Effectiveness of Learning Single Words Versus Words in Collocations and the Role of Imageability

Charles M. Mueller

Researchers of L2 vocabulary acquisition generally agree that incidental learning, especially through extensive reading (for a review, see Nakanishi, 2015), is effective for acquiring new words and consolidating learning (Nation & Webb, 2011; Webb, 2007a), with some researchers (e.g., McQuillan & Krashen, 2008) even arguing that this is all that is needed to develop an adequate vocabulary. Yet strong arguments have also been put forth that incidental learning needs to be augmented by a program of deliberate study (Cobb, 2007; Grabe, 2009; Webb & Nation, 2017).

A key issue when developing an effective approach to deliberate study is the appropriate size of the linguistic unit to be targeted. Should learners attempt to memorize a single word or a word embedded in the context of a collocation, long phrase, sentence, or passage? Previous research has not provided a conclusive answer to this important question. The current study therefore compares the efficacy of paired associate learning of words in isolation and in the context of a collocation in an experiment in which the items are rigorously controlled but the tasks have high ecological validity. The effectiveness of vocabulary learning strategies is also likely to be affected by the type of word being studied. The study therefore includes, as an additional variable, imageability, as this has not been explored as a potential moderator variable in previous research on this issue. The paper's introduction reviews theoretical frameworks and previous empirical findings relevant to the word versus phrase issue. It then reports the results of two experiments that examine retrieval and retention of vocabulary studied in isolation or in collocations. The conclusion situates the results in the context of previous

findings and considers theoretical and pedagogical implications.

BACKGROUND

Many frameworks used to account for SLA vocabulary learning are based on observations that elaboration can facilitate memorization within certain contexts (Griffin, 1992, Ch. 4). The levels of processing theory of Craik and Lockhart (1972), for example, claims that tasks requiring the learner to engage in deeper analysis of a stimulus lead to a more durable trace. Yet as pointed out by the theory's detractors (e.g., Eysenck, 1978), the concept of depth may be too vague to be useful as a theoretical construct.

In SLA research on vocabulary learning, several popular frameworks have attempted to specify the concept of depth in ways conducive to operationalization and experimental testing. For example, Technique Feature Analysis (TFA), a framework proposed by Nation and Webb (2011), introduces multiple criteria to model effective word learning tasks. According to TFA, depth of processing in vocabulary learning can be evaluated in terms of five components: (1) noticing, (2) retrieval, (3) generation, (4) motivation, and (4) retention. These five features can be assessed through 18 criteria corresponding to the questions in Table 1. In scoring a task, a point can be given for each "yes" response, leading to a total score ranging from 0 to 18.

Table 1

Technique Feature Analysis (adapted from Nation & Webb, 2011, p. 7)

#	Criteria
	Motivation
1	Is there a clear vocabulary learning goal?
2	Does the activity motivate learning?
3	Do the learners select the words?
	Noticing
4	Does the activity focus attention on the target words?
5	Does the activity raise awareness of new vocabulary learning?
6	Does the activity involve negotiation?
	Retrieval
7	Does the activity involve retrieval of the word?
8	Is it productive retrieval?
9	Is it recall?
10	Are there multiple retrievals of each word?
11	Is there spacing between retrievals?
	Generation
12	Does the activity involve generative use?
13	Is it productive?
14	Is there a marked change that involves the use of other words?
	Retention
15	Does the activity ensure successful linking of form and meaning?
16	Does the activity involve instantiation?
17	Does the activity involve imaging?
18	Does the activity avoid interference?

Most of the 18 TFA criteria would seem to have little direct bearing on the issue of whether words should be learned in isolation or in contexts. An exception may be the TFA criteria (#15 to #18) listed under the heading "retention". It could be that learning a word in a context, whether this be a collocation, a sentence, or a longer stretch of text, leads to fewer mistakes in linking the form with its meaning (#15). This should be particularly true for words that are highly abstract (e.g., words with low imageability) and for words that vary somewhat

from their L1 counterparts in terms of their semantics, grammatical behavior, and collocational constraints. In some cases, learning words within a linguistic context may also lead to a clearer mental picture associated with a meaning (#17). For example, learning the English word *vicious* with an L1 gloss may not lead a learner to form an image of the word. On the other hand, learning the word *vicious* embedded in the collocation *vicious dog* seems more likely to lead a learner to generate a mental image. Learning words in contexts may also help learners avoid interference (#18). For example, a learner encountering two similar L2 words such as *vicious* and *terrible* may have trouble learning them due to interference (Tinkham, 1993; Waring, 1997), but the same learner may experience less confusion if they are acquired in collocations with distinct meanings (e.g., a *vicious dog* and *terrible weather*).

On the other hand, some of the criteria suggest advantages for learning words in isolation. If it is important that activities focus attention on target words (#4), inclusion of context may have the deleterious effect of moving the learner's attention away from the target word to other words or the general message of the text with the result that "individual words, including new words to be learnt, may go unnoticed" (Laufer & Shmueli, 1997, p. 91). This is especially true if the context is rich, such as a sentence or paragraph.

