an English phrase with an identical translation places hardly any learning burden on the learner.

2.3 Phrasal Verbs

This thesis examines a particular sub-class of phrase that is referred to as the *phrasal verb* (PV). I use the term in a broad sense, referring to not only verb-adverb constructions but also verb-preposition constructions as well as verb-adverb-preposition constructions. The *Longman Dictionary of Phrasal Verbs* (1983) defines a phrasal verb as an idiomatic combination of a verb and adverb, or a verb and a preposition (or verb with both adverb and preposition). Some linguists use the term *multiword lexical verbs* as a coverall to refer to verb-particle combinations (as phrasal verbs), verb-preposition combinations (as prepositional verbs) and verb-adverb-preposition combinations (as phrasal-prepositional verbs) (e.g., Biber, Johansson, Leech, & Conrad, Finegan, 1999). Below, I use examples to illustrate why some analysts argue for treating phrasal verbs separate from prepositional verbs.

As shown in sentences (1), multiword verbs are comprised of a verb (e.g., carry, sat, ran, float) and an adverbial particle (e.g., out, down, up).

1. (a) John carried out research.
(b) Bob sat down.
(c) Jane ran up her phone bill.
(d) The balloon floated up.

The examples also show the transitivity of the verb-adverbial particle combinations, like (1a) and (1c) are transitive and (1b) and (1d) are intransitive. Further, the second lexical item denotes the path of motion. As for their meanings, examples (1a) and (1c) are non-compositional in the sense that the meanings of the verbs cannot be guessed from the verbs alone. The *Longman Dictionary of Phrasal Verbs* (1983) regards (1a) and (1c) as idiomatic combinations. In contrast, the meanings of examples (1b) and (1d) can be arrived at on the basis of the meanings of their parts.

Many of the lexical items that follow the main verb in (1) also occur in constructions other than PVs (Lindner, 1984). They occur in prepositional verbs, where the second lexical is a preposition and phrasal-prepositional verbs, where the second word is an adverb, and the third is a preposition, as exemplified in the following sentences:

2. (a) You may rely on me.

(b) I'll get back to you.

(c) Jane ran up a hill.

(d) The cat jumped on the TV.

The examples illustrate that prepositional verbs admit idiomatic combinations, as in (2a) and (2b) and literal ones, as in (2c) and (2d) (Courtney, 1983). They also express the direction or path of an activity, as in (2c), or the location of an activity, as in (2d). There are two ways to analyze their structure. One is to view the main verb as separate from the prepositional phrase, which then acts as an adverbial. The other is to consider the verb plus preposition as a single unit. From this perspective, the noun phrase following the preposition is analyzed as the object of the prepositional verb (Biber et al., 1999).

As mentioned before, some of the second lexical items in prepositional verbs overlap with the ones in PVs. Some analysts have, therefore, divided these items into three groups. The ones that only occur in PVs are considered as adverbs (e.g., across, along, around, away, aside, back, forth, over through); those that only occur in prepositional verbs, as prepositions (e.g., above, after, at, against, for, from, into, of, to, with, without); those that occur in either class, as prepositional adverbs (e.g., about, by, down, in, off, on, out, under, up) (Biber et al., 1999; Bolinger, 1961, Lindner, 1984, Sroka, 1962). While some linguists distinguish PVs from prepositional verbs (e.g., Lindner, 1984), others employ the term PV to include both PVs and prepositional verbs.

There are several tests that are offered in the literature for distinguishing PVs from prepositional verbs, namely the position of the particle relative to the direct object, stress, and the insertion of an adverbial phrase. The one most discussed are the word order criterion. It is used to distinguish combinations which make up transitive PVs from those which constitute prepositional verb constructions. According to Bolinger (1971), a particle can either occur before or after the noun phrase, whereas a preposition must precede it. To exemplify this distinction, Lindner (1984) offered some of the following examples:

3. (a) He ran up a bill
   (b) He ran up a hill.
   (c) He ran a bill up.
   (d) *He ran a hill up.
   (e) He ran it up (bill).
   (f) *He ran up it (bill).

The combination *run up* occurs in (3a) and (3b) and both are grammatically correct. However, placing *up* after the noun phrase indicates that *run up* is a transitive PV in (3a) and (3c), and a prepositional verb in (3b), as (3d) is grammatically incorrect. Where the direct object is a pronoun, the particle must necessarily be unconnected from the main verb, as in (3e) and cannot remain as a unit, as in (3f). Where there is no noun phrase following the particle, as in (4a), (4b) and (4c), then the combination is clearly intransitive (Armstrong, 2004) and so it is a PV.

4. (a) He shut up.
   (b) I threw up.
   (c) I ate out.

The system underlying PVs is semantically complex. Some combinations have meanings that can be worked out from an understanding of the meanings of their individual words, while others express meanings that are seemingly unrelated to the ones expressed by their parts. *Threw up* occurs in the (5a) and (5b). Its reading is *to propel something with force through the air by movement of the arm and hand*, and *up* means *towards a higher place or position.* While these meanings are present in (5a), they are not in (5b) -- at least, not in the conventional use of the PV in the context of eating a meal and becoming sick as a result.

5. (a) I <u>threw</u> <u>up</u> the ball.
   (b) I <u>threw</u> <u>up</u> my dinner.

In (5b), *threw* does not involve the movement of something through the air using the arm and hand. It still seems to retain the notion of propelling an object (i.e., vomit), but, instead of performing this action with one's hands, the action is performed involuntarily using muscles in the stomach. The intricacies of deciphering the semantic contributions of the lexical items to the whole have been a topic of extensive debate in the field of linguistics (e.g., see Bolinger, 1976; Lindner, 1981; Morgan, 1997).

It is the tendency of most linguists to attempt to categorize PVs according to some semantic criteria. There seems to be some agreement among analysts that complex verbs can be classified as either exhibiting degrees of literalness or degrees of figurativeness (Celce-Murica & Larsen-Freeman, 1999; Morgan, 1997). In the first category, transparent constructions are those which are made up of component words that largely retain their basic meanings, as in (6a), (6b), and (6c). In these sentences, the lexical items *up, away* and *down* maintain their prepositional meaning although they function as particles.

6. (a) I stand up.

(b) I threw away the garbage.

(c) I fell down.

(d) Jill drank up the milk.

(e) Let’s hang out.

In the second category, figurative PVs are composed of verbs that maintain their literal readings while the particles contribute aspectual meanings (Celce-Murica & Larsen-Freeman, 1999), or as Morgan (1997) puts it, the particles take on a cognitive image schema in the form of a metaphor. The lexical item *up* in (6d) inherits the conceptual metaphor *completion is up* while the meaning of the verb remains literal (see Lakoff & Johnson, 1984; Morgan, 1997). Combined with the meaning of the verb, the PV *drank up* indicates that Jill finished drinking all the milk, not just some of it.

Some linguists have argued for a third category to account for constructions where the individual words both exhibit opaque meanings, rendering the PV as an idiom. In (6e), *hang out* means something like *to spend time with friends.* This meaning is indistinguishable from the basic meanings of its component words: *hang* means *to suspend something from above,* and *out* means to *moving or appears to move away from a particular place.* However, cognitive linguists have argued that even the ostensibly idiomatic-behaving verb phrase in (6e) is decipherable (see Morgan, 1997).

2.4 Importance of PVs in the English language

In the previous section, I looked at some of the key structural and semantic characteristics of PVs. In the present section, I discuss the importance of these verb phrases in the English language to show why they are important for L2 students to learn. The section begins by looking at research that has examined the frequency of PVs in corpora. It then discusses the issue of the active and productive nature of complex verbs.

Despite their difficulty, PVs have to be learned at some stage because they are common in the English language (e.g., Bolinger, 1971). Several linguists have documented the prevalence of these complex verbs (Biber et al., 1999; Gardner & Davies, 2007; Liu, 2011). However, these studies may not have fully captured the ubiquity of this class of verbs in part because they used pre-defined search parameters in their corpus analysis. For example, Gardner and Davies (2007) created a PV list composed of only the 20 most frequent verbs in English. As a result, their list does not account for the most common PVs that consist of verbs outside the top 20 (e.g., Liu, 2011). Further, these studies restricted their analysis to the British National Corpus, which deals exclusively with British English. Despite these two limitations, these studies provide solid evidence that demonstrates that PVs are an essential feature of the English language. As exemplified by Boulton (2008), an individual can expect to hear a complex verb more than once in every five minutes in conversation based on a speech rate of 120 words per minute, or to read one on every page and a half of fiction based on an average of 400 words per page of text.

Another reason for L2 learners to acquire PVs is that native speakers exhibit a propensity to coin new ones regularly (Darwin & Gray, 1999). Bolinger (1971) indicated that complex verbs constitute "an explosion of lexical creativeness that surpasses anything else in our language" (p. xi) and that new ones are continuously being added to the English lexicon. For example, the invention of the computer led to the production of dozens of PVs as well as allowing old ones to take on new meanings, as exemplified in the following narrative: "When we buy a new computer, we need to *set* it *up*. After that, we need to *log in* to gain access to our computer's ability. On the internet, we usually have to *sign in* to our email account to read our emails. However, before we can use our computers, we need to *hook up* the cables and then *power* it *up* by *turning* it *on*. However, if you have an older model, it may take longer to *boot up*. Then, we want to install software. To do this, we need to *load up* the program by *clicking on* various command buttons. Once it is installed, we can navigate on the screen using our mouse, which allows us to *scroll up* and to *scroll down*. When we find something that we want to save, we *click on* it. It is always a good idea to *back up* your files as this will *free up* space on your hard drive so you won't *run out of* space. One last thing to keep in mind, don't *give out* your password because criminals want to *hack into* your account and possibly want to *wipe out* all your data. Oh yes, after you *log off*, don't forget to *shut down* your computer once you’re done with it."

The fact that native speakers seem to have little trouble in understanding novel PVs suggests that perhaps not all of them are stored as holistic units but are stored as decomposable lexical items, which are brought together by a lexical rule of some sort when the combination is required (Armstrong, 2004). For example, although a native speaker may have never heard the verb plus particle in (7a), it is possible to work out its meaning, which seems to be that this person desires to increase his or her morning dose of breakfast fibre.

7 (a) I like to fibre up in the morning.
(b) I'm woodpeckering away at my computer.

Similarly, if one understands the characteristic behaviour of woodpeckers and the aspectual meaning of the particle *away* (to show that an action continues), then the meaning of the PV in (7b) can most likely be successfully deduced without having to have learned it as an idiom beforehand. Pedagogically-minded cognitive linguists have exploited this way of processing complex verbs by raising learners' awareness of the underlying semantic system of the particles (see below).

This section has attempted to provide reasons for why L2 students need to learn PVs. It demonstrated that complex verbs occur very frequently in English, and this implies that they are known and regularly used by native speakers and that they are an element of native-like communication. This section further indicated that PVs are continuously being created and that old ones inherit new meanings, and this is further evidence of the importance for L2 students to learn PVs.

2.5 Factors influencing the learnability of PVs

The section discusses the challenges L2 learners experience in learning PVs. These challenges are caused by several factors that stem from the intrinsic properties of PVs (intralexical) as well as from the typological differences (interlexical) between English and the learners' L1.

2.5.1 Intralexical factors

There are several inherent characteristics of PVs that make them extremely difficult for L2 students to learn. The ones discussed below are length, word order, the semantic weight of the component words, the separa.1bility of the particle from the verb, and polysemy.

The length of a word can affect the degree to which an individual can learn a word. Long words may be more difficult to learn than short words (e.g., Campoy, 2008; Hulme, Maughan, & Brown, 1991). To learn a long word, a student has to remember more letters and how to pronounce those letters (Nation, 2001). Since PVs are made up of two or three words, a student might find it more difficult to remember the component words than their single-verb counterparts. Although length is likely a factor, it is challenging to measure its effect on PV learning, because it cannot be isolated from other factors, such as word order.

The order in which the component words of PVs occur may affect which of these words receives greater attention. Because the verb comes before the particle, students' attention is initially on the verb before the particle. As a result, greater attention may be placed on the verb than the particle, which, in turn, may strengthen the association between the verb and the meaning of the PV than between the particle and the meaning of the PV. Therefore, students may find it easier to recall the verb of a PV than its particle. Further, the verb tends to be longer than the particle, making it likely to be more salient than the particle. Like the length factor, it is difficult to assess whether word order is a significant factor because it cannot be isolated from other variables which have been found to affect the learning of other types of vocabulary.

The semantic weight of the individual words may be a significant factor behind the difficulties students have with learning PVs. Semantic weight is a function of the use of a word in different contexts. The greater the number of the various contexts in which a term may appear, the less the semantic weight. It is as if some of the meaning is rubbed off whenever a word is used in a different context (Kent & Lancour, 1971). The weight of words is determined by how frequently they occur in the language. Particles occur more frequently than verbs, so they may have little lexical meaning or are ambiguous (Carnap, 1937). As a result, the meanings of the verbs seem to carry most of the meaning of PVs. For example, a student may be able to understand the basic, although incomplete, meanings of *drink up,* *scroll down, ask around*, for instance, based on an understanding of the verbs alone. Nevertheless, the most frequent PVs consist of high-frequency verbs, and this means, relatively speaking, that they are semantically less distinct than the PVs that consist of lower frequency verbs.

The grammatical peculiarity of PVs is perhaps a major factor behind the difficulties L2 students face in learning PVs (Armstrong, 2004; Liu, 2011). As stated above, a particle can be decoupled from its transitive verb (Joe tore up the contract / Joe tore it up), but it cannot be separated from its intransitive verb (My car broke down on the street / *My car broke on the street down). The optional and obligatory separation of the component parts has the potential to confuse L2 students (Celce-Murica & Larsen-Freeman, 2016). For example, an L2 student might say I ran an old friend into yesterday rather than I ran into my friend yesterday perhaps because they are unaware that the separation of some PVs is not permissible (Azar, 1989, p. 2). De Cock's (2006) analysis of learner corpora (International Corpus of Learner English and Louvain International Database of Spoken English Interlanguage) showed that the most frequent error made by students involved using transitive PVs as intransitive ones as well as using intransitive PVs as transitive ones.

Another major obstacle in the learning of PVs is that each one has multiple meanings (Cornell, 1985). *Collins Cobuild Dictionary of Phrasal Verbs* lists 20 meaning entries for the phrasal verb *go down* and another eight for the phrasal verb *go down with*. *The Oxford Dictionary of Phrasal Verbs* (1993) lists approximately 114 complex verbs involving the headword come (come up, come down, come by, etc.), 20 of which are different entries for the combination come out. The semantic relation between these senses can be complicated for a non-native speaker to pick out, as exemplified in (8).

8. (a) Is John coming out?

(b) John came out.

(c) When does it come out?

The contexts surrounding the PVs are most likely too minimal to help learners decode the different meanings of come out. However, from a native speakers' perspective, they may be obvious: (8a) means to leave a place in order to do something, like play a game; (8b) means that John admitted his true sexual orientation; (8c) means to elicit information about something from someone. Changes in the meaning of PVs due to context can seriously hinder L2 students' ability to learn them (Armstrong, 2004). The fact that each complex verb has on average six meanings, which may or may not be related, presents many challenges for learners (Gardner & Davies, 2007)

The levels of idiomaticity are perhaps the most significant intra-lexical factor that affects students' ability to learn PVs. As discussed above, a complex verb may be used idiomatically in one context and literally in another context, as exemplified in (9a). One way to read the sentence is as a statement on the whereabouts of a man: he is no longer in a house; he has exited it. Another interpretation is as an announcement about a man revealing his sexual orientation.

9. (a) He came out.

(b) He passed away.

Other PVs have meanings that are extremely difficult to deduce based on the meanings of the component words. Some linguists might regard (9b) as an idiom, as the *MacMillan Dictionary* (www.macmillandictionary.com) states it means to die — and it is the only entry for it. While (9a) and (9b) are most likely extremely difficult for students to learn, (9a) is probably more laborious to acquire simply because one needs to firstly determine which of the twenty-one meanings of *come out* is used (see *The Longman Dictionary of Phrasal Verbs*, 1983), whereas, this task is much easier in (9b) as *pass away* has only two distinct non-compositional meanings.

2.5.2 Interlexical factors

Complicating matters even more, the absence of PVs in the learners' L1 is a major factor that can significantly hinder the learning process (Darwin & Gray, 1999). It is easier to learn any vocabulary if an equivalent of it exists in the learners' L1. Learners seem to have little trouble learning concrete objects in an L2 as the referent remains the same and only the word used to express it changes. The challenges of learning PVs are quite different from learning the names of objects or even common verbs such as *walk* and *talk* and *sit* and *eat*. PVs are present in Germanic languages and are rare in non-Germanic languages. Consequently, these two language systems encode motion events differently (Slobin, 2005). Germanic languages, such as English and Dutch and Swedish, encode the path of motion onto prepositions and conflate the manner of motion onto the main verb. In contrast, non-Germanic languages, such as Japanese and Hebrew and French, encode the path of motion onto the verb and put details about the manner of motion into other lexical items (e.g., Talmy, 2000). To illustrate this contrast, Yoshitomi (2006) showed that the PV *tiptoe out* translates into Japanese as *shinobiashi-de dete-iku* which literally means *on-tiptoe out-go* or *go out on tiptoe.* The differences between Germanic and non-Germanic languages means that English students with non-Germanic L1s are more likely to use awkward English translations of PVs or to avoid their use altogether (Dagut & Laufer, 1985).

Proficiency appears to be an additional factor that complicates PV acquisition. L2 students with higher levels of proficiency seem to experience less difficulty in understanding complex verbs than students with lower levels of proficiency (Liao & Fukuya, 2002). However, this may have to do less with proficiency level than about when students actually begin to learn these verb phrases explicitly. Many PV textbooks are designed for students with a proficiency level above intermediate (e.g., McCarthy & O'Dell, 2004; Gairns & Redman, 2011). Therefore, it is assumed that students with advanced proficiency levels received greater focused exposure to PVs than students with lower proficiency levels.

Summary

This section has shown that PVs are a peculiar subclass of phrasal constructions. Several factors were identified as potential obstacles to learning PVs. These factors were divided into intralexical and interlexical. While intralexical factors influence learning despite the students' L1s, it is assumed that interlexical factors are likely to hinder the learning of students with non-Germanic L1s greater than students with Germanic L1s. The evidence reported above indicates that students with non-Germanic L1s are more likely to express actions events using single verbs rather than PVs because no equivalent structure exists in their L1. This section clearly illustrates the need for an investigation into methods to help students overcome the difficulties they are certain to face in learning PVs.

2.6 Research on non-native speakers' PV avoidance behaviour

The previous section has shown that a plethora of factors can impede L2 learners' acquisition of PVs. Given these variables, learners might be inclined to use single verb substitutes instead of these verb phrases. This section reviews studies that have investigated the extent to which students have opted to use single verbs over PVs and discusses key factors that have been proposed to account for their preference to use single verbs.

Six studies have examined the PV avoidance behaviour of students from a range of L1 backgrounds, including Arabic, Chinese, Dutch, Hebrew, Italian, Russian and Swedish. Two studies used native speakers' preference for PVs as a baseline to compare non-native speakers' preference (e.g., Dagut & Laufer, 1985; Liao & Fukuya, 2002). One study compared two groups of learners with distinct L1s (Hebrew and Swedish; Laufer & Eliasson, 1993), while another examined a heterogeneous group (Arabic, Chinese, Italian and Russian; Siyanova & Schmitt, 2007). The remainder recruited L2 learners with the same L1 (Becker, 2014; Dagut & Laufer, 1985; Liao & Fukuya, 2002).

These studies gauged learners' preference to use PVs or their single verb counterparts using a series of elicitation tasks (Becker, 2014; Dagut & Laufer, 1985; Hulstijn & Marchena, 1989; Laufer & Eliasson, 1993; Liao & Fukuya, 2002; Siyanova & Schmitt, 2007). The elicitation tasks consisted of multiple-choice, translation and memorization. The multiple-choice task presented a sentence with a gapped space, and next to the gapped sentence was a PV and its single verb counterpart. Students were asked to select one of the options and insert it into the gapped space. The translation task presented an English sentence with a gapped space followed by L1 translations of the missing PV or its single-verb counterpart. Students were requested to translate one of the options and insert it into the blank. For the memorization task, a sentence was written in English and the learners' L1. After reading both sentences, the English sentence was presented again but without the verb. At the end of the sentence was the L2 translation of the missing verb. Students were asked to translate the verb into English and insert it in the gapped space. Slight variations in these tasks were used in each study.

Dagut and Laufer (1985) and Hulstijn and Marchena (1989) recruited Hebrew- and Dutch-speaking learners of English, respectively. While Hebrew is a non-Germanic L1 without PVs, Dutch is a Germanic L1 with PVs. On all the elicitation tasks, Dagut & Laufer (1985) found that students overwhelmingly avoided PVs in favour of their single verb counterparts. In contrast, Hulstijn & Marchena (1989) found that while intermediate learners tended to avoid the use of PVs, advanced learners were more likely to prefer them over their single verb counterparts. Dagut & Laufer (1985) attributed their findings to the absence of PVs in the learners L1, while Hulstijn & Marchena (1989) claimed that learners avoided PVs when they seemed “too Dutch-like” (p. 241).

Based on the findings of Dagut and Laufer’s (1985) study and Hulstijn and Marchena’s study (1989), Laufer and Eliasson (1993) compared the preference of using PVs by learners with two distinct L1s: Swedish which has a PV category and Hebrew which does not. Using the elicitation tasks, the authors showed that students of both language groups did not avoid using semi-transparent PVs more than literal ones but did avoid using the ones less semantically transparent. However, significantly fewer PVs were used by the Hebrew-speaking learners than the Swedish-speaking learners. They concluded that the main factor behind PV avoidance was L1-L2 differences.

Liao & Fukuya (2002) further investigated whether PV avoidance was mainly due to L1-L2 differences or to the different proficiency levels of learners. A group of Chinese-speaking learners of English of intermediate and advanced proficiency levels was given the elicitations tasks with literal and figurative PVs. The authors found that intermediate and advanced learners showed a greater preference for literal PVs than for figurative ones on the translation task but not on the other tasks. Further, intermediate learners showed a tendency to avoid PVs. Based on these findings, the authors attributed PV avoidance behaviour to the proficiency level of the learners as a sign of the lack of exposure to English in use.

Siyanova and Schmitt (2007) approached the issue of PV avoidance differently. They had a heterogeneous group of non-native and native speakers of English show their preference for using PVs or single-word verbs in gapped sentences on a 6-point Likert-type scale questionnaire. Although both groups showed a preference to use PVs over single word verbs, the native speakers showed a greater inclination than the non-native speakers. The authors also demonstrated that a factor that affects PV usage is exposure to English. Non-native speakers who had spent longer than 12 months in an English-speaking country were less likely to use single-word verbs than those who had spent less time. However, they suggested that greater exposure to English is not a straightforward solution to increasing non-native speakers' preference to use PVs over single-word verbs.

Recently, Becker (2014) investigated the PV avoidance issue using the standard multiple-choice and translation elicitations tasks but also included a story re-tell task. In this retell task, students first read an L2 translation of an L1 story that contained PVs and then retold the story in English. Compared to the multiple-choice task, fewer PVs were used in the retell task. Better performance on the multiple-choice task might have been caused by the presentation of the PVs as one of the options, whereas no option was given in the retell task.

In short, previous studies have shown that L2 learners are more inclined to use single verbs than PVs. Learners' PV avoidance behaviour most likely stems from whether PVs exist in their L1. Also, their proficiency level seems to play a role. There are, however, some limitations to the studies reported above. First, many studies collected data through controlled tasks such as multiple-choice rather than data elicited through oral or written production (except for Becker, 2014). Further, they tapped learners' preference rather than language use, and some studies did not include a native group as a baseline for comparison. Despite these limitations, there is little reason to doubt that L2 learners are most likely to eschew phrasal verbs in favour of their single verb counterparts.

2.7 Research on PV learning conditions

There has been some work into how to facilitate PV learning using a cognitive linguistics approach. The basic idea of this approach is that an understanding of the nature of semantic transparency or motivation behind phrasal verbs results from an understanding of metaphor, metonymy and conventional knowledge (Kövecses & Szabó, 1996, p. 345). The application of this approach typically involves grouping phrasal verbs according to the conceptual metaphor that they manifest, so, for example, the phrasal verbs *feel up, cheer up, buck up* and *eat up, chew up, give up* belong to the conceptual metaphor HAPPY is UP and COMPLETION IS UP, respectively. The teacher then explains the orientational metaphors of the particles of the phrasal verbs to students. Five studies have investigated variations of the cognitive linguistic approach in L2 classrooms (Kövecses & Szabó, 1996; Boers, 2000, Experiment 3; Condon, 2008; Yasuda, 2010; Strong, 2013). Let us review each one in turn.

Kövecses & Szabó (1996), Boers (2000, Experiment 3), Condon (2008), Yasuda (2010), and Strong (2013) utilized a semantic analysis to help students to acquire knowledge of complex PVs. In these studies, the participants' mother tongue lacked the PV system

(French: Boers, 2000; Condon, 2008; Hungarian: Kövecses & Szabó, 1996; Japanese: Yasuda, 2010; Strong, 2013). Previous studies have demonstrated that speakers of non-Germanic languages find PVs difficult to learn and, as a result, prefer to avoid their use. For Kövecses and Szabó (1996), the target PVs consisted of the particles *up* and *down*, for Boers (2000, Experiment 3) and Condon (2008), *up, out, in* and *down*, for Yasuda (2010), *up, out, into, down off*, and for Strong (2013), *up, out, down,* and *off*.

The researchers evaluated the effectiveness of their teaching methods with a comparison condition. The control required students either to read a PV with an L2 paraphrase of its respective meaning (Kövecses and Szabó, 1996; Condon, 2008; Yasuda, 2010) or with L1 explanations provided in dictionaries (Boers, 2000). A final test was administered to measure the effects of the treatments. The final test required students to fill in gapped spaces in sentences with particles (Kövecses & Szabó, 1996; Yasuda, 2010; Strong, 2013) or with intact PVs (Boers, 2000; Condon, 2008). Half of the blanks spaces were to be filled in with items taught, while the other half were to be filled in with items that were not taught but shared the same metaphoric theme as the taught ones.

The results of the studies showed that the teaching methods led to better performance on the final test on the taught PVs. Kövecses and Szabó (1996), Yasuda (2010) and Strong (2013) also observed a facilitation effect on the untaught PVs. This finding supports the possibility of the transferability of the underlying orientational metaphors of taught PVs to untaught ones. In contrast, Boers (2000) and Condon (2008) did not find evidence to support this view. There are two possible reasons for this. The first might be related to the participants' mother tongue. Students in the studies by Boers (2000) and Condon (2008) were both speakers of French, whereas, in the other studies, the students were speakers of other non-Germanic L1s. However, it is unclear how the students' first language might be a significant factor. The second reason regards the design of the final test. Boers (2000) and Condon (2008) had students select the correct PVs from a list and insert them into the gapped spaces, whereas the other studies had their students choose the missing particles from a list. The presentation of intact PVs may have interfered with participants' ability to evaluate the semantic contributions of the particles. As a result, they may have resorted to their own methods by which to understand PVs.

The cognitive linguistics approach examined in these studies provides valuable insights into how to effectively enhance students' knowledge of PVs. However, to date, cognitive linguistic-inspired materials for PV teaching are far from abundant. As a result, teachers will, therefore, be more likely to turn to what is readily available, which is mainstream ESL/EFL course textbooks. The question then becomes (a) what do these mainstream resources provide as PV instruction and (b) how effective are they?

2.8 Summary

Part I of the literature review discussed phrases in general and PVs in particular. It showed that PVs are an extremely complex type of verb that can cause considerable confusion for L2 learners due to structural and semantic as well as interlexical factors. Regardless of these complications, PVs are an important type of verb phrase in the English language. The reason is that they occur very frequently in language and this implies that PVs are a key element in native speakers' competency and performance. Further, native speakers seem to be constantly creating new PVs and the old ones seem to be inheriting new meanings. Therefore, few would disagree that L2 student should acquire a large stock of PVs to their lexicon. However, when given the opportunity to use a PV or a single-verb counterpart, L2 learners, especially those with non-Germanic L1s, appear more inclined to favour the use of single-verb counterparts. Although little research has been carried out to explore how to teach and learn PVs effectively, to date, most of the experimental studies that have looked into the effects of a teaching approach based on cognitive linguistics insights into metaphor theory and image schemata and have revealed positive effects. For the teaching approach to work successfully, teachers need first to obtain a depth of knowledge of the metaphoric language. However, many non-native English teachers, who have themselves struggled with understanding PVs, may find such an approach unappealing. Fortunately, alternative methods to facilitate PV learning exists. One popular source is general ESL/EFL course textbooks which contain various PV learning procedures. Despite the fact that countless teachers and students around the globe use such textbooks, as far as I am aware, no experimental study has sought to investigate their effectiveness. To fill this gap in research, the present thesis carried out three experimental studies. Part II of the literature review looks more closely at the attributes of the learning procedures that were examined.

Part II: Learning Conditions

2.9 Introduction

Part II of the literature review discusses the conditions of learning that are under investigation in this thesis. In this thesis, three studies examine treatments that are intended to help students to encode, to retain, and to subsequently retrieve PVs from memory. The elements involved in these treatments and the treatments themselves are discussed in Part II. Section 2.10 discusses the encoding processes and theories of acquisition. Section 2.11 looks at where encoded memories are stored. After that, Section 2.12 discusses retrieval processes followed by a discussion on theories of retrieval (Section 2.13). Next, research on the effects of testing (Section 2.14) and pretesting (Section 2.15) is discussed. Section 2.16 looks at studies that have compared the effects of testing and pretesting on learning. Section 2.17 looks at key factors that influence the benefits associated with the testing effect and the pretesting effect. Next, the rationale for the further research is discussed (Section 2.18). The last section (Section 2.19) summarizes Part II of the literature review.

2.10 Encoding processes

In this thesis, three studies examine treatments that require students to do three things: (1) to encode (2) to retain, and (3) to retrieve PVs from their memories. In this section, I discuss what is involved in (1). Encoding is a cognitive process that attempts to put target material into memory (Wingfield & Byrners, 1981). Ever since James' (1890) early attempts to describe memory, psychologists have been interested in understanding the mechanisms behind encoding in order to explain why some target items get encoded while others do not. One of the most influential theories posited on this topic was by Craik and Lockhart (1972). According to their Levels of Processing (LOP) model, the way information is processed substantially influences the degree to which that information is encoded and retained in memory, with deeper levels of processing resulting in greater learning gains. For example, the authors claimed that a word can be learned based on the way it sounds and how it is spelt as well as on what it means. However, processing of a word's meaning is seen as being deeper than processing its sound, which, in turn, is deeper than processing its spelling.

To test the predictions of the LOP model, Craik & Tulving (1975) carried out several experiments. In these experiments, tasks were used to induce participants to process words at different levels of depth. To elicit shallow processing, they made structural decisions on whether a word was written in capital letters or phonemic decisions on whether it rhymed with a specific word. Deep processing was prompted by asking participants to make semantic decisions on whether a word fits the context of a specific sentence. After the tasks, they received either a recognition test or a recall test. The results of these tests showed that performance was higher on the words that were processed deeply than on the ones that were processed shallowly. The general conclusion was that memory of a word was best when it was processed at a semantic level, next best at a phonological level and worst at the surface level.

There are several limitations to the Levels of Processing theory. The first limitation was reported in one of Craik and Tulving's (1975) experiments. According to the theory, two semantic decision tasks should elicit the same depth of processing and result in the same amount of learning gains. However, the authors found that judging the fit of a word in a long, complex sentence (e.g., The old man hobbled across the room and picked up the valuable ___ from the mahogany table), as opposed to a short, simple sentence (e.g., He dropped the ___), resulted in better memory. The second limitation was the unequal amount of time it took to complete the different decision tasks. Semantic decision tasks often took longer to complete than the phonological and surface tasks. The additional time on task was considered to be a factor behind the advantage of semantic tasks over the other tasks. Third, Craik and Lockhart (1972) did not explain why deep processing is more effective than shallow processing. Fourth, Eysenck and Keane (2010) claim that retention of information is not only affected by the depth at which it is processed, but also by the distinctiveness of that information. Fifth, the theory considered the relationship between the depth of encoding and learning gains but failed to take into account the relationship between encoding and the effects of how that information is subsequently retrieved from memory. Morris, Bransford, and Franks (1977) showed that if semantic information is irrelevant in a final test, the semantic decision task will tend to result in poorer performance than the structural decision tasks.

2.11 Memory storehouses: Episodic and semantic

In this section, I discuss (2) — retention. Retention means that a target item is maintained in a way that allows for it to be subsequently accessed, as it has not undergone any significant decay to its representation in the memory. Some target items might be easier to retrieve than others depending on whether they are stored in episodic or semantic memory. According to Tulving (1972), episodic memory contains encoded (autobiographical) experiences and events as well as temporal-spatial relationships among those experiences and events. However, traces in episodic memory are not perfectly preserved copies of past experiences and are instead incomplete recordings that are susceptible to corruption. Retrieval of a memory trace is suspected to change the contents of that retrieved trace.

The second memory storehouse discussed by Tulving (1972) is semantic memory. Traces in this semantic memory are considered permanent knowledge of things such as the meaning of words and the extensive collection of facts about the world. Words and concepts and the relationship between them are thought to be organized hierarchically. In contrast to episodic memory, semantic memory does not store experiences, and information retrieved from semantic memory and does not change its contents. The loss of knowledge probably does not occur unless there is physical damage to the brain. Information pulled from semantic memory is less vulnerable to decay compared with episodic memory. Further, it may be called upon unconsciously to ground new information with something old and familiar. Last, traces in semantic memory are separate but not mutually exclusive from traces in episodic memory.

I draw attention to these two storehouses of memory because researchers (e.g., Karpicke & Zaromb, 2010) have suggested that different conditions of learning may tap different memory storage systems. A condition of learning that administers a test after study of a target item is said to require a student to retrieve that target item from episodic memory — the experience in which it was learned in the input material. In contrast, a condition of learning that requires a student to take a test before study of a target item is said to tap (if any) semantic memory because no prior learning episode preceded the test. Since the experimental studies in this thesis examine these two types of learning conditions, how students arrive at a response on a test may be a factor that influences their performance. Having said that, this thesis does not set out to explore this factor.

2.12 Retrieval processes: Recognition and recall

In this section, I look at (3) — retrieval. Simply put, retrieval is the ability to remember a past experience. It can be subdivided into recognition memory and recall memory. Recognition is the ability to identify a previously encountered stimulus based on either recollection of contextual information or solely on familiarity (Eysenck & Keane, 2010). The distinction between these two elements is exemplified in the following scenario. On a multiple-choice test, a student is asked to select the correct answer among alternatives. The student first scans the choices and then favours one above the rest but is unsure whether it is actually the right answer. At this stage, the student demonstrates familiarity with the correct answer. After some consideration, the student concludes that the favoured choice is, in fact, the correct response. Here, recollection replaces familiarity of the target word. As for recall memory, it involves the ability not only to recognize a target word but also to produce it from memory when it is not presented (free recall) or in response to a cue (cued-recall).

Although recognition and recall involve remembering something from memory, it is unclear whether they are two distinct processes or features of a single process. According to a single-stage model, basically similar mental processes govern recognition and recall performance, and any difference between them are minor (Wingfield & Byrners, 1981). Memories have different strengths, and only stronger memories exceed a certain threshold, which makes them accessible for free recall. Weaker memories are likely below this threshold of accessibility. Recognition is a low threshold process (it is capable of detecting even the weakest of memories), while free recall is a high threshold process (Kornell, Bjork, & Garcia, 2011). In contrast to the single-stage model, the two-stage model proposes that recognition and recall are basically different processes because recall is assumed to contain certain retrieval processes not present in recognition (Tulving, 1976). According to the model, recall processes first generate a list of items from memory and then the correct response must be selected from this generated list. Recognition involves only the stage of response selection (not item generation).

Some relatively recent neurological research suggests that similar brain areas are associated with retrieval and recall processes. Eysenck and Keane (2010) reported a study by Staresina and Davachi (2006) that showed how successful performance on different types of retrieval tests was associated with increased activation in a part of the brain (i.e., left hippocampus and left ventrolateral prefrontal cortex) which was also activated during encoding. Activation was strongest in a recall test and weakest in the recognition test. This finding suggests that successful recall involves processes in addition to those involved in recognition, which seems to support the two-stage theory.

Regardless of the differences between the two models, both predict superior performance on recognition tests than on recall tests. The single-stage model holds that anything which influences recognition performance should similarly influence recall performance since both are presumed to be measures of the same mental process. However, the two-stage model explains what influences recall performance does not necessarily influence recognition performance. For example, Tulving & Thomson (1973) found that participants could recall target words but could not recognize them. Similarly, Shepard (1967) found that recognition scores on rare words were superior to recognition scores on common words, suggesting that the effect of word-frequency on recognition performance is the opposite of that on recall. In recent years, there seems to be greater support for the two-stage model than the single-stage model.

2.13 Theories of retrieval on learning

Retrieval is viewed as a key factor in the process of information acquisition (Roediger & Butler, 2011). The study of the role of retrieval on learning has been a prominent research area in the cognitive psychology field (Rowland, 2014) but less so in the applied linguistics field. Although many retrieval theories have been proposed (Karpicke et al., 2014), the most important and influential ones are retrieval effort (Bjork, 1975), elaborative retrieval (Carpenter, 2012) and episodic context (Karpicke et al., 2014). Let us review each of these theories in turn.

The *retrieval effort* model proposes that retrieval strengthens the representation of the target item in the memory. The degree to which it is strengthened, however, depends on the amount of effort that is involved during retrieval. Effortful retrieval improves acquisition more than effortless retrieval. The retrieval effort model has influenced many studies on retrieval in the field of cognitive psychology. Yet, despite this influence, it does not provide any apparatus capable of measuring the amount of effort one exerts during retrieval, although some researchers have considered the time it takes to remember a target word as a measure of retrieval effort, with slower response times taken as an indication of greater effort and faster ones suggesting less effort.

The *elaborative retrieval* model is another account of retrieval. It holds that retrieval is likely to be successful when a target item is somewhat related to the retrieval cue (e.g., table-chair) and unsuccessful when the target item is completely unrelated to the retrieval cue (e.g., book-spoon). This model has been used to account for the findings of studies that have compared closely related word-pairs (e.g., bus-car) with weakly related word-pairs (e.g., plate-dinner). It claims that a retrieval cue strongly associated with a target item is most likely already well-established in the memory. Therefore, little, if any, elaboration takes place between these items. However, a retrieval cue and a target item weakly associated forces an individual to create links between these items, usually with items semantically related to both, which creates routes that can aid in subsequent retrieval. Therefore, the magnitude of retrieval depends, to a large extent, on the number of semantically related items connecting the retrieval cue to the target item. Compared to the retrieval effort model, which posits difficulty as the underlying factor, elaboration puts forward the idea that the primary mechanism of retrieval is the stimulation of a semantic network composed of the retrieval cue, the target item and several words related to both.

Although some researchers view the elaboration model as superseding the retrieval effort theory (Carpenter, 2012), it fails to account for some basic findings. Elaboration might trigger cue overload (Nairne, 2002). Cue overload occurs when the memory system cannot differentiate between the retrieval cue, the target item and the semantically related items that together form an elaborative network. Thus, when a retrieval cue is present, it results in massive overload to the system as it attempts to determine which of the items in the network is the target item and which are not (Karpicke et al., 2014). The result is a decrease in the likelihood of successful retrieval.