Other considerations, unrelated to the 18 criteria, also suggest advantages for learning words in isolation. It is accepted that word learning, rather than occurring all at once, proceeds in a piecemeal fashion as learners acquire more words and deeper knowledge of the words they know. According to Dóczi and Kormos (2015), learners typically begin by learning a word's part of speech, written form, spoken form, and meaning (elements required for receptive mastery of a word) before going on to learn the word's grammatical behavior, use in collocations, other word forms, and other senses, (elements needed for successful production). If this is true, learning a word in context, which in addition to word meaning

and form provides some information on the word's typical grammatical and syntagmatic behavior, may impose too great a burden on the learner during initial word learning. Specifically, paired associate learning of words in isolation would typically focus exclusively on the receptive elements in the Dóczi and Kormos model, whereas paired associate learning of a word in a collocation or richer linguistic context would represent an attempt, probably unwarranted, to develop deeper word knowledge during initial learning.

PREVIOUS EMPIRICAL RESEARCH

At least a dozen studies have examined the effect of the provision of context on word learning. These studies will be reviewed, beginning with those with the most tenuous relevance to the current study and then gradually moving to studies with greater relevance.

Dempster (1987) conducted five experiments comparing the effects of learning words without context (i.e., definitions only) or with the definitions supplemented by either one or three example sentences. The experiments also examined the effects of massed or spaced repetitions, finding positive effects for spacing. All five experiments found null effects for learning condition (learning words in isolation or context) in spite of the fact that Exp. 5 used two measures (filling in a sentence context with the word or writing a sentence with the word) that should have favored the context group. While Dempster's studies are laudable for their methodological rigor, they have only limited relevance to the research question addressed in the current study as they involved participants learning L1 vocabulary. Even if it is assumed that this type of learning involves the same processes as L2 word learning, the type of words that tend to be targeted (i.e., L1 words unknown to educated native speakers) tend to be words conveying somewhat complex ideas (e.g., *apocryphal*) or words with complex register and

genre specifications (e.g., *loggia*), making them atypical of words targeted by most L2 learners.

Turning to L2A studies, Prince (1996) conducted a study of L1-French learners of English. In the between-subjects comparison, participants studied in a no-context condition (definition only) or in an incidental context condition (encountering the word used in a sentence context). As an immediate posttest measure, half of the participants in each learning condition took a translation test (either L1 to L2 or vice versa) and half took a fill-in-the-blank test. Results showed that learning individual L2 words through paired associate learning was more effective in terms of word recall than requiring learners to determine the meaning of words through contexts. A much older study by Seibert (1930) compared similar conditions and also reported that the no-context condition was superior. Unfortunately, the Prince and Seibert studies are of limited applicability to the current inquiry since the treatments given in the no-context and context conditions differed across several dimensions. Specifically, learners studying words in context were also tasked with determining the meaning of the words without recourse to an L1 gloss.

Laufer and Shmueli (1997) examined learners' memorization of words (1) in isolation, (2) in a minimal context, (3) in a text context, and (4) in elaborated context, in comparison with a "control group" (which studied the words as they wished as homework). In the four between-subject conditions, half of the words appeared with an L1 gloss and half with an L2 gloss. The treatment task for the groups differed in some important ways. The first two groups ("isolation" and "minimal context") were told to focus on the words, whereas the latter two groups ("text" and "elaborated") were given comprehension questions that focused on the general meaning of the text. However, a consolidation task that involved a fill-in-the-blanks exercise was provided to all groups so that even the incidental learning groups had their attention drawn to the target words at one point during

the treatment. In an apparent attempt to offset the focus on information irrelevant to learning the target words in the "text" and "elaborated" condition, participants in these conditions were given 55 minutes to do the task, much longer than the 10 minutes given to the "isolation" and "minimal context" groups. Results showed better learning of words in isolation or minimal contexts and for words with an L1 gloss. Unfortunately, the results have only limited bearing on the inquiry of the current study. After all, it is not clear whether the poorer vocabulary retention of the latter two groups is related to the context in which the words were imbedded or to learners' focus during the task. It could be that learners asked to focus on general understanding of the text spent much less time focusing on word meaning, especially if the word in question was not deemed crucial to completing the comprehension questions (cf. Peters, 2012).

Griffin (1992, Experiment 4) examined the effectiveness of vocabulary learning in an experiment in which the learning condition (single words or words in context) and assessment type (cued sentence completion) were treated as between-subject factors, and time (an immediate posttest and a posttest at a five-day delay) was treated as a within-subjects factor. The participants either saw a single L2 target word with its L1 translation or the bolded L2 target word embedded in a sentence with a translation of the entire sentence (and with the L2 translation of the target word bolded). Results showed advantages for context-based learning, especially for higher level learners who were assessed with a generation (L1 to L2) task.

A study by Webb (2007b), perhaps the most methodologically rigorous among the studies reviewed, compared Japanese EFL students' paired associate learning of words in isolation or with the addition of a one-sentence example sentence. To ensure that participants did not have partial knowledge of the target words, the selected L2 target forms were replaced with disguised forms (i.e., nonce words). Participants were not aware that the target words were not real

English words. Both groups spent eight minutes learning the words. At the end of the treatment, participants completed a surprise test battery consisting of ten tests focused on five aspects of word knowledge (orthography, paradigmatic association, meaning and form, syntagmatic association, and grammatical functions). The results for the two groups on the ten measures were virtually identical with no statistical differences.

In a study perhaps most relevant to the current experiment, Kasahara (2011) looked at paired associate learning of single words and words in two-word collocations among Japanese EFL students. In a between-subjects comparison, he found that the collocation group performed better on both an immediate and delayed posttest. The interaction effect showed that the decline in learning for the collocation group on the delayed posttest (one week after treatment) was also less than that of the no-context group. Kasahara posited, as a possible explanation, the encoding specificity principle (Thomson & Tulving, 1970).

An interesting feature of Kasahara's design was that the testing required participants learning words in isolation to write the L1 definition of the target word whereas it required the context group to write the L1 definition of the collocation. One methodological concern is that the provision of the collocation on the test may facilitate recall of a difficult word. To take just one example from his materials, his group learning words in isolation studied the word *lapse* in isolation and was then asked to translate this word. The collocation group learned the word *lapse* by studying the collocation *memory lapse* and was then, at testing, asked to translate this collocation. In this case, the presence of the same collocate (i.e., *memory*) during study and testing should facilitate recall of the target *lapse*. However, it is possible that the group learning *lapse* in isolation would also be better able to translate the word (and the resulting collocation) if they encountered it in combination with the word *memory*. His design also does not clarify whether the collocation group would recognize the unknown target word (e.g., *lapse*) if it

were encountered in isolation.

In a more recent experiment, Kanayama and Kasahara (2016) compared paired associate learning of single words with the same words learned in a two-word or a three-word collocation. Participants (Japanese EFL college students) were told to study the list of target words (in isolation, with a preceding English adjective, or with a preceding English adverb and adjective, depending on the condition) for six minutes. The collocating English words were words the participants were likely to know. On the immediate posttest, all groups were only tested on the target words, which were, as in Webb's (2007b) study, disguised forms. The two context groups were then also tested on the words with the collocates provided before the blank on each item. A delayed posttest on only the single word targets was given a week after the treatment. Unlike in the Kasahara (2011) study, the group learning the target words in isolation significantly outperformed the other groups, whereas the two-word collocation group outperformed the three-word group on both the posttest and the posttest with the original collocating words provided. The scores of the three groups converged on the delayed posttest.

The studies discussed in the literature review have been summarized in Table 2. The amount of context provided in the treatment in each condition is shown in five columns, whereas the provision of a definition in the treatment condition is indicated by a "D". If the treatment involved incidental learning, it was marked with an "I". An asterisk indicates that the treatment condition was found to be superior. Most studies have found either no differences between the conditions or an advantage for learning words in isolation. Only two studies (Griffin, 1992, Exp. 4; Kasahara, 2011) have found advantages for learning words in context.

As can be seen, most research has compared paired-associate learning of isolated words with a similar task supplemented by a sample sentence or sentences. Most of the dependent measures have involved translation of the target

words into the L1 or translation of the L1 equivalent into the target word, but many studies (especially, Webb, 2007b) have also used other measures to assess other types of word knowledge. Only two studies have looked at words learned in collocations and most studies have involved only immediate posttests or delayed posttests that were given a week or less after the treatment.

Table 2

Studies Comparing Learning of Words in Isolation with Context-Based Learning

Study	Context** and Learning Type***					Target		Test	
	N	C	S	S+	E	L1/L2	real/nonce	Translation or Definition	Context-Based
Dempster (1987, Exp. 1)	D		D	D		L1	real	X	
Dempster (1987, Exp. 2)	D		D	D		L1	real	X	
Dempster (1987, Exp. 3)	D			D		L1	real	X	
Dempster (1987, Exp. 4)	D			D		L1	real	X	
Dempster (1987, Exp. 5)	D			D		L1	real		X
Prince (1996) high-level participants	D*		I			L2	real	X	X
Prince (1996) low-level participants	D*		I			L2	real	X	X
Seibert (1930)	D*		D			L2	real	X	X
Laufer & Shmueli (1997)	D*		D	I	I	L2	real	X	
Webb (2007b)	D*		D			L2	nonce	X	X
Griffin (1992, Exp. 4) high-level participants	D		D*			L2	real		X
Griffin (1992, Exp. 4) low-level participants	D		D			L2	real		X
Kasahara (2011)	D	D*				L2	real	X	
Kanayama & Kasahara (2016)	D*	D x2				L2	nonce	X	

* Most effective condition in the experiment

** Context: N = no context (words in isolation), C = collocation, S = sentences, S+ = multiple sentences, E = elaborated text

*** Learning type: D = deliberate learning through an L1 translation or definition, I = incidental learning

One possible reason for different results in the reviewed research may be related to word type. Based on the presence of certain features, some words may be more learnable as either an isolated word or within a richer context. A possible candidate for such a relevant feature would be imageability. De Groot (1992) has shown that words with low imageability take longer for bilinguals to translate and slightly longer to recognize. In SLA research, imageability has been found to influence the learning of L2 idioms (Steinel, Hulstijn, & Steinel, 2007). Ellis and Beaton (1993) suggest that these effects reflect the fact that imageability confers meaning. The effect of imageability has thus been examined, as an additional moderator variable, in the current experiments.

EXPERIMENT 1

Method

Previous research has often compared the study of word in isolation with the study of words in collocations, sentences or multiple sentences. The underlying assumption has often been that if context is good, more is likely to be better. However, the use of rich context introduces some methodological conundrums. Initial word learning through definitions is probably best achieved using a simple L1 translation since this is more likely to develop into a stable L2 entry that can later be elaborated during subsequent encounters with a word. In the context conditions, reading such a definition accompanied by an example sentence or set of sentences will presumably require more time. Yet if time on task is controlled in both conditions, participants studying the word in isolation are likely to have more time than necessary or, if the time is reduced, participants in the context condition are likely to have insufficient time to read through the materials. For this reason, studies (e.g., Laufer & Shmueli, 1997) involving sizable differences between the amount of content in the no-context and context conditions face

thorny methodological problems in controlling the effects of time on task. The current study therefore examines the learning of target items through an L1 gloss, comparing the effectiveness of learning words in isolation with learning words embedded in collocations. These two conditions are amenable to experimentation as both tasks presumably require similar amounts of time.