Given the limitations of the past retrieval theories, a new model of retrieval, *the episodic context account*, has been proposed in recent years (Karpicke, Lehman, & Aue, 2014). It is based on four assumptions: episodic context, contextual reinstatement, context updating and restriction of the search set. The episodic context assumption is that the temporal and environmental context of a learning event gets encoded along with the target item. The contextual reinstatement assumption holds that retrieval is dependent on the extent to which a retrieval cue in the present can reconstruct a target item stored in the memory. The context updating assumption sees knowledge enrichment as a result of an original memory trace getting updated with new episodic information that occurs when the original memory trace is retrieved in a different episodic context. The restriction of the search set assumption postulates that retrieval is possible only when the retrieval cue restricts the search for the target item in the memory from other potential responses (Raaijmakers & Shiffrin, 1981). That is, a target item is successfully retrieved when the retrieval cue uniquely specifies it.

2.14 The testing effect

It is common for tests to be used to assess students' knowledge of target words before and after treatment. However, tests do much more than gauge learning (via performance), they also directly affect learning. Studies have demonstrated that retrieval is an effective memory mnemonic, as it can strengthen one's memory of a target item and makes forgetting less likely to occur (Roediger & Butler, 2011; Roediger & Karpicke, 2006). This phenomenon has been referred to as the *retrieval effect* or the so-called *testing effect* (Roediger & Butler, 2011).

The testing effect can occur as a result of any method that requires an individual to remember something that was previously learned. To be able to remember a past experience implies that that experience was encoded into episodic memory. Herein, I refer to the encoding event as a study trial and the retrieval event as a test trial. Together, they constitute the study-test paradigm that researchers use to engender the testing effect. The purpose of a study trial is to present a target item in a way that helps a student to learn it. The most common way is via word pair, where a student studies the target word in relation to another single word, which may be a synonym, an antonym or a weakly semantically related word. A less common way is to embed a target word in a short sentence or a longer text, although recently, a growing number of studies have begun to explore this type of study trial. The purpose of a test trial is to use some of the information in the study trial as a cue to elicit a retrieval of the studied item from memory. The vast majority of studies have examined the effects of decontextualized study-test conditions, where a student studies a word pair and then recalls one of the words given the other as the cue. The benefits of the study-test paradigm are often compared to the benefits of the study-study paradigm (see below).

The study-test paradigm is often compared to conditions of learning that involve an individual to read and reread a target word because restudying is regarded as the most common and effective mode of learning. The traditional view of learning holds that our memories work in much the same way as recording devices, where if we read over the same material many times, it eventually writes itself on our memories (e.g., Bjork & Bjork, 2011). This view regards tests not as learning events but rather as neutral events that function as devices to only measure learning (Hays, Kornell, & Bjork, 2013; Roediger & Butler, 2011). However, an abundance of studies has demonstrated how faulty this perspective is by showing that taking a test enhances learning and retention better than simply repeatedly studying. Unfortunately, Bjork & Bjork (2011) argue that the effectiveness of tests as learning events remains largely underappreciated, in part because testing is typically viewed as a vehicle of assessment, not a vehicle of learning.

Research on the study-test paradigm has tended to investigate the effects of multiple study and test trials on learning and retention (e.g., Carpenter & DeLosh, 2005; Cull, 2000; Kang et al., 2014; Karpicke & Bauernschmidt, 2011; Karpicke & Roediger, 2007; Logan & Balota, 2008; Pyc & Rawson, 2007). For example, Nakata (2015) compared the effects of different schedules of spacing a target item across study and test trials and found that expanding spacing as opposed to equal spacing led to greater final test performance. As valuable as this research is to optimizing vocabulary learning via retrieval practice, it does not shed light on the archetypal study-test strategy in educational materials, where the testing effect is thought to be engendered using a single vocabulary test, quiz, or exercise.

Regardless of materials designers' praise for their materials, educational researchers have raised doubts over whether only a single retrieval attempt could result in a beneficial effect on learning. However, studies have demonstrated that trying to remember a target word even one time after it was studied can lead to substantial learning gains. Smith, Roediger, and Karpicke (2011) examined and compared two versions of a study-test condition and a study-study condition. The participants in this study were 36 undergraduate native English speakers who learned 90 word pairs of a category name and items in that category (e.g., vegetable—cucumber). All conditions had the same initial step: participants studied the word pairs for a short time. The second step differed between the conditions. In the study-test condition, participants were shown the category word followed by the first two letters of the other word (e.g., vegetable—cu___) and asked to either remember the missing word (covert retrieval) or to type it in a text box next to the retrieval cue (overt retrieval). In the study-study condition, they were asked to reread the intact word pairs (e.g., vegetable-cucumber). Final test performance showed that there was hardly any difference between the two versions of the study-test condition and that both versions led to a 20% advantage over the study-study condition.

The beneficial effect of the study-test condition was further supported by a study carried out by Kornell, Klein, and Rawson (2014). The participants were undergraduate native English speakers required to learn 60 word pairs consisting of related words (e.g., sharp—razor). The first step in the study-test and the study-study conditions were the same: participants studied the word pairs for a limited amount of time. The second step of the study-test condition required them to remember the missing vowels of one of the words they had previously studied when presented with a fragment (e.g., fence—ch_ _n). To complete the fragment test, participants had to type the missing letters into a text box. The second step of the study-study condition required participants to reread the intact word pairs. The results of a final test showed that performance was greater in the study-test condition than in the study-study condition by approximately 18% percentage points.

While the study-test paradigm has received considerable attention in the field of memory research, only a single study in the field of applied linguistics has sought to investigate how it can facilitate new L2-L1 word learning. Barcroft (2007) examined and compared a retrieval condition with a restudy condition. The participants were 44 native English speakers who were learning Spanish. Both conditions had the same initial step. Participants were asked to study L2 word-picture pairs displayed on a TV screen for a short amount of time. The second step of the retrieval condition presented the picture without its corresponding L2 word for six seconds, and, during this time, participants were asked to covertly retrieve the missing word. A third step was included in this study. It re-presented the L2 word-picture pairs to the participants for six seconds. For the restudy condition, the second step required the participants to restudy the same L2 word-picture pairs for the same amount of time that was given to complete the retrieval condition. Performance on final tests showed that greater learning gains were obtained in the retrieval condition than in the restudy condition.

2.14.1 Limitations of previous studies and rationale for further research

The studies reviewed above provide strong evidence that the testing effect can occur as a result of even a single retrieval attempt. Although Smith, Roediger, and Karpicke (2011) and Kornell, Klein, and Rawson (2014) show that the study-test paradigm enhances the learning of associations between known words, they do not claim that it can also positively affect learning new form-meaning connections. It is Barcroft's (2007) study that sheds light on this issue, as he looked at the learning of new words. However, the author's study suffers from some methodological issues that may weaken his reported findings. He administered a test before the treatment and three tests afterward. While the pre-and-post-test design is commonly used by researchers in the field of applied linguistics (e.g., Nakata, 2015; Webb, 2007), it is not regularly used by psychologists in the field of memory research for the reason that it confounds the very phenomenon that they wish to investigate — the testing effect. Taking a test before a treatment, whether or not a student provides the correct target word, is a learning event according to the pretesting effect (see below). Likewise, taking multiple tests after the treatment count as learning events as each prior test influences performance on a later test. To eliminate this confound, the three studies in this thesis carried out a norming study, which is an innovative method by which to estimate participants' knowledge of the target words without exposing them to the words before the treatment.

2.15 The pretesting effect

A phenomenon related to the testing effect is the *pretesting effect* or the so-called *generation effect* (Slamecka & Graf, 1978). It is the finding that taking a test without any prior opportunity to study the target word enhances learning. The pretesting effect is triggered using the test-study paradigm, where a person attempts to remember a target item before they are shown it. Basically, the test-study paradigm can consist of the same study trial and test trial of the study-test paradigm, except the presentation of the trials is reversed. To induce a response in the test trial, the treatment asks a participant to think of a word based on a generation rule, which requests, for example, the synonym of *table* or the antonym of *black*. Like the testing effect, the pretesting effect directly challenges the traditional view of learning as involving encoding processes only.

Several studies have supported the beneficial effects of pretesting. In a seminal paper, Slamecka and Graf (1978) carried out five studies. In their first study, 28 undergraduate native English speakers learned 100 word pairs in a test-study condition and a study condition. The test-study condition presented a word and the initial letter of another word (e.g., rapid-f) with a generation rule to help participants generate the missing word. The study condition presented the word and the other word together (e.g., rapid-fast). Final recognition test performance showed that the test-study condition led to superior learning gains compared to the study condition. The advantage of the test-study condition was attributed to the fact that participants generated 94% of the missing words correctly. The remaining four studies in the paper also demonstrated the advantage of taking a test without any prior study opportunity over studying the target words.

2.15.1 Test errors

However, in educational contexts, students are highly unlikely to generate a response correctly on a test without having studied the answer beforehand. For example, an L2 student might be presented with an L1 word and asked to produce its L2 synonym. If the L2 counterpart was never learned or if it cannot be remembered at that exact moment, the student's response is bound to be erroneous. What effect does this erroneous generation have on learning? Several studies have sought to examine the extent to which test errors help or hinder learning.

2.15.2 Type of test errors

Test errors are not all alike. Kornell, Hays, and Bjork (2009) classified test errors as either *unsuccessful retrieval* or as *errorful generation*. Unsuccessful retrieval occurs when a person cannot remember something that was learned in the past. For example, a student is asked to name the capital of New Zealand. Although the student learned it in geography class, he cannot bring it to mind for some reason. However, he can remember certain formal features of the capital's name (It begins with "Well…"), eliminate names of cities that are definitely not it (I know it is not Auckland or Christchurch), conjure up a mental image of New

Zealand's geography (I'm certain it is in the North Island and not in the South Island), and even remember that the capital shares its name with a town in England and with a kind of steak dish (It is Beef Well…). Despite the effort the student invests in trying to remember the name of New Zealand's capital, it eludes him. However, failure to remember Wellington does not necessarily imply that the student does not know that it is the capital's name. It is possible that its representation in memory was blocked or interfered with for some reason. Unsuccessful retrieval has been the focus of several recent studies (e.g., Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2013; Huelser & Metcalfe, 2012). They are discussed further below.

Errorful generation occurs when a person provides the wrong answer on a test because it was never learned (e.g., Potts & Shanks, 2014; Warmington & Hitch, 2014; Warmington, Hitch, & Gathercole, 2013). For example, a student skipped the day in geography class where the teacher told the class that Wellington is the capital city of New Zealand. When asked on a test, the best this student can do is to think of names of cities in New Zealand that he is familiar with in the hope that one of them might be the correct answer (I know Auckland and Christchurch. Auckland is bigger so it must be the capital). In other words, without any pre-existing knowledge, the student takes a blind guess, which turns out, unsurprisingly, to be wrong. To help the student to learn the answer, a teacher might supply the answer key. The presentation of the capital's name in the study trial after the test trial constitutes the first opportunity to learn the correct answer. There has been considerably less research on errorful generation than on unsuccessful retrieval.

2.15.3 Studies on generation failures

Studies have shown that unsuccessful retrieval enhances learning better than simply studying the correct answer. Kornell, Hays, & Bjork (2009) had undergraduate students learn weakly associated word pairs (e.g., olive-branch). In one experiment (Experiment 4), the test-study condition asked participants to guess a related word to a test cue (e.g., olive-?) within eight seconds. After a response was generated, the test cue and the target word were displayed for five seconds. The test-study condition was compared to a study-study condition in which the participants simply studied the test cue and target word together for 13 seconds. A final test was administered five minutes after the treatment. In a different experiment (Experiment 5), the authors replicated the same test-study and study-study treatments, except the final test was given 24 hours later. In both experiments, although the test-study condition nearly always resulted in test errors during learning, final test performance was greater in the test-study condition than in the study-study condition.

Grimaldi and Karpicke (2012) carried out a partial replication of the test-study paradigm examined by Kornell et al. (2009). They had 32 undergraduate native English speakers learn 30 semantically related word pairs (e.g., jelly-bread). In the test-study condition, participants guessed a word related to the test cue (e.g., jelly-?) and then received the test cue and the target word as corrective feedback (e.g., jelly-bread). In the study-study condition, they studied the test cue and the target word together for the same amount of time it took to complete the test-study condition. Even though 94% of initial test responses were erroneous in the test-study condition, final test performance indicated that the test-study condition led to greater learning gains than the study-study condition.

The beneficial effect of unsuccessful retrieval or guessing is also supported by Huelser and Metcalfe's (2012) study, which is a partial replication of Kornell et al.'s (2009). Sixty undergraduate native English speakers learned 30 semantically related word pairs. The test-study conditions presented a test cue for five seconds and then showed the test cue and the target word for another five seconds. The test-study condition was compared to two versions of a study-study condition. One version presented the test cue and the target word for five seconds; the other, for 10 seconds. Despite the fact that only 3% of guesses on the test trial in the test-study condition were correct, failing to provide the right answer resulted in superior final test performance compared to studying the correct answer for a short and long time.

Although the above studies provide valuable information regarding the pretesting effect resulting from unsuccessful retrieval or guessing in the test-study paradigm, they do not indicate whether the same benefit can occur as a result of errorful generation in the test-study paradigm. We may doubt that students would be able to learn from their mistakes, especially when their errors are unrelated to the correct answer. However, some studies have found that errorful generation can indeed result in the pretesting effect. Kornell et al. (2009, Experiment 1) had 25 undergraduate native English speakers learn the answers to 20 fictional trivia questions (e.g., What is the last name of the person who panicked America with his book Plague of Fear?). Since it was impossible for the participants to know the answers to the trivia questions, their errors on a test would be unrelated to the correct answers. In the test-study condition, participants were asked to type an answer to a trivia question within eight seconds, after which the answer and the question were displayed for five seconds. In the study condition, they studied the question-answer pair for only five seconds. In another experiment (Experiment 2), the authors replicated the test-study condition and only added to the study-study condition an additional eight seconds to equal the time on task between the conditions. In both experiments, although participants in the test-study answered none of the fictional questions correctly, final test performance was greater in the test-study condition than the study-study condition in Experiment 1 while no difference was found between the conditions in Experiment 2. The authors suggested that the test-study condition was less successful in Experiment 2 due to between-participants difference as performance was lower in Experiment 2 than in Experiment 1.

While Kornell et al. (2009) induced errorful generation in a test-study condition using participants' L1, other researchers have used words from other languages. The learning of novel vocabulary represents a rather more realistic learning scenario for students learning an L2 (e.g., Potts & Shanks, 2014). If students make a mistake on a vocabulary test, quiz or exercise, they are making a genuine error, not simply failing to guess what the experimenter had in mind at the time, as was the case in the studies by Grimaldi and Karpicke (2012) and Huelser and Metcalfe (2012).

Potts and Shanks (2014, Experiment 2b) examined the effects of errorful generation, in which students made incorrect guesses because the test material was entirely new to them. The authors had 24 undergraduate native English speakers without any knowledge of Euskara learn 60 L1-L2 word pairs using two versions of a test-study condition and a study-study condition. The choice test-study condition presented an L2 word and three L1 lures. The generate test-study condition displayed the L2 word as a test cue. The study-study condition supplied the L2-L1 word pairs. Participants did not possess any knowledge of the L2 words, but they managed to select 31% of the L1 synonyms correctly in the choice test-study condition, which is at chance but failed to generate any response correctly in the generate test-study condition. Final recognition test performance was similar between the test-study conditions, while the generate test-study condition outperformed the study-study condition.

While L2 vocabulary learning involves forming connections between L2 forms and L1 meanings, it also includes creating associations between L2 words which may or may not be compositional. A student may know two L2 words but may not be aware that their co-occurrence results in the formation of a semantic unit and that substitution of one of the component words may not be permissible or the meanings of the parts are distinct from the meaning of the whole. Researchers have only begun to investigate this line of research using the test-study paradigm (e.g., Boers, Demecheleer, Coxhead, & Webb, 2014; Boers, Dang, & Strong, 2017; Stengers & Boers, 2015).

Boers et al. (2014) investigated and compared the effects of several versions of the test-study condition (which in applied linguistics parlance is vocabulary learning exercises) on the learning of L2 verb-noun collocations. In trial 1, 19 students studied 20 collocations using an insert-the-verb exercise and an insert-the-collocation exercise. The insert-the-verb exercise presented a list of verb choices above sentences with deleted verbs. To complete the exercise, participants were required to select the verbs that fit the sentences and insert them into the gapped spaces. The insert-the-collocation exercise displayed a list of collocations choices above sentences with deleted collocations. Participants were asked to select the collocations that fit each sentence and insert them into the gapped spaces. Corrective feedback was subsequently given after students handed in their test papers. Although the authors do not provide information on the students' performance on the test-study condition, it is assumed that it was very poor given their performance on the pre-test and final test. The number of correctly recalled verb responses on the final test was negligible. Further, the difference between the two test-study exercises was not statistically significant.

Boers et al. (2014) carried out three additional trials comparing different variants of a test-study exercise. In trial 2, the insert-the-collocation exercise of trial 1 was compared to an underline-the-verb exercise. The new condition presented a verb juxtaposed with two lures in a sentence. Participants were asked to underline the verb that fit the sentence. After handing in their test papers, the students received corrective feedback. Like trial 1, trial 2 did not find any statistically significant difference between the two test-study conditions. Trial 3 compared the exercises of trial 1 and trial 2. Like Trial 1 and 2, the difference in learning gains between the test-study conditions was not statistically significant and was negligible. The last trial examined a connect test-study condition along with the insert-the-verb, underline-the-verb, and insert-the-collocation exercises. The connect exercise presented verbs in the left-hand column with nouns in the right-hand column of the test paper. To complete the test, students had to match the nouns with their respective verbs to form verb-noun collocations. Again, the learning gains were negligible between the test-study exercises.

The researchers analyzed participants post-test responses with their pre-test responses to determine the effect of the exercises on learning the collocations. They noticed that certain test-study exercises affected some participants ability to remember on the post-test the collocation they correctly produced on the pre-test. Instead of producing the same correct response, it was found that they produced one of the lures presented in the treatment. For example, in trial 4, 25% of test trial errors were reproduced on the post-test. The results demonstrated that the test-study conditions, which are common in textbooks, cause a certain amount of proactive interference.

Boers, Dang, and Strong (2017) carried out a partial replication of the study by Boers et al. (2014). They compared the effects of three test-study exercises on learning verb-noun collocations by a group of EFL learners. The select-the-verb condition and the select-the-intact-phrase exercise were similar to the select-the-verb exercise and the select-the-collocation exercise examined in Boers et al. (2014). In addition to these two test-study exercises, the authors also examined a first-letter-of-the-verb exercise, in which the first letter of a verb was followed by an underlined space in a sentence. Unlike the other exercise, this one did not provide students with the correct answer. Instead, they had to recall it from their memory, if it was possible. The gapped sentences of the exercises were accompanied by L1 translations to help students in providing the correct response. After students completed the exercises, corrective feedback was given. The number of correct responses in a verb gap-fill post-test was greatest in the select-the-intact-phrase exercise and followed by the first-letter-of-the-verb exercise. Test performance was lowest in the select-the-verb exercise.

As indicated by Boers, Dang, and Strong (2017), there are some notable limitations to the study by Boers et al. (2014), which might raise concern over whether the trials were powerful enough to demonstrate differences between the test-study exercises. First, like Barcroft’s (2007) study, Boers et al. (2014) used a pre-test to assess prior knowledge of the target items. As mentioned above, pre-testing might confound the results of the trials because pre-testing tends to result in the pre-testing effect, even if participants do not provide the correct responses. Recognizing the methodological problems associated with pre-testing, Boers, Dang, and Strong (2017) normed the target items using a parallel group of students matched as closely as possible with the participants in the treatment. Second, the number of participants was fewer than was needed to carry out a parametric test, so the non-parametric tests that were used might not be powerful enough to observe meaningful differences between the exercises.

2.16 The testing effects vs. The pretesting effect

On the surface, the study-test paradigm seems to share many similarities with the test-study paradigm. First, both paradigms consist of a study trial that requires a student to learn a target item. Second, they include a test trial that instructs a student to try to remember the target word. In short, the trials in the retrieval-inducing conditions seem to be indistinguishable from the trials in generation-inducing conditions.

Despite these similarities, there are some meaningful differences between testing conditions and pretesting conditions. First, the presentation order of the trials is reversed. In a testing condition, the study trial comes before the test trial, whereas, in a pretesting condition, the study trial comes after the test trial. Second, in the pretesting condition, the study trial may also provide feedback on whether the test response was correct or incorrect, but in the testing condition, the study trial does not provide this information. Last, the study-test procedure is more likely to tap episodic memory than semantic memory to retrieve a target word. In contrast, the generation of a response without any prior knowledge of the target word may rely more heavily on information in semantic memory rather than episodic memory (Karpicke & Zaromb, 2010; Kornell, Klein, & Rawson, 2014).

There are two main reasons to compare the effects of retrieval-inducing techniques with generation-inducing techniques. First, Karpicke and Zaromb (2010) indicated that many psychologists tend to associate the testing effect with the pretesting effect. Second, while pedagogically-minded researchers overwhelmingly advocate for the use of retrieval strategies to improve learning, teachers seem to favour the use of generation strategies in the classroom (Mayer, 2008). Based on these reasons, any distinction between the study-test condition and the test-study condition would have not only theoretical implications but also practical implications for learning in educational contexts.

Studies have demonstrated that retrieval and generation have different effects on learning. Karpicke and Zaromb (2010) had 60 native English speakers learn 40 semantically related word pairs (e.g., love-heart). The retrieval- and generation-inducing conditions both began by having students read (not study) a word pair under incidental learning conditions. Approximately five minutes later, they were shown fragments of the target word (e.g., heart-l_v_). The retrieval condition required participants to recall the word they had previously studied, whereas the generation condition told them to type the first word that came to mind that was related to the test cue and successfully completed the fragment. Although test trial performance was the same across the conditions, final test performance was greater in the retrieval condition than in the generation condition.

A semi-replication of Karpicke and Zaromb's (2010) study was carried out by Kornell, Klein, and Rawson (2014). They had 68 native English speakers learn 60 related word pairs. The retrieval condition required participants to study a target item before attempting to recall the missing vowels of the fragment (e.g., fence-ch_ _n). The generation condition consisted of only the fragment test trial. The results of the final test showed that learning gains were greater in the retrieval condition than the generation condition.

Warmington, Hitch, and Gathercole (2013) examined how 49 children would learn ten novel name-object pairs using a study-test condition and a test-study condition. The study-test condition was structured to foster error-free retrieval. The children were shown an object and told the first letter of the object as well as the name itself (e.g., This object's name begins with the letter P. It's a prot. Can you say prot?). In contrast, the test-study condition resulted in errorful generation, as the children were shown the first letter of the object and the object itself and asked to guess its name. (e.g., This object's name begins with the letter P. Can you guess its name?). Performance on the final test was higher in the error-free condition than in the errorful condition. Warmington and Hitch (2014) replicated this study with adults and administered an immediate and delayed post-test. The results showed that the error-free retrieval condition led to greater learning and retention compared to the errorful generation condition.

Although studies have demonstrated that retrieval leads to better learning gains than generation, the difference might be related to the methods that were used to induce these effects. Karpicke and Zaromb (2010) and Kornell, Klein, and Rawson (2014) included a study trial and a test trial for their retrieval conditions but only a test trial for their generation conditions. Therefore, the treatments were unbalanced. In contrast, Warmington and associates (2013, 2014) compared retrieval and generation conditions consisting of identical trials. One possible reason Karpicke and Zaromb (2010), and Kornell, Klein, and Rawson (2014) did not include a study trial in their generation treatments was that they expected participants to generate the majority of responses successfully (and they did), whereas Warmington and colleagues (2013, 2014) anticipated that participants generated responses would be incorrect, making it important for students to receive the correct answers. Regardless of whether the generation condition included a study trial or not, performance was poorer in the generation conditions than in the retrieval conditions in all three studies. Further research is required to examine whether this trend occurs for other lexical items besides semantically related word pairs or picture-name pairs, such as semantically complex verb phrases.

2.17 Factors affecting learning in retrieval and generation conditions

This section discusses some of the moderating variables that have been found to influence the effects of retrieval and generation learning conditions. First, it discusses the separation of the study and test trials from one another. Then, it looks at the materials-to-be-learned in relation to retrieval and generation conditions.

2.17.1 Separating study and test trials

The separation of repeated presentations of a target word has a beneficial effect on learning and retention (Nakata, 2015). Research has examined the effects of massed and spaced learning. Massed learning refers to the absence of any interruptions between two study trials of a target item. For example, a participant studies a word and then immediately studies it again. In contrast, spaced learning refers to a separation between the two study trials by intervening items or some length of time. For example, a participant studies a word and then studies other words before restudying the initial word again. The spacing effect refers to enhanced learning from spaced learning rather than from massed learning (Cepeda, Pashler, Vul, Wixted & Rohrer, 2006). In a review of 317 experiments on distributed practice over the past 100 years, Cepeda, Pashler, Vul, Wixted, and Rohrer (2006) reported that spaced learning improves acquisition and retention more than massed learning does.

Regardless of the advantage of spaced learning, some psychologists have expressed concern over whether the separation of a study trial and a test trial might lead to problems that may be difficult to resolve. The argument is that spaced learning increases the failure rate of recalling a word on a test trial and this then might obstruct subsequent learning, whereas massed learning increases the hit rate of success. This argument reinforces the emerging "errorless learning" movement, which advocates massed learning, where a person recalls a target word immediately after it was studied.

In the errorless learning literature, a plethora of studies has demonstrated that errorless retrieval improves learning. For example, Baddeley and Wilson (1994) taught participants a list of five-letter words using a recall task in which the first two letters of each word were presented as a cue to retrieve the missing letters. In the errorless learning condition, immediately after the presentation of the retrieval cue, the answer was given, and participants were asked to repeat the answer. Because there was no gap between the presentation of the target item and the demand to recall it, hardly any errors were produced on the test. Despite the fact that the errorless condition seemed to reflect rote rehearsal, the authors found that it led to substantial learning gains.

While most research on the spacing effect has been done on retrieval conditions, Hays, Kornell, and Bjork (2013) carried out a study on spacing using generation conditions. The authors attempted to determine whether immediate or delayed feedback after a test had a positive effect on learning. In experiment 1, the participants were 70 undergraduate students learning weakly semantically associated word pairs (e.g., frog-pond). Two generation conditions were compared. Both conditions tested the word pairs (e.g., frog-?) and then presented the word pairs (e.g., frog-pond). The critical difference between them was that in the massed generation condition, the test trial and the study trial were consecutive while in the spaced generation condition, the two trials were separated by other items (and approximately 9.5 minutes). The final test was the same as the initial test (e.g., frog-?). The authors only looked at items that participants did not produce correctly on the initial test. Thus, they examined failed generations in relation to immediate and delayed feedback. These two generation conditions were compared to two restudy conditions. In both restudy conditions, participants studied a word pair and then studied the word pair again. The key difference between the two conditions was that one was spaced while the other was massed. The findings showed that recall on the final test was greater in the massed generation condition than in the massed restudy condition. In contrast, the spaced restudy condition outperformed the spaced generation condition. This finding suggests that spaced and massed learning have the reverse effect on generation learning conditions compared to restudy conditions. With respect to the generation conditions, massed learning enhanced learning much more than did spaced learning. This finding is inconsistent with most research on the spacing effect. In experiment 2, the same generation conditions were compared, and the finding was similar to the first experiment: the massed generation condition led to better performance on a final test than did the spaced generation condition. Overall, the pattern of results in both these experiments is extraordinary. The authors suggest that a failed test response facilitates subsequent learning but the potentiation dissipates as time separates the failed test from the subsequent presentation (p. 292). This finding implies that delaying feedback is substantially less beneficial than immediate feedback.

Grimaldi and Karpicke (2012, Experiment 3) also examined whether delaying feedback affected learning compared to immediate feedback after a failed test response. Forty-four undergraduate native English speakers learned 30 word pairs in a generation condition and a read condition. In the delayed generation condition, the test trial and the study trial was separated by 29 filler items, whereas in the immediate generation condition, the two trials were not separated by any filler items. During learning, in both generation conditions, only six per cent and 5 per cent of responses were answered correctly in the immediate generation condition and delayed generation condition, respectively. These correct responses were removed from further analysis in order to determine the effects of test errors on learning. The findings of a post-test indicated that the immediate generation condition resulted in better learning than the delayed generation condition. The authors interpreted this finding to mean that the benefits of making an error on a test are short-lived. That is, the active process of searching for the answer during learning has a positive effect on encoding the corrective feedback, but this cognitive state fades quickly, and if corrective feedback is delayed, it is more difficult to encode the correct answer into memory.

A similar study by Kornell (2014) replicated the procedure used by Grimaldi and Karpicke (2012) to examine the effect of immediate versus delayed feedback following an unsuccessful generation attempt to a retrieval cue. In experiment 1, 23 participants studied weakly related word pairs in three learning conditions: test-study, test-delay-study and study-only. In the test-study condition, the test trial was followed immediately by the study trial. In the test-delayed-study condition, the test trial was separated from the study trial by 10 items. An average of 4 minutes elapsed between the test trial and the study trial. The study-only condition presented the word pair intact. As the findings of the studies reported above, final test performance was greatest in the test-study condition. It was followed by the study-only condition and lowest in the test-delayed-study condition, although it was found that the difference between the study-only and test-delayed-study conditions was not statistically significant.

Overall, the consensus from the articles just reviewed (i.e., Grimaldi & Karpicke, 2013; Hays et al., 2013; Kornell, 2014) is that when people learn word pairs, guessing before studying is helpful only if corrective feedback follows immediately after the error. There is at least one experiment that produced a different result. Kornell (2014, Experiment 2) used meaningful trivia questions instead of word pairs as the target stimulus material. Thirty-one participants were asked to learn 35 trivia questions in the three learning conditions examined in experiment 1 (see above). Participants only produced the correct answer 2.6% of the time during learning. These correct responses were removed from the analysis of the final test. The analysis showed that final test performance was highest in the test-study condition followed by the test-delay-study condition, and was lowest in the study-only condition. This result diverges from prior studies that used word pairs as stimuli material (Grimaldi & Karpicke, 2012; Hays et al., 2013). Based on this finding, the author suggests that when stimulus material is semantically rich in content, delayed feedback can enhance learning better than when the stimulus material is less meaningful. Word pairs represent decontextualized learning materials because, for the participants, the connection between the words may seem arbitrary, whereas, in rich sentential contexts, the word is meaningful.

2.17.2 Types of target items

Research has, for the most part, examined retrieval and generation conditions on the learning of word pairs comprised of related words and unrelated words, including associations between new and familiar words (Potts & Shanks, 2014) as well as picture-word pairs (Barcroft, 2007; Warmington & Hitch, 2014). However, less attention has been given to the learning of words longer than a single lexical item. On this account, we cannot confidently presume that the benefit of these conditions on the learning of single words can be extended to the learning of phrases. There are two key reasons for this. First, the amount of cognitive effort required to retrieval two or more words (that comprise a phrase) from memory might be more demanding than is needed to retrieve a single word from memory. Second, it is possible to design a retrieval condition in which the study trial presents the verb and particle of a PV, while the test trial presents the verb and not the particle. In such a condition, a participant studies both the verb and the particle and then restudies the verb and retrieves the particle. A generation condition can also be designed similarly, by presenting the verb and not the particle in the test trial and then introducing the verb and the particle in the study trial. In such a condition, a participant studies the verb in the test trial and then restudies the verb in the study trial while studying the particle for the first time. Theoretically, knowledge should be enhanced more for the particle than the verb because the particle was retrieved while the verb was just studied. However, the problem with this assumption is that the verb and the particle form a semantic unit rather than being two unrelated words, even though their literal meanings do not contribute to the meaning of the PV. It is possible, therefore, that learning one of the words (e.g., verb) affects the learning of the other word (e.g., particle).

The trials by Boers et al. (2014) shed some light on this scenario with respect to generation conditions. In the insert-the-verb condition, participants studied the noun twice, on in the test trial and another time as corrective feedback, whereas they retrieved the verb and studied it in the corrective feedback. Theoretically, this condition should have enhanced knowledge of the verb compared to the noun. In contrast, in the insert-the-collocation, participants retrieved the verb that was connected with the noun in the test trial and then studied the verb-noun collocation in the corrective feedback. This condition should have enhanced knowledge of both the verb and the noun. Because the verb was retrieved in the insert-the-verb condition and the insert-the-collocation condition, the number of correct responses on the final verb gap-fill test did not differ between the two conditions. Unfortunately, because the final test only required participants to produce the verb, it is unclear how the generation conditions affected the learning of the noun, which was just studied in the insert-the-verb condition and retrieved in the insert-the-collocation condition.

2.18 Rationale for further research

In the field of memory research, a plethora of studies has investigated the testing effect and the pretesting effect. Recently, studies have begun to investigate how learning is affected when a student fails to provide a correct response on a test. Most studies have focused on unsuccessful retrieval, where a learner has studied the target item before but is unable to recall it at the time of the test. Less research has been carried out on errorful generation, where a student has never learned the target item but must still provide a response on a test. In the field of applied linguistics research, some researchers have investigated the testing effect. However, they used a paradigm that differs from the common one that students most likely use and can be found in any ESL/EFL course textbook — where a student is given one opportunity to retrieve a new lexical item. As far as I am aware, Boers and associates (2014, 2015, 2017) are the only researchers who have examined the effects of pretesting on learning L2 vocabulary. As Chapter 3 reveals, textbooks make use of several variants of the study-test and the test-study paradigms. Boers and associates have looked at some of the test-study procedures, but several other deserve attention as well. Further, there are several study-test procedures that have not been examined. PV textbooks make use of testing and pretesting procedures, but no study has examined their effectiveness. One of the aims of the present thesis is to explore the effects of some of these procedures.

Furthermore, this thesis is interested in the nature of PV learning. It is assumed that it involves forming associations between (1) form and meaning (e.g., *hang out = to spend time with friends*) and between (2) verb and particle (*hang* + *out).* Some of the experiments in cognitive psychology, using L1 participants, resemble (1) using new words (e.g., Potts & Shanks, 2014) and fictional trivia questions (e.g., Kornell et al., 2009), and others resemble (2) using known words (e.g., Kornell, Klein, & Rawson, 2014). Both types of experiments are relevant to the ones carried out in this thesis because students are asked to associate PVs with their meanings in study trials and to recall the PVs when cued with their meanings in test trials (e.g., ____ — *to spend time with friends*). Also, they are asked to reproduce the verb-particle associations (word-word associations = _[hang]_ + _[out]_).

2.19 Summary of part II

Part II of the literature review discussed learning processes and learning methods. The learning processes discussed were encoding and retrieval. The importance of retrieval was underscored in the discussion on retrieval theories. Next, empirical studies that have examined the testing effect were reviewed. After that, the pretesting effect, a related phenomenon to the testing effect, was discussed. The review of the empirical studies showed that guessing on a test can affect the way the target item is subsequently encoded into memory. It also showed that making errors on the test can have positive and negative effects. The discussion then turned to factors that can influence the extent to which retrieval-oriented and generation-oriented conditions can influence learning and retention. Finally, Part II showed that most of the research conducted on testing and pretesting had been carried out in the field of memory research on semantically related or unrelated word pairs. In contrast, only a few studies in the field of applied linguistics have explored the effects of pretesting on phrases, and, as far as I am aware, no study has applied the study-test paradigm to learning phrases. There is a need to explore further the effects of the study-test paradigm and the test-study paradigm for learning phrasal such as PVs because these paradigms are common in general ESL/EFL course textbooks, as revealed in Chapter 3.

Chapter 3 - Textbook analysis

3.1 Introduction

Materials designers of ESL/EFL course books claim (in their promotional statements on the back cover of their books or their publishers' website) to offer effective methods of learning vocabulary which includes phrases. They also claim that countless teachers and students around the globe use their materials. Based on these declarations, it is surprising that little research has been carried out to identify what these methods are and how effective they are. A few years ago, Boers et al. (2014) examined a sample of textbooks of intermediate and upper-intermediate levels in order to identify types of exercises used for learning phrases. While they identified several different kinds, they did not consider whether the exercises were preceded by exemplar material (i.e., study trial) or followed by corrective feedback (i.e., study trial). Moreover, they identified only exercises that required students to recognize phrases (recognition), but it is assumed that these books also include exercises that ask students to produce phrases (recall). The present textbook analysis aims to investigate further the conditions of learning phrases offered in general ESL/EFL course textbooks. I adopted the same procedure Boers et al. employed for identifying phrase-focused exercises. After that, I determined whether the target phrases in the exercises were presented in some kind of input material before the exercises and some kind of corrective feedback after the exercises. Other features of the conditions of learning were also recorded. The textbook analysis was conducted on 52 textbooks of various titles and proficiency levels. The large sample size made it possible to notice trends in the kinds of conditions that are used for learning phrases across different titles and proficiency levels. The data collected here may be of value to researchers interested in examining and comparing the effectiveness of various types of phrase-learning conditions in textbooks.

This chapter has different sections. Section 3.2 presents the research questions guiding the textbooks analysis. Section 3.3 describes the methodology. After that, the results are provided in Section 3.4. Last, Section 3.5 provides the rationale for the experimental studies carried out in this thesis.

3.2 Research questions

An analysis was carried out on the phrase-learning procedures in general ESL/EFL course textbooks. The analysis aimed to find information about these procedures in order to answer the following research questions:

1. What is the most common test trial (or exercise) to learn phrases?
2. What is the most common study trial (or input material) to learn phrases?
3. Which learning methods occur more often: retrieval (study trial before test trial) or generation (study trial after test trial)?
4. How many phrases are to be learned in a single exercise?
5. Are the study trials and the test trials contextualized or decontextualized?

3.3 Methodology

This section presents the methodology of the textbook analysis. First, how the ESL/EFL course textbooks were selected is explained. Second, the method by which phrase learning procedures were identified is discussed.

3.3.1 Selection of textbooks

There are countless contemporary ESL/EFL course textbooks available around the globe for teachers and students to use. Because analyzing the treatment of phrases in all these books is beyond the scope of the present study, I delimited my selection to nine popular titles with books ranging from beginner to advanced levels. My selection was based upon the textbooks available to use by English teachers at the English Language Institute (ELI) at Victoria University of Wellington. The university purchased for the ELI teachers the following popular titles: *Cutting Edge, English Result, Global, New English File, New Headway, New Inside Out, New Total English, Speak Out,* and *Straight Forward*. I collected a total of 52 textbooks to analyze (see Appendix IX).

The textbooks that I selected were the most recent editions available in The School of Linguistics and Applied Language Studies' Resource Room. Each of the nine titles consists of books at the advanced and upper-intermediate levels, intermediate and pre-intermediate levels, elementary and beginner levels, except no advanced level book was for *English Result*, and *Speakout* had a starter level book instead of a beginner level one. The textbook analysis was carried out on the entire selection of books as well as at each of the different proficiency levels. The purpose of the analysis at each level was to determine whether materials designers provide the same type of phrase-focused conditions of learning at the lower levels as they do at the higher levels as well as whether they incrementally increase the number of these conditions from the lower to the higher proficiency levels.

3.3.2 Identification of phrase-focused learning conditions

The textbook analysis involved three main steps. The first step was to identify exercises on learning phrases. Following the procedure employed by Boers et al. (2014), each textbook was manually screened for exercises that required students to deal with strings of words. For the most part, this step was relatively easy, as many exercises indicated that the vocabulary focus was on phrases of some kind. The features of each exercise were recorded in an Excel file, including the instructions, the terms used to refer to the target "phrases," the number of phrases-to-be-learned, and whether it required students to recognize a phrase given in the exercise or to produce one from memory. I recorded all these features because they are all factors that can affect how well students learn phrases from the exercises.