Participants. The participants (n = 46) were Japanese-L1 EFL students (all first-year) from two intact basic English classes at a large public university in Japan. They could be described as roughly at the B1 level of the Common European Framework of Reference (Council of Europe, 2001). Only data from participants who were present for all phases of the experiment were included in the analysis.

Materials. The target vocabulary consisted of 72 monosyllabic English words (Appendix A) that the participants were unlikely to know, selected from a list of words rated for imageability in Cortese and Fugett (2004). After selecting a pool of candidate target words, a native speaker of Japanese reviewed the words to ensure that none had obvious Japanese cognates. Technical terms that would require participants to learn both the word and a new concept were avoided. Based on the Cortese and Fugett (2004) norms, 36 of the items had low imageability (hereafter, LI) and 36 had high imageability (hereafter, HI). Nonsignificant results on Shapiro-Wilk tests (p = .274 for LI items and p = .253 for HI items) and visual examination of Q-Q plots confirmed that both LI and HI items had a normal distribution. The mean imageability ratings for LI and HI items were based on the reported participant averages in Cortese and Fugett (2004). Descriptive statistics for the items are shown in Table 3.

Table 3

Imageability of Target Items Used in Experimental Measures and Treatment

LI Item Imageability		HI Item Imageability	
M (*SD*)	Range	*M* (*SD*)	Range
2.17 (0.33)	1.40—2.80	6.00 (0.32)	5.50—6.70

To determine whether the two sets of items had significantly different imageability ratings, an ANOVA was conducted, treating Imageability Rating as the independent variable. Examination of box plots and Levene's statistic confirmed that the assumption of homogeneity of variances was met, $F(1, 70) = .017$, $p = .896$. The ANOVA showed that the imageability of the HI items was significantly higher than that of the LI items, $F(1, 70) = 2602.90$, $p < .001$.

A pretest ($\alpha = .80$) was created with 75 items (36 LI items, 36 HI item, and three distractors). The distractors were multi-syllabic two-word combinations (e.g., *abstract concept*) that were slightly easier than the target items. Two distractors appeared as the first and last item on the test and the other in the middle of the test. To ensure that placement on the test did not influence results, the LI and HI items appeared in a pseudo-random order with an alternating distribution so that each LI item was followed by a HI item and vice versa. Except for the distractors, which required participants to translate the collocation, the test required participants to supply only the target English word based on a Japanese translation. This format was adopted so that the learning of items in the four conditions could be assessed using identical criteria. To prevent participants from supplying possible translations other than the target word, boxes showing the number of letters of the target word were provided along with one or more letters in the target word. The beginning of the test provided directions along with an example item. The posttest ($\alpha = .87$) and delayed posttest ($\alpha = .85$) were similar with the same criteria used for placement of distractors and target items but with the items in a different order.

The study materials consisted of a hand-out showing a Japanese word with its English translation (i.e., the target items). Half of the target words appeared in isolation and half appeared as part of a collocation. In the directions and in verbal instructions in class, participants were told to learn all 75 words on the hand-out as homework in preparation for a vocabulary test. They were informed that the test would require them to supply the English word and that the collocation items would require them to supply the entire collocation. The directions were designed to encourage participants to focus on the entire collocation during their study of the words. The actual test presented the target words only in isolation. Only the distractor items appeared in collocations.

In the analysis, the 72 target items were divided into four blocks differentiated by imageability rating and study condition. A quarter (18) of the items were LI items studied as single words (hereafter, Single condition); a quarter were LI items studied as collocations (hereafter Collocation condition); a quarter were HI items studied in the Single condition, and a quarter were HI items studied in the Collocation condition. The assignments of words to conditions were counterbalanced between the two intact classes so that the words studied in the Single condition in one class were studied in the Collocation condition in the other class, and vice versa. On the study materials, words from the four blocks appeared in a pseudo-random order so that the proportion of words in each of the four conditions (i.e., LI Single, LI Collocation, HI Single, HI Collocation) was the same on each quarter of the test.

The collocations were selected based on an examination of the most frequent collocating items in the COCA corpus (Davies, 2008-). It was assumed that use of frequent collocations that paired the target word with a word that the participants were likely to know would foster maximal learning in the Collocation condition. It was felt that the semantic congruency between the words in these collocations would be greater and that this would facilitate recall of an unknown word in a

collocation (cf. Bein, Livneh, Reggev, Gilead, & Goshen-Gottsein, 2015; Ouyang, Boroditsky, & Frank, 2017). For example, if the word *alms* is learned initially as part of the collocation *alms for the poor*, the collocating words *for the poor* should help the learner recall the correct meaning of *alms*.

Procedure and Analysis. Participants were given the pretest with 25 minutes to respond. They then received a list of 75 words to study and were informed that they would have a graded test on the words in two weeks. Two weeks after the pretest, they took the posttest (also 25 minutes). Six weeks after the posttest, participants took an unannounced delayed posttest (also 25 minutes). To ensure that review of the tests or knowledge of their scores did not affect the results, participants did not receive test results or scores until the data collection was complete.

Test results are reported for scores based on a partial credit model as well as for scores based on dichotomous scoring (i.e., giving points only for completely correct answers). For the partial credit tally, the number of letters provided in the correct position in the word was divided by the number of missing letters in the word. For example, Item #11 on the posttest, targeting *claw*, required the letters C, L, and W in the appropriate positions (see Figure 1). If the participant responded with C, R, and W, the second letter was counted wrong, resulting in a score of 66.6% for that item.

11　つめ（獣）

		a	

Figure 1. Example test item.

Results

The descriptive statistics for the Exp. 1 participants' (n = 46) scores on the LI items using a partial credit model are shown in Table 4. Participants studied 18 of the items as single words and 18 as collocations.

Table 4

Scores (Using Partial Credit Scoring) on Low Imageability Items Learned as Single Words or As Collocations in Exp. 1

Learning Condition					
Single			Collocation		
Pretest	Posttest	Delayed Posttest	Pretest	Posttest	Delayed Posttest
M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Range	Range	Range	Range	Range	Range
0.3 (0.5)	15.0 (3.8)	4.8 (3.4)	0.2 (0.4)	14.5 (4.6)	3.8 (3.2)
0.0—2.0	5.0—18.0	0.0—12.7	0.0—1.2	2.0—18.0	0.0—13.6

The pretest scores indicated that participants had virtually no knowledge of the target words prior to the intervention. As can be seen in Table 4, participants' posttest gains for the LI items were nearly identical (around 80%) regardless of the learning condition (i.e., whether the vocabulary was learned as single words or in phrases). As would be expected, there was a sharp decline in scores on the delayed posttest to around a quarter of the items, with slightly lower retention of words learned as phrases.

Participants' scores on the high imageability items are shown in Table 5.

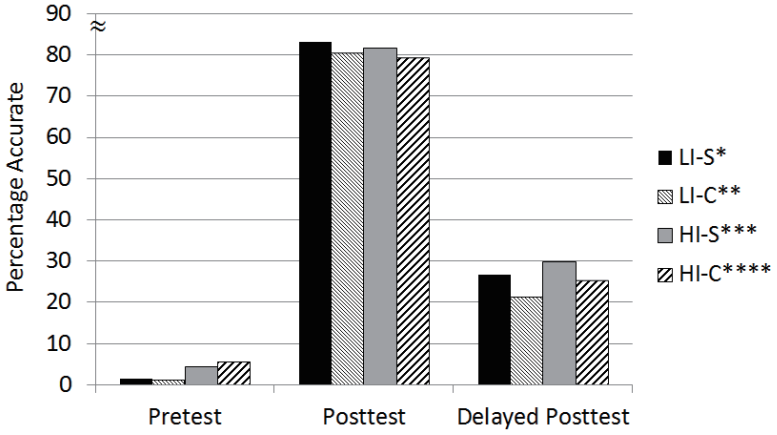
Table 5

Scores (Using Partial Credit Scoring) on High Imageability Items Learned as Single Words or As Collocations in Exp. 1

Learning Condition					
Single			Collocation		
Pretest	Posttest	Delayed Posttest	Pretest	Posttest	Delayed Posttest
M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Range	Range	Range	Range	Range	Range
0.8 (1.0)	14.7 (4.2)	5.4 (3.5)	1.0 (0.9)	14.3 (4.3)	4.6 (3.6)
0.0—5.0	3.2—18.0	0.3—16.0	0.0—3.2	2.0—18.0	0.0—15.8

As can be seen, the pretest to posttest gains for the high imageability items were slightly lower than those for the low imageability items, but the retention (operationalized in terms of delayed posttest scores) was higher, especially for the target items learned as single words. The results using dichotomous scoring were virtually identical to the results using partial credit, so for the sake of brevity, these results have been omitted here.

Percentage accuracy, using partial credit scores, for the four conditions is shown in Figure 2. As the figure indicates, participants generally learned over three-quarters of the target words but only retained about a quarter of the words later. Retention for words in the four conditions was similar with a slight numerical advantage for words learned as single items.



* Low imageability items learned as single words

** Low imageability items learned as phrases

*** High imageability items learned as single words

**** High imageability items learned as phrases

Figure 2. Percent accuracy for LI and HI words learned as single words or collocations.

The experimental design involved three independent variables that served as within-subjects factors. Time was observed at three levels (pretest, posttest, and delayed posttest). Imageability was observed at two levels (LI and HI). Learning Condition was also observed at two levels (single word versus collocation). To determine whether these three independent variables had a significant effect on the dependent variable (i.e., test scores using partial credit scoring), a three-way repeated measures (RM) ANOVA was conducted in SPSS Version 25.

To determine whether the assumption of normality was met, Q-Q plots and histograms were examined. The assumption of normality was not met, so the data were submitted to a log10 transformation. These data were normal and were thus used for the statistical analysis. Other assumptions of an RM-ANOVA, using the transformed data, were confirmed. Comparison of box plots for the data according to factor and an examination of covariances in the residual SS and cross-products matrix indicated that they had equal variances. The residuals also had a normal distribution and equal variance. Mauchly's test indicated that the assumption of sphericity was met for the main effect of Time, ($\chi2(2) = 2.28$, $p = .320$, the interaction of Time and Imageability ($\chi2(2) = 3.06$, $p = .216$, and the interaction between Time and Learning Condition, ($\chi2(2) = 2.84$, $p = .241$. The assumption was not met for the three-way interaction between Time, Imageability, and Learning Condition, ($\chi2(2) = 6.00$, $p = .050$, so the Greenhouse-Geisser corrected values are reported for the three-way interaction.