The second step was to identify whether input materials preceded the exercises. This step was more time consuming than the first one. It required reviewing all the input materials that preceded a phrase-focused exercise in order to determine whether the phrases-to-be-learned had occurred beforehand. I examined the input material looking for whether the target phrases were textually enhanced or whether the instructions informed students that the text contained the target phrases or whether a gloss for the target phrases accompanied the text. There is a possibility that some of the input materials may have presented the target phrases without the use of any attention-raising device, but I did not consider these instances because I was focused on explicit, as opposed to incidental, conditions of phrase-learning. It should be noted, however, that the instructions of some exercises referred students to input material in which the phrases were not textually enhanced. These instances do not reflect incidental vocabulary learning because the exercises remind students of where they had encountered the target phrases. For example, in Inside Out: Intermediate (Kay & Jones, 2000), the instructions state the following: “Match words and phrases from the two columns to make seven expressions from the text on page 51” (p. 52). This example was therefore counted as an explicit phrase-focused exercise. After identifying that a target phrase occurred in input material before the exercise, I recorded whether the input material presented the phrase in context. A phrase that occurred in a sentence or a longer text was recorded as contextualized input material, whereas a phrase that occurred without any contextual support was counted as decontextualized input material.

The third step involved re-reading the instructions to determine whether they asked students to check over their answers using an answer key of some sort. If an exercise did require this, it was recorded as providing corrective feedback. The results of the textbook analysis are presented below.

3.4 Results

This section presents the results of the textbook analysis. Subsection 3.4.1 reports on the terms used to describe strings of words. After that, the types of test materials that were identified in the textbooks are described (subsection 3.4.2). The next two subsections provide information on the frequency of the retrieval and generation procedures that were found in the textbook corpus (sub-section 3.4.3) and on the frequency of these procedures at each proficiency level (sub-section 3.4.4). The results section then describes the types of input materials that were identified in the textbook corpus (sub-section 3.4.5) followed by how many phrases and the treatment of phrases in test materials (sub-section 3.4.6). After that, the frequency of the learning techniques per textbooks is presented. Finally, a summary of the main findings is provided.

3.4.1 Terms used to refer to strings of words

Phrase (67%) was the most common term used to express strings of words in exercises. Other terms were also used but occurred much less often, such as *expression* (12%), *phrasal verb/multiword verb* (8%), *collocation* (4%), and *idiom* (3%). The least common term was *multiword unit* (1%).

3.4.2 The test materials

Table 1 shows the results of the textbook analysis on the types of exercises (hereby referred to as test trials) along with the treatments of phrases in those test trials. Of the 1,660 test trials recorded in the excel file, decontextualized ones, 56%, occurred more frequently than contextualized ones, 44%. Additionally, test trials induced recognition processing 67% of the time while inducing recall processing only 33% of the time. Further, phrases were treated as whole units, 69%, more than they were to be reassembled from individual words, 31%. The most typical configuration of these three elements was decontextualized recognition tests treating phrases as whole units, 33%, and the least typical was contextualized recall tests treating phrases as separate words. Examples of the test trials presented in Table 1 are provided below.

Table 1

Test materials on phrases that occurred in 52 general ESL/EFL course textbooks

Test trials	Occurrence
Decontextualized recognition test trial treating phrases as intact wholes	549
Contextual recognition test trial treating phrases as intact wholes	509
Decontextualized recognition test trial treating phrases as separate words	228
Contextualized recognition test trial treating phrases as separate words	155
Decontextualized recall test trial treating phrases as separate words	107
Decontextualized recall test trial treating phrases as intact wholes	45
Contextualized recall test trial treating phrases as intact wholes	35
Contextualized recall test trials treating phrases as separate words	32

Decontextualized recognition test trial treating phrases as intact wholes

The most common test trial was decontextualized, and it treated phrases as intact units. Example 1 gives an illustration of it. Underlined spaces follow phrases, and the correct answer is presented with lures. The instructions direct students to choose the synonym and write it down in the underlined space. Although the correct response is provided, the challenge for students is to select the right synonym and not one of the others. Even if the wrong phrase was selected, a student might eventually realize whether the response was incorrect because the other synonyms might not match the other phrases.

Example 1

Please fill in the blank space with a word of the same meaning.

return, approach, make good progress

- coming along ____
- get back ____
- go up to ____

Contextual recognition test trial treating phrases as intact wholes

The second most common test trial presents a phrase in a sentence and treats it as a whole unit (see Example 2a). Above sentences with gapped spaces are phrases. The instruction requires students to select the phrases that fit the sentences. The context of the sentences helps students to make their choices. Although placing the phrases together makes it difficult to decide which phrase belongs to which sentence, the chances of selecting the incorrect answer is less than phrases set with distractors because distractors purposefully mislead students into thinking they have chosen the correct answer. Example 2b is a variant of Example 2a. It presents a phrase separated by distractors with forward-slashes in sentences. The instruction asks students to select the phrase that fits the sentence. Here, the distractors mislead students into choosing the wrong answer and prevents them from self-correcting in the same way that is possible in Example 2.

Example 2a

Fill in the correct form of one of the phrasal verbs from the box.

Bring up, held up, fill out

- It isn't easy to ____ children nowadays.
- ____ this application form and mail it in.
- Three masked gunman ____ the Security Bank this afternoon.

Example 2b

Choose the best answer to complete the sentences.

- After a while we all *sat down / stood up / fell down* to eat.
- Someone *fell down / stood up / went into* in the middle of the hall and asked a question.
- She *went into /sat down / got up* and walked across the room.
- She *laid the book down / moved the book up / ran the book out* on the table.

Decontextualized recognition test trial treating phrases as separate words

The third common test trial does not present any contextual support but rather lists the nodes and collocates of phrases into two separate columns (See Example 3). The instruction asks students to reconstitute the broken up phrases by matching the nodes with their respective collocates.

Example 3

Match the words in the left column with the words in the right column. It is possible to form more than one phrasal verb.

bring	out
chew	off
nod	up

Contextualized recognition test trial treating phrases as separate words

The fourth common test trial uses supportive context to help students reconstitute broken up phrases. As shown in Example 4a, a list of node words of phrases is presented above sentences missing these nodes. The instruction informs students to select the nodes that fit the sentences. Example 4a is similar to Example 2a in that it provides students with contextual support in choosing their answers. There are three variants of this test trial. Example 4a shows an exercise which pairs particles of PVs with distractors in sentences. Similar to Example 2b, the instruction in 4a asks students to choose the correct response, which they can do by circling or underlining their answer. Example 4c randomly presents the words in a sentence and separates them with forward-slashes. The instruction tells students to reorder the words correctly, and in the process, they also reassemble phrases. The last variant of this test trial is presented in Example 4d. It fragments a sentence at the phrase, so the verb of a PV is in one fragment, and its particle is in the other. The instruction tells students to rejoin the sentences, and doing so correctly requires knowledge of the PVs.

Example 4a

Complete the sentences with the missing verbs from below.

turn, used, chop

- Sam ___ up too much time on the first exam question.
- Guess who ___ up at midnight last night!
- Please could you ___ up these onions for me?

Example 4b

Choose the best answer to complete the sentences.

- After a while we all sat down / up / on to eat.
- Someone stood down / up / into in the middle of the hall and asked a question.
- We turned off / in / over the lights before anyone could see us.
- She laid the book down / up / out on the table.

Example 4c

Rearrange the words and phrases into the sentence

- You’d/out/cold/It’s/scarf!/so/put/on/better/a
- Try/I/these/fit?/make/Can/jeans/to/sure/on,/they
- a/got/so/I’ve/tonight/I’m/to/up./date/dress going

Example 4d

Match the sentence halves in column 1-3 with the sentence halves in column A-C.

1. I’m going to dress	A. off the meeting until tomorrow.
2. We asked the boss to put	B. out what he was saying.
3. I couldn’t make	C. up as a pirate for the party.

Decontextualized recall test trial treating phrases as separate words.

The fifth common test trial was decontextualized, and it treated phrases as words-to-be reassembled (see Example 5). It presented synonyms followed by the initial word of the phrases. Next to the phrases were underlined spaces. The instruction informs students to fill in the missing last word of the phrase based on knowledge of the synonym and the first word of that phrase. Unlike the Examples 1-4, it does not provide the answers for students to choose from. Instead, they are expected to either have knowledge of it or to take a blind guess.

Example 5

Please fill in the blank spaces with missing words of the phrase

- Cancel = put __
- Respect = look __
- Invent = make __

Decontextualized recall test trial treating phrases as intact wholes

The sixth common test trial was also decontextualized, and it treated phrases as intact wholes. It is perhaps the most difficult test trial because, as shown in Example 6, students are required to produce the phrases from memory given a one-word definition. If students are not familiar with the phrases, then they have to take a blind guess.

Example 6

Please fill in the blank spaces with matching phrases

- Cancel = ____
- Respect = ____
- Invent = ____

Contextualized recall test trial treating phrases as intact wholes

The seventh test trial uses context to help students recall the missing phrases. Example 7a shows gapped sentences and the instruction asks students to produce the phrases. A variant of this test trial is illustrated in Example 7b. It also consists of sentences with underlined spaces, but students are given the first letter of the phrase as additional support in helping them in producing the missing phrase. Like Examples 5 and 6, to respond, students must have knowledge of the phrase or take a blind guess.

Example 7a

Write down the missing phrasal verbs in the underlined spaces.

- The Prime Minister has decided to ______ after 10 years in office.
- We heard the bomb ____ from the hotel where we checked in.
- Large companies sometimes ____ smaller ones
- I can't ____ if it's a woman or man, because the person is too far away.

Example 7b

Fill in the blank space with phrasal verbs. The first letter is already provided for you.

- L__ __ your answer before the end of the exam.
- The thunder w__ __ the children.
- You may t__ __ the page now and read the exam questions.

Contextualized recall test trial treating phrases as separate words

The eighth common test trial also requires students to produce the missing phrases. Example 8a presents a sentence with missing verbs of phrases. The context is supposed to help students to remember them, but if they cannot, then they need to take blind guesses and write their responses in the underlined spaces. Example 8b is a variant of this test trial. It presents sentences with bolded phrases which do not fit the sentences. The instruction asks students to correct the phrases by producing the right ones.

Example 8a

Use your knowledge and write down the missing verb of the phrasal verbs in spaces provided.

- He ____ out of college and became a mechanic.
- The lecture was so boring that I ____ off.
- She ____ in to stop the argument.

Example 8b

Correct the phrasal verbs in these sentences.

- Where were you **grown up**?
- It took the passengers a long time to **come out** the plane.
- The explorer **started up** on a journey to Asia.
- After many years, he **ended out** having travelled around the world.

3.4.3 The input materials

As part of the textbook analysis, phrases that occurred in study trials prior to test trials were registered. Table 2 shows that of the 1,660 test trials identified, only 470, or 28%, were preceded by study trials. Phrases occurred in spoken (on the companion CD) and in written texts (on the pages of the textbook). In the spoken medium, phrases were embedded in context, while in the written medium, they were either in context or devoid of context. Combined, contextualized spoken and written texts account for 51% of the exemplar materials that preceded the test trials, while decontextualized written texts made up 40% of the study trials.

The textbook analysis also showed that study trials were either visible to students while doing the test trials or on pages that did not make the study trials accessible to students while doing the test trials. In some instances, when the study trials and the test trials were on the same page, the test trial instructions asked students to cover the study trial with their hand to avoid copying the answers.

Table 2

Study trial types

Study trial	Total
Contextualized (written & spoken)	286 (61%)
Decontextualized	184 (39%)

3.4.4 Treatment of phrases

As discussed above, test trials required students to do different things with phrases. Some required students to reassemble phrases by either matching them together or recalling one of the missing words. Others asked students to choose the correct phrase or to recall it from memory. The most common treatment of phrases was as intact wholes in decontextualized test trials, 36%. The next most common test trials treated phrases as intact wholes in contextualized materials, 33%. The two least common test trials presented phrases as separate words in decontextualized, 20%, and contextualized, 11%, materials. Concerning recognition-based exercises, of the 1,660 test trials, 64% treated phrases as whole units, and 23% treated them as separate words. Concerning recall-based test trials, of all the recorded exercises, only 5% required recall of phrases as whole units and 8%, as separate words.

3.4.5 The most common learning procedure in the textbook corpus

Table 3 shows the percentage of the learning methods that consisted of study trials before test trials (retrieval learning methods) and those that were made up of study trials after test trials (generation learning methods) that were identified in the textbooks. The table shows that a large majority of phrase-learning involves generation (72% to be exact). Learners are expected to either know the phrase before the test in order to provide the right answer or if they are unfamiliar with it, to take a blind guess. After providing their answers, the students are directed to evaluate their responses against the answer key located in the appendix of the textbooks.

Table 3

Total number of retrieval and generation exercises in the sample

Learning method	Total
Retrieval learning method	470 (28%)
Generation learning method	1,190 (72%)

3.4.6 Frequency of learning procedures at different proficiency levels

Table 4 provides the percentage of retrieval and generation learning methods occurring in each of the six proficiency levels. The interesting trend to report in this table is the fairly consistent proportion of retrieval learning against generation learning at each of these levels. Generation accounted for approximately 70% of phrase learning conditions while retrieval accounted for around 30%. Where there was a slight increase in generation learning conditions was at the pre-intermediate level and a small decrease at the intermediate level.

Table 4

Total number of retrieval and generation learning methods per proficiency level

Course book level	Retrieval	Generation
Beginner	27 (28%)	68 (72%)
Elementary	34 (28%)	122 (72%)
Pre-Intermediate	69 (21%)	254 (79%)
Intermediate	133 (33%)	267 (67%)

Upper-Intermediate	83 (30%)	191 (70%)
Advanced	124 (30%)	288 (70%)

3.4.7 Frequency of learning procedures in each course textbook series

Table 5 reports the percentage of retrieval-oriented and generation-oriented learning procedures identified in each course textbook series. The distribution of learning conditions appears fairly consistent across the different series. In each, a greater percentage of generation-oriented learning procedures (72%) than retrieval-oriented learning procedures (28%) were found. It should be noted, however, that generation-oriented learning procedures were predominately used in New English File (92%). Further, a greater number of retrieval- and generation-oriented learning procedures were identified in Speak Out, and the least number of these procedures was found in New Headway.

Table 5

Total number of retrieval- and oriented phrase learning procedures in each course textbook series

Course book series	Retrieval	Generation
Cutting Edge	52 (23%)	171 (77%)
English Result*	38 (33%)	76 (67%)
Global	73 (37%)	126 (63%)
New English File	16 (8%)	192 (92%)
New Headway	15 (22%)	54 (78%)
New Inside Out	58 (36%)	105 (64%)
New Total English	82 (37%)	138 (63%)
Speak Out	100 (33%)	210 (67%)
Straight Forward	36 (23%)	118 (77%)

* *Note.* No advanced level

3.4.8 Frequency of test materials at each proficiency level

Table 6 presents the eight different proficiency levels and the percentage of phrase-learning procedures in each. Unsurprisingly, the rate of phrase-focused procedures gradually increased from the beginner to the advanced levels, except for an abrupt decrease at the upper-intermediate level. This finding suggests that as learners advance in their proficiency level, vocabulary learning gradually shifts to focus on learning the relationship between words.

Table 6

Total number of phrase-focused exercises per proficiency level

Course book level	Exercises
Beginner	95 (6%)
Elementary	156 (9%)
Pre-Intermediate	320 (19%)
Intermediate	402 (24%)
Upper-Intermediate	275 (17%)
Advanced	412 (25%)

3.4.9 The average number of phrases in test trials

The last finding to report from the textbook analysis is the average number of phrases included in test trials. The average was 6.94 with a range from two to 20. Thus, students are expected to learn and retain knowledge of around seven phrases in a single exercise.

3.5 Rationale for the experimental studies

The textbook analysis identified 1,660 phrase-focused exercises in 52 ESL/EFL course textbooks, with the vast majority fostering generation learning. It also uncovered many different features comprising these methods and provided information on their distribution across different proficiency levels and textbook series. While Boers et al. (2014) have investigated the effectiveness of some of the recognition-based generation learning conditions identified here, I know of no study that has examined the retrieval via recall conditions and the generation via recall conditions to learn phrases. This gap in research partly motivated the investigation into the different version of the retrieval and generation procedures that were examined to facilitate PV knowledge in the three experimental studies in this thesis. These are reported in the following chapters.

Chapter 4 - Study 1

4.1 Introduction

Chapter 2 indicated that a tendency of L2 learners is to favour the use of single-word verbs compared to PVs for the reason that they are an extremely complex type of verb phrase to learn. It also showed a need for additional research to investigate the effects of methods to facilitate the learning of PVs because the majority of studies have only explored applications of the cognitive linguistics approach. In Chapter 3, I carried out a textbook analysis to determine what types of procedures materials designers advocate for the teaching and the learning of phrases in general and PVs in particular. The results of this analysis identified a number of similarities and differences between phrase-learning conditions. The next three chapters look at some of the features of these conditions.

The present chapter has six main sections. Section 4.2 presents the research questions of Study 1. Section 4.3 describes the methodology of this study. The learning conditions and the procedure are described in Sections 4.4 and 4.5, respectively. Section 4.6 then describes how the test responses were marked. Section 4.7 looks at the results of the analysis. The main findings of the study are summarized in Section 4.8, and Section 4.9 concludes the chapter by presenting the rationale for the next study.

4.2 Research questions

Study 1 was designed to answer the following questions:

1. Which learning procedure enhances short-term and long-term recall memory of PVs the best: study-test vs. test-study?
2. Which learning procedure enhances recognition memory of PVs the best: study-test vs. test-study?

3. Which learning procedure improves knowledge of the individual words of the PVs the best: study-test vs. test-study?
4. Which learning procedure causes greater interference on short-term and long-term memory of PVs: study-test vs. test-study?

4.3 Methodology

This section presents the methodology of the present study. First, it describes the L2 students used as participants and the PVs used as target items in the study. Next, it justifies the way the target items were selected. After that, it provides details about the features of the learning methods and outlines the procedure of the study. In this study, the independent variable was treatment with two levels: retrieval and generation. The dependent variable was scores on post-tests.

4.3.1 Participants

The participants were 199 Japanese-speaking EFL learners recruited from five parallel second-year English classes at a Japanese university in Hiroshima. Participants had studied English in Japan for at least 7 years prior to the treatment, and none had travelled to an English-speaking country before the study. All participants were familiar with the 2,000 most frequent words in English, as measured by the Vocabulary Levels Test (VLT) (Schmitt, Schmitt, & Clapham, 2001). The average score was 27.5 out of 30 (*SD* = 1.2). Sixty-five participants took part in a norming study while the others were randomly assigned to one of two treatment groups. All participants signed a consent form to participate in this study (see Appendix I)

4.3.2 Target items

Twenty-four PVs were selected as target items and were unknown to all participants in the norming study (see Appendix II). They were selected from a range of materials, including phrasal verb frequency lists (e.g., Biber et al., 1999; Gardner & Davies, 2007; Garnier & Schmitt, 2015; Liu, 2011), course textbooks (McCarthy & O'Dell, 2004; Gairns & Redman, 2011), L2 studies (Dagut & Laufer, 1985), and paper dictionaries (Courtney, 1983; Sinclair, 1990) and online dictionaries (www.macmillandictionary.com; www.dictionary.cambridge.org). On the basis of these materials, the paraphrases of meanings of the target items were constructed by the experimenter and another native English speaker and reviewed by a non-native speaker. Using Lextutor (www.lextutor.com), the BNC-COCA-25-word list was used to analyze the frequency of the individual words of the target items and the ones in the explanations of the definitions. The results placed approximately 85% in the first frequency-band, 12% in the second frequency-band, and three percent in the third frequency-band. The 24 target items were divided into four sets of six: A, B, C, and D. The decision on the number of target items to include in each set was based on the finding of the textbook analysis which showed that test trials (or exercises) require students to deal with around 6-8 phrases.

4.3.3 Norming study

One week before the treatment a norming study was carried out on a sample of the population to determine whether others with the same lexical profile would possess knowledge of the target items. Fifty candidate target items were selected from the materials listed above. To assess participants' knowledge of the target items, they were shown only the initial letter of a PV and a paraphrase of its meaning. On the basis of this retrieval cue, they were asked to fill in the missing letters of the PV. Of the ones none of the participants answered correctly, twenty-four were selected as target items.

The preferred method of selecting the target items was to use the norming study compared to the standard pre-and-post-test design. The benefit of the norming study is that participants in the treatment groups do not encounter any of the target items and the explanations of their definitions beforehand, which avoids causing the testing effect. The limitation of it is that there is a possibility that a small minority of participants in the treatment groups possess knowledge of the target items. However, I argue that the advantages of norming the target items outweigh its disadvantages for the reason that using the pre-and-post-test procedure poses a greater threat to internal validity. Pretesting has the potential to result in the pretesting effect, which may facilitate subsequent learning, or cause proactive interference, which may hinder subsequent learning. A plethora of studies in memory research has demonstrated the effects of pretesting. Another reason to avoid the use of the pre-and-post-test design was that the pretesting effect is the very phenomenon that is under investigation in the present study. Its use, therefore, would naturally confound the results of the generation condition (not to mention the retrieval condition) on the learning of target items.

4.4 Learning conditions

The present study investigated the effects of two conditions of learning. I describe their design below. Rather than the traditional pen-and-paper method of data collection used in applied linguistics, I decided to administer the treatments and collect the data using a software program (Qualtrics), which has the added benefit of controlling a number of potentially confounding variables, such as unequal time spent during learning. Further, it allows for the provision of instantaneous corrective feedback, which is a key element of one of the treatments. It should also be noted that after finishing each set of target items, participants completed a 10-minute distractor task before the post-test (see below for more details).

As mentioned above, the current study examines and compares two treatments on the learning of PVs. Both treatments are composed of the same two learning events. One is a study trial, and the other is a test trial. In the retrieval condition, the study trial came before the test trial; in the generation condition, test trial came before the study trial which also presented corrective feedback. Further details of these two treatments are discussed below.

4.4.1 Retrieval condition

The study trial displayed a PV accompanied by a paraphrase of its meaning. This form-meaning association stayed on the screen for 15 seconds (e.g., *hang out — to spend time with friends*). During this time, participants attempted to learn the target item. Once time expired, the study trial was automatically replaced by the test trial, which presented the initial letter of the PV followed by an underlined space, which, in turn, was followed by the explanation of the target item's definition (e.g., *h____ — to spend time with friends*). During this time, the participants had 15 seconds to recall the missing PV and to type their response in a text box located under the blank space. Each target item was treated in the same way and one at a time (see Appendix III for a sample of the retrieval condition).

4.4.2 Generation condition

The test trial displayed the first letter of a PV followed by an underlined space, which, in turn, was followed by the explanation of the PV's meaning. The instructions informed participants to think of the missing PV and then to type their guess in the text box located under the underlined space. Like those in the retrieval condition, participants in the generation condition had a maximum of 15 seconds to complete the test trial. Afterwards, the study trial appeared on the screen showing the complete form-meaning connection of the target item, and it also provided feedback on the participant's response by indicating whether it was right or wrong using a green ✔ or a red **X**, respectively. Participants had 15 seconds in which to recognize whether their answer was correct or not and to study the correct association (see Appendix IV).

4.5 Procedure

This study consisted of four phases: practice, treatment, distractor, and post-test

4.5.1 Practice phase

The practice phase occurred one week before the learning phase. It had three main purposes. One was to familiarize students with the conditions of learning, so as to avoid any issues that may arise due to unfamiliarity with the use of the program. The second was to estimate how much time students needed to complete each trial of the treatments, so time on task can be controlled in the treatment. The third was for students to pre-learn the individual words of the target items and the ones used in the paraphrases of their meanings, so as to ensure that comprehension difficulties with these words was not a factor in learning the PVs.

At the beginning of class, students were sent a link to their computers. The link opened the program that presented the retrieval and generation conditions. The items to be learned in the practice phase were collocations rather than PVs because the focus here was on learning to use the conditions and the technology in which they were presented and not on learning the items per se. The program recorded how much time it took participants to complete each trial. In the practice retrieval condition, participants spent approximately 8 seconds on the study trial and 4 seconds on the test trial. In the practice generation condition, they spent around 10 seconds on the study trial and 11 seconds on the test trial. Given these figures, it was decided that 15 seconds should be sufficient for students to complete each trial in both treatments in the treatment phase of the study.

4.5.2 Treatment phase

The treatment phase started at the beginning of each of the five classes. The teachers in each of these classes gave the same set of instructions, which was for students to follow the instructions provided on their computers. Upon clicking on the link, the participants were randomly but equally assigned to one of the two experimental conditions. The first set of items in each of the conditions were fillers. The set of target items (A, B, C, and D) were counterbalanced in each learning condition by participant. In the retrieval condition, participants had 15 seconds to study a target item before they were presented with a retrieval cue to elicit a memory of that target item and had 15 seconds to produce a response. In the generation condition, participants had 15 seconds to produce a response to the (retrieval) cue and then 15 seconds to review their response in light of the corrective feedback. After each set of target items was a distractor task followed by a post-test.

4.5.3 Distractor phase

For the distractor phase, participants were presented with L1 trivia questions and two digit additions (e.g., 48-19=?). These questions do not require participants to use English. Therefore, there is no chance of them accidentally using any of the individual words in the treatment. If they did use any of these words, then this would count as a learning event. The distractor task was long — lasting 10 minutes, which should have been a sufficient amount of time to flush their short-term working memory of residual traces of the PVs.

4.5.4 Post-test phase

A cued-recall test was used to assess learning immediately after the treatment and retention after a one-week delay. The test presented the initial letter of a target item followed by an underlined space, which, in turn, was followed by the paraphrase of the PV's meaning. To complete the test, participants had to type the missing PV in a text box below the underlined space. The target items on the test were reversed the order in which they were presented in the treatment.

A recognition test was also used to assess retention. It was administered after the delayed cued-recall post-test. The test presented the target items in a left-hand column and the explanations of their definitions in the right-hand column. To complete the test, participants had to drag and drop the meanings next to their respective forms. One set of target items was presented on the screen at a time, and the PVs in each set were randomized.

4.6 Counting scores

The cued-recall test was scored in two ways. A full point was awarded when participants produced the verb and the particle correctly. Spelling mistakes on the verbs were accepted as long as they did not form a new word. However, spelling mistakes on the particles were not accepted because an error on one particle might form another particle (e.g., on vs. in).

Replicated errors on the cued-recall test were counted. A replicated error is an incorrect response on the test trial that a participant reproduces on the final test. A replicated PV error is when a participant reproduces the same verb and particle error; a replicated verb error is when a participant produces the same verb error; a replicated particle error is when a participant gives the same particle error.

A more in-depth error analysis was conducted on the nature of replicated errors produced by participants in the generation group. First, it examined the relationship between a replicated response on one of the items in relation to the type of response given on the other item. For example, in the test trial, a participant produces the verb and the particle incorrectly, and then on the final test, this individual recalls the same verb error and produces the particle correctly. This type of performance shows that the corrective feedback had a positive effect on learning the particle but not on learning the verb. The error analysis considered 35 possible scenarios, and they are described below.

XX,XX. On the test trial, a participant errs on the verb (X) and the particle (X). On the post-test, that individual replicates the verb error (X) and the particle error (X) instead of recalling the correct verb (C) and the correct particle (C) presented in the corrective feedback. The "X" indicates a mistake. The "C" designates a correct response. The "," separates test trial verb and particle responses with post-test verb and particle responses. Thus, the XX,XX means that on the test trial, a participant failed to produce the correct verb and particle responses and replicated the same incorrect verb and particle responses on the post-test.

XX,CX. On the test trial, a participant errs on the verb (X) and the particle (X). On the post-test, that individual recalls the correct verb (C) but replicates the particle error (X). Thus, following the test trial, the participant learned the right verb response but not the particle response.

XX,XA. On the test trial, a student errs on the verb (X) and the particle (X). On the post-test, that student replicates the verb error (X) and produces a different particle error (A) from the one generated in the test trial. Therefore, the student retained knowledge of the original verb error but not of the particle error. The corrective feedback did not positively affect the learning of the correct verb and particle responses.

XX,XC. On the test trial, a student errs on the verb (X) and the particle (X). On the post-test, that individual replicates the verb error (X) while correctly recalling the particle response presented in the corrective feedback. Thus, the feedback helped the student learn the correct particle but did not affect learning the right verb.

XC,CX. On the test trial, a participant errs on the verb (X) but not on the particle (C). On the post-test, that individual produces the right verb response but forgets the right particle response. The corrective feedback helped the student learn the correct verb response but may have interfered with the recall of the correct particle response.

XB,CX. On the test trial, a student incorrectly generates the verb response and leaves the particle response blank. On the post-test, that individual produces the correct verb response but not the particle response. The corrective feedback facilitated learning of the verb but not of the particle.

XC,XC. On the test trial, a participant errs on the verb response but successfully generates the particle response. On the post-test, that person replicates the verb error and correctly recalls the particle response. The corrective feedback had a positive effect on strengthening the memory of the particle response but not the verb response.

XC,AX. On the test trial, a member of the generation group generates the wrong verb response but the correct particle response. On the post-test, this individual produces an incorrect verb response different from the one produced on the test trial and forgets the right particle response. The corrective feedback did not facilitate learning the correct verb response and failed to strengthen the memory of the right particle response.

CX,XA. On the test trial, the student generates the correct verb response but not the particle response. On the post-test, this individual forgets the right verb response and produces a particle error different from the one generated on the test trial. The corrective feedback did not strengthen the memory of the correct verb response, nor did it aid in learning the proper particle response.

CX,CX. On the test trial, a participant successfully generates the verb response but not the particle response. On the post-test, this individual correctly recalls the verb response but replicates the particle error. The corrective feedback had a positive effect on strengthening the memory of the verb response but not of the particle response.

XX,AX. On the test trial, a student in the generation group generates the wrong verb and particle responses. On the post-test, he or she produces a different verb error than the one produced in the test trial and replicates the particle error. The correct feedback did not positively affect learning the correct verb and particle responses.

CX,XC. On the test trial, a participant produces the correct verb response but not the particle response. On the post-test, the student fails to recall the right response of the verb but provides the right particle response. The corrective feedback did not help to learn the correct verb response, but it did have a positive effect on learning the particle response.

XX,XB. On the test trial, the participant failed to generate the correct verb and particle responses. On the post-test, this individual also did not produce the right verb response and provided a particle error that was different from the one given in the test trial. The corrective feedback did not have any positive effect on learning the correct verb and particle responses.

XC,XX. On the test trial, a participant errs on the verb response but produces the correct particle response. On the post-test, this individual replicates the same verb response and produces an incorrect particle response. The correct feedback did not help to strengthen the memory of the particle response correctly generated in the test trial and had no effect on the learning of the verb response.

CX,XX. On the test trial, a member of the generation group produces the correct verb response but not the particle response. On the post-test, this member forgets the right verb response and replicates the particle error. The correct feedback did not aid in entrenching the right verb response in memory, and it did not have any effect on the particle error.

XB,AX. On the test trial, a student incorrectly generates the verb response and leaves blank the particle response. On the post-test, this individual produces an error different from the one produced in the test trial and provides an incorrect particle response. The corrective feedback did not help to correct the participant's mistakes.

XC,XB. On the test trial, an individual fails to produce the correct verb but gives the right particle. On the post-test, this person recalls the same verb error and does not provide a particle response. The corrective feedback did not help to consolidate the correct particle response the individuals produced during learning.

XC,XA. On the test trial, a participant produces the wrong verb but correct particle. On the post-test, this person recalls the same incorrect verb but fails to remember the correct particle. The corrective feedback did not help the student to remember the right particle response.

XB,XX. On the test trial, a participant produces the wrong verb and does not provide a particle response. On the post-test, this person recalls the same verb error and produces an incorrect particle response. The corrective feedback did not help the student learn the correct verb and particle combination.

XB,XC. On the test trial, a participant does not produce the correct verb and does not provide a particle response. On the post-test, this individual replicates the verb error but correctly provides the particle. The corrective feedback assisted the learner to acquire the particle response but not the verb response.

XB, XB. On the test trial, a participant did not produce the verb correctly and did not produce any particle response. On the post-test, this individual repeated the same verb error and again did not provide any particle response. The corrective feedback did not help the student learn the correct verb and particle combination.

XX,BX. On the test trial, a student does not produce the correct verb and particle correctly. On the post-test, the same person does not provide any verb response and replicates the same particle error. The corrective feedback did not help the student to learn the correct verb and particle combination.

XC,BX. On the test trial, an individual produces the verb incorrect but the particle correctly. On the post-test, the student does not provide a verb response and fails to recall the particle response. The corrective feedback did not help the learner to remember the correct particle response.

XB,XA. On the test trial, a student does not generate the correct verb and fails to provide a particle response. On the post-test, this individual repeats the same verb response and produces an error that is different from the one generated in the test trial. The corrective feedback was not effective in helping the student learn the correct verb and particle combination.

XB,BX. On the test trial, a student does not produce the correct verb and fails to provide a particle response. On the post-test, the individual does not give a verb response and provides an incorrect particle response. The corrective feedback did not assist the student in learning the correct verb + particle combination.

CX,XB. On the test trial, a student produces the correct verb response but not the particle response. On the post-test, the student fails to recall the right response of verb and does not provide a particle response. The corrective feedback did not help the individual to remember the correct verb response.

CX,BX. On the test trial, a learner produces the correct verb but does not provide a particle response. On the post-test, this student does not provide any verb response and reproduces the particle error. The corrective feedback did not help the student to remember the correct verb response.

CX,AX. On the test trial, a student produces the correct verb response but not the particle response. On the post-test, this learner fails to remember the right response of verb and recalls the same particle error. The corrective feedback did not help the student to remember the correct verb response.

BX,XX. On the test trial, an individual does not provide any verb response and fails to produce the correct particle response. On the post-test, the same person fails to recall the correct verb and repeats the same particle error. The corrective feedback did not help the student to learn the verb + particle combination.

BX,XC. On the test trial, a participant does not provide any verb response and incorrectly generates a particle response. On the post-test, this person produces an incorrect verb response but recalls the correct particle response. The corrective feedback helped the student to learn the right particle response.

BX,XB. On the test trial, a student does not provide any verb response and fails to produce the correct particle response. On the post-test, this person does not give the right verb response and provides no particle response. The corrective feedback had no positive effect on helping the student o learn the PVs.

BX,XA. On the test trial, a student does not provide a verb response and gets the particle wrong. On the post-test, the student fails to produce the right verb response and gives a particle error that is different from the one generated in the test trial. The corrective feedback had no positive effect on helping the student to learn the PV.

BX,CX. On the test trial, a participant does not provide any verb response and incorrectly generates the particle response. On the post-test, this person correctly recalls the verb response but reproduces the erroneous particle response. The corrective feedback had a positive effect on learning the correct verb response.

BX,BX. On the test trial, an individual does not provide a verb response and incorrectly produces the particle response. On the post-test, this person again leaves the verb response blank and recalls the same erroneous particle response. The corrective feedback did not have any positive effect in helping the student learn from the errors.

BX,AX. On the test trial, a student does not provide a verb response and fails to provide a correct particle response. On the post-test, the student produces a verb error and replicates the particle error. The corrective feedback did not have any positive effect on learning the PV.

4.7 Results

This section provides descriptive and inferential statistics on correct responses on the test trial, the post-tests and on the number of replicated errors. The first set of descriptive statistics is of the treatment groups' full marks on the recall post-tests. The second set is on their performance on the recognition post-test. The third set is on their half marks on the recall post-tests. The last set is on replicated errors. Following the descriptive statistics, the inferential statistics are reported. The first set of models were performed on full marks; the second set, on half marks. The last analysis was performed on replicated errors.

4.7.1 Descriptive statistics on test trial, recall and recognition post-tests

Table 7 shows the total number of correct responses on the test trial, the recall post-tests, and the recognition post-test. On the test trial, participants in the retrieval group correctly recalled 98% of the target items, whereas those in the generation group did not correctly guess any of them. The fact that the generation group did so poorly is evidence that the target items were unknown to them and most likely to those in the retrieval group, which supports the results of the norming study.

The percentage of correct responses on the immediate recall post-test was greater in the retrieval condition (70%) than in the generation condition (57%). However, there was a significant amount of attrition between the time of the immediate post-test and the delayed post-test for both treatment groups. The loss in knowledge for the retrieval group (70% to 19%) was 51 percentage points and for the generation group (57% to 18%), 39 percentage points. The number of correct responses produced by participants on the recognition test was slightly greater in the retrieval condition (49%) than in the generation condition (47%)

Table 7

Correct responses on the test trial, recall post-tests and recognition post-test

Group	Test trial	Recall immediate	Recall delayed	Recognition
Retrieval	1,199	848	237	600
Generation	0	696	217	571

Note: The total number of possible responses on the recall and recognition post-test for each group was 1224.

Table 8 shows the number of correct verb and particle responses on the recall post-tests. With respect to verbs, on the immediate post-test, for the retrieval group, performance was better on verbs (87%) than on particles (72%). Similarly, for the generation group, participants recalled more verbs (80%) than particles (62%). On the delayed post-test, verbs were recalled more than particles for the retrieval group (39% and 26% respectively) and for the generation group (31% and 26% respectively).

The table also shows that on the immediate and delayed post-test, the number of correct verb responses was greater in the retrieval group (87% and 39% respectively) than in the generation group (80% and 31% respectively). The number of correct particle responses was also greater in the retrieval group (72%) than in the generation group (62%) on the immediate post-test, but not on the delayed post-test (26% and 26% respectively).

Table 8

Correct verb and particle responses on the recall post-tests

	Immediate		Delayed	
Group	Verb	Particle	Verb	Particle
Retrieval	1059	884	481	314
Generation	976	762	388	315

Note: The total number of possible responses on the recall for each group was 1224.

Table 9 provides the percentage of replicated errors on the recall post-tests by participants in the generation group. No error analysis was conducted on the retrieval group because participants in this group produced so few test trial errors. The percentage of replicated PVs errors was greater on the immediate post-test (24%) than on the delayed post-test (11%). Similarly, for replicated verb errors, they occurred more often on the immediate post-test (34%) than on the delayed post-test (12%). Likewise, for replicated particle errors, 27% were found on the immediate post-test and only 10% on the delayed post-test. On the immediate post-test, the percentage of replicated errors was greater for verbs than for particles, while on the delayed post-test, the percentage of replicated errors was relatively the same.

Table 9

Replicated PV, verb and particle errors on the immediate and delayed recall post-test by the generation group

Errors	Immediate	Delayed
PV	124 (24%)	116 (11%)
Verb	58 (34%)	121 (12%)
Particle	21 (27%)	100 (10%)
Other	325 (61.5%)	670 (66.5%)

Note: Total errors on the immediate post-test was 528 and on the delayed post-test was 1,007.

4.7.2 Mixed effects logistic regression models on the recall post-tests: Full marks

Mixed effects logistic regression modelling (in R version 3.2.3 with lm4 package; Bates, Bolker, & Walker, 2014) was used to analyze the scores on the recall and recognition post-tests. The mixed effects logistic regression is an extension of a binary logistic regression that predicts membership of only two categorical outcomes: the number of 1's and 0's. Laplace approximation fitted all mixed logit models for obtaining restricted maximum likelihood estimates.

The odds ratio is provided with each mixed logit model. The odds are defined as the probability of an event occurring divided by the probability of it not occurring (Field, Miles, & Field, 2012). In other words, the odds of obtaining a correct response on a post-test are the probability of obtaining a correct response divided by the probability of not obtaining a response. For example, if the odds of an event are 2.0, this success is 2 times as likely as a failure (Agresit, 2007). With the odds ratio, the odds of success in one treatment are compared to the odds of success in another treatment (Agresit, 2007). If the odds ratio is greater than 1.0, then a success is more likely in the first treatment. Each reported odds ratio is presented with its 95% confidence interval. In addition to mixed logit modelling, z-tests for one proportion were run on the proportion of errors on the post-tests that were either the same ones generated during learning or different from the ones generated during learning.

Prior to running mixed effects logistic regression on the recall post-tests and the recognition post-test, R was used to check the assumption of normality using Shapiro test as well as by plotting the data according to Studentized residuals, Hat values and Cook's distance to identify outliers and influential cases. The analysis showed that data were normally distributed and that there was no significant outliers or influential cases.

The first step in the model was to analyze the number of correct responses in the immediate and delayed recall post-tests. Participants' group membership, VLT scores, and class identification were included in the model as fixed effects, while participants nested by learning method and target items nested by participant were counted as random effects. Next, to find the best fit model, all the predictors were initially included: treatment, VLT scores, class identification. A backwards stepwise approach was employed (Field, 2013) by removing the fixed effects in a stepwise fashion. If the effects of the removal of the predictors were not statistically significant and did not improve the fit of the model, they were removed.

Table 10 shows a summary of the fixed effects estimates for the retrieval group based on the intercept of the generation group. Each treatment in the table includes an estimate of fixed effect, standard error (*SE*), the z-value, the p-value, the odds ratio and the 95% confidence interval for the odds ratio. If the p-value was over 0.05, the odds ratio and the 95% confidence interval for the odds ratio were not calculated, unless otherwise specified.

The first step in the analysis was to compare the performance of the treatment groups. Table 10 provides the best fit repeated measures model on correct responses of the treatment groups. The odds of the retrieval group obtaining a correct response were 2.154 times the odds of the generation group obtaining a correct response. The model also recorded a significant interaction between time and treatment ($p < 0.0001$), which indicates that attrition was greater for the retrieval group than for the generation group.

Table 10

Mixed effects logistic regression repeated measure analysis on the recall post-tests

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	0.400	0.234	1.710	0.087	1.492	0.943, 2.362
Retrieval	0.767	0.134	5.712	< 0.0001	2.154	1.655, 2.803
R within	-2.377	0.115	-20.614	< 0.0001	0.092	0.074, 0.116
Interaction	-0.778	0.165	-4.710	< 0.0001	0.459	0.332, 0.634

Note: The generation group is the intercept. *R within* refers to performance across the post-tests for the retrieval group. *Interaction* refers to treatment interaction.

Having established that there was a large difference in the odds between the treatment groups across the recall post-tests, the second step was to look for differences between the treatment groups on the individual post-tests. Table 11 shows the best fit model on the immediate post-test. The odds of obtaining a correct response by a participant in the retrieval group were 2.348 times the odds of obtaining a correct response by a participant in the generation group. Thus, the retrieval group was much more likely than the generation group to supply the correct PV response.

Table 11

Mixed effects logistic regression model on the recall post-test: Summary of fixed effects on immediate post-test

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	0.3814	0.252	1.513	0.13	1.464	0.893, 2.400
Retrieval	0.8537	0.195	4.376	< 0.0001	2.348	1.602, 3.441

Note: The generation group is the intercept.

Table 12 shows the best fit model on the treatment groups' correct delayed post-test responses. The difference between the treatment groups was found to be not statistically significant ($p = 0.694$).

Table 12

Mixed effects logistic regression model on recall post-test: Summary of fixed effects on delayed post-test

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-2.144	0.294	-7.270	< 0.0001	-	-
Retrieval	-0.094	0.239	-0.393	0.694	-	-

Note: The generation group is the intercept.

4.7.3 Mixed effects logistic regression model on the recognition post-test: Full marks

Table 13 shows a summary of the best fit mixed effects logistic regression model on correct recognition post-test responses. The model indicated that the difference between the treatments was not statistically significant ($p = 0.267$).

Table 13

Mixed effects logistic regression model on recognition post-test

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.011	0.184	0.061	0.952	0.988	0.689, 1.418
Retrieval	0.099	0.090	1.109	0.267	1.105	0.926, 1.318

Note: The generation group is the intercept.

4.7.4 Mixed effects logistic regression models on the recall post-test: Half marks in each condition

The next set of models were performed on the treatment groups' half marks (verb and particle responses). The fixed and random effects that fitted the best models on full marks were the same fixed and random effects that fitted the models on half marks.

4.7.4.1 Retrieval: Verb vs. particle

Table 14 shows the best fit repeated measures model for the retrieval group's performance on verb and particle responses. The odds of a participant in this group correctly recalling a verb was 3.495 times the odds of that participant correctly recalling a particle. The model also showed that the interaction between time and performance on the verb and particle responses was not statistically significant ($p = 0.571$). Additional models were run on the individual post-tests. On the immediate post-test (see Table 15), the odds of the verb being recalled successfully were 3.013 times the odds of the particle being recalled successfully. However, on the delayed post-test (see Table 16), the difference between successful recall of the verb and the particle was found to be not statistically significant ($p = 0.227$), although the odds of correctly recalling the verb were 1.915 times the odds of recalling the particle.

Table 14

Retrieval group: Verb vs particle: Fixed effects repeated measures summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-1.535	0.388	-3.956	< 0.0001	0.215	0.100, 0.460
Verb	-1.251	0.388	3.220	0.001	3.495	1.632, 7.487
Interaction	-0.209	0.370	-0.565	0.571	0.811	0.392, 1.675

Note: The particle is the intercept.

Table 15

Retrieval group: Verb vs particle: Fixed effects on immediate post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	1.163	0.359	3.235	< 0.01	3.200	1.581, 6.474
Verb	1.103	0.379	2.907	0.003	3.013	1.432, 6.339

Note: The particle is the intercept.

Table 16

Retrieval group: Verb vs particle: Fixed effects on delayed post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-1.490	0.457	-3.260	0.001	0.225	0.091, 0.551
Verb	0.649	0.538	1.206	0.227	1.915	0.666, 5.507

Note: The particle is the intercept.

4.7.4.2 Generation: Verb vs. particle

Table 17 shows the best fit repeated measures model on correct verb and particle responses by participants in the generation group. The odds of a participant in this group correctly recalling a verb were 1.137 times the odds of that participant correctly recalling a particle. The model also indicated an interaction between time and recall of verbs and particles ($p <$ 0.0001), showing a greater loss in knowledge of verbs than particles. Given the statistically significant difference reported in the repeated measures model, further models were performed on the individual post-tests. On the immediate post-test, the odds of successfully recalling a verb were 2.692 times the odds of recalling a particle (see Table 18). However, on the delayed post-test, the different scores between these two items were found to be not statistically significant ($p = 0.963$) (see Table 19).

Table 17

Generation group: Verb vs. particle: Fixed effects repeated measures summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-1.168	0.097	-11.988	< 0.0001	0.310	0.256, 0.376
Verb	0.319	0.093	3.405	< 0.001	1.375	1.145, 1.653
Interaction	0.631	0.134	4.693	< 0.0001	1.880	1.444, 2.448

Note: The particle is the intercept.

Table 18

Generation group: Verb vs particle: Fixed effects on immediate post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	0.5930	0.262	2.260	0.023	1.809	1.082, 3.026
Verb	0.9905	0.292	3.384	< 0.001	2.692	1.516, 4.778

Note: The particle is the intercept.

Table 19

Generation group: Verb vs particle: Fixed effects on delayed post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-1.3274	0.359	-3.692	< 0.001	-	-
Verb	-0.0226	0.490	-0.046	0.963	-	-

Note: The particle is the intercept.

4.7.5 Mixed effects logistic regression models on the recall post-tests: Comparison of treatments on half-marks

The last set of models were carried out on the verb and particle responses between the treatment groups to observe whether success on an item was affected by the learning condition. The first set of models was performed on the treatment groups' performance on the verb. The second set of models was conducted on these groups' scores on the particle. The fixed effects that fitted the first two sets of mixed effects logistic regression models were the same fixed effects that fitted the models on half marks. However, the random effects were different. For random effects, the participants were nested by treatment, and the verb and the particle responses were nested by participant.

4.7.5.1 Verb: Retrieval vs. generation

Tables 20 and 21 present the best fit models for the treatment groups' correct scores on verb responses on the immediate and delayed post-tests respectively. On the immediate post-test, the odds of the retrieval group successfully recalling a verb were 1.740 times the odds of the generation group (see Table 20). The odds also favoured the retrieval group on the delayed post-test as their odds of correctly producing a verb were 1.565 times the odds of the generation group (see Table 21).

Table 20

Treatment groups' performance on verb: Fixed effects on immediate post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	1.589	0.179	8.864	< 0.0001	4.900	3.448, 6.963
Retrieval	0.554	0.114	4.840	< 0.0001	1.740	1.390, 2.178

Note: The generation group is the intercept.

Table 21

Treatment groups' performance on verb: Fixed effects on delayed post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-1.149	0.272	-4.215	< 0.0001	0.316	0.185, 0.540
Retrieval	0.448	0.098	4.569	< 0.0001	1.565	1.291, 1.898

Note: The generation group is the intercept.

4.7.5.2 Particle: Retrieval vs. generation

Tables 22 and 23 reports the best fit models on the treatment groups' correct particle scores on the immediate and delayed post-tests respectively. On the immediate post-test, the odds of obtaining a correct particle response in the retrieval condition were 1.740 times the odds of obtaining a correct particle response in the generation condition. However, on the delayed post-test, the difference in particle scores between the retrieval group and the generation group was found to be not statistically significant ($p = 0.900$).

Table 22

Treatment groups' performance on particle: Fixed effects on immediate post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	0.411	0.295	1.390	0.164	1.508	0.844, 2.694
Retrieval	0.554	0.095	5.792	< 0.0001	1.740	1.443, 2.100

Note. The generation group is the intercept.

Table 23

Treatment groups' performance on verb: Fixed effects on delayed post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-1.397	0.390	-3.578	< 0.001	-	-
Retrieval	-0.012	0.101	-0.125	0.900	-	-

Note. The generation group is the intercept.

4.7.6 Error analysis on the recall post-tests

Based on the test trial errors of the treatment groups, error analysis was carried out on the generation group but not on the retrieval group. The reason is that the retrieval group produced so few test trial errors to provide any meaningful data on the effects of interference on learning, whereas the generation group's test trial responses were all errors.

Replicated errors on the immediate post-test

The generation group produced 528 immediate post-test errors of which 24% were replicated test trial PV errors, and 76% were other errors. A z-test for one proportion showed that the likelihood of producing replicated PV errors was less than producing other errors ($z = 20.032$, $p < 0.0001$, 95% CI = 20.42, 30.88). The generation group incorrectly produced 248 verb responses on the immediate post-test of which 34% were replicated test trial verb errors. They also erroneously generated 462 particle responses of which 24% were replicated test trial particle errors. A z-test for one proportion indicated that the likelihood of recalling the same verb error was less than producing a random error ($z = 30.575$, $p < 0.0001$, 95% CI = 29.96, 38.22). Similar results were found for replicated particle errors ($z = 23.195$, $p < 0.0001$, 95% CI = 23.26, 31.00).

Replicated errors on the delayed post-test

On the delayed post-test, participants in generation group produced 1,007 errors of which 11% were replicated test trial PV errors. They incorrectly recalled 836 verb responses of which 12% were replicated test trial verb errors. The participants erroneously produced 909 particle responses of which 10% were replicated test trial particle errors. The likelihood of recalling the same PV response, the verb response and the particle response were less than randomly producing an error response ($z = 8.736$, $p < 0.0001$, 95% CI = 9.13, 13.10; z =

10.192, $p < 0.0001$, 95% CI = 10.06, 14.17; z = 7.280, $p < 0.0001$, 95% CI = 8.22, 12.02 respectively).

Comparison of replicated verb and particle errors

The next set of analyzes were carried out to determine whether participants in the generation group were more likely to produce replicated verb or particle errors on the post-tests. On the immediate post-test, a z-test indicated that participants were more likely to produce replicated verb errors (72%) than replicated particle errors (21%) (z = 23.324, $p < 0.0001$, 95% CI = 60.76, 81.52). Similarly, on the delayed post-test, the z-test showed that the likelihood of recalling replicated verb errors (55%) was greater than replicated particle errors (45%) (z = 34.105, $p < 0.0001$, 95% CI = 48.19, 61.68).

Replicated errors on one item compared to response type on the other item

Table 24 shows the results of the detailed error analysis on the relationship between test trial errors and post-test errors. XX,XX was found to be the most common type of scenario not only on the immediate but also on delayed post-test. This finding suggests that participants were more likely to replicate a verb error and a particle error together rather than replicate a verb error with a different response or replicate a particle error with a different response.

Table 24

Error analysis on the cued-recall post-tests

Scenario	Immediate	Delayed
xx,xx	33	88
xx,cx	32	10
xx,xa	31	51
xx,xc	19	9
xc,cx	18	10

xb,cx	13	5
xc,xc	10	27
xc,ax	7	29
cx,xa	7	14
cx,cx	7	23
xx,ax	6	17
cx,xc	4	5
xx,xb	3	2
xc,xx	3	9
cx,xx	3	9
xb,ax	2	6
xc,xb	1	1
xc,xa	1	6
xb,xx	1	10
xb,xc	1	1
xb,xb	1	4
xx,bx	0	0
xc,bx	0	0
xb,xa	0	0
xb,bx	0	0
cx,xb	0	1
cx,bx	0	0
cx,ax	0	0
bx,xx	0	0
bx,xc	0	0
bx,xb	0	0
bx,xa	0	0
bx,cx	0	0
bx,bx	0	0
bx,ax	0	0

Summary of results

The present study sought to investigate and compare the effects of two conditions on the learning of PVs. The retrieval condition had participants study a PV and then attempt to recall it. The generation condition had participants take a guess on the PV given its definition and then to study the correct answer. Learning was measured after a 10-minute distractor task and retention was assessed after a one-week delay. The results of the immediate post-test showed that performance was greater for participants in the retrieval condition than those in the generation condition. However, the advantage of the retrieval condition was short-lived as no difference was found between the conditions of learning on the delayed post-test. The results also indicated that the rate of attrition over a one-week period was very high for participants in both conditions, but higher for those in the retrieval group. The results of the recognition post-test showed that performance did not differ between the two groups. Analysis of the individual words of the PVs showed that performance was greater on verb than particle for both groups on the immediate post-test. However, the loss of knowledge over a one-week period was higher for verbs than for particles, and this rate of attrition resulted in no difference in recall of the verbs and the particles on the delayed post-test. Considering performance on the verbs, learning gains were higher for participants in the retrieval condition than those in the generation condition on the immediate and delayed post-tests. Considering performance on the particles, learning gains were higher for participants in the retrieval condition than those in the generation condition on the immediate post-test, but performance was relatively the same on the delayed post-test. Finally, the generation condition resulted in a substantial amount of proactive interference, which tended to be higher on the verb than the particle.

Chapter 5 - Study 2

5.1 Introduction

The previous chapter has shown that retrieval, as opposed to generation, produces significantly better short-term learning of PVs, although no such advantage was found in long-term retention. Study 2 further explores the effects of retrieval and generation on PV learning and retention. Study 1 and the present study differ in design in three critical areas. First, the current study uses a different type of test trial. In Study 1, the test trial presented the *initial letter of a PV* followed by an underlined space and a paraphrase of its meaning (e.g., *h___ - to spend time with friends*). In the present study, the difference is that the test trial displays the *verb of a PV* (e.g., *hang__ - to spend time with friends*). Second, the present study employs a different type of cued-recall final test compared to the one used in Study 1. In the former study, the final test showed participants the first letter of the PV next to a blank and its meaning (e.g., *h___ - to spend time with friends*). In Study 2, the test trial shows participants an underlined space and the meaning of the PV (e.g., ___ *- to spend time with friends*). Third, the present study considers an additional moderating variable: the number of items in a single study episode. More specifically, it looks into two versions of a retrieval condition. In the first version, the study trial and the test trial occur consecutively. In the second one, participants receive the test trial after 13 other items. The current study also examined two versions of the generation condition: one where the test trial and the study trial occur back-to-back; another where the two trials are separated by 13 other items. The results of the present study should shed light on the following issues. First, how learning a PV is affected when only one of the component words is retrieved while the other is simply studied. Second, how the learning of one item is affected by the simultaneous learning of other items. Finally, whether a retrieval-based treatment enhances PV knowledge better than a generation-based one.

Similar to the previous chapter, the current chapter has six main sections. Section 5.2 presents the research questions of Study 2 while Section 5.3 presents its methodology. Section 5.4 and 5.5 describe the learning conditions and the procedure, respectively. Section 5.6 then explains how the test responses were scored. Section 5.7 describes the results of the analysis. The main findings of the study are summarized in Section 5.8. Section 5.9 concludes the chapter with the rationale for the next study.

5.2 Research questions

Study 2 was guided by the following research questions:

1. Which retrieval condition enhances learning: immediate or delayed?
2. Which generation condition enhances learning: immediate or delayed?
3. Which condition is more effective: immediate retrieval versus immediate generation?
4. Which condition is more effective: delayed retrieval versus delayed generation?
5. Which condition results in greater interference from treatment errors?

5.3 Methodology

This section begins with a description of the participants in this study. It is then followed by a discussion on the target items used in the study. The independent variable was Treatment with four levels: immediate retrieval, delayed retrieval, immediate generation, delayed generation. The dependent variable was Performance on two kinds of post-tests: recall and recognition.

5.3.1 Participants

The study was conducted during a semester in five intact intermediate EFL classrooms in an English program in a university in Japan. Although the five classes were taught by different instructors (one instructor taught three, another taught two), they followed the same instructional goals and curriculum. The learners were 145 second-year students who agreed to participate in the present study. Their average raw score on Version 2 of the VLT at the second 1,000-word level was 26.2/30 (SD=1.5). This average score shows that participants had receptive knowledge of around 1,700 of the 2,000 most frequent English words and that they should have little trouble understanding the words at this level. The participants were randomly but equally divided into four groups: 29 students in the immediate retrieval group; 29 students in the delayed retrieval group; 29 students in the immediate generation group; 29 students in the delayed generation group; and, 29 students in the norming group. By the time of the study, none of the participants had studied English outside Japan.

5.3.2 Target items

Twenty-eight PVs were selected as target items and divided into two sets of 14 pairs. Prior to the experimental study, 29 students (who did not take part in the experiment reported here) were randomly selected from the sample population to sit a cued-recall-of-form test (the same type of test used to assess learning following the treatment, see below) on the items to gauge the extent to which the students participating in the experiment might possess knowledge of the PVs. None of the students' responses were correct. This finding was taken as an indication that the students in the experiment would also have no productive knowledge of the target items prior to the experiment. The use of this method to norm the target items was preferred over the use of the standard pre-test method for the reason that a pre-test has the potential to produce the pre-testing effect, which is the very phenomenon that we have set out to investigate.

The PVs were comprised of verbs and adverbial particles. Twenty-six of the verbs were monosyllabic and two were disyllabic (*figure, brighten*). According to the BNC-COCA CORE-4-word list, the verbs occurred in different 1,000-level frequency bands: 20 verbs in the first frequency band (*back, brighten, call, catch, chip, figure, get, give, hang, head, hold,* *make, open, own, pass, pop, run, stick, turn*), seven verbs in the second frequency band (*boil, brush, crack, dive, rip, screw wrap*), and one verb in the third frequency band (*nod*). All adverbial particles were monosyllabic except for one, which was disyllabic (*away*). They all occurred in the first 1,000 frequency band (*down, on, in, out, away, up, off*). Sixty-three words used to paraphrase the meanings of the PVs occurred in the first 1,000 frequency band and eight in the second 1,000 frequency band (*asleep, create, feature, improve, popular, secret, skill, unintentionally*).

5.4 Treatment

As mentioned above, the present study investigated four learning conditions: study-test retrieval, study-delay-test retrieval, test-study generation, and test-delay-study (see Appendix VI). These conditions are described below. All conditions were administered on computers using the software program Qualtrics. After completing each set of target items, participants in the treatment groups were given a filler task and a post-test. The design of the conditions is summarized below:

Treatment	*Time interval between trials on a single PV*
Study-test retrieval	Study trial -> 0 minutes -> Test trial
Study-delay-test retrieval	Study trial -> 6.5 minutes -> Test trial
Test-study generation	Test trial -> 0 minutes -> Study trial
Test-delay-study generation	Test trial -> 6.5 minutes -> Study trial

5.4.1 Study-test retrieval condition

The study-test retrieval condition consisted of a 15-second presentation of a PV and its respective meaning (e.g., *hang out – to spend time with friends*) before a 15-second interval to complete a particle cued-recall test (e.g., *hang __ – to spend time with friends*). The instructions of the test asked participants to first think back to the PV in order to recall the missing particle and then to type the particle in a text box below the underlined space.

5.4.2 Study-delay-test retrieval condition

The study-delay-test retrieval condition consisted of a two-minute-and-10-second study presentation of 14 PVs with their corresponding meanings (15 seconds per PV) before a two-minute-and-10-second interval to complete a particle cued-recall test on those PVs. The instructions of the study-delay-test retrieval condition were the same as the instructions of the study-test retrieval condition.

5.4.3 Test-study generation condition

The test-study generation condition consisted of a 15-second interval to complete a particle cued-recall test on a PV before a 15-second presentation of the corrective feedback. The instructions of the exercise asked participants to type the missing particle in the text box below the blank space or to take a guess if they did not know it. Corrective feedback was as follows: If a participant correctly produced the missing particle, a green ✔ was displayed as feedback, whereas, if the response was erroneous, a red ✘ and the correct particle was placed next to it.

5.4.4 Test-delay-study generation condition

The test-delay-study generation condition consisted of a two-minute-and-10-second interval to complete a particle cued-recall test on 14 PVs before a two-minute-and-10-second presentation of the corrective feedback. The type of instructions and the type of corrective feedback of the test-study generation condition was the same type of instructions and type of corrective feedback of the test-delay-study generation condition.

5.5 Procedure

This study consisted of four phases: practice, treatment, distractor, and post-test.

5.5.1 Practice phase

One week before the treatment phase was the practice phase. The purpose of it was to introduce and familiarize participants with the online program that they would be using in the treatment phase. The design of the retrieval and generation methods were the same. However, instead of the target item, non-target items were used, such as verb-noun collocations. The non-target items were selected by the participants' English instructors. The participants were asked to follow the instructions of the treatments. Their English instructors addressed any problems they had with the use of the program. The participants studied the non-target items using the treatments. The reason was that they would be randomly assigned to one of the four treatments in the treatment phase. The next step was for the participants to learn the individual words of the vocabulary that was used in the treatments. These words, along with other words selected by their English instructors, were presented with their L2 meanings. Participants studied the word pairs without any time restriction.

5.5.2 Treatment phase

The treatment was carried out in the participants' classrooms during regular class hours. The EFL teachers sent a link of the treatments to their students' desktop computers. Once clicked on, the link randomly but equally assigned each participant to one of the four treatment groups. The program first asked students to provide bio-data, such as their names and ages and whether they studied English outside Japan. After submitting this information, the treatment began with a set of practice filler items. The next set of items were the target items from either set A or B (The sets were counterbalanced by participant). Participants in the retrieval conditions had 15 seconds to study the PVs in the study trial and 15 seconds to retrieve them in the test trial. Similarly, those in the generation conditions had the same amount of time to complete the test trial and to review the correct answers in the study trial. After each set, the participants completed a 10-minute distractor task followed by a post-test. In other words, a distractor task and a post-test followed set A, and a distractor task and a post-test followed set B.

5.5.3 Distractor phase

For the distractor phase, participants were presented with L1 trivia questions and two digit additions (e.g., 48-19=?). These questions do not require students to use English. Therefore, there is no chance of them accidentally using any of the individual words that were used in the treatments. If they did use any of these words, then this would count as a learning event. The distractor task was long — lasting 10 minutes, which should have been a sufficient amount of time to flush their short-term working memory of residual traces of the PVs.

5.5.4 Post-test phase

Two direct tests were used to measure the effects of the treatment on learning and retention. The first was a cued-recall-of-form test. It consisted of paraphrases of definitions of phrasal verbs accompanied by underlined spaces (e.g., ______ - *to spend time with friends*). Participants were asked to type the phrasal verbs corresponding to the displayed meanings in text boxes below the blanks. No corrective feedback was given. A cued-recall-of-form test was administered after a distractor task that was given after each of the two sets of target items. The participants took the test again after a delay of seven days.

The second direct test measured recognition of form. It presented the phrasal verbs in the left-hand column and the paraphrases of their meanings in the right-hand column. Their order in the columns was randomized. Participants were instructed to drag-and-drop the definitions that corresponded with the phrasal verbs using their computer mouse. The recognition-of-form test was administered only after the delayed cued-recall-of-form test in order to avoid a testing effect, which may occur from retrieving the same target items from memory repeatedly.

5.6 Counting scores

The cued-recall-of-form test was scored in two ways. A full score was awarded for a response in which participants produced the correct verb and the correct particle of a phrasal verb. A half score was awarded for a response in which participants only produced either the verb or the particle of a phrasal verb correctly. I used half score marking to observe which treatment affected the learning of the parts of the phrasal verbs

Additionally, I was interested to investigate the extent to which errors made in the treatment affected the learning and the retention of the phrasal verbs. To measure the effects of error-making, participants' treatment errors were compared to their cued-recall-of-form test errors. Treatment errors duplicated on the post-tests were counted separately from the treatment errors that were not duplicated on the post-tests.

5.7 Results

This section provides descriptive and inferential statistics on the correct scores on the post-tests as well as on the replicated errors on the post-tests. The first set of descriptive statistics is on the treatment groups' full marks on the recall post-tests. The second set of analysis is on the treatment groups' performance on the recognition post-test. The third set is on their half marks on the recall post-tests. Following the descriptive statistics, the inferential statistics are reported. The first set of models were performed on full marks. The second set was carried out on half marks. The last analysis was on replicated errors.

5.7.1 Descriptive statistics on test trial, recall and recognition post-tests

Table 25 shows the total number of correct particle responses on the test trial, the recall post-tests, and the recognition post-test. On the test trial, some participants (*n* = 4) in the generation groups demonstrated prior knowledge of the particles. As for participants in the retrieval condition, I cannot tell if the same held true. Their performance was strikingly different from the other participants in the same group. They were, therefore, regarded as outliers and removed from any subsequent analysis.

Table 25

Total number of correct responses in the test trial, recall and recognition post-tests

Group	Test trial	Immediate Recall	Delayed Recall	Recognition
ST	784	585	107	496
SDT	667	553	66	369
TS	146	257	67	399
TDS	171	301	49	429

Note. The total number of possible responses on the test material and on each recall post-test was 812 for each group. ST stands for study-test, SDT for study-delay-test, TS for test-study, and TDS for test-delay-study.

Table 26 reports the revised figures on the test trial, the recall post-tests, and the recognition post-test. On the test trial, the percentage of correct answers was higher in the study-test condition (97%) than in the study-delay-test condition (82%). Both the generation conditions demonstrated no prior knowledge of these target items.

On the immediate recall post-test, among the four treatment groups, performance was greatest in the study-test condition, 72%, followed by the study-delay-test condition, 68%, which, in turn, was followed by the test-delay-study condition, 31%, and was lowest in the test-study condition, 27%.

There was a significant amount of attrition between the immediate post-test and the delayed post-test. The loss in knowledge for the study-test group (72% to 13%) was 59 percentage points, for the study-delay-test group (68% to 8%) was 60 percentage points, for the test-study group (27% to 6%) was 22 percentage points, and for the test-delay-study group (31% to 5%) was 26 percentage points.

Recognition post-test performance was highest in the study-test condition (61%) followed by the test-delay-study condition (49%), which was followed by the test-study condition (49%). Performance on this test was lowest in the study-delay-test condition (45%).

Table 26

Total number of correct responses in the test material, recall and recognition post-tests

Group	Test material	Immediate Recall	Delayed Recall	Recognition
ST	784	585	107	496
SDT	667	553	66	369
TS	0	197	43	354
TDS	0	227	39	387

Note. The total number of possible responses on the test material and on each recall post-test was 812 for the retrieval groups and 728 for the generation groups.

Table 27 shows the total number of half marks on the recall post-tests for each treatment group. With respect to verbs on the immediate post-test, the largest percentage of correct responses was recorded for the study-delay-test group (78%) followed by the study-test group (76%), which, in turn, was followed by the test-delay-study group (57%). The lowest percentage of correct responses was registered for the test-study group (47%). With respect to verbs on the delayed post-test, performance was greatest in the study-test condition (18%), followed by the test-study condition (15%) and then the study-delay-test condition

(13%) and was lowest in the test-delay-study condition (12%). These scores suggest a flooring effect.

With respect to particles on the immediate post-test, the largest percentage of correct responses was recorded for the study-test group (78%), followed by the study-delay-test group (72%) and then the test-delay-study group (48%). The lowest percentage of correct responses was registered for the test-study group (42%). With respect to particles on the delayed post-test, the study-test group produced the largest number of correct responses (21%). The study-delay-test group produced the second largest number of correct responses (19%). The test-study and the test-delay-study groups produced about the same number of correct responses (18%, 18% respectively).

Comparison of verb and particle performance on the immediate post-test showed that all treatment groups, except for the study-test group, recalled a larger number of correct verb responses than correct particle responses. In contrast to this pattern of results, on the delayed post-test, all the treatment groups recalled a larger number of correct particle responses than correct verb responses.

Table 27

Total number of correct scores on the verb and the particle responses on the immediate and delayed recall post-tests

	Verb		Particle	
Groups	Immediate	Delayed	Immediate	Delayed
ST	620	145	636	169
SDT	631	104	588	155
TS	345	106	306	128
TDS	412	85	349	131

Note. The total number of possible responses on the test material and on each recall post-test was 812 for the retrieval groups and 728 for the generation groups.

5.7.2 Mixed effects logistic regression models on the recall post-tests: Full marks

Like Study 1, the present study analyzed the data using mixed effects logistic regression modelling. The first step in the analysis of the data was to use R to check the assumption of normality. Shapiro test was used, and the data were plotted according to Studentized residuals, Hat values and Cook's distance to identify outliers and influential cases. Based on this analysis, the assumption of normality was satisfied.

The first model analyzed the number of correct full responses on the immediate and delayed recall post-tests. Participants' group membership, VLT scores, and class identification were included in the model as fixed effects. For random effects, the participants were nested by treatment, and the target items were nested by participants. To find the best fit model, all the predictors were initially included: treatment, VLT scores, class identification. Following a backwards stepwise approach, the fixed effects were incrementally removed. If the removal of a predictor was not statistically significant and did not improve the fit of the model, it was removed. This approach revealed that participants' VLT scores and class identification did not improve the model.

5.7.2.1 Study-test vs. Study-delay-test

Table 28 shows the best fit repeated measure model for the retrieval groups' performance on the recall post-tests. The odds of obtaining a correct target item in the study-test condition were 1.385 times the odds of obtaining a correct target item in the study-delay-test condition. No interaction was reported between the two retrieval conditions. Because the repeated measures model found that the difference between the retrieval conditions was statistically significant, separate models were run on the individual post-tests (i.e., immediate and delayed recall post-tests).

Table 28

Mixed effects logistic regression model on the retrieval groups' recall post-test scores

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	0.802	0.265	3.021	< 0.01	2.231	1.325, 3.756
ST	0.326	0.132	2.454	0.014	1.385	1.067, 1.798
Within	-3.466	0.164	-21.141	< 0.0001	0.031	0.022, 0.043
Interaction	0.272	0.207	1.311	0.189	-	-

Note. The delayed retrieval group is the intercept. ST means immediate retrieval group.

Table 29 shows the best fit model for the retrieval conditions on the immediate post-test. The odds of the study-test group recalling the correct answer were not statistically different from the odds of the study-delay-test group recalling the correct answer (p = 0.171). The best fit model in Table 30 reports that the difference between the retrieval groups on the delayed post-test was statistically significant (p = 0.049). The odds of the study-test group successfully recalling a target item were 2.442 times the odds of the study-delay-test group.

Table 29

Mixed effects logistic regression model on the retrieval groups' immediate recall post-test scores

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.628	0.170	-3.690	< 0.0001	0.533	0.381, 0.744
ST	0.275	0.201	1.367	0.171	1.316	0.887, 1.952

Note. The intercept is the SDT group.

Table 30

Mixed effects logistic regression model on the retrieval groups' delayed recall post-test scores

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-4.120	0.497	-8.283	< 0.0001	0.016	0.006, 0.043
ST	0.892	0.453	1.968	0.049	2.442	1.003, 5.943

Note. The intercept is the SDT group.

5.7.2.2 Test-study vs Test-delay-study

The next model was run on the data from the generation conditions. Table 31 shows a repeated measures model for the generation conditions. The difference between the test-study group and the test-delay-study group was found to be not statistically significant (p = 0.131). However, the model indicates that there was a sharp loss in knowledge in both the generation conditions (p < 0.0001) over a one week period. Because the model did not reveal any statistically significant difference between the generation conditions, no further models were performed on the individual recall post-tests.

Table 31

Mixed effects logistic regression model on the retrieval groups' recall post-test scores

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.2065	0.3332	0.620	0.535	0.813	0.423, 1.562
TS	0.4256	0.2820	1.509	0.131	0.653	0.375, 1.135
Within	-2.4095	0.1967	12.249	< 0.0001	0.089	0.061, 0.132
Interaction	-0.3834	0.2680	1.431	0.153	1.467	0.867, 2.481

Note. The TDS group is the intercept.

5.7.2.3 Study-test vs. Test-study

The next set of models were carried out to compare the study-test condition with the test-study condition. Table 32 shows that, in the best fit model, the odds of obtaining a correct response in the study-test condition were 7.456 times the odds of obtaining a correct response in the test-study condition. The model also indicates a significant interaction between the two conditions ($p < 0.0001$). Given the significant difference between the treatment groups, further models were run on the individual post-tests.

Table 32

Repeated measures best fit model on the study-test condition and the test-study condition

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.916	0.2685	3.413	0.0006	0.399	0.236, 0.676
ST	2.009	0.1394	14.409	< 0.0001	7.456	5.673, 9.799
Retention	-2.078	0.1928	10.781	< 0.0001	0.125	0.085, 0.182
Interaction	-1.220	0.2360	5.172	< 0.0001	0.294	0.185, 0.468

Note. The test-study condition is the intercept.

Table 33 shows the best fit model on the immediate post-test for the study-test condition and the test-study condition. The odds of participants in the study-test group obtaining a correct target item were 3.987 times the odds of participants in the test-study group producing a correct target item. Table 34 reports that on the delayed post-test, the odds of recalling a target item in the study-test condition were 3.534 times the odds of recalling a target item in the test-study condition.

Table 33

Mixed effects logistic regression model on the study-test condition and test-study condition on the immediate recall post-test

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-1.746	0.217	-8.034	< 0.0001	0.174	0.113, 0.267
ST	1.383	0.230	5.993	< 0.0001	3.987	2.536, 6.268

Note: The test-study condition is the intercept.

Table 34

Mixed effects logistic regression model on the study-test condition and test-study condition on the delayed recall post-test scores

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-4.783	0.629	-7.605	< 0.0001	0.008	0.002, 0.028
ST	1.262	0.517	2.439	0.014	3.534	1.281, 9.750

Note: The test-study condition is the intercept.

5.7.2.4 Study-delay-test vs Test-delay-test

Table 35 shows the results of a repeated measures mixed effects logistic regression model performed on the scores of the study-delay-test group and the test-delay-study group. The odds of the study-delay-test condition resulting in a correct response were 3.770 times the odds of the test-delay-study condition resulting in a correct answer. The statistically significant difference between the two conditions prompted the analysis of the individual recall post-tests.

Table 35

Repeated measures best fit model on the study-delay-test condition and the test-delay-study condition

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.615	0.318	-1.930	0.0536	0.540	0.289, 1.009
SDT	1.327	0.145	9.098	< 0.0001	3.770	2.832, 5.018
Within	-2.304	0.191	-12.009	< 0.0001	0.099	0.068, 0.145
Interaction	-1.050	0.244	-4.297	< 0.0001	0.349	0.216, 0.564

Note. The test-delay-study group is the intercept.

Table 36 reports the best fit model on the immediate post-test. The odds of a participant in the study-delay-test group supplying the correct answer were 2.993 times the odds of a participant in the test-delay-study group supplying the correct answer. Table 37 provides the best fit model on the delayed post-test. The difference in delayed post-test performance between these two groups was found to be not statistically significant (*p* = 0.209).

Table 36

Mixed effects logistic regression model on the study-delay-test condition and the test-delay-study condition on the immediate post-test

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-1.753	0.245	-7.142	< 0.0001	0.173	0.106, 0.280
SDT	1.096	0.295	3.712	< 0.001	2.993	1.677, 5.340

Note: The test-delay-study group is the intercept.

Table 37

Mixed effects logistic regression model on the study-delay-test condition and the test-delay-study condition on the delayed post-test

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-4.5801	0.5927	-7.728	< 0.0001	0.010	0.003, 0.032
SDT	0.7381	0.5874	1.257	0.209	2.091	0.661, 6.615

Note. The test-delay-study group is the intercept.

5.7.3 Mixed effects logistic regression models on the recognition test: Full marks

On the delayed recognition post-test, performance was higher for the study-test group than for the study-delay-test group ($b = -0.868$, $z = -3.227$, $p < 0.01$). The study-test group also outperformed the test-study group ($b = -0.681$, $z = -2.530$, $p < 0.05$). The difference between the study-test group and the test-delay-study group was approaching significance ($b = -0.469$, $z = -1.744$, $p = 0.08$). The odds of the study-test group producing correct responses were 2.38 (95% CI: 1.4, 4.0) times the odds of the study-delay-test group, 1.96 (95% CI: 1.2, 3.3) times the odds of the test-study group and 1.24 (95% CI: 0.7, 2.1) times the odds of the test-delay-study group.

5.7.4 Mixed effects logistic regression models on the recall post-tests: Half marks of each condition

The next set of models were performed on the treatment groups' half marks (verb and particle responses). The first model was run on the study-test group and the second on the study-delay-test group. The third model was performed on the test-study group and the fourth on the test-delay-study group. The fixed and random effects that fitted the best models on full marks are the same fixed and random effects that fitted the models on half marks.

5.7.4 Study-test: Verb vs. particle

Table 38 shows the best fit repeated measures model for the study-test group's performance on verb and particle responses. The difference between these two items was found to be not statistically significant ($p = 0.308$). Consequently, separate analyses on the immediate and delayed recall post-tests were not performed.

Table 38

Study-test group's performance on verb and particle responses: fixed effects repeated measures summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	1.725	0.333	5.174	< 0.0001	-	-
Particle	-0.301	0.296	-1.018	0.308	-	-

Note. The verb is the intercept.

5.7.5 Study-delay-test: Verb vs. particle

Table 39 shows the best fit repeated measures model for the study-delay-test group's performance on verb and particle responses. The odds of a participant in the study-delay-test group successfully recalling a verb were 2.012 times the odds of that individual in the study-delay-test group correctly recalling a particle. On account of the statistically significant difference reported in the repeated measure model, separate models were run on the study-delay-test group's verb and particle performance on the immediate and delayed post-tests.