For the RM-ANOVA, a Type II sum of squares analysis was used, and the statistical analyses were conducted in SPSS Version 25. All effects are reported as significant at $p < .05$. For ease of interpretation, the 95% confidence intervals are shown in back-transformed values. As would be expected, there was a significant main effect of Time, $F(2, 90) = 579.84$, $p < .001$, $\eta p^2 = .93$, and also a significant main effect of Imageability, $F(1, 45) = 21.05$, $p < .001$, 95%CI: [0.10, 0.26], $\eta p^2 = .32$. This reflected poorer performance on LI items. The effect of Learning

Condition was small and only marginally significant, $F(1, 45) = 4.00$, $p = .052$, 95%CI[0.00, 0.12], $\eta p^2 = .08$; however, it should be noted that attenuated main effects are due in part to the experimental design since the effects of Learning Condition only come into play on the posttest and delayed posttest after the items have been studied. Therefore, the simple effects analysis (shown in Table 6), examining significance for Learning Condition for each level of the other independent variables, is more revealing. As can be seen, Learning Condition is only significant for the LI items on the posttest and for both the LI and HI items on the delayed posttest. In all cases, this reflects superior learning for words learned in isolation.

Table 6

Simple Effects Analysis for the Time, Imageability, and Learning Condition in Exp. 1

Pretest		Posttest		Delayed Posttest	
LI	HI	LI	HI	LI	HI
S* vs. C**	S* vs. C**	S* vs. C**	S* vs. C**	S* vs. C**	S* vs. C**
$p = .792$	$p = .150$	$p = .018$	$p = .227$	$p = .030$	$p = .028$

* The single word condition

** The collocation condition

There was a significant interaction effect between the level of Time and Imageability, $F(2, 90) = 18.94$, $p < .001$, $\eta p^2 = .30$. To break down this interaction, contrasts were performed on both the posttest and delayed posttest gains for the LI and HI items relative to the pretest. These contrasts revealed significant interactions when comparing imageability scores on the pretest and posttest relative to imageability scores on the pretest and delayed posttest, $F(1, 45) = 11.24$, $p = .002$, $\eta p^2 = .20$. Likewise, there was a significant contrast when comparing the imageability scores on the pretest and posttest relative to comparisons of posttest and delayed posttest scores , $F(1, 45) = 9.77$, $p = .003$, $\eta p^2 = .18$. The interactions reflect the finding that initial negative effects of

imageability on the pretest were overcome on the posttest (probably reflecting the effects of cramming), whereas imageability exerted a noticeably negative effect on retention of words, as assessed on the delayed posttest.

The interaction between Time and Learning Condition was also significant, $F(2, 90) = 7.31$, $p = .001$, $\eta p^2 = .14$. To break down this interaction, contrasts were performed. The contrasts revealed significant interaction when comparing the two Learning Conditions on the pretest and posttest relative to the contrast when comparing the pretest and delayed posttest, $F(1, 45) = 11.67$, $p = .001$, $\eta p^2 = .21$. Likewise, the contrasts revealed significant interactions when comparing Learning Conditions on the pretest and posttest relative to the same comparison of scores on the posttest and delayed posttest, $F(1, 45) = 5.49$, $p = .024$, $\eta p^2 = .11$. These interactions reflect the fact that the learning condition (single word versus collocation) exerted only a minor effect on initial learning (as assessed on the posttest) but exerted a stronger effect on retention as observed by higher gains for items learned as single words on the delayed posttest.

The interaction between Imageability and Learning Condition was not significant, $F(1, 45) = 1.22$, $p = .276$. Moreover, the three-way interaction between Time, Imageability, and Learning Condition, using Greenhouse-Geisser corrected values, was not significant, $F(1.77, 79.83) = 0.59$, $p = .539$.

In the pairwise contrasts, adjustments for multiple comparisons were made using a Bonferroni correction. Pairwise comparisons for the factor Time indicated that the posttest scores were significantly higher than pretest scores, $p < .001$, 95%CI: [7.9, 11.1], and delayed posttest scores, $p < .001$, 95%CI: [1.6, 2.8]. Delayed posttest scores were significantly higher than pretest scores, $p < .001$, 95%CI: [1.7, 2.9].

Discussion

Exp. 1 suggests that when learners initially study vocabulary using paired

associate learning prior to an announced and graded test, learning words in isolation versus in collocations provides only small advantages that only reach levels of significance for low imageability items. However, words learned in isolation appear to be remembered much better. The results thus agree with previous research that has found learning words in isolation to be superior. Initial word learning prior to a test appears to be little effected by the imageability of a word, but after being learned, words with low imageability appear to be forgotten more readily than words with high imageability.

EXPERIMENT 2

The Exp. 1 results confirm the findings of much previous research regarding advantages for learning words in isolation. Yet it could be that the benefits of learning words as collocations appear after more intense study of words in which paired associate learning is reinforced by tasks promoting more intense semantic processing. As mentioned in the introduction, Technique Feature Analysis (TFA) assumes that vocabulary tasks that involve generation will be more effective (Nation & Webb, 2011). A second experiment was therefore conducted that added an in-class sentence writing task in order to promote participants' greater engagement with the target items' semantics.

Method

Participants. The participants (n = 39) were first-year university students in an English literature and linguistics program at a private women's university in Japan from two first-year English writing classes. In terms of the Common European Framework of Reference (CEFR), they could be described as roughly at the low B1 level. They therefore had slightly lower proficiency than the participants in Exp. 1. They were motivated learners, many of whom aspire to

enter career fields requiring English.