Table 39

Comparison of the study-delay-test group's verb and particle performance: fixed effects repeated measures summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	1.6797	0.2931	5.730	< 0.0001	2.66522289	1.804, 3.937
Particle	-0.6994	0.2938	-2.380	0.0173	2.01257827	1.131, 3.579

Note. The verb is the intercept.

Table 40 reports the best fit model on the number of correct verb and particle responses given by participants in the study-delay-test group. The difference between these two items was trending toward significance (*p* = 0.058). The odds of successfully recalling a verb were 1.471 times the odds of successfully recalling a particle. Table 41 provides the best fit model on the delayed post-test, showing the difference between the verb and particle responses was not statistically significant (*p* = 0.292).

Table 40

Comparison of the study-delay-test group's verb and particle performance: fixed effects summary on immediate post-test

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	1.6249	0.2641	6.152	< 0.0001	3.450232	1.961, 6.068
Particle	-0.3865	0.2043	-1.892	0.0585	1.471778	0.986, 2.196

Note. The verb is the intercept.

Table 41

Comparison of the study-delay-test group's verb and particle performance: fixed effects summary on delayed post-test

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-2.6155	0.3131	-8.354	< 0.0001	-	-
Particle	0.5656	0.5362	1.055	0.292	-	-

Note: The verb is the intercept.

5.7.6 Test-study: Verb vs. particle

Table 42 reports the best fit repeated measures model for the test-study group's correct verb and particle responses. The difference between successful recall of the verb and the particle was found to be not statistically significant (*p* = 0.784). Further analyses on the individual post-tests were not performed.

Table 42

Test-study condition: verb vs particle: fixed effects repeated measures summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	0.0685	0.34719	-1.998	0.045691	-	-
Particle	0.094	0.34485	0.274	0.784236	-	-

Note. The verb is the intercept.

5.7.7 Test-delay-study: Verb vs particle

The best fit repeated measures model for the test-delay-study group on verb and particle responses is shown in Table 43. Like the test-study group, the difference between the number of verb and particle responses was found to be not statistically significant ($p = 0.065$), although it was showing a weak trend toward significance. Further analyses on the individual post-tests were not run.

Table 43

Test-delay-study condition: verb vs particle: fixed effects repeated measures summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.0685	0.2980	-0.230	0.8182	-	-
Particle	0.6080	0.3301	-1.842	0.0655	-	-

Note. The verb is the intercept.

5.7.5 Mixed effects logistic regression models on the recall post-tests: Half marks comparison between similar conditions

The last set of models were carried out on the verb and particle responses between the treatment groups to observe whether success on an item was affected by the learning condition. The first model was performed on the retrieval groups performance on the verb. The second model was carried out on their performance on the particle. The generation groups success on recalling the verb was analyzed in the third model and on recalling the particle was analyzed in the fourth model. The fixed effects that fitted the first two sets of mixed effects logistic regression models are the same fixed effects that fitted the models on half marks. However, the random effects were different. For random effects, the participants were nested by treatment, and the verb and the particle responses were nested by participant.

5.7.5.1 Verb: Study-test vs. Study-delay-test

Table 44 shows the best fit repeated measures model for the retrieval groups on verb responses. The difference between the groups was found to be not statistically significant (*p* = 0.881). The interaction between the groups was showing a trend toward significance (*p* = 0.052). Further analyses were not conducted.

Table 44

Study-test vs. Study-delay-test on verb performance: fixed effects repeated measures summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	1.49609	0.22151	6.754	< 0.0001	-	-
ST	-0.047	0.31767	-0.149	0.881	-	-
Interaction	0.44966	0.23148	1.943	0.0521	-	-

Note. The study-delay-test group is the intercept.

5.7.5.2 Particle: Study-test vs. Study-delay-test

Table 45 reports the best fit repeated measures model for the retrieval groups' performance on the particle. The table indicates that there was a trend toward significance (*p* = 0.059). The interaction was found to be statistically significant (*p* = 0.034). Because of the lack of significance, no further models were run on the individual post-tests.

Table 45

Study-test vs. Study-delay-test on particle performance: fixed effects repeated measures summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	1.0318	0.1413	7.302	< 0.0001	2.80621578	2.127, 3.701
ST	0.4329	0.2296	1.885	0.0594	1.54168341	0.982, 2.417
Interaction	-0.4054	0.1913	-2.120	0.0340	0.66669518	0.458, 0.969

Note. The study-delay-test group is the intercept.

5.7.5.3 Verb: Test-study vs Test-delay-study

Table 46 shows the best fit repeated measures model for the generation groups' correct verb responses. The difference between the two groups was found to be not statistically significant ($p = 0.204$), while the interaction was found to be significant ($p < 0.001$). Additional models on the individual recall post-tests were not performed.

Table 46

Comparison of generation groups' verb performance: fixed effects repeated measures summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.0935	0.2842	-0.329	0.742158	-	-
TS	-0.4381	0.3451	-1.269	0.204269	-	-
Interaction	0.7931	0.2346	3.380	< 0.001	-	-

Note: The test-delayed-study group is the intercept.

5.7.5.4 Particle: Test-study vs Test-delay-study

The best fit repeated measures model on the generation groups' particle performance is shown in Table 47. The model indicated that the difference between the groups was not statistically significant ($p = 0.932$) nor was the interaction ($p= 0.164$). No additional models were run on the individual recall post-tests.

Table 47

Test-study vs. Test-delay-study on particle performance: fixed effects repeated measures summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.5347	0.25006	-2.139	0.0325	-	-
TS	-0.0244	0.28721	-0.085	0.9321	-	-
Interaction	0.28438	0.20442	1.391	0.1642	-	-

Note. The test-delay-study group is the intercept.

5.7.6.1 Verb: Study-test vs. Test-study

Tables 48-50 below give the figures for analyses of the correct verb responses on the post-tests by participants in the study-test group and the test-study group. Because the best fit repeated measures model indicated that the difference between the groups was statistically significant, follow up models were performed on the post-tests (see Table 48). It can be seen that the significant difference on the immediate post-test (see Table 49) and the delayed post-test (see Table 50) was in favour of the study-test group than the test-study group. The odds of the study-test group obtaining a correct verb response compared to the test-study group was very high on the immediate post-test (10.243) and lower on the delayed post-test (2.698).

Table 48

Study-test vs. Test-study on verb responses: fixed effects repeated measures summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.5316	0.1958	-2.715	0.006	-0.587	0.400, 0.862
ST	1.996	0.266	7.487	< 0.0001	7.361	4.365, 12.414
Interaction	-1.268	0.212	-5.977	< 0.0001	0.281	0.1845, 0.426

Note. The test-study group is the intercept.

Table 49

Study-test vs. Test-study on verb responses: fixed effects immediate post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.556	0.276	-2.016	0.044	0.573	0.333, 0.984
ST	2.326	0.375	6.206	< 0.0001	10.243	4.912, 21.358

Note. The test-study group is the intercept.

Table 50

Study-test vs. Test-study on verb responses: fixed effects delayed post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-2.641	0.376	-7.010	< 0.0001	0.071	0.034, 0.149
ST	0.868	0.321	2.698	0.007	2.383	1.268, 4.477

Note. The test-study group is the intercept.

5.7.6.2 Particle: Study-test vs Test-study

The best fit models on the correct particle responses by participants in the study-test group and in the test-study group are presented in Tables 51-52. Table 51 shows the repeated measures results, which indicates that there was a significant difference between the groups. On the immediate post-test (see Table 52), the study-test group outperformed the test-study group ($p < 0.0001$), but on the delayed post-test (see Table 53), the difference was not statistically significant ($p = 0.111$).

Table 51

Study-test vs. Test-study on particle performance: fixed effects repeated measures summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.559	0.141	-3.958	< 0.0001	0.571	0.433, 0.754
ST	2.023	0.229	8.815	< 0.0001	7.567	4.825, 11.868
Interaction	-1.731	0.199	-8.668	< 0.0001	0.176	0.119, 0.261

Note. The test-study group is the intercept.

Table 52

Study-test vs. Test-study on particle performance: fixed effects immediate post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.635	0.281	-2.257	0.024	0.529	0.305, 0.919
ST	2.381	0.368	6.472	< 0.0001	10.823	5.261, 22.264

Note. The test-study group is the intercept.

Table 53

Study-test vs Test-study on particle performance: fixed effects delayed post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-2.304	0.364	-6.384	< 0.0001	-	-
ST	0.423	0.265	1.595	0.111	-	-

Note. The test-study group is the intercept.

5.7.6.3 Verb: Study-delay-test vs Test-delay-study

Tables 54-56 show the best fit models on the verb responses by the study-delay-test group and the test-delay-study group. The repeated measures best fit model (see Table 54) indicated that the difference between the two groups was statistically significant ($p = 0.002$). On the immediate post-test (see Table 55), the odds producing a correct verb response in the study-delay-test condition were 3.838 times the odds of obtaining a correct verb response in the test-delay-study condition. On the delayed post-test (see Table 56), the odds also favoured the study-delay-test group (2.192) compared to the test-delay-study group.

Table 54

Study-delay-test vs. Test-delay-study on verb performance: fixed effects repeated measures summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-2.606	0.307	-8.469	< 0.0001	0.073	0.040, 0.134
SDT	1.055	0.341	3.091	0.002	2.873	1.471, 5.611
Interaction	0.069	0.215	0.323	0.749	1.072	0.702, 1.637

Note. The test-delay-study group is the intercept.

Table 55

Study-delay-test vs. Test-delay-study on verb performance: fixed effects immediate post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.129	0.336	-0.385	0.699	0.878	0.454, 1.698
SDT	1.345	0.422	3.184	0.001	3.838	1.676, 8.784

Note. The test-delay-study group is the intercept.

Table 56

Study-delay-test vs. Test-delay-study on verb performance: fixed effects delayed post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-2.547	0.300	-8.474	< 0.0001	0.078	0.043, 0.141
SDT	0.785	0.261	3.008	0.002	2.192	1.314, 3.658

Note. The test-delay-study group is the intercept.

5.7.6.4 Particle: Study-delay-test vs Test-delay-study

Tables 57 to 59 show the results of the best fit models on the study-delay-test and the test-delay-study conditions on particle performance. Table 57 reports that the repeated measures model indicated a statistically significant difference between the treatment groups. Table 58 shows the immediate post-test results, which indicates that the odds of recalling the particle by participants in the study-delay-test condition were 6.978 times the odds of recalling the particle by those in the test-delay-study condition. On the delayed post-test (see Table 59), no difference was found between the two treatment groups (*p* =0.136).

Table 57

Study-delayed-test vs Test-delay-study on particle performance: fixed effects repeated measures summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.534	0.250	-2.139	0.032	0.585	0.358, 0.056
SDT	1.566	0.287	5.456	< 0.0001	4.790	2.728, 8.409
Interaction	-1.042	0.196	-5.316	< 0.0001	0.352	0.240, 0.517

Note. The test-delay-study group is the intercept.

Table 58

Study-delay-test vs Test-delay-study on particle performance: fixed effects immediate post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.782	0.353	-2.214	0.026	0.457	0.228, 0.914
SDT	1.942	0.444	4.375	< 0.0001	6.978	2.922, 16.663

Note. The test-delay-study group is the intercept.

Table 59

Study-delay-test vs Test-delay-study particle performance: fixed effects delayed post-test summary

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-2.247	0.318	-7.053	< 0.0001	-	-
SDT	0.358	0.240	1.491	0.136	-	-

Note. The test-delay-study group is the intercept.

5.7.6 Error analysis on recall post-tests

The percentage of replicated errors produced by participants in the treatments on the immediate and delayed recall post-tests (see Table 60) was analyzed using z-tests for one proportion, which compared the proportion of replicated errors with other errors. On the immediate post-test, the z-test showed that the study-delay-test condition was more likely to result in the production of other errors than replicated errors ($z = -4.78$, $p < 0.0001$; 95% CI: 0.29, 0.41). In contrast, the test-study condition was more likely to result in replicated errors than other errors ($z = 7.94$, $p < 0.0001$; 95% CI: 0.67, 0.71). Similarly, the test-delay-study condition was more likely to result in replicated errors than other errors ($z = 10.22$, $p < 0.0001$; 95% CI: 0.67, 0.76). Error analysis was not conducted on the study-test condition because so few errors were produced during learning.

On the delayed recall post-test, the z-tests indicated that the study-test condition and the study-delay-test condition were more likely to result in the production of other errors than replicated errors ($z = -25.12$, $p < 0.0001$; CI 95%: 0.01, 0.04; ($z = -17.79$, $p < 0.0001$; CI 95%: 0.15, 0.20, respectively). In contrast, the test-study condition and the test-delay-study condition were more likely to cause participants to produce replicated errors than other errors ($z = 9.12$, $p < 0.0001$; CI 95%: 0.63, 0.70; $z = 7.42$, $p < 0.0001$; CI 95%: 0.60, 0.67).

Table 60

Total number of replicated errors on the recall post-tests

Learning conditions	Immediate	Delayed
Study-test	0 (0%)	19 (5%)
Study-delay-test	91 (35%)	130 (17%)
Test-study	371 (67%)	497 (67%)
Test-delay-study	371 (73%)	484 (73%)

Note. Immediate indicates the number (and proportion) of errors repeated on the immediate post-test that were the same errors on the test trial. Delayed indicates the number (and proportion) of errors repeated on the delayed post-test that were the same errors on the test trial.

5.8 Summary of results

Study 2 sought to examine and compare the effects of immediate and delayed testing in a retrieval condition and immediate and delayed feedback in a generation condition. In the study-test condition, the two trials were consecutive, while in the study-delay-test condition, they were separated by other items. In the test-study condition, the two trials were presented back-to-back, while in the test-delay-study condition, they were separated by other items. After the treatments, a filler task and a recall post-test were administered. One week later, the same recall post-test was given along with a recognition post-test. The results of the norming study strongly suggested that the participants in all four treatment groups were unlikely to have productive knowledge of the target items prior to the treatment. However, some participants in the generation conditions did demonstrate knowledge, so these participants were removed.

The first set of analyses compared post-test performance between the retrieval groups and between the generation groups. The first analysis showed that the study-test condition resulted in better learning than the study-delay-test condition. The generation groups did not differ in their performance, according to the second analysis. The third analysis compared the number of correct responses produced by the study-test group and the test-study group. It indicated that learning was better for participants in the study-test group than for those in the test-study group on the immediate and delayed post-test. The fourth analysis showed that participants in the study-delay-test condition outperformed those in the test-delay-study condition on the immediate post-test but not on the delayed post-test.

The second set of analyses compared the treatment groups' recognition memory of the target item. It was found that recognition memory was greater for participants in the Study-test group than for those in the other treatment groups (with the difference approaching significance with the test-delay-study group).

The third set of analyses compared each groups' performance on verb and particle responses. The first and second analyses showed that for both retrieval groups, there was no difference in the percentage of correctly recalled verbs and particles, respectively. Similarly, the third and fourth analyses indicated that the difference in the percentage of correctly recalled verbs on the immediate and delayed post-tests were not statistically significant from the percentage of correctly recall particles on these post-tests between the generation groups, respectively.

The fourth set of analyses compared performance on the component words by the retrieval conditions with generation conditions. The first analysis indicated that a greater percentage of verb responses were recalled by participants in the study-test group than those in the test-study group. The second analysis showed that the study-test group recalled a greater percentage of particle responses on the immediate post-test but not on the delayed post-test relative to the test-study group. The third analysis showed that the study-delay-test group recalled a greater percentage of verb responses on the immediate and delayed post-test relative to the test-delay-study group. The fourth analysis indicated that performance on particle responses was best for the study-delay-test group on the immediate post-test but not on the delayed post-test.

Like Study 1, the generation conditions both resulted in a great deal of proactive interference. Because the study-test group and the study-delay-test group produced some errors during the trial test, they were compared to post-test errors. The analysis found that greater retroactive interference was caused by the study-delay-test condition than the study-test condition.

5.9 Rationale for the next study

So far, studies 1 and 2 have shown that PVs are learned better under conditions which induce retrieval rather than generation. The results of these studies may be due to the fact that the conditions were decontextualized in the sense that the PVs were not embedded in any context in the study trial and the test trial. As a result, students might have found it difficult to recall a PV (or one of its component words) based on the limited information provided in the retrieval on the test trial and on the final test. Perhaps by including context in the study trial, a student might find it easier to learn the PVs. Including context in the test trials and the final test might also help the student to remember the missing verb phrases. Another reason to investigate contextualized retrieval and generation conditions on the learning of PVs is that the textbook analysis indicated that many phrase-learning methods feature context. The aim of study 3 was to investigate this issue.

Chapter 6 - Study 3

6.1 Introduction

Studies one and two compared and examined the use of retrieval and generation conditions to facilitate PV learning and retention. Both procedures paired PVs with paraphrases of their definitions but did not provide any contextual support in the study trial and elicited recall of the PVs using the explanations of their definitions with or without other additional prompts (i.e., the first letter of the verb in Study 1 and the verb in Study 2). Thus, both the study trials and the test trials were devoid of context in these studies. Although the textbook analysis indicated that many phrase-learning methods are decontextualized, it also showed that many more of them embed phrases in sentential contexts. More specifically, approximately 61% of study trials provide contextual support for phrases in (written or spoken) texts, and around 45% of test trials make use of sentences. As far as I am aware, no study has examined the effectiveness of contextualized retrieval and generation conditions on the learning of PVs. Therefore, the primary aim of the present study is to examine and compare the learning benefits of these two conditions.

There are many ways to arrange a context to support the learning of a phrase. As observed in the textbook analysis, a phrase may be presented in a single sentence or in a longer text, where it may or may not be textually enhanced. The benefit of using an attention-drawing device (e.g., bold-typeface, underline, highlight, gloss) is to increase the likelihood of students noticing the target phrase (Kim, 2006). But its use is uncommon in the input materials of textbooks. This is most likely because the texts in which the phrases occur are often the main text of the unit that is used for learning all sorts of other language related issues (e.g., text comprehension, morphology, syntax).

While phrases may go unnoticed in the study trials of general textbooks, students are obviously aware of them in specialized ones as their titles clearly indicate the vocabulary focus: *Phrasal Verbs in Use* (McCarthy & O'Dell, 2004), *1000 Phrasal Verbs in Context* (Errey, 2007), *Idioms and Phrasal Verbs* (Gairns & Redman, 2011), and *Work on your Phrasal Verbs* (Flockhart & Pelteret, 2012). A PV is often presented with an explanation of its meaning and followed by an example sentence to illustrate its use. For example, Flockhart & Pelteret (2012) present a PV in large bold-typeface and provide a paraphrase of its meaning below it. On the next line of the page, the PV occurs in a sentence. McCarthy and O'Dell (2004) and Gairns and Redman (2011) often use a grid and place a PV in the left-hand column, its definition in the center, and a sentence illustrating its use in the right-hand column. The sentential context is pedagogical rather than natural (e.g., Beck, McKeown, & McCaslin, 1983). Pedagogical context, or what Schouten-van Parreren (1989) referred to as a "pregnant context," promotes vocabulary acquisition from context. It seems the idea is that the richer the context that surrounds a PV in a text, the easier it is for the student to infer what it means from the context and to remember it later. The same idea seems to be behind the use of contextualized test trials (i.e., exercises). The context helps a student to retrieve a PV encountered in a similar context presented in the study trial.

The present study adopted McCarthy and O'Dell's (2004) study trial format rather than Flockhart and Pelteret's (2012). Although both materials designers essentially provide the same information, McCarthy and O'Dell use a layout that is clearer and easier to follow. The test trial format is a common gap-fill sentence. It does not only occur in McCarthy and O'Dell's and Flockhart and Pelteret's books but also in every textbook analyzed in this thesis.

6.2 Research questions

Study 3 was guided by the following research questions:

1. Which learning condition enhances knowledge of PVs: contextualized retrieval vs. contextualized generation?
2. Which learning condition causes greater interference: contextualized retrieval vs. contextualized generation?

6.3 Methodology

This section begins with a description of the students selected as participants and the PVs selected as target items. It is then followed by a description of the learning conditions examined in the present study. Next, the procedure of the study is explained before there is a discussion on the type of post-test that was used to measure learning.

6.3.1 Participants

One-hundred-and-seventy-three third year university EFL students in Japan agreed to participate in the present study. However, the data of three of them were disregarded because of either truancy on the delayed post-test or technical difficulties during the collection of their data. Consequently, 170 participants remained. Their raw score on Version 2 of the VLT at the second-1,000-word level was 28/30 (SD = 2.2), indicating that they had receptive knowledge of around 1,800 of the 2,000 most frequent English words and that they should have little trouble in understanding the words used at this level. One-hundred-and-forty participants were assigned to either the retrieval group or the generation group; 30 were allocated to the norming group.

6.3.2 Target items

Fourteen PVs were selected as target items. They were divided into two sets of seven (Set A and Set B). These target items were selected from the norming study that was carried out in

Study 1. For each target item, three dialogues were created by the experimenter and a second native speaker. These dialogues were very similar to one another. The reason was to ensure that the meaning of a PV did not differ between the three dialogues. For example, the following dialogues were created for the PV *hang out: "Hey, Yuki. If you're not busy after work, do you want to hang out?" "Mike, I have some free time after 5.30 pm, do you want to hang ___?" "I think I'll be free later this evening, shall we hang___?"* These dialogues were used in the study trial, the test trial and the final test, respectively. In addition to the dialogue, the study trial presented an explanation of the PV's meaning: *to spend time with friends.* This meaning was not provided with the dialogues in the test trial and the final test. The frequency of the vocabulary used in the dialogues was analyzed using Vocabprofile (www.lextutor.ca). The program indicated that over 96% of the individual words occurred below the third 1,000 frequency band of the BNC-COCA-25-word list. The other words were proper nouns. The EFL teachers of the participants in this study stated that these words would most likely be familiar to the students (see Appendix VII).

6.4 Learning conditions

This section presents the design of the two learning conditions that were examined in experimental study 3 (see Appendix VIII).

6.4.1 Contextualized retrieval

The retrieval condition reflects closely with the design of some contextualized phrase-learning procedures in textbooks. The study trial is modelled from the exemplary input material used by McCarthy and O'Dell (2004). Using a three by seven table, the left-hand column presented the forms of the PVs, the centre column presented the explanations of their meanings, the right-hand column presented the written dialogues between two speakers to illustrate the uses of the PVs. The instruction of the study trial asked participants to learn the

PVs. The test trial presented dialogues similar to the ones used in the study trial. In the dialogues, the particles of the PVs were replaced by underlined spaces. Below the gapped dialogues were text boxes. The instruction of the test trial informed participants to recall the particles from the PVs presented in the study trial and to type their responses in the text boxes.

6.4.2 Contextualized generation

The generation condition differed from the retrieval condition only in terms of the presentation order of the trials. More specifically, the test trial preceded the study trial. The instruction of the test trial asked participants to guess the missing particles in the underlined spaces. They were also informed that they were going to be presented with the study trial afterward. The study trial displayed the correct PVs in dialogues similar to the ones used in the test trial. Since the test trial was no longer visible, participants had to mentally compare their test trial responses with the ones displayed in the study trial.

6.5 Procedure

So far, the Methodology section discussed the participants, target items and treatments. This section (Section 6.5) describes the administration of the treatments and the data collection procedure. The study had four phases: practice, treatment, distractor, and post-test. The data collection was approved by the Human Ethics Committee of Victoria University of Wellington, New Zealand.

6.5.1 Practice phase

Like Study 2, a practice phase preceded the treatment phase by one week. It introduced the participants to the online program used to present the learning conditions in the treatment phase. The objective was for students to be familiar with the program before the treatment phase rather than to learn the vocabulary per se. The students' English instructors selected the vocabulary students learned in the practice phase, none of which were PVs. The teachers also answered any problems that students might have with the program during this time. After practicing with the software, the students studied the individual words of the vocabulary in the treatment phase using electronic flashcards which paired L1 words with L2 translations. During this time, a random sample of students sat the norming test. It involved filling in gapped dialogues with missing particles. The norming test was exactly the same as the treatment post-test. After students studied the vocabulary and completed the norming test, the teachers began with their daily lesson plans.

6.5.2 Treatment phase

The treatment phase occurred one week after the practice phase. At the start of class, students received a link on their computers. Clicking on it randomly but equally assigned them to either the retrieval or the generation condition. The teachers informed their students to read the instructions and to ask any questions before they began. They also told the students that they were going to sit a final test after the treatments, so they should attempt to learn the PVs. For participants in the retrieval condition, they received the study trial first. After they felt confident that they could recall the PVs, they clicked on the next button on their screen which replaced the study trial with the test trial. For participants in the generation condition, they were given the test trial first. Once they finished typing their answers in the text boxes, they clicked on the next button on their screen which then displayed the study trial. Since the participants could not review their test trial responses, they were asked to make mental corrections of their mistakes with the correct PVs presented in the study trial. After participants in both treatments finished each set of PVs, the software program presented a ten-minute distractor task followed by a fill-in-the-blanks test. The items in each set were randomized, and the sets were counterbalanced across participants.

6.5.3 Distractor phase

The distractor phase was presented after Set A and after B. It asked participants to answer L1 trivia questions and two-digit additions (e.g., 85-9?) for 10 minutes. The purpose of this phase was to flush residual traces of the PVs from short-term memory.

6.5.4 Post-test phase

A sentential particle gap-fill test was used to measure learning and retention. It was similar in design to the test trial used in the treatment conditions. In sentences, the particles of PVs were replaced by underlined spaces. Below the spaces were text boxes, and students were asked to fill them in with the missing particles. The post-test was administered immediately after each set of PVs and again after a delay of seven days.

6.6 Counting scores

A full point was awarded when the particle response was correctly produced. No half point was given for spelling mistakes because an error on one particle might form another particle (e.g., on vs. in). In addition to counting correct responses, replicated errors were counted. This procedure involved identifying whether an error on a post-test was the same error produced on the test trial in the treatment. If it was found, it was labelled as a replicated error. More specifically, if an immediate post-test error was found to be the same error in the test trial, it was labelled as a replicated error. Similarly, if a delayed post-test error was found to be the same error in the test trial, it was called a replicated error.

6.7 Results

This section provides descriptive and inferential statistics on the correct scores on the post-tests as well as on the replicated errors on the post-tests. The first set of descriptive statistics is of the treatment groups' full marks on the recall post-tests. The next set is on replicated errors on these tests. Following the descriptive statistics, the inferential statistics are reported. The first set of models were performed on full marks. The last analysis was on replicated errors.

6.7.1 Descriptive statistics on test trial scores and correct recall post-tests scores and replicated errors

Table 61 shows the total number of correct responses on the test trial and the recall post-tests. On the test trial, participants in the generation group correctly produced 13% of the answers. This score indicates that they either had prior knowledge of the target items or were lucky in guessing. The target items correctly produced were removed from the analysis in order to observe the effect of errorful generation on learning.

Table 61

Total number of correctly recalled responses on the test trial and on the recall post-tests

Group	Test trial	Immediate	Delayed
Retrieval	931	699	539
Generation	127	602	405

Note. Scores are out of 980.

Table 62 presents the total number of correct responses on ten target items on the test trial and the recall post-tests. On the test trial, 98% of responses were answered correctly by participants in the retrieval group while no responses were produced correctly by those in the generation group. On the immediate post-test, performance was greater for the retrieval group, 66%, than for the generation group, 57%. There was a significant amount of attrition over a one-week period for both treatment groups. The loss in knowledge for the retrieval group (66% to 50%) was approximately 16 percentage points and for the generation group (57% to 34%) was around 23 percentage points.

Table 62

Total number of correctly recalled responses on the test trial and on the recall post-tests

Group	Test trial	Immediate	Delayed
Retrieval	686	465	353
Generation	0	402	240

Note. Scores are out of 700

Table 63 provides the number of replicated errors on the immediate and delayed post-tests produced by participants in the generation group. Because the retrieval group produced so few errors during the test trial, no error analysis was performed on their data. A greater percentage of replicated errors were found on the delayed post-test than on the immediate post-test.

Table 63

Total number of replicated errors on the recall post-tests

Group	Immediate	Delayed
Generation	120 (42%)	213 (46%)

Note. Total number of errors on the immediate post-test was 298 and on the delayed post-test was 460.

6.7.2 Mixed effects logistic regression models on the recall post-tests

Like the previous studies, the present study analyzed the data using mixed effects logistic regression modelling. The first step in the analysis of the data was to use R to check the assumption of normality. Shapiro test was used, and the data were plotted according to

Studentized residuals, Hat values and Cook's distance to identify outliers and influential cases. Based on this analysis, the assumption of normality was satisfied.

The first model analyzed the number of correct responses on the immediate and delayed recall post-tests. Participants' group membership, VLT scores, and class identification were included in the model as fixed effects. For random effects, the participants were nested by treatment, and the target items were nested by participants. To find the best fit model, all the predictors were initially included: treatment, VLT scores, class identification. Following a backward stepwise approach, the fixed effects were incrementally removed. If the removal of a predictor was not statistically significant and did not improve the fit of the model, it was removed. This approach revealed that participants' VLT scores and class identification did not improve the model.

The first step in the analysis was to compare the performance of the treatment groups. Table 64 shows the best fit repeated measures model for the correct responses of the treatment groups. The odds of the retrieval group obtaining a correct response were 1.572 times the odds of the generation group producing a correct response. An interaction between treatment and time (immediate to delayed post-tests) was trending toward significance (p = 0.055).

Table 64

Mixed effects logistic regression repeated measure analysis on the recall post-tests

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	0.4014	0.229	1.750	0.080	1.493	0.952, 2.342
Retrieval	0.4526	0.159	2.838	0.004	1.572	1.150, 2.149
R within	-1.1305	0.120	-9.411	< 0.0001	0.322	0.255, 0.408

Interaction	0.3223	0.168	1.917	0.055	1.380	0.992, 1.919

Note: The generation group is the intercept. *R within* refers to performance across the post-tests for the retrieval group. *Interaction* refers to treatment interaction.

Having established that there was a difference between the treatment groups, the second step was to look for differences between the treatment groups on the individual post-tests. Table 65 shows the best fit model on the immediate post-test. The difference between the treatment groups was statistically significant (p = 0.008). The odds of a participant in the retrieval group producing a correct response were 1.586 times the odds of a participant in the generation group producing a correct response. Table 66 reports the best fit model for the delayed post-test responses. The odds of the retrieval group providing the correct answer were 2.121 times the odds of the generation group providing the right response.

Table 65

Mixed effects logistic regression model on the recall post-test: Summary of fixed effects on immediate post-test

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	0.4018	0.246	1.628	0.103	1.494	0.921, 2.424
Retrieval	0.4614	0.175	2.628	0.008	1.586	1.124, 2.237

Note: The generation group is the intercept.

Table 66

Mixed effects logistic regression model on recall post-test: Summary of fixed effects on delayed post-test

Group	Estimate	SE	z-value	p-value	Odds ratio	Odds ratio 95% CI
Intercept	-0.7052	0.238	-2.957	0.003	0.494	0.309, 0.788
Retrieval	0.7523	0.138	5.423	< 0.0001	2.121	1.616, 2.784

Note: The generation group is the intercept.

6.7.3 Error analysis on recall post-tests

The error analysis was carried out on the generation group's post-test mistakes. On the immediate post-test, 289 incorrect responses were produced, of which 120 errors were the same ones generated in the test trial. In other words, 42% of post-test errors were replicated treatment errors. A z-test for one proportion was run to analyze whether the difference between the percentage of other errors differed statistically from replicated errors on the immediate post-test. The results showed that other errors were more likely to be produced than replicated errors ($z = -3.367$, $p < 0.05$; CI 95%: 0.358, 0.464). On the delayed post-test, participants produced 460 incorrect responses of which 213 were replicated errors. In other words, 46% of post-test mistakes were the same ones produced during the test trial. A z-test showed that participants were as likely to produce replicated errors as they were to produce other errors after a one-week delay ($z = -1.597$, $p > 0.05$; CI 95%: 0.424, 0.519).

6.7.4 The fate of successful generation on learning

Some participants in the test trial had produced the correct answer. It was unclear whether these responses demonstrated prior knowledge of the target items or were, in fact, lucky guesses. The fate of these responses was traced on the immediate and delayed post-tests. Of the 174 correct test trial responses, it was found that ten were subsequently produced correctly on the immediate post-test and five on the delayed post-test. This shows that 94% of the correct test trial responses were erroneously recalled on the immediate post-test. This figure increased to 97% by the time of the delayed post-test. This finding suggests that over 90% of correctly generated responses on the test trials were most likely lucky guesses.

6.8 Summary of results

Study 3 sought to examine and compare the use of contextual support with retrieval and generation conditions to enhance knowledge of PVs. The results of the test trial showed that some participants in the generation group correctly filled in the missing particles, despite the findings of the norming study. It was unclear whether these participants had prior knowledge of the target items or were lucky in guessing. Although the target items were removed from the analysis, I decided to track whether the participants recalled them on the post-test correctly or incorrectly. The tracking showed that more than 95% of the correct test trial responses were incorrectly produced on the post-test, which suggests that they were nearly all lucky guesses.

The first mixed effects logistics regression model compared the performance of the treatment groups on the immediate and delayed post-tests. It showed that learning gains were higher for the retrieval group than for the generation group. These findings further support the findings of studies 1 and 2 that learning conditions that induce error-free retrieval are better than those that induce errorful generation. The error analysis indicated that the generation condition caused considerable proactive interference. Of all test errors on the immediate post-test 40% were replicated test trial errors, and on the delayed post-test, 46% were repeated test trial errors. The error analysis was not conducted on the retrieval group because so few errors were made during learning.

Chapter 7 - Discussion

This chapter provides a detailed analysis of the key research findings presented in Chapter 3, 4, 5, and 6. The first section (Section 7.1) discusses the phrase-learning procedures identified in the textbook analysis. After that, Section 7.2 looks at the results of Study 1, which investigated and compared the effects of a retrieval condition and a generation condition on learning PVs. Section 7.3 then looks at how the separation of the study trial and the test trial in a retrieval condition and a generation condition influenced learning and retention. Next, Section, 7.4 discusses the results of a contextualized retrieval condition and a contextualized generation condition on learning PVs. A general discussion of the findings is presented in Section 7.5. Sections 7.6 and 7.7 looks at implications for teaching and theory building, respectively. The final section (Section 7.8) discusses the limitations of the studies and suggestions for future research.

7.1 Discussion of results: Textbook analysis

This section discusses the results of the analysis that sought to identify common phrase learning techniques in ESL/EFL course textbooks. It determined that nearly all methods have in common at least two learning episodes: a study trial and a test trial. The order of their presentation in the books was either a study trial followed by a test trial or a test trial followed by a study trial. I referred to the presentation of the study trial before the test trial as a retrieval-oriented learning procedure and the study trial after the test trial as a generation-oriented learning procedure. The reason is that in the retrieval method, to complete the test trial, students need to think back to their first encounter with a target item in the study trial. Thus, they are retrieving an episodic memory of the target item. In contrast, in the generation method, to provide an answer on the test trial, students need to take a blind guess since they have never encountered the target item beforehand and only afterward do they receive the correct answer displayed in the study trial (i.e., answer key in the appendix of the book).

Thus, they are not retrieving an episodic memory of the target item as the learning event is non-existent. That is, no memory of the answer is available to access (unless the target items were encountered outside the textbooks). Seventy-two per cent of phrase learning techniques conformed to the generation condition.

The textbook analysis also identified two types of test trials. One involves recognition and the other recall. Recognition tests, like multiple-choice, present a question (cue) paired with the correct answer and some lures. It requires a student to attempt to retrieve from memory the meanings of the lures and the answer to reject the lures as incorrect and accept the answer as correct. There are some unique benefits to such kinds of recognition tests. First, it is possible to learn new information from the test (Marsh & Cantor, 2014). For example, a student might not know the correct answer but have some knowledge of the lures. Using a deduction strategy, the learner could eliminate the options that do not seem as plausible candidates, which leaves one as the potential correct answer. Feedback confirming that this choice was correct could facilitate learning of the target item that was not known before the test. Second, although a student might have learned the correct answer before the test, for one reason or another, it cannot be recalled from memory at that moment in time. Cognitive psychologists refer to this experience as *marginal knowledge* (Berger, Hall, & Bahrick, 1999). Displaying the answer with the test is one method by which to stabilize access to marginal knowledge and enhance learning (Marsh & Cantor, 2014). It cannot be achieved using recall tests.

Despite these benefits, the concern voiced by many researchers is that recognition tests that present lures along with the correct answer expose learners to incorrect responses. That is, multiple-choice tests require students to read and consider plausible wrong information.

Therefore, it is unsurprising that a multiple-choice test has some negative consequences on account of the lures. It increases the likelihood of incorrectly answering later questions with one of the lures. Studies have demonstrated that the probability of a student retrieving a lure of a test is greater than if it had not been available at all (Marsh, Agarwal, & Roediger, 2009; Roediger & Marsh, 2005; Odegard & Koen, 2007). More specifically, students might select the correct answer on the test but then retrieve the test lure on a subsequent test, and they may choose the lure on the test and then later retrieve the same lure (Butler, Marsh, Goode, & Roediger, 2006; Boers et al. 2014).

Since the textbook analysis identified 87% of test trials as recognition-based, it is important to evaluate possible solutions for eliminating the negative testing effect -- where learners pick up false knowledge. One is to make the lures implausible, which reduces the likelihood of students regarding the lures as correct answers. A counterargument is that students should not easily discard the lures as incorrect. Rather, the lures should be plausible. The reason is that students should consider why the correct answer is the right choice and why each lure is wrong (Little, Bjork, Bjork, & Angello, 2012). Therefore, plausible lures increase the retrieval effort involved in selecting the correct answer, whereas the use of implausible lures keeps at a minimum the retrieval effort involved in choosing the right answer. According to the retrieval effort theory, the design of a recognition test trial should increase desirable difficulty rather than reduce it.

Another solution is to reduce the number of lures in a multiple-choice test. When there are fewer choices, learners are less likely to endorse a lure (Butler, Marsh, Goode, & Roediger, 2006). However, increasing the number of lures also increases desirable difficulty, which should boost learning (Marsh & Cantor, 2014). Although there is evidence for and against the use of fewer as opposed to more lures, the decision to increase the number of lures must ultimately be based on whether learners can successfully overcome them. If increasing the number of lures on test results in an increase in the number of errors, then learning suffers. This information may be valuable for teachers who are familiar with the linguistic abilities of their students as they can base their choice on this information. However, how do materials designers make this decision? Although it was not investigated, perhaps the number of lures increases at the different proficiency levels of the textbooks -- fewer lures in the lower levels and more of them in the higher levels. Further research is needed to verify the hypothesis.

The next solution to overcome errors resulting from multiple-choice tests is to provide feedback. The textbook analysis indicated that feedback often follows a multiple-choice test in the form of an answer key located in the appendix of the book. However, not all multiple-choice tests include feedback, which suggests that in such cases, it is the teachers' responsibility to check the accuracy of their students' test responses. The answer key typically presents the correct answer (answer feedback) rather than whether the answer is right or wrong (verification feedback) and does not provide explanations of the answer (elaborative feedback). In multiple-choice tests, it appears that the content of the feedback is not very important (Marsh & Cantor, 2014). For example, although verification feedback provides minimal information, it helps to constrain the possible choices of which option is the correct one (whereas it seems to do little to help students learn the correct answer on a recall test). There is concern over leaving teachers to provide the feedback because it is uncertain whether they always do, which leaves students with no recourse to find the correct answers. Testing without follow-up feedback results in less learning than testing followed by feedback (Butler & Roediger, 2008). Another concern is whether or not teachers provide corrective feedback immediately after the test as they might assign it (quiz/exercise) as homework. The worry is that delaying feedback might entrench errors more deeply into memory than if they were immediately corrected (Baddeley & Wilson, 1994). However, Butler & Roediger (2008) indicated that, although both immediate and delayed feedback were equally effective at reducing the negative testing effect, delaying feedback may actually boost retention of correct responses.

As mentioned above, recall is the second test type. Unlike recognition tests which require a student to choose the answer, recall tests require a student to produce the answer from memory. All of the benefits associated with recognition tests also apply to recall tests. However, controversy surrounds whether taking a blind guess facilitates subsequent learning. One advantage of recall tests over recognition tests is that recall test trials typically enhance retention of a target item better than recognition test trials when learning is subsequently assessed using a recall test (e.g., Potts and Shanks, 2014). Another benefit is that since recall tests involve production, they help students to use the vocabulary better than recognition tests. Despite the advantages of recall tests, the textbook analysis identified that they make up only 13% of all test trials.

The textbook analysis also examined the treatment of phrases in test trials. It indicated that test trials present phrases as holistic units 69% of the time and as individual words 31% of the time. This finding may please educational researchers who have suggested that a piecemeal approach is less beneficial for learning phrases than a holistic approach. Boers et al. (2014) point out some of the problems associated with learning phrases via practicing connecting their parts. With regard to some verb-noun collocations, students might have trouble remembering the correct combination because the verb holds little meaning in itself which makes it seemingly replaceable with other similar verbs. For example, students may replace "make" in "make a mistake" with "do" to erroneously produce "do a mistake." Boers et al. (2014) and Boers, Dang, and Strong (2016) showed that students who worked with phrases did better than those who focused on reassembling phrases from their parts.

The textbook analysis further found test trials either providing contextual retrieval cues or none. Decontextualized test trials do not provide any sentential context to help students recover the missing target item from memory. Therefore, to provide a correct response, the student must have established a stable relationship between the retrieval cue and the target item, besides offering the meaning of a phrase or one of its words to prompt the other. In contrast, contextualized test trials provide priming information to help students recover the target item from memory. Contextualized test trials most likely foster retrieval of the correct answer better than decontextualized test trials. The reason is a retrieval cue with minimal priming information results in an extensive search process in the memory for the target item, and the broader the search process, the less likely for the recovery of the answer.

As mentioned above, textbooks feature more generation learning procedures than retrieval ones. Regardless of the proficiency level of the textbooks, approximately 70% promote phrase-learning via generation. Across the nine titles of course textbooks, the 3-to-1 distribution of generation to retrieval techniques was fairly representative. However, the New English file series stands out from the other titles because over 90% of phrase-learning techniques were generation-oriented. The findings of the studies in this thesis recommend that materials designers should be using retrieval-based, as opposed to generation-based, methods to facilitate PV acquisition.

The last finding to discuss is the use of the terms to refer to the string of words. Across the 52 textbooks, the most common designation was "phrase, " and the least common was "multiword unit." This finding suggests that materials designers seem less concerned with how to call a series of connected words and more interested in drawing attention to the fact that they represent a lexical phenomenon. Further, Boers, Dang, and Strong (2016) point out that "phrase" is most likely a term that is familiar to students. The use of it by materials designers shows an attempt to not confuse learners with the technical jargon used by researchers (e.g., idioms, collocations, formulaic sequences, binomials).

7.2 Discussion of results: Study 1

This section provides a detailed analysis of the key research findings of Study 1 with reference to the research for each of the research questions. Subsection 7.2.1 discusses the participants' performance on the recall final tests. Their performance on the recognition test is discussed in Subsection 7.2.2. After that, the effects of the treatment on learning the individual words are discussed (Subsection 7.2.4). Last, Subsection 7.2.5 looks at interference caused by the treatments.

7.2.1 Research question 1: Which learning procedure enhanced short-term and long-term recall memory of the PVs the best?

The first research question investigated and compared the effects of a retrieval condition and a generation condition on learning PVs. Mixed effects logistic regression models indicated that the retrieval condition resulted in greater learning gains compared to the generation condition on an immediate post-test. This finding demonstrates that short-term learning gains can be obtained better when a target item is studied and then successfully recalled on a test than when a test response is erroneously produced and feedback is given to correct the error. In other words, learning from errors is less of an effective strategy than preventing errors from occurring during learning. Although the retrieval group outperformed the generation group by approximately 13 percentage points, the benefit of the retrieval condition was short-lived as the difference between the treatment groups was not statistically significant on the delayed post-test. This result shows that forgetting occurred more rapidly in the retrieval condition than in the generation condition.

Concerning the results of the immediate post-test, the advantage of the retrieval condition corresponds to the findings of other studies. Squires, Hunkin, and Parkin (1997) found that retrieval, as opposed to generation, had a positive effect on learning novel associations between words by memory-impaired individuals. Similarly, Warmington and Hitch (2014) observed that the presentation of study material immediately before a test had a significant beneficial effect on learning novel names of pictures by children. Likewise, Karpicke & Zaromb (2010) reported that semantically related word pairs were acquired better in a retrieval condition than in a generation condition by undergraduate students.

Concerning the results of the delayed post-test, the significantly high attrition rate observed in Study 1 was also found in the study by Squires, Hunkin, and Parkin. They reported that learning in the retrieval condition was less stable than learning in the generation condition on a delayed post-test. In contrast, Warmington and Hitch indicated that the benefit of retrieval over generation persisted sometime after the immediate post-test. However, they measured long-term retention using a three-day post-test, while Study 1 used a one-week post-test. Thus, more forgetting is likely to occur on a one-week post-test than on a three-day post-test.

From a theoretical point of view, the advantage of the retrieval condition can be explained by the testing effect, which states that information retrieved from memory enhances its representation. Further, the presentation of the test trial immediately after the study trial significantly reduced the likelihood of students failing to recall the correct answer for the reason that immediate testing greatly minimized their attempts to consider alternative responses that have the potential to interfere with the ones they just studied (Baddeley & Wilson, 1994). Consequently, successful retrieval of a studied target item enhanced knowledge of that target item while unsuccessful retrieval of a studied target item did not. Therefore, according to this finding, it is important that retrieval be successful rather than unsuccessful to foster the formation of accurate memories.

On the delayed post-test, the sharp loss in knowledge in the retrieval condition can be explained by the retrieval effort theory, elaborative retrieval theory and the episodic context theory. All three theoretical models assume that immediate retrieval resembles too closely to rote rehearsal. Without a substantial lag between the study trial and the test trial, the retrieval effort theory proposes that the memory trace of the target item does not get strengthened because retrieval is too easy, while the episodic context theory contends that the representation of the target item does not get updated with new temporal-spatial contexts, which is key to consolidating new memories.

According to the elaborative retrieval theory, failure to subsequently recall a PV may be due to whether or not the explanation of its meaning provided a fertile context in which to encode it into memory (e.g., Carpenter, 2012). It seems more likely that pairing a verb phrase with an L1 explanation can establish a relationship with other related words and concepts more strongly and elaborately than can be achieved with the use of L2 paraphrases.

Latsanyphone and Bouangeune (2009) and Joyce (2015) showed that L2 students learned better when they studied L2 words with L1 translations than with L2 definitions. Thus, the use of paraphrases rather than L1 explanations may account for the poor performance of both treatment groups on the delayed post-test.

Looking more closely at the effects of the generation condition, participants in the generation group performed poorly most likely due to the same variables that affected the poor performance of those in the retrieval condition. However, the generation condition was less effective than the retrieval condition probably because students in the generation group failed to produce a single correct response in the test trial, even though they later received corrective feedback. Although past research has often found that test failures enhance subsequent encoding of the correct answer, this benefit seems only to occur when the error is semantically related to the cue and the target item. When there is a relationship between the error, the cue, and the target item, subsequent recall of the error may act as a mediator between the cue and the target item. In other words, a student recalls the mistake, which then primes retrieval of the correct answer. According to the elaborative retrieval theory, when there is no semantic relationship between the error with the cue and the target item, the error does not act as a mediator but rather interferes with subsequent encoding of the correct answer presented as feedback. In some circumstances, students might be more inclined to remember their incorrect answers over the correct ones. I analyzed the participants in the generation group's failed test trial responses, and none of them seemed related to the target item. For example, some participants in the generation group, when presented with the test trial *f___ - to understand something* produced f*ire off* and *fit up* as well as *file up* and *follow up*. These responses are semantically unrelated to the correct PV *figure out*. Thus, in many cases, the errors did not seem to function as mediators in helping to recall the PVs.

In contrast to the predictions of the elaborative retrieval theory, a study by Potts and Shanks (2014) demonstrated that even errorful generations (blind guessing), whereby there is little chance of a student producing a correct response related to the target item, enhanced learning compared to a restudy condition. The authors suggested that this finding might have occurred on account of students processing the correct answer in the feedback more effortfully than guessing a response in the test trial. However, the benefit of generation was relative to restudying, which is typically viewed as a less effective learning strategy. If Study 1 included a restudy condition, a benefit of a failed generation might have been found -- this is purely speculation, of course. It would, therefore, be a good idea for future research to compare not only retrieval and generation conditions but also include restudy conditions, for comparative purposes, on the learning of PVs.

7.2.2 Research question 2: Which learning procedure enhanced recognition memory of PVs the best?

Previous studies have demonstrated that testing can affect learning. The testing effect has been the primary reason researchers have measured retrieval conditions using typically a single final test rather than many different types of criterial tests. Warmington and Hitch (2014) administered a criterial recognition test only as a delayed post-test instead of as an immediate post-test. According to the researchers, they did this because “it would have provided participants with an additional learning opportunity, thus making it difficult to disentangle the effects of this additional presentation from that of the learning method on delayed performance” (p. 585). For the same reason, Study 1 (as well as Study 2) did not administer a recognition test as an immediate post-test but did use it as a delayed post-test.

The second research question examined whether the treatments would differentially affect recognition memory of the target items after a one-week period following the interventions. Mixed effects logistic regression analysis was carried out on participants' correct recognition test responses. The results showed a two percent advantage for the retrieval group over the generation group. However, the difference was found to be not statistically significant. This finding is similar to the one reported by Warmington and Hitch. They did not observe any difference in performance between participants in a retrieval (errorless) group and those in a generation (errorful) group. They explained that on a recognition test, students could demonstrate knowledge of a target item on the basis of partial information. In Study 1, to complete the recognition test, students had to match the PVs with their respective meanings. Doing so does not seem to require the representations of the PVs to be well-specified in memory. Knowledge of the verb or even the particle might be sufficient to stimulate the memory of the PV it belongs to. In stark contrast, a well-specified memory trace is necessary to retrieve the PV correctly on a recall test.

7.2.3 Research question 3: Which learning procedure enhanced knowledge of the individual words of the PVs the best?

Past research has shown that not all information gets encoded equally into memory (Tulving, 1972). It is, therefore, expected that acquisition of the individual words of PVs might occur at different rates. That is, it is possible for the verb to be acquired quicker than the particle or vice versa. Several factors may play a role in learning the component words of phrases (Boers et al., 2014; Nesselhauf, 2003). One key factor is whether the semantic weight of the individual words is equal or unequal. Verbs are widely assumed to carry more semantic weight than particles, which offer orientational meanings. Thus, students might associate the PVs with their verbs more than their particles. In this study, participants' responses to the individual words of the PVs were scored. The mixed effects logistic regression models showed a greater number of correct verb responses than correct particle responses produced by participants in both treatment groups on the immediate post-test. This finding, therefore, suggests that the L2 students associated PVs with their verbs more than their particles.

There may be additional factors contributing to this finding. One is the order of the component words. Because verbs come before particles, the likelihood of forming connections between the verbs and the meanings of the PVs might be greater than forming links between the particles and the PVs' meanings. An alternative account is that the format of the recall post-test facilitated retrieval of the verb more than the particle. The reason is the recall post-test displayed the initial letter of the verb but not the particle. The first letter most likely constrained the search for the verb in memory while the absence of the initial letter of the particle did not limit the search process. In other words, due to the format of the recall post-test, students might have found it easier to remember the verb compared to the particle.

As for the delayed post-test, performance was greater on the verbs than the particles, but the difference was not statistically significant according to a mixed effects logistic regression analysis on correct individual responses. This finding shows that the benefit of the initial letter of the verb is less beneficial as the target item fades from memory. However, the fact that the initial letter of the verb might have influenced performance on the immediate post-test prevents us from proposing that the verbs of PVs may be easier to recall than their particles. Limitations of this study and suggestions for future research are discussed further below.

Mixed effects logistic regression models were also run to compare the treatment groups’ performance on the individual words. Concerning the verb, the retrieval group outperformed the generation group on both the immediate and delayed post-tests. Concerning the particle, the correct recall was greater for the retrieval group than for the generation group on the immediate post-test but not on the delayed post-test. The interesting finding here is with respect to the nonsignificant difference found between the treatment groups on the particle on the delayed post-test. The fact that the retrieval condition was effective in the other instances suggests that particles may receive less of a boost from retrieval than verbs do when it comes to long-term retention.

7.2.4 Research question 4: Which learning procedure caused greater interference on short-term and long-term memory of PVs?

Some researchers, particularly in the field of errorless learning, have voiced concerns regarding the extent to which one can learn from self-generated errors. They have argued that errors may interfere with the subsequent learning of the correct answer. For retrieval conditions, this implies that a failed retrieval may interfere with access to the memory of the correct response studied before the test. For generation conditions, it means that the errorful generation may interfere with encoding the correct response presented in the corrective feedback.

Error analysis was carried out on the treatment groups' post-test error responses in relation to their test trial error responses. Because the retrieval group produced so few errors during the test trial in the treatment, error analysis was not carried out on their post-test responses. However, because the generation group only produced incorrect responses during the test trial, test trial errors were compared to post-test mistakes. The error analysis indicated that replicated errors accounted for 24% of all errors on the immediate post-test and 11% of all errors on the delayed post-test. This finding shows that even after a one-week period, students continued to regard some of their original test trial errors as the right answers even though they studied the correct responses after having generated the errors. This finding is rather surprising because it means that students formed a false association in one sense between the incorrect verb and the meaning of the PV and in the other sense between the incorrect particle and that meaning of the PV. Further, they created a false association between the wrong verb and the particle with the meaning of the PV. Based on the results of this error analysis, errorful generation seems to disrupt subsequent encoding of the correct answer and even prevents access to it on a later criterial test.

Error analysis of the individual words showed a much greater percentage of replicated verb errors, 72%, compared to replicated particle errors, 21% on the immediate post-test. On the delayed post-test, while the percentage of replicated verb errors decreased, 55%, it increased for replicated particle errors, 45%. One explanation for this finding is that a very small set of particles make up PVs (Darwin & Grey, 2007). Therefore, the likelihood of taking a random guess on the delayed post-test and having that guess be the same one generated on the immediate post-test is much greater than if it was for verbs.

Replicated verb errors occurred more often than replicated particle errors. Given the fact that verbs precede particles in PVs, students might have spent more time and cognitive effort in thinking of a verb response on the test trial than in contemplating on a particle response. As a result, the association between the verb error and the cue was likely stronger than that between the particle error and the cue. Moreover, the link between the verb and the particle was weaker than the connection between the verb and the cue. Therefore, based on these findings, it seems like a good idea for students to avoid making errors during learning because once they produce them, the errors can be difficult to unlearn.

An in-depth by-item error analysis was also performed across all responses participants produced on the test trial and the post-tests. This error analysis considered all incorrect post-test responses in which one of the individual errors was a replication of the error generated on the test trial. The by-item error analysis indicated that on both the immediate and delayed post-test, the most common outcome (of all 33 outcomes) was XX,XX. This means that on the test trial, a member of the generation group incorrectly generated a verb response and a particle response. On the post-test, this individual replicated the verb error and the particle error. This finding suggests that for some students, they were more likely to recall on the post-test, the same verb and particle error together rather than only remember one of them. Thus, the corrective feedback did not have any influence to undo the adverse effect of errorful generation on the test trial. Further, the by-item error analysis suggests that these errors may have become fossilized in the learners' mental lexicon.

7.3 Discussion of results: Study 2

This section provides a detailed analysis of the key research findings of Study 2. The results of the study are also discussed in relation to previous research studies. Subsection 7.3.1 discusses the effect of the lag between the study trial and the test trial of the retrieval condition. Subsection 7.3.2 looks at the effect of the time interval between the test trial and study trial of the generation condition. Next, the treatment which enhanced PV retention the best is discussed (Subsection 7.3.3) After that, the discussion is on comparing the effects of the delayed retrieval and then delayed generation conditions. Subsection 7.35 looks at the effects of treatments on performance on the individual words. The last subsection discusses which treatment caused greater amounts of interference.

7.3.1 Research question 1: Does the lag between a study trial followed by a test trial improve PV learning?

The first research question investigated whether the timing between a study trial and the test trial would affect the benefit of retrieval. There is evidence to suggest that the delaying of a test trial after a study trial facilitates learning. However, the evidence from these studies is typically confounded by the fact that the retrieval conditions consisted of multiple test trials rather than a single one.

Mixed effects logistic regression models indicated that recall post-test scores were greater in the study-test condition than in the study-delay-test condition. This finding suggests that the benefit of taking a test right after study of a target item is that it facilitates short-term and long-term knowledge acquisition of PVs compared to waiting for approximately 6.5 minutes to take a test after study of a target item. More specifically, even though immediate testing has been claimed to foster rote rehearsal, the situation may not be so straightforward when the target items are extremely complex to acquire. Therefore, this finding challenges the view that separation of the two learning episodes has a greater positive effect than presenting closely together.

7.3.2 Research question 2: Does the lag between a test trial followed by a study trial improve PV learning?

The results of the mixed effects logistic regression models conducted on the Test-study condition and the Test-delay-condition are strikingly different from the comparison between the Study-test condition with the Study-delay-test condition. Recall of the target items did not differ between participants in the Test-study condition and those in the Test-delay-study condition. This suggests that delayed feedback had no positive effect on enhancing the learning of PVs.

The absence of a significant benefit in the separation of a test trial from a study trial in the generation condition has been observed in past research (Hays, Kornell, & Bjork, 2013; Grimaldi & Karpicke, 2012; Kornell, 2014). It might be argued that 13 other to-be-learned items or the approximately 6.5 minutes that separated the study trial from the test trial may have been too long for participants to recall the target item on the test trial. This argument might also be levelled against Hays, Kornell, and Bjork (2013) who separated the trials by around 9.5 minutes and Grimaldi and Karpicke (2012) who inserted 29 filler items between the two trials. However, Kornell (2014) only separated the study trial from the test trial by 10 filler items or approximately 4 minutes, which brings the two trials closer together than the present study and the studies by Hays, Kornell, and Bjork (2012) and Grimaldi and Karpicke (2013). Despite bringing the two learning events closer together, Kornell observed that immediate testing after a failed test response resulted in greater learning gains than delayed testing after a failed test response. While the present study did not find an advantage for delayed testing like the previous studies, it did not find that it was less effective than immediate testing.

Although the delayed corrective feedback was suspected to lead to greater learning gains compared to immediate corrective feedback, this did not occur. This finding may be accounted for by the elaborative retrieval theory (Grimaldi & Karpicke, 2012), which suggests that when a student is asked to produce a response but cannot retrieve the correct answer, the semantic network associated with the retrieval cue (in this case, the meaning of the PV and the words that comprise it as well as the words used in the paraphrase of its definition) is primed and this priming enhances learning during immediate feedback.

However, this priming dissipates over time and is no longer active when delayed feedback occurs (Kornell, 2014).

7.3.3 Research question 3: Which condition enhanced PV knowledge the best?

Although the Study-test condition and the Test-study condition in Study 2 very closely resemble the retrieval condition and the generation condition examined in Study 1, there were notable differences with respect to the format of the test trial and the final test, so these two conditions were compared in Study 2. Based on the transfer-appropriate processing theory (Morris, Bransford, & Franks, 1977), it was assumed that the similarity between the test trial and the final test in Study 1, both of which presented the paraphrase of a PV's meaning along with its initial letter, would make it easier to recall the target item compared to the differences between the test trial and the final test in Study 2. In this study, the paraphrase of a PV and its verb were presented in the test trial while the paraphrase of the PV was presented only in the final test. It was assumed that because the verb was processed on the final test differently to the way it was learned in the trial test, performance on the verb would be lower than performance on the particle, which was processed on the final test in the same way it was learned in the trial test. Despite these design differences between studies 1 and 2, the hypothesis remained the same: the study-test condition would enhance learning better than the test-study condition.

Similar to the results of Study 1, a mixed effects logistics regression model indicated that final recall was greater in the Study-test condition than in the Test-study condition on both the immediate post-test and the delayed post-test. This finding is further evidence that generating an error on a test, even though it is followed by corrective feedback, is less beneficial in learning PVs. However, a significant interaction was reported in the model, which suggests that attrition was greater in the Study-test condition than in the Test-study condition. This finding suggests that error-free retrieval has a positive effect on short-term learning but contributes less to long-term retention, as found in Study 1.

7.3.4 Research question 4: Which condition is more effective: delayed retrieval versus delayed generation?

Mixed effects logistic regression modelling was also performed to compare the effects of the Study-delay-test condition with the Test-delay-study condition. The learning gains were greater in the Study-delay-test condition than in the Test-delay-study condition on the immediate post-test, but no difference was found between the two conditions on the delayed post-test. The advantage of the Study-delay-test condition is most likely related to the Study-test condition, in that few errors in the test trial had a positive effect on performance. However, the advantage dissipated over a one-week period perhaps due to the same factor that caused a high attrition rate in the Study-test condition but also due to the delaying of the test trial after a study trial, which was found to be less effective than administering a test trial right after a study trial.

7.3.5 Research question 5: Which component word benefits most from the conditions?

Study 2 was also interested in the effects of the treatments on learning the individual words of the PVs. I assumed that performance on the individual words would differ. The reason is that the design of the treatments required students to study the verb and the particle in the study trial and to study the verb and recall the particle in the test trial. In other words, students only studied the verb while they studied and retrieved the particle. The different ways of processing the component words of the PVs may affect the participants' ability to recall them on the final test. Past studies required separate conditions to compare retrieval and restudy of a target item, but Study 2 conflates these two conditions.

The study-test condition had participants read the verb and the particle in the study trial before they restudied the verb and attempted to recall the particle in the test trial. Past research has demonstrated that restudy of a word is less effective than recall of it. Therefore, it was predicted that final test performance would be greater for the recalled component word than the restudied one. The findings showed that although particle performance was slightly higher than verb performance on the immediate and delayed post-tests, the difference was not statistically significant. This finding suggests that the retrieval effect may be influenced by the relationship between the word studied and the one retrieved. In past research, the studied word and the retrieved one were never the components of a phrase.

Like the study-test condition, participants in the study-delay-test condition read the verb and particle in the study trial and reread the verb and attempted to recall the particle in the test trial. However, they completed the test trial only 6.5 minutes after reading the PV in the study trial. In contrast to the results of the study-test condition, participants in the study-delay-test condition were more likely to correctly recall the verb than the particle on the immediate post-test, although this difference faded over a one-week period. This finding is unexpected because it was assumed that retrieval would enhance learning and retention better than restudying. Instead, the findings show that delayed restudying of the verb is more effective than delayed retrieval of the particle. This result may be due to the different characteristics of the word studied vs. the one retrieved.

The test-study condition required participants to study the verb and recall the particle and then immediately to reread the verb and to study the particle for the first time. The test-delay-study condition was the same except a 6.5-minute interval separated the trials. Despite the lag between these two conditions, mixed effects logistic regression modelling showed performance on the final test between the verb and particles in both conditions did not differ. This result suggests that rereading a verb is as effective as failing to recall a particle followed by corrective feedback. However, since the verb and particle are related to each other in the sense that they represent a single PV meaning, it is unclear whether final test performance was affected by this relationship.

Overall, the results of the item-analysis demonstrate that retrieval of the particle item of a verb phrase does not lead to greater learning gains compared to only studying the verb item. This finding may be due to factors related to the characteristics of the individual words and the semantic relationship between the verb and the particle when they express a PV meaning. Because Study 2 did not require participants to retrieve the verb and study the particle, it is difficult to be certain whether the finding is due to the characteristics of individual words of the PVs. It would have been insightful if the study required students to recall the verb and study the particles on some target items while recalling the particle and studying the verb on other target items.

7.3.6 Research question 6: The effects of making errors during learning

One of the most harmful effects of a test error is for it to be regarded as the correct answer even after corrective feedback is given. One way to prevent test errors is to have students study the target item immediately before taking a test. However, the lack of a delay between the study trial and the test trial resembles rote rehearsal, which has been found to be ineffective. Delaying the test, therefore, is presumed to increase the positive effect of retrieval even though the likelihood of producing a test error is greater than immediate testing. However, this study showed that the best performance occurred in the study-test condition than in the study-delay-test condition and that the study-test condition resulted in much less retroactive interference than the study-delay-test condition. Error analysis indicated that on the immediate post-test, zero percent of final test errors were the same as test trial errors in the study-test condition, whereas 35% of final test errors were replicated errors in the study-delay-test condition. On the delayed post-test, replicated errors were greater in the study-delay-test condition than in the study-test condition. Therefore, the study-test condition not only led to greater learning gains compared to the study-delay-test condition, but it also reduced the harmful effects of interference in fossilizing incorrect knowledge.

Participants in the test-study condition replicated 67% of their test trial errors and those in the test-delay-study condition reproduced 73% of them. Therefore, proactive interference was very high in both treatment conditions. One explanation to account for this finding is that greater cognitive effort was involved in generating a response on the test trial (even though it turned out to be incorrect) than in studying the answer in the corrective feedback. As a result, the memory trace of the error was more robust than the memory trace of the correct response. Since both traces are associated with the same retrieval cue, the stronger memory trace is more likely to be remembered than the weaker one.

Overall, the findings of the error analysis in Study 2 resembles the results of the error analysis in Study 1. Taking a test and producing an error on that test, even though the correct answer is subsequently provided in corrective feedback, causes greater proactive interference than does taking a test and failing to recall the correct answer that was previously studied cause retroactive interference.

7.4 Discussion of results: Study 3

This section discusses the findings of Study 3 with reference to each of the research questions. The results are also discussed in relation to previous research studies.

Research question 1: Which learning condition enhanced knowledge of PVs the best?

The first research question asked whether PV acquisition could be enhanced via contextualized retrieval and generation conditions. Past research on L2 vocabulary acquisition has looked at incidental vocabulary acquisition and direct (intentional) vocabulary learning in context. The former refers to picking up vocabulary as a by-product of reading or listening, and the latter to learning words with the intention to encode them in memory. Study 3 focussed on intentional PV learning for two main reasons. First, in ESL/EFL course textbooks, students are aware of the vocabulary-to-be-learned in the test trial. Second, confidence in using PVs is unlikely to occur through picking them up incidentally as some avoidance studies have shown that even advanced professional non-native speakers continue to struggle to use PVs and prefer to avoid their use altogether (Siyanova & Schmitt, 2007).

There are several benefits to embedding a target item in context compared to pairing it with a definition. First, the words in a sentence can help students to infer the meaning of the target item. Most L1 vocabulary acquisition takes place in such a manner rather than through a definition as is the case with decontextualized learning. Second, the semantic relationship and the collocation with other words may also be learned in a sentential context. Students can notice the words surrounding the target item and how they express meaning when they co-occur. It seems unlikely for this information to be gained in paired associates. Third, students may come to learn how to use the target item through meeting it in contexts. Fourth, verb phrases seen in context shows how they function and collocate with other words, whereas this is unlikely to occur through decontextualized learning (Webb, 2007). Thus, various aspects of PV knowledge may be acquired through learning in context. It was, therefore, assumed that the use of contextualized retrieval and generation conditions in Study 3 would facilitate PV acquisition better than the use of decontextualized retrieval and generation conditions examined in Studies 1 and 2.

The analysis of test trial responses indicated that participants in the retrieval group successfully recalled 98% of the particles after the PVs were read in the study trial. Unexpectedly, some participants in the generation group had also correctly generated 13% of the particles on the test trial even before they were given the study trial, which suggests that these PVs might have been known to them, despite the findings of the norming study. There are two plausible reasons why these particles were answered correctly in the generation condition. One is that since there are a finite number of particles, and some tend to occur more often than others in PVs (Garnier & Schmitt, 2015), participants were lucky in guessing. The second is that they had knowledge of the PVs prior to the study. The fate of these PVs answered correctly on the test trial in the generation condition were tracked on the post-tests to determine whether they were produced correctly or erroneously. If they were again answered correctly, then this would count as a likely demonstration of prior knowledge, whereas, if they were answered incorrectly, then this would indicate that they were lucky guesses on the test trial. The tracking indicated that 94% and 97% of correct test trial responses were incorrectly recalled on the immediate post-test and the delayed post-test, respectively. This strongly suggests that nearly all of the correctly answered particles on the test trial were, in fact, lucky guesses. Nevertheless, the PVs were removed from the analysis and the data was run on 10 rather than 14 target items.

As mentioned above, the mixed effects logistic regression models indicated that the contextualized retrieval condition resulted in better learning than the contextualized generation condition did. Moreover, the rate of attrition was slightly greater for participants in the generation condition than for those in the retrieval condition. More specifically, while the participants in the retrieval condition showed a loss in knowledge of 16 percentage points, those in the generation condition forgot 23% of the target item they successfully recalled on the immediate post-test. The generation group's loss in knowledge is similar to the rate of attrition recorded for the test-study condition, 22%, and the test-delayed-study condition, 26%, in Study 2, while it was significantly less than that of the generation group, 39%, in Study 1. The loss in knowledge of the retrieval group in Study 3 was much lower than the 59% recorded for the study-test condition and the 60% found for the study-delay-test condition in Study 2 and the 51% recorded for the retrieval condition in Study 1.

Research question 2: Which learning condition caused greater interference?

I assumed that the generation condition would result in a considerable amount of proactive interference based on the results of studies 1 and 2. As expected, many of the errors on the post-tests were, in fact, the same errors participants had generated on the test trial. This finding suggests that context does not immunize students from regarding some of their failed generations as correct responses any better than paired-associates. In other words, one of the dangers of taking a test "cold," so to speak, is that errors on that test may be difficult to unlearn. Many researchers in the field of errorless learning have voiced concern regarding making mistakes during learning. They claim that errors can do more to hinder rather than to help in the acquisition of knowledge. Therefore, they have advocated for the use of procedures that minimize the likelihood of students making errors during learning (Warmington & Hitch, 2014). One of these procedures involves students studying the target item before the test. The retrieval condition in Study 3 was such an attempt to minimize the harm that may arise from recalling an incorrect response on a test. Because participants in the retrieval group produced so few errors in the test trial, post-test errors were not affected by test trial errors. However, in this study, it seems that one of the costs of lessening the chances of interference from errors is that knowledge of the target items dissipated quickly over a one-week period.

7.5 General Discussion

This section provides a general discussion of the results of the three studies carried out in this thesis by addressing three main questions.

7.5.1 Which treatment was best for enhancing recall memory of PVs?

All three studies administered a final recall test, but the cues used in it differed in each study. In Study 1, the recall test presented the initial letter of a PV and a paraphrase of its meaning (e.g., h___ - to spend time with a friend). In Study 2, it displayed only the explanation of a PV's definition (e.g., ___ - to spend time with a friend). In Study 3, the recall test showed the verb of a PV without its particle in dialogue (e.g., Speaker A: Hi, Toko. What are you doing tomorrow? Do you want to hang___?). We cannot compare the performance of the participants' verb responses in the recall tests across the three studies. The reason is the recall test in Study 1 presented the first letter of the verb as a retrieval cue, while in Study 2, no hint was provided, and Study 3 provided the verb as one of the retrieval cues. However, in all three studies, the recall test did measure knowledge of the particle. Therefore, we can compare particle performance across the studies. Yet, we should keep in mind that the presentation of the verb or the initial letter of it as a retrieval cue may prompt retrieval of the particle in different ways. For example, *hang__* may prompt *out* more easily than *h___* or ___.

For the retrieval conditions on the immediate post-test, the percentage of correct particle responses was 72% in the decontextualized retrieval treatment in Study 1, 78% and 72% in the study-test and the study-delay-test treatments, respectively, in Study 2, and 66% in the contextualized retrieval treatment in Study 3. For the generation conditions on the immediate post-test, the percentage of successfully recalled particles was 62% in the decontextualized generation treatment in Study 1, 42% and 48% in the test-study and the test-delay-study treatments, respectively, in Study 2, and 57% in the contextualized generation treatment in Study 3. Averaging the learning gains of the retrieval groups across the studies showed that retrieval enhanced knowledge of 72% of the target PVs, which is significantly greater than the 52% in learning gains obtained from generation. The finding that there was an advantage for all three versions of a retrieval condition over all three versions of a generation condition is consistent with the results of past research on this topic (Baddeley & Wilson, 1994; Warmington & Hitch, 2014). Thus, there is growing evidence to suggest that testing can have a positive effect on learning when students produce the correct response on the test, but it can also have a less beneficial effect when students fail to produce the right answer on the test, even though the correct answer is subsequently provided.

Although past research has shown retrieval to be more effective than generation on learning, hardly any of these studies measured how they affected long-term retention. One of the aims of the present study was to investigate whether the benefits of retrieval persisted over a span of time. A one-week period may be a sufficient amount of time for forgetting to occur, leaving only strong memory traces formed during the treatment. The use of a one-week delayed post-test allowed students to demonstrate the strength of their memories of the PVs. For the retrieval groups on this post-test, the percentage of correct particles recalled was 26% in the decontextualized retrieval condition in Study 1, 21% in the study-test condition and 19% in the study-delayed-test condition in Study 2, and 50% in the contextualized retrieval condition in Study 3. For the generation groups, the percentage of successfully recalled particles was 26% in the decontextualized generation condition in Study 1, 18% and 18% in the test-study and the test-delay-study conditions, respectively, in Study 2, and 34% in the contextualized generation condition in Study 3. An average of the participants' performance on the delayed post-test showed that those in the retrieval condition retained knowledge of approximately 29% of the target items and those in the generation condition preserved the knowledge of 24% of the target items in recall memory. The comparison showed that the retrieval condition led to slightly better retention of the target items compared to the generation condition. However, a comparison between immediate and delayed post-test performance showed a steep loss in the ability to recall the target items over a one-week period.

Many factors may have contributed to the very high attrition rate recorded for both treatment groups. First, studies 1 and 2 provided participants with only 15 seconds in the study trial to learn an incredibly complicated type of verb phrase that they would likely prefer to avoid using (e.g., Dagut & Laufer, 1985). Study time was, therefore, brief and probably insufficient to allow for students to establish connections between the new target item and other ones in their mental lexicon, resulting in the formation of a partial representation of the target item. A target item partially represented in memory most likely cannot be recalled successfully as recall demands that it be well-specified. Eliminating the time constraint did help to increase retention, particularly for participants in the retrieval condition than for those in the generation condition as demonstrated in Study 3, but other more significant factors may also be responsible for their greater rate of retention, such as contextual support.

7.5.2 Which treatment was best for enhancing recognition memory of PVs?

A recognition post-test was administered to participants in the treatment groups in Study 1 and Study 2 but not in Study 3. To avoid confounding the results of the recall test, the recognition test was administered only after the delayed recall test (e.g., Warmington & Hitch, 2014). On the recognition test, for the retrieval conditions, the percentage of correct responses was 49% in the retrieval condition in Study 1, and 61% and 49% in the contextualized study-test condition and the study-delay-test condition in Study 2, respectively. For the generation conditions, the percentage of right answers was 47% in the contextualized generation condition in Study 1, and 49% and 53% in the test-study condition and test-delay-study condition in Study 2, respectively. On average, the retrieval conditions enhanced recognition memory of the target items by 53% while the generation conditions improved this memory by 50%.

It is unsurprising that both treatment groups yielded greater recognition memory than recall memory as recognition tests usually result in higher scores of retention than do recall tests. According to Tulving (1968), the superiority of recognition over recall is likely due to two factors. The first is the number of options from among which the right answer is to be selected. The second is the amount of information retained in memory needed for the identification and production of the learned target items. What is surprising is that between these two treatment conditions there was a small difference in scores on the recognition test in comparison to the large difference in scores on the recall test. Such a finding suggests that despite the negative effect of making errors on creating recallable memories, errorful generation was much less of an issue in creating recognition memories (Warmington & Hitch, 2014).

7.5.3 Which treatment caused greater interference?

Making errors during learning may affect subsequent memory of the correct answer. Interference theory suggests that study of a target item but failure to recall it on an initial test may cause retroactive inference, where the error, as opposed to the correct answer previously studied, is recalled on a later test. It also holds that an incorrect response on a test followed by study of the correct answer may cause proactive interference, where the erroneous test response, rather than the feedback, is recalled on a final test. Interference caused by retrieval learning and generation learning is undesirable because it can be difficult to unlearn errors once they are made. To minimize the chances of making an error during learning, the retrieval condition in Study 1, as well as the study-test condition and the study-delayed-test condition in Study 2, had participants do the study trial before the test trial. In contrast, to maximize the likelihood of making mistakes during learning, the generation condition in Study 1, as well as the test-study condition and the test-delay-study condition in Study 2, had students do the test trial before the study trial. As expected, on the test trial, hardly any responses were incorrectly recalled by students in the retrieval conditions, whereas, every response was erroneously generated by those in the generation conditions. Therefore, the possibility of interference from test trial errors was much greater in the generation conditions than in the retrieval conditions.

7.5.4 What is the verdict on using errorful generation as a learning strategy?

According to the elaborative retrieval theory, taking a test can facilitate learning even when a test response is incorrect, as long as the test taker receives corrective feedback. However, mistakes only benefit learning when they are semantically associated with the correct answer and the retrieval cue. Therefore, learners should avoid making errors that are unrelated to the target item as they might harm rather than help retention.

However, Potts and Shanks (2014) demonstrated that test errors unrelated to the correct answer have a positive effect on retention, which challenges the assumption of the elaborative retrieval theory. The authors indicate that participants most likely processed the corrective feedback deeply in order to suppress recovery of the initial test response error. This finding, however, has not been replicated in other studies. Boers et al. (2014) observed that trial-and-error learning procedures led to only minimal learning gains (5-10%). Similarly, Stengers and Boers (2015) did not find an advantage for a different kind of trial-and-error method in comparison to an error-free learning condition.

The difference between the findings reported by Potts and Shanks (2014) and those by Boers et al. (2014) and Stengers and Boers (2015) may be because of differences between their experimental design. Potts and Shanks (2014) presented the study trial immediately after the test trial. In contrast, Boers et al. (2014) and Stengers and Boers (2015) administered the study trial after the test trial after a much longer period. While a student was tested on one target item and then read the correct answer in Potts and Shanks (2014) study, a student was tested on several target items and only then received corrective feedback on those items in the studies by Boers et al. (2014) and Stengers and Boers (2015). Study 2 looked at whether the lag between the test trial and the study trial was a factor in retention. Mixed effects modelling on the performance of the test-study group and the test-delay-study group showed that the lag was not a factor in acquiring knowledge of PVs.