Procedure and Materials. Participants took a pre-test on the same 72 target items used in Exp. 1 using the same test forms with the same counterbalancing of items between classes. They were then informed that there would be a graded test on these vocabulary items and that there would also be a graded homework assignment in which they would need to write a sentence using each of the target words (in isolation or within a collocation depending on the condition). A week after the pre-test, they were also given one 45-minute class period in a computer room to work on the homework. They submitted the sentences using an online site (https:/www.quia.com). Most were unable to complete the homework during this time and therefore finished it outside of class. Two weeks later they took the posttest. Four weeks later, they took an unannounced delayed posttest. Participants did not receive feedback on their sentences or any of the tests prior to the conclusion of the experiment. Only data from participants who completed all the tests as well as the sentence writing task were included in the analysis. The target items and test forms were identical to those used in Exp. 1.

Results

The descriptive statistics for the Exp. 2 participants' scores on the LI items using a partial credit model are shown in Table 7. Participants studied 18 of the LI items as single words and 18 as collocations.

Table 7

Scores (Using Partial Credit Scoring) on Low Imageability Items Learned as Single Words or As Collocations in Exp. 2

Learning Condition					
Single			Collocation		
Pretest	Posttest	Delayed Posttest	Pretest	Posttest	Delayed Posttest
M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Range	Range	Range	Range	Range	Range
0.1 (0.3)	10.6 (5.2)	3.6 (2.7)	0.2 (0.3)	8.8 (5.3)	2.8 (2.8)
0.0–1.0	1.0–18.0	0.0–9.3	0.0–1.0	0.0–17.8	0.0–11.7

The pretest scores indicated that participants had virtually no knowledge of the target words prior to the intervention. As can be seen in Table 7, their posttest gains for the low imageability items were slightly higher for words learned in isolation. As would be expected, delayed posttest scores fell sharply, with lower retention of words learned as phrases. While lower, the pattern of scores resembled those in Exp. 1

Participants' scores on the high imageability items are shown in Table 8.

Table 8

Scores (Using Partial Credit Scoring) on High Imageability Items Learned as Single Words or As Collocations in Exp. 2

Learning Condition					
Single			Collocation		
Pretest	Posttest	Delayed Posttest	Pretest	Posttest	Delayed Posttest
M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Range	Range	Range	Range	Range	Range
0.4 (0.5)	10.9 (5.2)	4.1 (3.1)	0.3 (0.5)	9.7 (5.1)	2.9 (2.5)
0.0–1.5	2.0–18.0	0.0–13.3	0.0–1.5	0.5–17.8	0.0–9.7

As can be seen, the pretest to posttest gains for the HI items roughly mirrored the results for the LI items. Results for the items calculated using dichotomous scoring (points given only for items that are completely correct) were virtually identical to the results using partial credit scoring, so for the sake of brevity, only the results based on partial credit scores are reported here. The descriptive results shown in terms of percentage accuracy are shown in Figure 3. As can be seen, the Exp. 2 participants showed poorer learning of LI items relative to HI items on both the posttest and delayed posttest. Moreover, words learned in isolation were learned better than words learned as collocations.

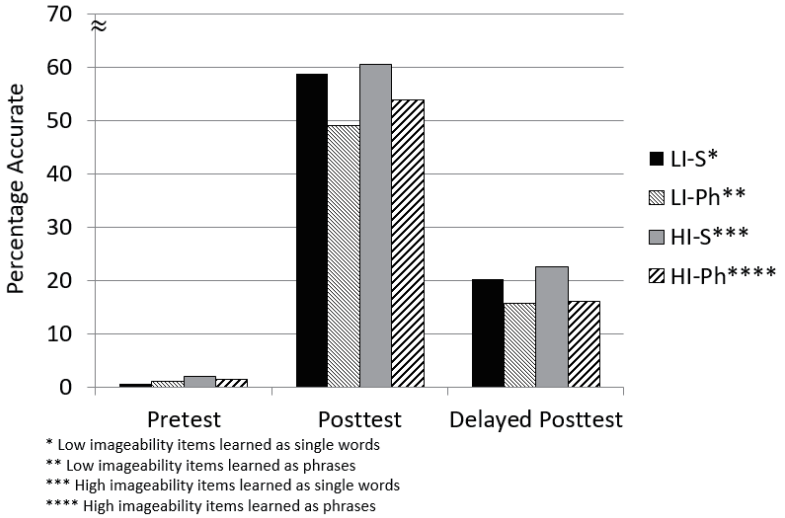


Figure 3. The percentage accuracy on the three vocabulary tests with scores categorized in terms of the target words' imageability and the learning condition.

As in Exp. 1, an RM-ANOVA was conducted in SPSS Version 25 and Type II sum of squares analysis was used. All effects are reported as significant at $p < .05$. As in Exp. 1, the assumption of normality was not met, so the data were

subjected to square root transformations as this succeeded in normalizing the data. These data were used for the statistical analyses. To ease interpretation, confident intervals have been reported using back transformed values.

Assumptions related to normality, homoscedasticity, and independence of observations were met. Mauchly's test indicated that the assumption of sphericity was met for Time, ($\chi 2(2) = 3.33$, $p = .189$, and the interaction between Time and Learning Condition, ($\chi 2(2) = 0.56$, $p = .756$, and was narrowly met for the three-way interaction between Time, Imageability, and Learning Condition, ($\chi 2(2) = 5.19$, $p = .075$; however, the assumption was not met for the interaction between Time and Imageability, ($\chi 2(2) = 12.43$, $p = .002$. Greenhouse-Geisser correction values are thus reported for both the three-way interaction and for the Time and Imageability interaction.