Another factor that might account for the conflicting findings of the studies discussed above is the use of context in the generation procedure. Potts and Shanks (2014) used a cued-recall test format to elicit retrieval of the target item, whereas Boers et al. (2014) employed a contextualized multiple-choice test, where students select the correct answer from among lures and insert it into gapped sentences. A fill-in-the blanks sentence format was used as the test trial by Stengers and Boers (2015). Studies 1 and 2 in this thesis employed a cued-recall test format similar to the one used by Potts and Shanks (2014), and Study 3 used a contextualized trial-and-error procedure comparable to the one employed Stengers and Boers (2015) than to the one used by Boers et al. (2014). Comparison of the results of studies 1, 2 and 3 showed hardly any difference between the generation conditions on short-term learning, but the contextualized generation condition led to better long-term learning than both the decontextualized generation conditions. Taking into account that Boers et al. (2014)

and Stengers and Boers (2015) administered their post-test two weeks after the treatment may explain the generation groups' poor performance.

Further, the type of criterial test used may also account for the differences in the findings of these studies. Potts and Shanks (2014) used a criterial post-test which was the same format as one employed in the test trial. Boers et al. (2014) used a fill-in-the-blanks sentence criterial test, which differs from the multiple-choice format employed in the test trial. For the criterial test, Stengers and Boers (2015) used fill-in-the-blanks-in-a-sentence format, which is similar to the one they used in the test trial, except sentential context differed between the two. According to transfer-appropriate processing theory (Morris, Bansford, & Franks, 1977), performance is likely to be greater on a final test when it is similar to the format of the learning condition. This theory might explain why the performance of students was highest in the generation condition in the study by Potts and Shanks (2014), next best in Stengers and Boers (2014) and worst in Boers et al. (2014).

The last main factor to consider is the type of target items examined in these studies. Potts and Shanks (2014) used Euskara single-words as target items, whereas Boers et al. (2014) and Stengers and Boers (2015) used English verb-noun collocations as target items. While one type of vocabulary may be more challenging to learn than the other, the difficulty in learning these items may be regarded as irrelevant to the effectiveness of the generation condition because effectiveness is typically gauged against a comparison condition. That is, despite the poor performance of students in the generation condition, poorer performance of students in a comparison condition should be observed if we are to assume that generation is an effective strategy learning strategy.

In short, the factors that most likely explain the conflicting results of these studies is the time the researchers administered the final test and the similarities or differences between in the cue used to elicit retrieval in the test trial and the final test. I expand on these factors below. First, it seems easier to recall something sooner rather than later because, without subsequent exposure to the target item, its trace in memory inevitably fades over time. Potts and Shanks (2014) seem to be measuring short-term learning gains, whereas Boers et al. (2014) and Stengers and Boers (2015) appear to be measuring long-term learning gains. The studies in this thesis showed that learning gains were greater on the immediate post-test than on the delayed post-test. Thus, students in Boers et al. (2014) and Stengers and Boers (2015) most likely found it challenging to recall the target items because of the two weeks which passed since they learned it. Similarly, students in the studies in this thesis probably found it more difficult to recall the target items after a one-week period. Second, the three studies used different formats to induce generation on the test trial and the post-test. Potts and Shanks (2014) employed a test trial that demanded recall or recognition and a post-test that elicited recognition. In contrast, Boers et al. (2014) used a test trial that required students to recognize a target item and a post-test they had them recall it, while Stengers and Boers (2015) used a test trial that elicited recall and a post-test that also elicited recall. In the three studies in this thesis, the format of the test trial and main criterial test induced recall processes. Therefore, for these reasons, the effects of pre-testing had a different effect in all these studies. Overall, the negative effects of pre-testing seem to outweigh its positive effects on learning phrases.

7.6 Some implications for teaching

The present thesis showed that L2 students have considerable difficulties in acquiring PV knowledge and have, as a result, taken to avoiding their use in favour of using near synonymous single-word verbs (Dagut & Laufer, 1985). But, it is important that students acquire PVs because they are common (Darwin & Gray, 1999) in the English language and instrumental to language fluency (Boulton, 2008). To help students add these complex verb phrases to the lexicon, the studies in this thesis explored three versions of a retrieval procedure and a generation procedure. The findings of these studies largely support the view that one learns better by errorlessly recalling a studied target item than by attempting to learn from one's errors. However, the main limitation of the errorless retrieval approach is that learning gains seem to fade quicker than the learning gains that resulted from effortful generation. This finding, however, is not unexpected given the challenges that surround PV acquisition, especially for those L2 learners whose L1 system lacks these multiword verbs. Undoubtedly, PV learning would benefit more greatly from continuous retrieval practice, where each subsequent retrieval attempt is slightly more difficult than the previous one (Nakata, 2015).

On the basis of the results of the studies in this thesis, some suggestions can be made about how to design PV learning procedures in general ESL/EFL course textbooks. First, as it was indicated in the textbook analysis, a large majority (72% to be exact) of learning procedures are generation-based, which were found in all three studies to be less effective than retrieval-based learning procedures. This suggests that although some learning gains may come from generation procedures, the disadvantages of one making mistakes during learning may outweigh the advantages of learning from one's mistakes. There are two reasons for this. First, feedback on a failed test response does not always override the memory of the self-generated error, which can then result in proactive interference. As the error analysis indicated, in some cases over 70% of final test errors were in fact the same errors students had produced on the test trial, which shows that the memory of the error, which was actively generated, was stronger and, thus, more memorable than the memory of the correct response, which was passively encoded as feedback. Even when the corrective feedback presented the answer in richer contextual support than what was offered in the test trial, it did not prevent proactive interference. This shows that it can be difficult to unlearn errors once they are made. Consequently, it seems a good idea for materials designers to foster PV learning using retrieval-based procedures rather than generation-based ones.

Second, some textbooks ask students to complete a test trial by finding and copying the answer in the study trial. For example, in McCarthy and O’Dell’s phrasal verb textbook (2004), the test trial is often presented on the opposite page facing the study trial. The instructions of the test trial ask students to refer to the phrases in the study trial to complete the test trial. Such learning procedures raise the question of whether copying is a form of retrieval. Since learners most likely do not access the target item from memory, it seems retrieval processes are not involved. This form of learning has been found to result in the same amount of learning that comes about through generation learning (Stengers & Boers, 2015). Although materials designers cannot always help but place a study trial next to a test trial given the space limitation in the textbook, it is good to see that some make attempts to foster retrieval by asking students to cover the study trial with their hand while tackling the test trial. It would be a good idea if all test trials stated this in their instructions.

Third, educational researchers debate whether phrase learning is enhanced when students reassembled phrases from their component words or process them as wholes. For example, Watson (1997) proposed that by practicing joining words to form phrases, students' knowledge of them improves. However, others have suggested against this practice. They have, instead, advocated for students to learn phrases without breaking them apart to prevent them from making common errors, such as "do a choice" instead of "make a choice"(Boers et al., 2014; Boers, Dang, & Strong 2016; Lewis, 2000; Wray & Fitzpatrick, 2010). Although the aim of this thesis was not to compare discrete phrase learning vs. holistic phrase learning, we may, nevertheless, speculate by comparing Study 2, which consisted of holistic learning, with Study 3, which involved discrete learning, while acknowledging differences between the target items and the design of these studies. Taking into account overall learning gains across the post-tests, discrete retrieval of PVs (58%) resulted in superior performance compared to holistic retrieval (42%). This comparison, therefore, shows that targeting a problem item in a verb phrase, such as the particle rather than the verb, may facilitate learning better than not drawing attention to it during learning.

Fourth, in general ESL/EFL course textbooks, there is often a considerable amount of time between the presentation of a phrase in the study trial and the retrieval of that phrase in the test trial. The reason is that after a study trial, there are many other exercises and activities (e.g., grammar, writing, and text comprehension) that students are required to do before the phrase test trial. Thus, retrieval of a phrase is often not immediately elicited after it was learned. However, the time interval (as well as the number of intervening items learners deal with) between the study trial and the test trial appears to have a significant influence on retention. As Study 2 found, it is better when retrieval is immediate than when it is delayed. On the basis of this finding, it is recommended that materials designers attempt to have students retrieve a studied target item sooner rather than later. However, with respect to generation-based procedures, it doesn’t seem to make a difference whether the feedback is immediate and delayed.

Fifth, although decontextualized and contextualized conditions were not compared in a single study in this thesis, I can, nevertheless, make teaching recommendation based on the trends observed in these studies. The use of contextual support in the retrieval condition in

Study 3 was slightly less effective than the decontextualized retrieval conditions in Studies 1 and 2. Similar findings have been reported by researchers who have examined explicit vocabulary learning from context and explicit vocabulary learning devoid of context (Seibert, 1930; Griffin, 1992; Dempster, 1987; Laufer & Shmueli, 1997; Webb, 2007). This suggests that adding contextual support (a single dialogue between two individuals) has little added value in establishing a recallable memory trace of a PV. However, it is plausible that the contextualized retrieval condition contributed to the acquisition of other aspects of lexical knowledge, but it was not measured in any of the studies examined in this thesis.

7.7 Some implications for theory building

The theories discussed in the literature review can partially explain the findings of the treatments in all three studies. According to the retrieval effort theory, the difficulty involved in retrieving a target item from memory affects its representation in the memory, with effortful retrieval strengthening a memory trace more than effortless retrieval. The retrieval condition in Study 1 and the study-test condition in Study 2 fostered error-free retrieval by eliciting recall immediately after students studied a target item. While they minimized the risk of making errors, it was probably easy for students to remember the target item. As a result, participants recalled only a small amount of target items after a one-week delay. The study-delay-test condition in Study 2 also attempted to foster error-free learning but at the same time endeavoured to increase retrieval difficulty by inserting a 6.5-minute lag to recall the target item after a student studied it. However, the study-delay-test condition did not enhance learning better than the study-test condition. In fact, on the recognition test, performance was worse for participants in the study-delay-test condition than for those in the study-test condition. If we accept that a 6.5-minute lag increases retrieval effort more than no lag, the retrieval effort theory cannot fully account for the findings of Study 2.

The retrieval effort theory may partially account for the results of the generation conditions in all the studies. While the effort involved in retrieval is necessary to facilitate retention, it is also important for the target item to be generated successfully. This idea is expressed as desirable difficulty (Bjork & Bjork, 2011). It holds that retrieval must be hard but not too hard to foster retention. For participants in the generation groups, producing a response on a test of a target item considered to be extremely difficult to learn seemed to violate the notion of desirable difficulty. As a result, performance was poorer (at least on the immediate post-test) for students in the generation conditions than for those in the retrieval conditions.

The poor performance of the generation groups may also be accounted by the elaborative retrieval theory. According to this theory, a test error can enhance learning as long as it is semantically related to the retrieval cue and the target item. Since participants had no prior knowledge of the form-meaning connection of the PVs, their test trial errors were unrelated to the correct answer. As a result, the test trial error most likely hindered rather than helped to encode the correct response into memory.

All retrieval theories discussed in the literature review hold that retrieval, as opposed to restudying, leads to better learning and retention. Past research often compared a retrieval condition with a restudy condition and showed the retrieval condition to be superior to the restudy one. Study 2 conflated these procedures into one. Participants restudied the verb of a PV while recalling the particle. Contrary to the notion of the testing effect and past research showing a benefit of retrieval over restudy, Study 2 showed that recall of the particle item did not result in better performance than restudying of the verb item in any of the treatments. This finding may be because participants did not recall a target item separately from restudying the other. Instead, encoding and retrieval processes were likely active at the same time, and this interaction may have affected the learning in a way that differs if the two processes were activated separately.

7.8 Limitations and recommendations for future research

This section discusses the limitations of the studies examined in this thesis. Additionally, it provides recommendations for future research.

Limitations

This thesis used an innovative method to calculate participants' prior knowledge of the target items. Boers, Dang, and Strong (2016) first used it in their study. It involves norming potential target items with a parallel group of participants with similar L2 linguistic abilities, lexical knowledge, and backgrounds to the participants in the treatment groups. The advantage of such a norming study is that it circumvents the need to administer a pre-test, which has the potential to confound the results of the treatment by causing a pre-testing effect. Though, there is a slim chance that students in the treatment may demonstrate knowledge of the target items that learners in the norming study did not. Of course, there are alternative to the norming study. One is to use target items that students are either highly unlikely to have ever encountered before the treatments, such as extremely low-frequency words or words from another language. Moreover, researchers may use words that students have never seen before, such as nonsense words. Both these alternatives are valid methodologically as they ensure participants possess no prior knowledge of the target items. However, they do not appear to be valid ecologically, as learning PVs is a genuine issue for many L2 learners who most likely have knowledge of the individual words but not of the meaning that emerges when they co-occur together in a particular order. Thus, although some participants generated correct responses on the test trial, they were dropped from the analysis.

A further limitation was that the treatment groups pre-studied the individual words of the target items while the norming group did not. As a result, even though it was assumed that participants had prior knowledge of the individual words, this difference may have unbalanced the norming group with the treatment groups. However, Study 1 showed that none of the participants in the norming group demonstrated knowledge of the target items, which validates the results of the norming group. Nevertheless, it would have been preferable if the norming group had pre-learned the individual words of the PVs before taking the norming test.

Another limitation regards the size of the sets of target items in each study. Ideally, the size of the sets should have been larger than what they were to collect more data points to analyze. However, I conducted the studies during students' regular English lessons in their L2 classroom. I, therefore, had a limited amount of time to run the experiment and collect the data. Study three used the least number of target items. The reason for this is that pilot testing indicated that to collect data on 28 target item, I would need longer than the time I was allotted to collect the data. Since each set consisted of 7 target items, I had to drop two sets to keep my study within the given time frame.

In retrospect, to allow comparison between studies, it would have been a good idea to administer the same type of post-test. It was difficult to compare the results of the studies because different post-tests were used in each study. Study 1 provide the initial letter as a retrieval cue along with the paraphrase of the PV’s definition. The way participants recalled the PV on this post-test may be different from the way they recalled the PV in Study 2, which only presented the paraphrase of the PV’s definition.

A final notable limitation was assessing the same target items on the immediate post-test and delayed post-test. Researchers in the field of memory research ordinarily assume that retesting the same items results in the testing effect. One way to avoid the testing effect is to not administer a delayed post-test on the same target items that were tested on the immediate post-test. As a result, participants do not take a test on a target item more than once. This procedure is preferable to the one I employed in this thesis.

Recommendations for future research

Very little research has examined the effects of pre-testing and testing on the learning of L2 phrases. Therefore, there remains much research to do on this topic. I suggest three areas of future research that are worthy of investigation.

First, Study 2 conflated a restudy condition with a retrieval condition. More specifically, a participant studied a verb and a particle in the study trial and restudied the verb but produced the particle in the test trial. The final test measured the extent to which a student was able to recall both the verb and the particle. This design was set up to compare the effects of encoding processes on the verb with retrieval processes on the particle. There is plenty of evidence that retrieval of a word is superior to study of that word. However, the findings of Study 2 do not fully support this view. To further explore this issue, a study should also reverse the conditions so that a student retrieves the verb and only restudies the particle. This research would reveal whether the pre-testing effect and the testing effect is more effective on verbs than particles or the other way around and if there is no effect perhaps it is due to the relationship between the individual words as they represent a single semantic unit.

Second, the studies in this thesis used a test trial (or exercise) that induced recall. Although recall test trials are one of the types of test trials identified in the textbook analysis, the more common one elicits recognition. Recognition test trials, like multiple-choice, include the answer with several lures. It would be interesting to investigate how the lures influence phrase learning. A researcher may compare a multiple-choice with many lures with one with a few lures. Further, one may examine how the semantic relationship between the lures and the correct answer affects learning and retention. In one condition, the lures are unrelated, and in another one, they are related. How a student processes the lures while choosing the correct answer may influence retention of the correct answer. Additionally, it would be interesting to compare recognition-based retrieval conditions with recall-based ones to determine which one enhances phrase learning the best.

Third, in all the studies in this thesis, the conditions only required students to retrieve or generate the target item once during learning. Since there is little doubt that repeated exposures have a positive effect on learning, future research should design retrieval and generation conditions that offer more than a single retrieval/generation attempt. Relatedly, it would be interesting to compare phrase learning using a retrieval condition followed by a generation one vs. a generation condition followed by a retrieval one. Perhaps following a generation condition with a retrieval attempt, one may undo the adverse effects of making errors during learning. For additional ideas on future areas of phrase learning research, the textbook analysis in this thesis offers plenty of learning conditions that deserve exploration.

Chapter 8 - Conclusion

This chapter presents, firstly, a summary of the key findings of the research, followed by a consideration of pedagogical implications for teachers and materials designers, as well as recommending implications for future research. The limitations of the study are assessed subsequently. The chapter concludes with a brief summary of the preceding sections.

The primary objective of this thesis was to investigate and compare study-test conditions and test-study conditions on the learning and retention of PVs. The secondary aim of the thesis was to identify the conditions of PV learning available in general ESL/EFL course textbooks. The studies were carried out at English language learning classrooms at a different university in Japan on native Japanese speakers with an intermediate level of English proficiency. Qualtrics was used to collect data and to administer the treatments, which were timed. Although pretests were substituted by norming studies to avoid the pretesting effect from confounding the results of the studies, some participants in the treatments produced the responses correctly, which may be interpreted as a demonstration of prior knowledge or lucky guessing.

A preliminary finding of this research was that retrieval-oriented conditions improved learning better than generation-oriented conditions. In Study 1, immediate recall post-test performance was greater for the retrieval condition than the generation condition but delayed recall post-test results showed hardly difference in the rate of retention. Further delayed recognition post-test showed that learning gains did not differ between the conditions. Analysis of the individual parts indicated that performance (on the immediate and delayed post-test) was greater on verbs than particles for both the retrieval condition and the generation condition. Error analysis showed errorful generation led to the replication of errors and that replicated verb errors were more common than replicated particle errors. Study 2 showed an overall advantage for retrieval of PVs than generation of PVs. Performance on the immediate and delayed cued recall post-tests was greatest in the study-test condition. Similarly, delayed recognition post-test performance was greatest in the study-test condition. Analysis of the individual parts showed that verbs were remembered better than particles on the immediate post-test (except for the study-test condition), but verbs were remembered worse than particles on the delayed post-test. Error analysis indicated the likelihood of reproducing the same error produced during learning was greater in the generation conditions than the retrieval conditions and that replicated errors. The findings of Study 3 resemble the findings of studies 1 and 2. The retrieval condition led to greater memory of the PVs than the generation condition. It also indicated that retrieval led to greater retention than generation. Analysis of replicated errors indicated that nearly 50% of post-test errors were, in fact, the same errors students in the generation group made during learning. Preceding these studies was a textbook analysis which identified that the most common type of conditions for learning phrases are generation-oriented, which the studies in this thesis have shown are least effective for enhancing knowledge of PVs.

All the findings of the studies in this thesis are original, as no previous research examined and compared retrieval-inducing conditions with generation-inducing conditions on the learning of PVs. Although a textbook analysis on phrase-focused exercises was carried out in a previous study (Boers et al., 2014), it was on a small number of textbooks and used a different method to identify the condition of phrase learning.

The results of the present thesis have confirmed that retrieval methods enhance knowledge of PVs better than generation methods. Clearly, using retrieval-oriented methods, where a student must try to remember something learned in the past, foster the production of accurate memories, while generation-oriented methods have the undesirable side-effect of impeding the unlearning of self-generated errors. Recognizing the potential harm that may result from generation procedures on learning would benefit materials designers and classroom teachers, not to mention students. Prior knowledge of which method is more beneficial for PV acquisition can assist materials designers in creating more effective PV exercises.

In the L2 classroom, the teacher could help ensure that students are given exemplar material that presents PVs with their meanings and with examples sentences. Students should then be informed that after study, they will sit a low-stakes test, quiz or exercise in order to direct their attention to encoding the complex verbs into memory. The test should follow immediately after students have studied the PVs because forgetting is likely to occur very soon afterward. Successful retrieval of a studied PV on the test suggests that the memory of that verb phrase has received a boost, making it slightly stronger. However, the next opportunity to retrieve the complex verb should be earlier than a week, preferably with several minutes and then days of the initial retrieval attempt.

It is also suggested that materials designers structure their conditions of learning to induce retrieval rather than generation. For materials designers, the most obvious advantage of generation conditions is that they are economical: study materials can be presented in the form of an answer key in the appendix of the textbooks, where they occupy little space. In contrast, for retrieval conditions, the study material must come before the exercise and cannot be presented in the same format as an answer key. As a result, exemplar materials occupy a considerable amount of space in textbooks. Given the results of the textbook analysis and the findings of the studies in this thesis, materials designers should focus more on using retrieval-oriented methods than generation-oriented methods.

The studies in this thesis provide valuable information on the effects of a single retrieval attempt on learning as well as how errors can either hinder or help learning. However, the conditions examined in the studies represent a limited set of learning procedures that resemble the conditions of learning identified in the textbook analysis. Further research should consider the other formats uncovered in the textbook analysis in order to determine which method promote PV and phrase learning the best.

CHAPTER 9 - References

Abel, B. (2003). English idioms in the first language and second language lexicon: a dual representation approach. *Second Language Research, 19*(4), 329-358. doi:10.1191/0267658303sr226oa

Abel, M., & Bauml, K. H. (2014). The roles of delay and retroactive interference in retrieval-induced forgetting. *Memory Cognition, 42*(1), 141-150.

Anderson, J. R. (1972). FRAN: A simulation model of free recall. *Psychology of learning and motivation, 5*, 315-378.

Anderson, J. R., & Reder, L. M. (1979). An elaborative processing explanation of depth of processing. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, M., Bjork, E., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review, 7*(3), 522-530.

Anderson, N. D., & Craik, F. I. (2006). The mnemonic mechanisms of errorless learning. *Neuropsychologia, 44*(14), 2806-2813. doi:10.1016/2006.05.026

Arnold, K. M., & McDermott, K. B. (2013). Free recall enhances subsequent learning. *Psychonomic Bulletin Review, 20*(3), 507-513. doi:10.3758/s13423-012-0370-3

Atkinson, R. C., & Shiffrin, R. M. (1968). Chapter: Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2). New York: Academic Press.

Azar, B. (1989). *Understanding and using English grammar*. New Jersey: Prentice Hall Regents.

Baddeley, A. (2010). Working memory. *Current Biology, 20*(4), 136-140. doi:10.1016/j.cub.2009.12.014

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia, 32*(1), 53-68.

Barcroft, J. (2000). *The effects of sentence writing as semantic elaboration on the allocation of processing resources and second language lexical acquisition.* (PhD Thesis), University of Illinois at Urbana-Champaign.

Barcroft, J. (2002). Semantic and structural elaboration in L2 lexical acquisition. *Language learning, 52*(2), 323-363. doi 10.1111/0023-8333.00186

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning, 57*(1), 35-56. doi: 10.1111/j.1467-9922.2007.00398.x

Barcroft, J. (2015). Can retrieval opportunities increase vocabulary learning during reading? *Foreign Language Annals, 48*(2), 236-249. doi:10.1111/flan.12139

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

Becker, T. (2014). Avoidance of English phrasal verbs: Investigating the effect of proficiency, learning context, task type, and verb type. *Asian Journal of English Language Teaching, 24*, 1-33.

Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(4), 553-563. doi:10.1037/0278-7393.13.4.553

Bertsch, S., Pesta, B., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*(2), 201-210.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. *Language and Computers, 26*, 181-190.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of written and spoken English*. Harlow: Pearson Education Ltd.

Birjandi, P., Alavi, S. M., & Najafi Karimi, S. (2015). Effects of unenhanced, enhanced, and elaborated input on learning English phrasal verbs. *International Journal of Research Studies in Language Learning, 4*(1). doi:10.5861/ijrsll.2014.772

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The loyola symposium* (pp. 123-144). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimarmura (Eds.), *Metacognition* (pp. 185-205). Cambridge: The MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35-67). Hillsdale, NJ: Erlbaum.

Boers, F. (2000). Metaphor awareness and vocabulary retention. *Applied Linguistics, 21*(4), 553-571. doi:DOI 10.1093/applin/21.4.553

Boers, F. (2013). Cognitive Linguistic approaches to teaching vocabulary: Assessment and integration. *Language Teaching, 46*(02), 208-224. doi:10.1017/S0261444811000450

Boers, F., Dang, T., & Strong, B. (2017). Comparing the effectiveness of phrase-focused exercises, A partial replication of Boers, Demecheleer, Coxhead, and Webb (2014). *Language Teaching Research*, 362-380.

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effects of exercises on verb-noun collocations. *Language Teaching Research, 18*(1), 54-74. doi:10.1177/1362168813505389

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: putting a Lexical Approach to the test. *Language Teaching Research, 10*(3), 245-261. doi:10.1191/1362168806lr195oa

Boers, F., Eyckmans, J., & Stengers, H. (2006). Motivating multiword units: Rationale, mnemonic benefits, and cognitive style variables. *EUROSLA Yearbook, 6*, 169-190. doi: 10.1075/eurosla.6.11boe

Boers, F., Eyckmans, J., & Stengers, H. (2007). Presenting figurative idioms with a touch of etymology: more than mere mnemonics? *Language Teaching Research, 11*(1), 43-62.

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

Boulton, A. (2008). Looking for empirical evidence of data-driven learning at lower levels. In B. Lewandowska-Tomaszczyk (Ed.), *Corpus linguistics, computer tools and applications: State of the art*. Frankfurt: Peter Lang.

Bradshaw, G., & Anderson, J. R. (1982). Elaborative encoding as an explanation of levels of processing. *Journal of Verbal Learning and Verbal Behavior, 21*(2), 165-174. doi: 10.1016/S0022-5371(82)90531-X

Brown, H. D. (2006). *Principles of language learning and teaching* (5 ed.). Chicago: Pearson Education.

Brown, P. C., Roediger, H. L., III., & McDaniel, M. A. (2014). *Make it stick*: Harvard University Press.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*(4), 273-281. doi:10.1037/1076-898X.13.4.273

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(4-5), 514-527. doi:10.1080/09541440701326097

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604-616. doi:10.3758/MC.36.3.604

Butler, D. C., & Peterson, D. E. (1965). Learning during "extinction" with paired associates. *Journal of Verbal Learning and Verbal Behavior, 4*(2), 103-106. doi:10.1016/S0022-5371(65)80092-5

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.

Cacciari, C., & Tabossi, P. (2014). *Idioms: Processing, structure, and interpretation*. New York: Psychology Press.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563-1569. doi:10.1037/a0017021

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279-283. doi:10.1177/0963721412452728

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*(2), 268-276.

Carr, F., & Eales, F. (2005). *New cutting edge upper intermediate: Student's book*. Harlow: Pearson Education.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633-642.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49*(2), 231-248. doi:10.1016/S0749-596x(03)00061-5

Celce-Murcia, M., & Larsen-Freeman, D. (2016). *The Grammar Book: An ESL/EFL Teacher's Course* (3 ed.). Boston: MA: Cengage Heinle.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin, 132*(3), 354-380. doi:10.1037/0033-2909.132.3.354

Chan, T.-P., & Liou, H.-C. (2005). Effects of web-based concordancing instruction on EFL students' learning of verb–noun collocations. *Computer assisted language learning, 18*(3), 231-251.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds Ratios in epidemiological studies. *Communications in statistics - Simulation and Computation, 39*(4), 860-864. doi:10.1080/03610911003650383

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 coursebook: Student's book*. Oxford: MacMillan.

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

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

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

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

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

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

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

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

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

Clare, L., & Jones, R. S. (2008). Errorless learning in the rehabilitation of memory impairment: a critical review. *Neuropsychological Review, 18*(1), 1-23. doi:10.1007/s11065-008-9051-4

Condon, N. (2008). How cognitive linguistic motivations influence the learning of phrasal verbs. In F. Boers & S. Lindstromberg (Eds.), *Cognitive linguistics approaches to teaching vocabulary and phraseology*. New York: Mouton de Gruyter.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics, 32*, 45-61. doi:10.1017/S0267190512000074

Cooper, T. C. (1999). Processing of idioms by L2 learners of English. *TESOL Quarterly, 33*(2), 233-262. doi: 10.2307/3587719

Cornell, A. (1985). Realistic goals in teaching and learning phrasal verbs. *International Review of Applied Linguistics in Language Teaching, 23*(1-4), 269-280. doi: 10.1515/iral.1985.23.1-4.269

Courtney, R. (1983). *Longman dictionary of phrasal verbs*. Cambridge: Longman.

Cowie, A. P. (1998). *Phraseology: Theory, analysis, and applications*: OUP Oxford.

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

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.

Criss, A. H., & Shiffrin, R. M. (2004). Context noise and item noise jointly determine recognition memory: a comment on Dennis and Humphreys (2001). *Psychology Review, 111*(3), 800-807. doi:10.1037/0033-295X.111.3.800

Crutcher, R., & Healy, A. (1989). Cognitive operations and the generation effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(4), 669-675. doi: 10.1037/0278-7393.15.4.669

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*(3), 215-235. doi: 10.1002/(Sici)1099-0720(200005/06)14:3<215::Aid-Acp640>3.0.Co;2-1

Cunningham, S. (2007). *New cutting edge advanced: Student's book.* Harlow: Pearson Education.

Cunningham, S., & Moor, P. (2005). *New cutting edge elementary: Student's book*. Harlow: Pearson Education.

Cunningham, S., & Moor, P. (2007). *New cutting edge pre-intermediate: Student's book.* Harlow: Pearson Education.

Cunningham, S., & Moor, P. (2009). *New cutting edge intermediate: Student's book*. Harlow: Pearson Education.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research, 28*(3), 369-382. doi:10.1177/0267658312443651

Dagut, M., & Laufer, B. (1985). Avoidance of phrasal verbs: A case for contrastive analysis. *Studies in Second Language Acquisition, 7*(01), 73-80. doi:10.1017/S0272263100005167

Darwin, C. M., & Gray, L. S. (1999). Going after the phrasal verb: An alternative approach to classification. *TESOL Quarterly, 33*(1), 65-83. doi: 10.2307/3588191

De Cock, S. (2006). Learners and phrasal verbs. 35. Retrieved from http://www.macmillandictionaries.com/MED-Magazine/February2006/35-Phrasal-Verbs-Learners.htm

De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). *An automated approach to the phrasicon of EFL learners* (S. Granger Ed.). London: Addison Wesley Longman.

Diana, R. A., & Reder, L. M. (2005). The list strength effect: A contextual competition account. *Memory and Cognition, 33*(7), 1289-1302.

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: a three-component model. *Trends Cognition Science, 11*(9), 379-386. doi:10.1016/j.tics.2007.08.001

Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *Journal of Educational Research, 75*(5), 309-313.

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

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

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

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University.

Ellis, R. (1997). *SLA Research and Language Teaching*. Oxford: Oxford University.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse, 20*(1), 29-62.

Evans, J. J., Wilson, B. A., Schuri, U., Andrade, J., Baddeley, A., Bruna, O., Taussik, I. (2000). A comparison of "errorless" and "trial-and-error" learning methods for teaching individuals with acquired memory deficits. *Neuropsychological Rehabilitation: An International Journal, 10*(1), 67-101. doi:10.1080/096020100389309

Evans, V., & Green, M. (2006). *Cognitive Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Eysenck, M., & Keane, M. (2010). *Cognitive psychology: A student's handbook* (Sixth ed.). New York, NY: Psychology Press.

Fazio, L. K., & Marsh, E. J. (2010). Correcting false memories. *Psychological Science, 21*(6), 801-803. doi:10.1177/0956797610371341

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London: Sage Publications Inc.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955 *Studies in Linguistic Analysis. Special Volume, Philological Society* (pp. 1-32). Oxford: Blackwell.

Fitzer, S. C., Caldwell, G. S., Clare, A. S., Upstill-Goddard, R. C., & Bentley, M. G. (2013). Response of copepods to elevated pCO2 and environmental copper as co-stressors--a multigenerational study. *PLoS One, 8*(8), e71257. doi:10.1371/journal.pone.0071257

Folse, K. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly, 40*(2), 273-294.

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. *Researching pedagogic tasks: Second language learning, teaching, and testing*, 75-93.

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

Ganji, M. (2011). The best way to teach phrasal verbs: Translation, sentential contextualization or metaphorical conceptualization? *Theory and Practice in Language Studies, 1*(11), 1497-1506. doi:10.4304/tpls.1.11.1497-1506

Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly, 41*(2), 339-360.

Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research, 19*(6), 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

Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory and Cognition, 8*(2), 149-156.

Gillund, G., & Shiffrin, R. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*(1), 1-67. doi: 10.1037/0033-295x.91.1.1

Glover, J. A. (1989). The" testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392.

González, R. A. (2010). Making sense of phrasal verbs: A cognitive linguistic account of L2 learning. *AILA Review, 23*, 50-71. doi:0.1075/aila.23.04ale

Gordon, L. T., Thomas, A. K., & Bulevich, J. B. (2015). Looking for answers in all the wrong places: How testing facilitates learning of misinformation. *Journal of Memory and Language, 83*, 140-151. doi:10.1016/j.jml.2015.03.007

Graf, P. (1981). Reading and generating normal and transformed sentences. *Canadian Journal of Psychology-Revue Canadienne De Psychologie, 35*(4), 293-308. doi:10.1037/h0081193

Graf, P. (1982). The memorial consequences of generation and transformation. *Journal of Verbal Learning and Verbal Behavior, 21*(5), 539-548. doi: 10.1016/S0022-5371(82)90764-2

Granger, S. (1998). *Prefabricated patterns in advanced EFL writing: Collocations and lexical phrases*. Oxford: OUP.

Grant, L., & Bauer, L. (2004). Criteria for re-defining idioms: Are we barking up the wrong tree? *Applied Linguistics, 25*(1), 38-61.

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(3), 371.

Griffith, D. (1976). The attentional demands of mnemonic control processes. *Memory and Cognition, 4*(1), 103-108. doi:10.3758/BF03213261

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory and Cognition, 40*(4), 505-513. doi:10.3758/s13421-011-0174-0

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory and Cognition, 37*(4), 801-812. doi:10.1037/a0023219

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

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

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

Handcock, M., & McDonald, A. (2010). *English result: Upper-intermediate student's book*. Oxford: Oxford University Press.

Haslam, C., Hodder, K. I., & Yates, P. J. (2011). Errorless learning and spaced retrieval: how do these methods fare in healthy and clinical populations? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 432-447. doi:10.1080/13803395.2010.533155

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics, 4*(2), 237-258.

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 and Cognition, 39*(1), 290-296. doi:10.1037/a0028468

Hirshman, E., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 484-494. doi:10.1037/0278-7393.14.3.484

Howard, M. W., & Kahana, M. J. (2002). When does semantic similarity help episodic retrieval? *Journal of Memory and Language, 46*(1), 85-98.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19*(1), 24-44. doi: 10.1093/applin/19.1.24

Hu, H. C. M., & Nassaji, H. (2016). Effective vocabulary learning tasks: Involvement Load Hypothesis versus Technique Feature Analysis. *System, 56*, 28-39. doi:10.1016/j.system.2015.11.001

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory and Cognition, 40*(4), 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*(3), 241-255.

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language learning, 51*(3), 539-558.

Humphreys, M. S., Wiles, J., & Dennis, S. (1994). Toward a theory of human memory: Data structures and access processes. *Behavioural and Brain Sciences, 17*(04), 655-667.

Hunkin, N. M., Squires, E. J., Parkin, A. J., & Tidy, J. A. (1998). Are the benefits of errorless learning dependent on implicit memory? *Neuropsychologia, 36*(1), 25-36.

Irujo, S. (1986). Don't put your leg in your mouth: Transfer in the acquisition of idioms in a second language. *TESOL Quarterly, 20*(2), 287-304.

Jackendoff, R. (1995). The boundaries of the lexicon. In C. Everaert, E. van der Linden, A. Schenk, & R. Screuder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 133-165). Hillsdale: NJ: Erlbaum.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17*, 649-667.

Jacoby, L. L., Debner, J. A., & Hay, J. F. (2001). Proactive interference, accessibility bias, and process dissociations: Valid subject reports of memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(3), 686.

Jarrold, C., Tam, H., Baddeley, A. D., & Harvey, C. E. (2011). How does processing affect storage in working memory tasks? Evidence for both domain-general and domain-specific effects. *Journal of Experimental Psychology: Learning, Memory and Cognition, 37*(3), 688-705. doi:10.1037/a0022527

Jessen, F., Heun, R., Erb, M., Granath, D. O., Klose, U., Papassotiropoulos, A., & Grodd, W. (2000). The concreteness effect: evidence for dual coding and context availability. *Brain Language, 74*(1), 103-112. doi:10.1006/brln.2000.2340

Johns, E. E., & Swanson, L. G. (1988). The generation effect with nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(1), 180.

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

Jonsson, F. U., Kubik, V., Sundqvist, M. L., Todorov, I., & Jonsson, B. (2014). How crucial is the response format for the testing effect? *Psychology Research, 78*(5), 623-633. doi:10.1007/s00426-013-0522-8

Kane, J. H., & Anderson, R. C. (1978). Depth of processing and interference effects in the learning and remembering of sentences. *Journal of Educational Psychology, 70*(4), 626-635.

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*(6), 1544-1550. doi:10.3758/s13423-014-0636-z

Kang, S. H., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*(4-5), 528-558. doi:10.1080/09541440601056620

Kang, S. H. K., Lindsey, R. V., & Mozer, M. C. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval? *Psychonomeic Bullet Review, 21*(6), 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*(5), 1250-1257. doi:10.1037/a0023436

Karpicke, J. D., Blunt, J. R., Smith, M. A., & Karpicke, S. S. (2014). Retrieval-based learning: The need for guided retrieval in elementary school children. *Journal of Applied Research in Memory and Cognition, 3*(3), 198-206. doi:10.1016/j.jarmac.2014.07.008

Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review, 24*(3), 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.

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*(2), 151-162. doi:10.1016/j.jml.2006.09.004

Karpicke, J. D., & Roediger, H. L., 3rd. (2008). The critical importance of retrieval for learning. *Science, 319*, 966-968. doi:10.1126/science.1152408

Karpicke, J. D., & Roediger, H. L. I. (2008). The critical importance of retrieval for learning. *Science, 319*, 966-968.

Karpicke, J. D., & Smith, M. A. (2011). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*(1), 17-29. doi:10.1016/j.jml.2012.02.004

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62*(3), 227-239. doi:10.1016/j.jml.2009.11.010

Kay, H. (1955). Learning and retaining verbal material. *British Journal of Psychology, 46*(2), 81-100.

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

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

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

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

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

Ke, Y. (2016). A bi-axis model for profiling English phrasal verbs for pedagogic purposes. *TESOL Quarterly*. doi:10.1002/tesq.345

Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research, 12*(3), 365-386. doi:10.1177/1362168808089922

Kecskes, I. (2008). Dueling contexts: A dynamic model of meaning. *Journal of Pragmatics, 40*(3), 385-406.

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.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language learning, 58*(2), 285-325.

Kinnell, A., & Dennis, S. (2011). The list length effect in recognition memory: An analysis of potential confounds. *Memory & Cognition, 39*(2), 348-363. doi:10.3758/s13421-010-0007-6

Knight, J. B., Hunter Ball, B., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language, 66*(4), 731-746. doi:10.1016/j.jml.2011.12.008

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65*(2), 85-97. doi:10.1016/j.jml.2011.04.002

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 35*(4), 989-998. doi:10.1037/a0015729

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(1), 283-294. doi:10.1037/a0037850

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning. *Psychology of learning and motivation, 65*, 183-215. doi:10.1016/bs.plm.2016.03.003

Kövecses, Z., & Szabó, P. (1996). Idioms: A view from cognitive semantics. *Applied Linguistics, 17*(3), 326-355.

Krashen, S. (1977). The monitor model for adult second language performance. In M. Burt, H. Dulay, & M. Finocchioaro (Eds.), *Viewpoints on English as a second language* (pp. 152-161). New York: Regents.

Krashen, S. (2013). *Second language acquisition: Theory, applications and some conjectures*. Mexico City, Mexico: Cambridge University Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Laufer, B. (2000). Avoidance of idioms in a second language: The effect of L1-L2 degree of similarity. *Studia Linguistica, 54*(2), 186-196.

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.

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics, 29*(4), 694-716. doi:10.1093/applin/amn018

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language learning, 54*(3), 399-436. doi: 10.1111/j.0023-8333.2004.00260.x

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

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 6*(2), 647-672. doi:10.1111/j.1467-9922.2010.00621.x

LeBlanc, D. C. (2004). *Statistics: concepts and applications for science*. Sudbury: Jones & Bartlett Learning.

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review, 120*(1), 155-189. doi:10.1037/a0030851

Lessard-Clouston, M. (1993). Catching on: Understanding phrasal verbs for ELT. *ELI Teaching: A Journal of Theory and Practice, 15*(5-9).

Lewis, M. (1993). *The lexical approach: The state of ELT and the way forward.* Hove, England: Language Teaching Publications.

Liao, Y., & Fukuya, Y. (2002). Avoidance of phrasal verbs: The case of Chinese learners of English. *Second Language Studies, 20*(2), 71-106.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning, 65*(S1), 185-207. doi:10.1111/lang.12117

Lindner, S. J. (1981). *A lexico-semantic analysis of English verb particle constructions with out and up.* (PhD), University of California, San Diego.

Liu, D. (2010). Going beyond patterns: Involving cognitive analysis in the learning of collocations. *TESOL Quarterly, 44*(1), 4-30. doi:10.5054/tq.2010.214046

Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly, 45*(4), 661-688. doi:10.5054/tq.2011.247707

Loaiza, V. M., McCabe, D. P., Youngblood, J. L., Rose, N. S., & Myerson, J. (2011). The influence of levels of processing on recall from working memory and delayed recall tasks. *Journal of Experimental Psychology: Learning, Memory and Cognition, 37*(5), 1258-1263. doi:10.1037/a0023923

Maclin, A. (1987). *Reference guide to English: A handbook of English as a second language* (second ed.). New York: Holt, Rinehart and Winston.

Marsh, E. J., Roediger, H. L., 3rd, Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomics Bullet Review, 14*(2), 194-199.

Mart, C. (2012). How to teach phrasal verbs. *English Language Teaching, 5*(6). doi:10.5539/elt.v5n6p114

Martinez, R., & Murphy, V. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly, 45*(2), 267-290. doi:10.5054/tq.2011.247708

Matlock, T., & Heredia, R. R. (2002). Understanding phrasal verbs in monolinguals and bilinguals. *Advances in Psychology, 134*, 251-274.

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

McDaniel, M. A., & Masson, M. E. (1977). Long-term retention: When incidental semantic processing fails. *Journal of Experimental Psychology: Human Learning and Memory, 3*(3), 270-281. doi: 10.1037//0278-7393.3.3.270

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(2), 371-385. doi:10.1037/0278-7393.11.2.371

McDaniel, M. A., & Waddill, P. J. (1988). A contextual account of the generation effect: A three-factor theory. *Journal of Memory and Language, 27*, 521-536.

McElroy, L. A., & Slamecka, N. J. (1982). Memorial consequences of generating nonwords: Implications for semantic-memory interpretations of the generation effect. *Journal of Verbal Learning and Verbal Behavior, 21*(3), 249-259.

McKoon, G., & Ratcliff, R. (1989). Semantic associations and elaborative inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(2), 326.

Melton, A. (1963). Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavior, 2*(1), 1-21. doi: 10.1016/S0022-5371(63)80063-8

Middleton, E. L., & Schwartz, M. F. (2012). Errorless learning in cognitive rehabilitation: a critical review. *Neuropsychological Rehabilitation: An International Journal, 22*(2), 138-168. doi:10.1080/09602011.2011.639619

Morgan, P. S. (1997). Figuring out figure out: Metaphor and the semantics of the English verb-particle construction. *Cognitive Linguistics, 8*(4), 327-357. doi: 10.1515/cogl.1997.8.4.327

Morris, C., Bransford, J., & Franks, J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519-533. doi: 10.1016/S0022-5371(77)80016-9

Nairne, J. S. (2002). The myth of the encoding-retrieval match. *Memory, 10*(5-6), 389-395. doi:10.1080/09658210244000216

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*(4), 677-711.

Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? *Studies in second language acquisition, 38*(03), 523-552.

Nassaji, H., & Tian, J. (2010). Collaborative and individual output tasks and their effects on learning English phrasal verbs. *Language Teaching Research, 14*(3), 397-419. doi:10.1177/1362168810375364

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P., & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Boston: MA: Heinle, Cengage Learning.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223-242. doi: 10.1093/applin/24.2.223

Nesselhauf, N. (2005). *Collocations in a learner corpus* (Vol. 14): John Benjamins Publishing.

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

O'Brien, E. J., Shank, D. M., Myers, J. L., & Rayner, K. (1988). Elaborative inferences during reading: Do they occur on-line? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 410.

Oxenden, C., & Latham-Koenig, C. (2006). *New English file: Intermediate student's book*. New York: Oxford University Press.

Oxenden, C., & Latham-Koenig, C. (2008). *New English file: Upper-intermediate student's book*. New York: Oxford University Press.

Oxenden, C., & Latham-Koenig, C. (2008). *New total English: Upper-intermediate student's book*. Harlow: Pearson Education.

Oxenden, C., & Latham-Koenig, C. (2009). *New English file: Beginner student's book*. New York: Oxford University Press.

Oxenden, C., & Latham-Koenig, C. (2010). *New English file: Advanced student's book*. New York: Oxford University Press.

Oxenden, C., Latham-Koenig, C., & Seligson, P. (2006). *New English file: Pre-intermediate student's book*. New York: Oxford University Press.

Oxenden, C., Latham-Koenig, C., & Seligson, P. (2006). *New English file: Elementary student's book*. New York: Oxford University Press.

Oxenden, C., Latham-Koenig, C., & Seligson, P. (2006). *New English file: Elementary student's book*. New York: Oxford University Press.

Page, M., Wilson, B. A., Shiel, A., Carter, G., & Norris, D. (2006). What is the locus of the errorless-learning advantage? *Neuropsychologia, 44*(1), 90-100. doi:10.1016/j.neuropsychologia.2005.04.004

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1051-1057. doi:10.1037/0278-7393.29.6.1051

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191-225). New York: Longman.

Pellicer-Sánchez, A. (2015). Learning L2 collocations incidentally from reading. *Language Teaching Research*, 1-22.

Peters, E. (2012). Learning German formulaic sequences: The effect of two attention-drawing techniques. *The Language Learning Journal, 40*(1), 65-79. doi:10.1080/09571736.2012.658224

Potts, R. (2013). *Memory interference and the benefits and costs of testing.* UCL (University College London).

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of experimental Psychology: general, 143*(2), 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*(8), 1917-1927.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*(4), 437-447.

Pye, G. (1996). *Don't give up, look it up! Defining phrasal verbs for the learner of English.*

Raaijmakers, J., & Shiffrin, R. (1981). Search of associative memory. *Psychological Review, 88*(2), 93-134. doi: 10.1037//0033-295x.88.2.93

Rajaram, S., & Roediger, H. L. (1993). Direct comparison of four implicit memory tests. *journal of experimental psychology learning memory and cognition, 19*(4), 765-765.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 163-178.

Ravenscroft, J., Liakata, M., Clare, A., & Duma, D. (2017). Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements. *PLoS One, 12*(3), e0173152. doi:10.1371/journal.pone.0173152

Richards, J., & Bohlke, D. (2011). *Four corners*. Cambridge: Cambridge University Press.

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15*(3), 243-257. doi:10.1037/a0016496

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

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences, 15*(1), 20-27. doi:10.1016/j.tics.2010.09.003

Roediger, H. L., III, & Karpicke, J. (2006). The power of testing memory: basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210. doi:10.1111/j.1745-6916.2006.00012.x

Rott, S. (1999). The effects of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in second language acquisition, 21*(04), 589-619.

Rudzka-Ostyn, B. (2003). *Word power: Phrasal verbs and compounds*. Berlin: Walter de Gruyter GmbH & Co.

Schachter, J. (1974). An error in error analysis. *Language learning, 24*(2), 205-214.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207-217.

Schmitt, N. (2013). Formulaic language and collocation. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*: Blackwell Publishing Ltd.

Schmitt, N., & Redwood, S. (2011). Learner knowledge of phrasal verbs: A corpus-informed study. In F. Meunier, S. D. Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 173-209): John Benjamins Publishing Company.

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

Schooler, J., Foster, R., & Loftus, E. (1988). Some deleterious consequences of the act of recollection. *Memory & Cognition, 16*(3), 243-251. doi: 10.3758/Bf03197757

Schouten-van, P. (1989). Vocabulary learning through reading: Which conditions should be met when presenting words in texts? In P. Nation & R. Carter (Eds.), *Vocabulary acquisition* (pp. 75-85). Amsterdam: Free University Press.

Schunk, D. (2012). *Learning theories: An educational perspective* (Vol. 6). Boston: Pearson Education Inc.

Shiffrin, R., Ratcliff, R., Murnane, K., & Nobel, P. (1993). TODAM and the list-strength and list-length effects: comment on Murdock and Kahana (1993a). *Journal of Experimental Psychology: Applied, 19*(6), 1445-1449.

Side, R. (1990). Phrasal verbs: Sorting them out. *ELT journal, 44*(2), 144-152.

Sinclair, J. (1991). *Corpus, concordance, collocation*: Oxford University Press.

Siyanova, A., & Schmitt, N. (2007). Native and nonnative use of multi-word vs. one-word verbs. *IRAL-International Review of Applied Linguistics in Language Teaching, 45*(2), 119-139.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27*(2), 251-272. doi:10.1177/0267658310382068

Siyanova-Chanturia, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics, 36*(5), 549-569. doi:10.1093/applin/amt054

Slamecka, N., & Fevreiski, J. (1983). The generation effect when generation fails. *Journal of Verbal Learning and Verbal Behavior, 22*(2), 153-163. doi: 10.1016/S0022-5371(83)90112-3

Slamecka, N., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592-604.

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language, 26*(6), 589-607. doi:10.1016/0749-596x(87)90104-5

Slobin, D. I. (2005). Linguistic representations of motion events: What is signifier and what is signified. In C. Maeder, O. Fischer, & W. Herlofsky (Eds.), *Iconicity inside out: Iconicity in language and literature* (Vol. 4, pp. 307-322). Amsterdam: John Benjamins.

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory, 22*(7), 784-802. doi:10.1080/09658211.2013.831454

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*(6), 1712. doi:10.1037/a0033569

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

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

Soars, J., & Soars, V. (2011). *New headway: Elementary student's book*. New York: Oxford University Pres.

Soars, L., & Soars, J. (2008). *New headway: Advanced student' s book*. New York: Oxford University Press.

Soars, L., & Soars, L. (2003). *New headway: Intermediate student's book*. New York: Oxford University Press.

Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions. *Language learning, 63*(1), 121-159. doi:10.1111/j.1467-9922.2012.00730.x

Staresina, B. P., & Davachi, L. (2006). Differential encoding mechanisms for subsequent associative recognition and free recall. *Journal of Neuroscience, 26*(36), 9162-9172. doi:10.1523/JNEUROSCI.2877-06.2006

Stengers, H., & Boers, F. (2015). Exercises on collocations: A comparison of trial-and-error and exemplar-guided procedures. *Journal of Spanish Language Teaching, 2*(2), 152-164.

Stengers, H., Boers, F., Housen, A., & Eyckmans, J. (2010). Does 'chunking' foster chunk-uptake. In S. De Knop, F. Boers, & A. De Rycker (Eds.), *Fostering language teaching efficiency through cognitive linguistics* (pp. 99-117). New York: Walter de Gruyter GmbH & Co

Storm, B. C. (2011). The benefit of forgetting in thinking and remembering. *Current Directions in Psychological Science, 20*(5), 291-295. doi:10.1177/0963721411418469

Strong, B. (2013). A cognitive semantic approach to L2 learning of phrasal verbs. *The Language Teacher, 37*(5).

Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior, 18*(5), 523-534.

Taevs, M., Dahmani, L., Zatorre, R. J., & Bohbot, V. D. (2010). Semantic elaboration in auditory and visual spatial memory. *Frontiers in Psychology, 1*, 228. doi:10.3389/fpsyg.2010.00228

Talmy, L. (2000). *Toward a cognitive semantics* (Vol. 2): MIT press.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*(1), 84-119. doi: 10.1093/applin/17.1.84

Trebits, A. (2009). The most frequent phrasal verb in English language EU documents — A corpus-based analysis and its implications. *System, 37*(3), 470-481. doi:0.1016/j.system.2009.02.012

Tulving, E. (1972). Episodic and semantic memory 1. *Organization of Memory. London: Academic, 381*(4), 382-404.

Tulving, E. (1974). Recall and recognition of semantically encoded words. *Journal of Educational Psychology, 102*(5), 778-787. doi: 10.1037/h0036383

Tulving, E. (1986). What kind of a hypothesis is the distinction between episodic and semantic memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(2), 307-311. doi:10.1037/0278-7393.12.2.307

Tulving, E. (1991). Memory research is not a zero-sum game. *American Psychologist, 46*, 41-42.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53*(1-15), 1-25. doi:10.1146/annurev.psych.53.100901.135114

Tulving, E., & Osler, S. (1968). Effectiveness of retrieval cues in memory for words. *Journal of Experimental Psychology, 77*(4), 593-601.

Tulving, E., & Thomson, D. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*(5), 352-373.

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review, 19*(5), 899-905. doi:10.3758/s13423-012-0276-0

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15*(2), 130-163.

Warmington, M., & Hitch, G. J. (2014). Enhancing the learning of new words using an errorless learning procedure: evidence from typical adults. *Memory, 22*(5), 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*(3), 456-465. doi:10.1016/j.jecp.2012.10.007

Webb, S. (2012). Learning vocabulary in activities. In H. P. Widodo & A. Cirocki (Eds.), *Innovation and creativity in ELT methodology*. New York: Nova Science Publishers, Inc.

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning, 63*(1), 91-120. doi:10.1111/j.1467-9922.2012.00729.x

Wood, D. (2002). Formulaic language acquisition and production: Implications for teaching. *TESL Canada Journal, 20*(1), 01-15.

Wood, D. (2006). Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *Canadian Modern Language Review, 63*(1), 13-33. doi: 10.1353/cml.2006.0051

Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *The Canadian Journal of Applied Linguistics, 12*(1), 39.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications*: Bloomsbury Publishing.

Wyss, R. (2003). Putting phrasal verbs into perspective. *TESOL Journal, 12*(37), 38.

Yasuda, S. (2010). Learning phrasal verbs through conceptual metaphors: A case of Japanese EFL learners. *TESOL Quarterly, 44*(2), 250-273. doi:10.5054/tq.2010.219945

Yoshitomi, A. (2006). The use of phrasal verbs by Japanese learners of English: Implications from storytelling data. In A. Yoshitomi, T. Umino, & M. Negishi (Eds.), *Readings in second language pedagogy and second language acquisition: In Japanese context*. Amsterdam: John Benjamin's Publishing.

APPENDICES

Appendix I

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON

Participant Consent Form

Research Project Title
An Examination of Exercises Designed for Learning Prepositional and Phrasal Verbs: The Effectiveness of Trial-and-Error and Error-Free Practice.

Researcher
Brian Strong, School of Linguistics and Applied Language Studies, Victoria University of Wellington

I have been given and have understood an explanation of this research project. I have had an opportunity to ask questions and have them answered to my satisfaction.

I understand that I may withdraw myself (or any information I have provided) from this project, without having to give reasons, by e-mailing brian.strong@vuw.ac.nz.

I understand that any information I provide will be kept confidential to the researcher and their supervisor, the published results will not use my name, and that no opinions will be attributed to me in any way that will identify me.

I understand that the data I provide will not be used for any other purpose or released to others.

I understand that, if this interview is audio recorded, the recording and transcripts of the interviews will be erased within 2 years after the conclusion of the project. Furthermore, I will have an opportunity to check the transcripts of the interview.

Please indicate (by ticking the boxes below) which of the following apply:

☐ I would like to receive a summary of the results of this research when it is completed.

☐ I agree to this interview being audio recorded.

Signed:

Name of participant:

Date:

Appendix II

Experimental study 1: Target phrasal verbs

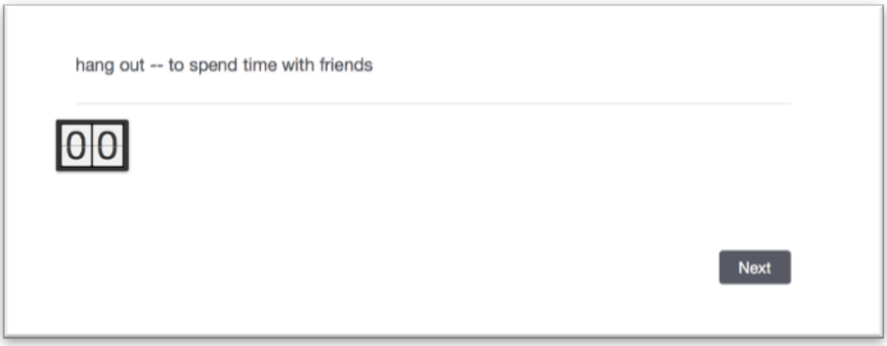
Phrasal verb	Definition
Back down	To decide not to do something
Catch on	Something becomes popular quickly
Dive in	To start eating food
Figure out	To understand something with difficulty
Get out	A secret becomes known unintentionally
Hang out	To spend time with friends
Hold up	To make something late
Make up	To create a story
Nod off	To fall asleep unintentionally
Own up	To admit doing something wrong
Pop in	To visit for a short time
Rip off	To charge too much money
Run out	To use all of something
Screw up	To make a mistake
Boil down	To give the most important information
Chip in	To give some money
Crack on	To continue to do something quickly
Give in	To accept that you cannot win
Brush up	To improve your skill
Open up	To talk about your personal feelings
Pass away	To die

Wrap up	To finish something
Head off	To go somewhere
Turn off	To lose interest in something

Appendix III

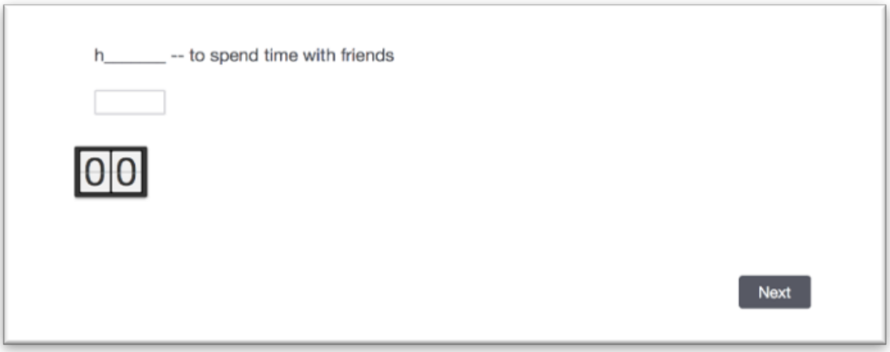
An example of the retrieval condition used in experimental study 1.

In snapshot 1, participants were presented with a phrasal verb and the paraphrase of its meaning. A countdown timer was displayed below the phrasal verb. In the treatment, the timer was set to 15 seconds. In snapshot 2, participants were asked to recall the phrasal verb given its first letter and the paraphrase of its meaning. Their responses were typed in the text box.



hang out -- to spend time with friends

Snapshot 1



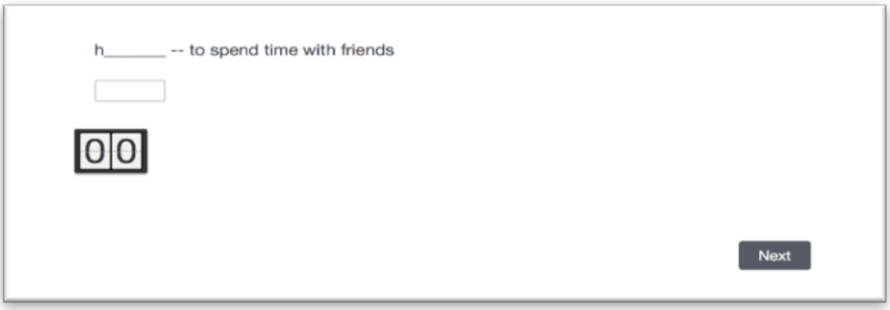
h______ -- to spend time with friends

Snapshot 2

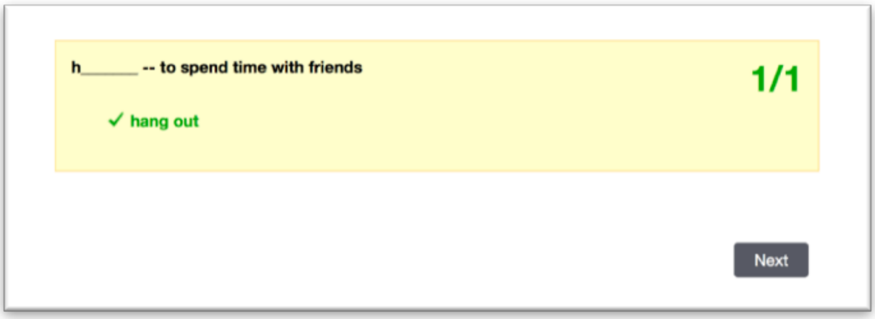
Appendix IV

An example of the generation condition used in experimental study 1

In snapshot 3, participants were asked to guess the phrasal verb given its first letter and the paraphrase of its meaning. They were to type in their response in the text box. A countdown timer was displayed below the text box. In snapshot 4, if participants produced the correct answer, their response was highlighted in green, and a green check mark was presented next to it. In snapshot 5, if participants produced the wrong answer, their response was highlighted in red, and a red X was placed next to it. Below their error, the correct response, highlighted in green, was displayed.

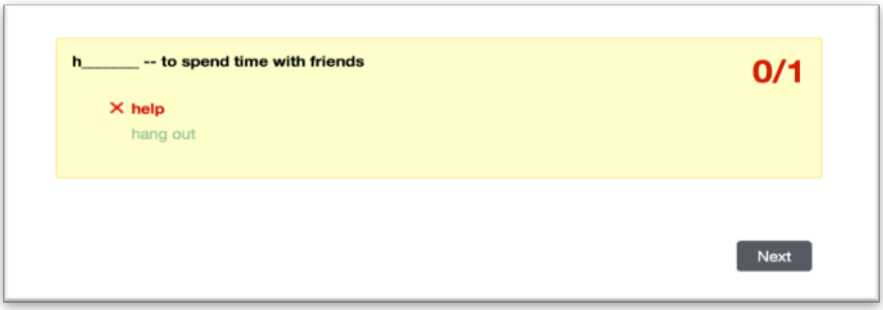

h______ -- to spend time with friends

Snapshot 3


h______ -- to spend time with friends

✓ hang out

Snapshot 4


h______ -- to spend time with friends

✕ help
hang out

Snapshot 5

Appendix V

Experimental study 2: Target phrasal verbs

Phrasal verb	Definition
Back down	To decide not to do something
Catch on	Something becomes popular quickly
Dive in	To start eating food
Figure out	To understand something with difficulty
Get out	A secret becomes known unintentionally
Hang out	To spend time with friends
Hold up	To make something late
Make up	To create a story
Nod off	To fall asleep unintentionally
Own up	To admit doing something wrong
Pop in	To visit for a short time
Rip off	To charge too much money
Run out	To use all of something
Screw up	To make a mistake
Boil down	To give the most important information
Brighten up	To become happier
Chip in	To give some money

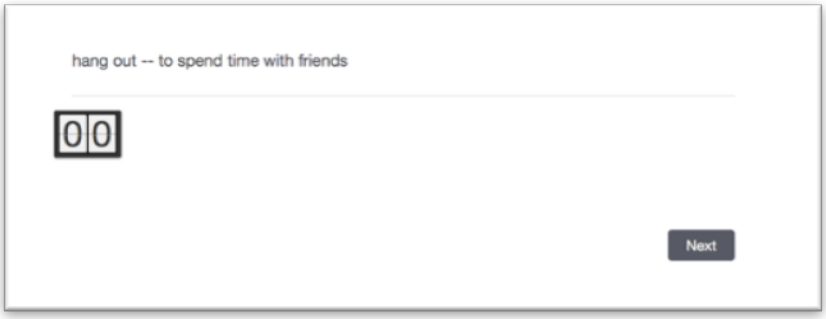
Crack on	To continue to do something quickly
Give in	To accept that you cannot win
Brush up	To improve your skill
Open up	To talk about your personal feelings
Pass away	To die
Call off	To decide that something will not take place
Stick out	To easily notice a feature of something
Show off	To act in a way to attract people's attention
Wrap up	To finish something
Head off	To go somewhere
Turn off	To lose interest in something

Appendix VI

Experimental study 2: study-test and study-delay-test conditions

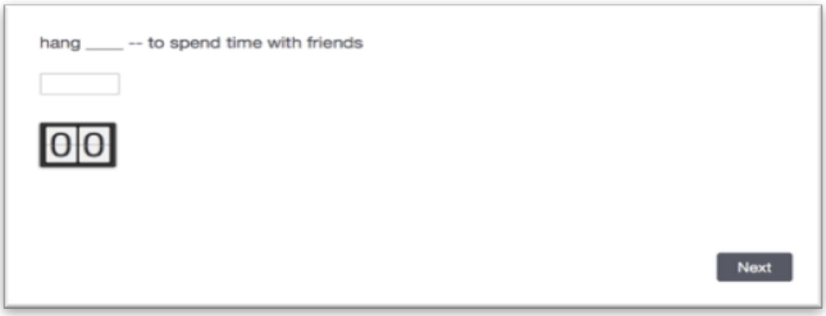
Study-test condition. In snapshot 1, participants were presented with a phrasal verb and the paraphrase of its meaning. A countdown timer was displayed below the phrasal verb. In the treatment, the timer was set to 15 seconds. In snapshot 7, participants were asked to recall the particle given the verb and the paraphrase of the meaning of the phrasal verb. Their responses were typed in the text box.

Study-delay-test condition. Participants were presented with 14 phrasal verbs and their paraphrases of their meanings. After they were asked to recall 14 particles given the verbs and the paraphrases of the meanings of the phrasal verbs



hang out -- to spend time with friends

Snapshot 6



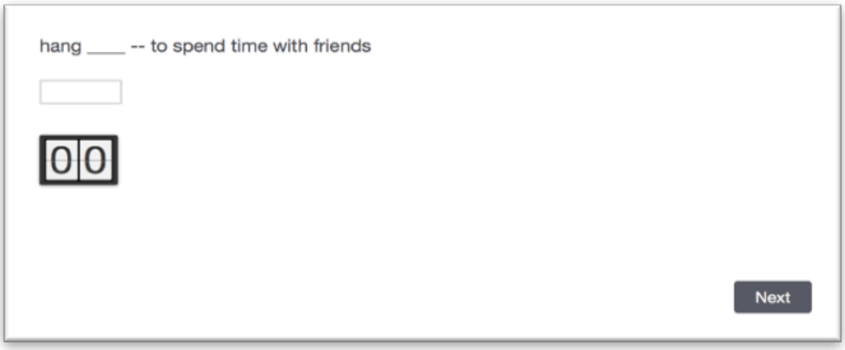
hang ____ -- to spend time with friends

Snapshot 7

Experimental study 2: Test-study and Test-delay-study condition

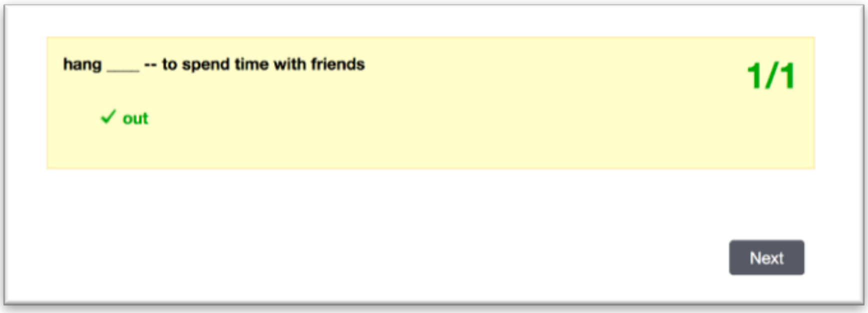
Test-study condition. Snapshot 8 illustrates an example the test trial. It shows the verb and the meaning of a phrasal verb. The instructions asked participants to provide the missing particle in the text box. Snapshot 8 presents an example of the study trial. It shows the missing particle along with the cue. Participants received this type of study trial if they produced the correct particle. Snapshot 8 illustrates an example of the type of study trial participants received if they did not produce the correct particle.

Test-delay-study condition. Instead of a single cue, participants were shown 14 cues in the test trial. Instead of a presentation of a study trial on a single PV, they were shown correct feedback on 14 PVs.



hang ____ -- to spend time with friends

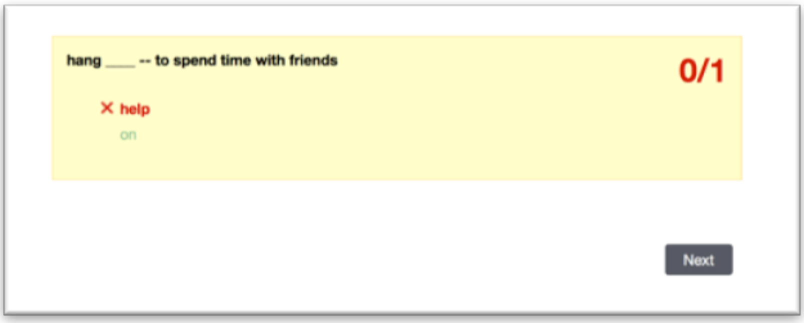
Snapshot 8



hang ____ -- to spend time with friends

✓ out

Snapshot 9



hang ____ -- to spend time with friends

0/1

✕ help

on

Snapshot 10

Appendix VII

Experimental study 3: Target phrasal verbs presented in the study trial of the retrieval condition and the feedback trial in the generation condition

SET A

1) *Catch on: something becomes popular quickly*
    a) Speaker A: Apple just released a new product called the Apple Watch. Do you think it'll catch on?
    b) Speaker B: Yes, I do. I'm actually planning to buy one. Aren't you?
2) *Run out: to use all of something*
    a) Teacher: Your time to complete the English test has run out. Please stop writing.
    b) Student: Oh, but I'm not finished. Could you give me 5 minutes?
3) *Hang out: to spend time with friends*
    a) Speaker A: Hey, Yuki. If you're not busy after work, do you want to hang out?
    b) Speaker B: I'm sorry Tomoko, but I'm not feeling well. How about tomorrow?
4) *Pass away: to die*
    a) Speaker A: How's your grandmother doing? I heard she was very sick.
    b) Speaker B: Yeah, she's in the hospital. Doctors think that she might pass away anytime soon.
5) *Brush up: to improve your skill*
    a) Speaker A: Mayu, why are you taking this English communication class? You already speak English like a native speaker.
    b) Speaker B: Thanks. But, it has been a long time since I spoke English. I want to brush up my English before I go to Canada.
6) *Head off: to go somewhere*
    a) Speaker A: Mr. Yamada, could you give this box to Ms. Nakata before you head off.

   b) Speaker B: No problem. Her office number is 204, right?

7) *Give in: to accept that you cannot win*

   a) Speaker A: My wife and son have been telling me that we need to buy a new car. I keep telling them that the car we have is fine.

   b) Speaker B: I think you should just give in. You really do need a new car.

SET B

1) *Open up: to start to talk about your personal feelings*

   a) Speaker A: You seem sad. Why don't you tell me what's wrong?

   b) Speaker B: Thanks, but I'm not ready to open up.

2) *Drop out: to leave something before it is finished*

   a) Speaker A: Too many students drop out of university after only one year.

   b) Speaker B: Yeah, I wonder why?

3) *Blow up: to become angry quickly*

   a) Speaker A: I'll blow up if you say bad things about my mother.

   b) Speaker B: Sorry. I won't tell any jokes about your mother.

4) *Chicken out: to not do something because you are scared*

   a) Speaker A: I want to try to surf, but I'm afraid of sharks in the water.

   b) Speaker B: Don't chicken out. Sharks won't attack you. Trust me!-

5) *Hold on: to wait for a short time*

   a) Speaker A: We'll hold on for another minute, then we have to leave.

   b) Speaker B: Don't worry. I'll be there very soon.

6) *Pop in: to visit for a short time*

   a) After the movie, why don't you and Ted come to my place?

   b) Maybe we'll pop in after the movie.

7) *Back up: to tell someone to repeat something they said*

   a) Speaker A: I found $500 in a bag on the street by my house!

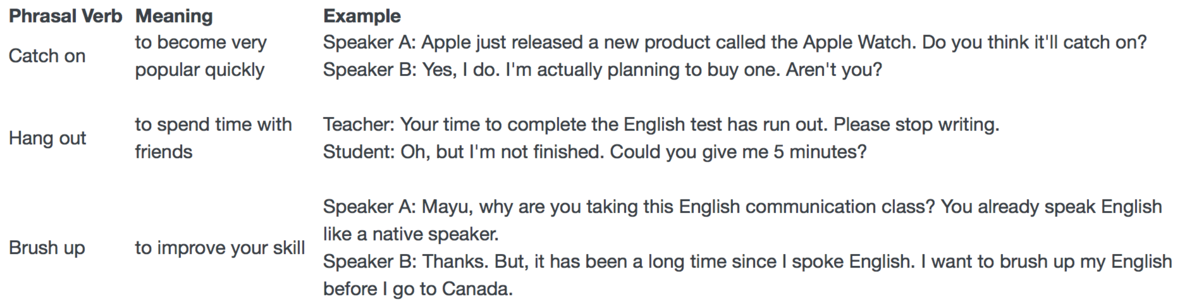
      Speaker B: Back up: Did you say you found money?

Appendix VIII

Example of the treatments:

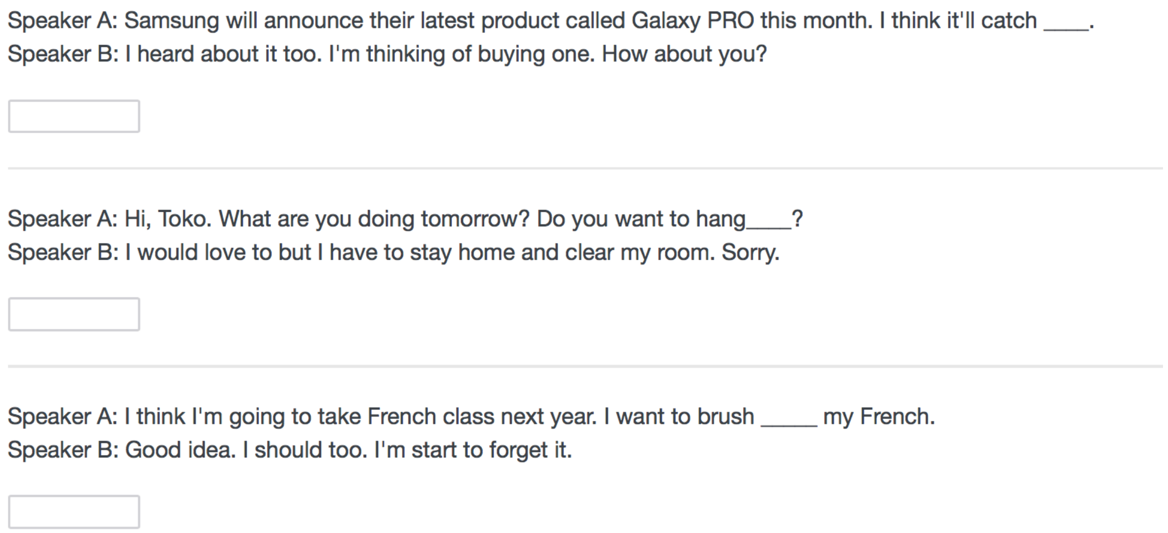
The retrieval group received the input trial before the test trial. The generation group received the test trial before the study trial.

Study trial



Phrasal Verb	Meaning	Example
Catch on	to become very popular quickly	Speaker A: Apple just released a new product called the Apple Watch. Do you think it'll catch on? Speaker B: Yes, I do. I'm actually planning to buy one. Aren't you?
Hang out	to spend time with friends	Teacher: Your time to complete the English test has run out. Please stop writing. Student: Oh, but I'm not finished. Could you give me 5 minutes?
Brush up	to improve your skill	Speaker A: Mayu, why are you taking this English communication class? You already speak English like a native speaker. Speaker B: Thanks. But, it has been a long time since I spoke English. I want to brush up my English before I go to Canada.

Test trial



Speaker A: Samsung will announce their latest product called Galaxy PRO this month. I think it'll catch ____.
Speaker B: I heard about it too. I'm thinking of buying one. How about you?

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.

Speaker A: I think I'm going to take French class next year. I want to brush _____ my French.
Speaker B: Good idea. I should too. I'm start to forget it.

Appendix IX

ESL/EFL course textbooks in the sample.

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

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

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

Handcock, M., & McDonald, A. (2010). *English result: Upper-intermediate student's book*. Oxford: Oxford University Press.

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

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

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

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

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

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

Cunningham, S. (2007). *New cutting edge advanced: Student's book*. Harlow: Pearson Education.

Cunningham, S., & Moor, P. (2005). *New cutting edge elementary: Student's book*. Harlow: Pearson Education.

Cunningham, S., & Moor, P. (2009). *New cutting edge intermediate: Student's book*. Harlow: Pearson Education.

Cunningham, S., & Moor, P. (2007). *New cutting edge pre-intermediate: Student's book*. Harlow: Pearson Education.

Carr, F., & Eales, F. (2005). *New cutting edge upper intermediate: Student's book*. Harlow: Pearson Education.

Oxenden, C., & Latham-Koenig, C. (2010). *New english file: Advanced student's book*. New York: Oxford University Press.

Oxenden, C., & Latham-Koenig, C. (2009). *New english file: Beginner student's book*. New York: Oxford University Press.

Oxenden, C., Latham-Koenig, C., & Seligson, P. (2006). *New english file: Elementary student's book*. New York: Oxford University Press.

Oxenden, C., Latham-Koenig, C., & Seligson, P. (2006). *New english file: Elementary student's book*. New York: Oxford University Press.

Oxenden, C., & Latham-Koenig, C. (2006). *New english file: Intermediate student's book*. New York: Oxford University Press.

Oxenden, C., Latham-Koenig, C., & Seligson, P. (2006). *New english file: Pre-intermediate student's book*. New York: Oxford University Press.

Oxenden, C., & Latham-Koenig, C. (2008). *New english file: Upper-intermediate student's book*. New York: Oxford University Press.

Soars, L., & Soars, J. (2008). *New headway: Advanced student's book*. New York: Oxford University Press.

Soars, J., & Soars, V. (2003). *New headway: Elementary student's book*. Oxford: Oxford University Pres.

Soars, J., & Soars, V. (2011). *New headway: elementary student's book*. New York: Oxford University Pres.

Soars, L., & Soars, L. (2003). *New headway: Intermediate student's book*. New York: Oxford University Press.

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

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

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. (2008). *New inside out: Elementary student's book*. Oxford: MacMillan.

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

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

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

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

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

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

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

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

Oxenden, C., & Latham-Koenig, C. (2008). *New total English: Upper-intermediate student's book*. Harlow: Pearson Education.

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

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

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

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

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

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

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

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

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

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

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

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

Blank page