Tests of the effects of the three within-subjects factors (Time, Imageability, and Learning Condition) on transformed scores revealed significant main effects for Time, $F(2, 76) = 260.28$, $p < .001$, $\eta p^2 = .87$. There were also significant main effects for Imageability, $F(1, 38) = 13.08$, $p = .001$, 95% CI [0.01–0.04], $\eta p^2 = .27$, reflecting lower scores for LI items. There was a significant effect for Learning Condition, $F(1, 38) = 23.47$, $p < .001$, 95% CI [0.12, 0.29], $\eta p^2 = .38$, indicating better performance on items learned as single words compared to items learned as phrases. As discussed in Exp. 1, the Learning Condition is irrelevant to the pretest, so its effects are better observed in simple effects analysis (shown in Table 9). As can be seen, the effects for Learning Condition were significant on both the posttest and delayed posttest, signifying better performance on words learned in isolation.

Table 9

Simple Effects Analysis for the Time, Imageability, and Learning Condition in Exp. 2

Pretest		Posttest		Delayed Posttest	
LI	HI	LI	HI	LI	HI
S* vs. C**	S* vs. C**	S* vs. C**	S* vs. C**	S* vs. C**	S* vs. C**
$p = .279$	$p = .195$	$p < .001$	$p = .006$	$p = .011$	$p < .001$

* The single word condition

** The collocation condition

There was also a significant effect for the interaction between Time and Learning Condition, $F(2, 76) = 9.02$, $p < .001$, $\eta p^2 = .19$. To break down this interaction, contrasts were performed comparing both the posttest and delayed posttest gains for the two learning conditions relative to the pretest. These revealed significant interactions, $F(1, 38) = 13.89$, $p = .001$, $\eta p^2 = .27$. This reflects the fact that items successfully learned as isolated words (as assessed on the posttest) were more subject to forgetting (34.5%, as assessed on the delayed posttest) relative to words learned as collocations (29.1% of which were forgotten). In other words, learning words in isolation appears to have a positive effect on initial learning but the acquired knowledge appears to be more subject to forgetting. Contrasts that compared the effect of Learning Condition between the pretest and posttest with its effect between the posttest and delayed posttest were not significant, $F(1, 38) = 0.41$, $p = .528$. The interaction between Time and Imageability, $F(1.6, 59.1) = 0.12$, $p = .835$, and the three-way interaction between Time, Imageability, and Learning Condition, $F(1.8, 67.2) = 0.11$, $p = .058$, using Greenhouse-Geisser correction values, were not significant.

Pairwise comparisons using Bonferroni adjustments showed that pretest scores were significantly lower than posttest scores, $p < .001$, 95% CI [-9.74, -5.92], and delayed posttest scores, $p < .001$, 95% CI [-2.79, -1.17]. Posttest scores were significantly higher than delayed posttest scores, $p < .001$, 95% CI [1.27, 2.77].

Discussion

The Exp. 2 results have some similarities with those of Exp. 1. In both experiments, learning words in isolation proved to be more effective, and imageability was associated with slightly less learning. Yet one striking difference was the effect of Learning Condition (i.e., superior performance for words learned in isolation) on initial learning (as measured by the posttest). Superior performance on words learned in isolation on the Exp. 2 posttest may reflect the impact of the in-class sentence writing task. It could be that this sort of generative task was less effective when targeting words in collocations. Another suggestive finding is the attenuated effects of learning condition on the delayed posttest. It could be that although the study of single words leads to much greater immediate gains, these gains are more fragile than the more painstakingly acquired knowledge based on study of words in collocations.

CONCLUSION

The current study suggests that studying words in isolation leads to much more initial learning compared to learning words in collocations, and that this advantage is observed even four to six weeks later. However, Exp. 2 tenuously suggests that this advantage may dissipate over time as words learned in isolation are forgotten at a higher rate. Kasahara (2011) also found greater declines in learning for words learned in isolation. A key question is whether testing learners over even longer intervals of time would show an advantage for learning words as collocations. Future research should therefore examine the two approaches to word learning using delayed posttests given months after initial learning.

The current study is in agreement with a number of studies showing advantages for words learned in isolation (Kanayama & Kasahara, 2016; Prince, 1996; Webb, 2007b). One interpretation of these studies is that initial word

learning should only target the formation of a fairly sparse lexical entry. Attempts to acquire multiple aspects of word knowledge simultaneously may lead to faulty encoding or the weak association of multiple cues (e.g., both the L1 translation and target language collocates) instead of the strong association of a single cue (i.e., the L1 translation).

While in agreement with much previous research, the current study has a number of worthwhile innovations that expand on the current state of knowledge. First, few studies in this area have examined the effect of the amount of context on word learning over long intervals of time. In light of the Exp. 1 findings, which show little effect of Learning Condition on immediate learning but a significant effect on retention, this is a concern. Second, the current study employs learning tasks with high ecological validity. In language teaching, learners' self-paced study of words using paired associate learning is a familiar practice that is probably common to most foreign language programs throughout the world. Third, the current study examines the potential moderating effects of imageability on the effect of Learning Condition. Future research should continue to employ designs that include such independent variables. Such designs may detect boundary conditions, showing that learning words in isolation is best suited for particular types of words. Even if the learning of single words is found to have a consistent advantage, such research may shed light on key variables that are responsible for the effect.

In light of the high ecological validity of the current study, several clear pedagogical recommendations can be put forward. Vocabulary study that rapidly establishes a lexical entry with bare specifications (e.g., the sound or spelling of the word in a single form, a single key meaning conveyed through an L1 equivalent of the target word, and the word's part of speech) is probably most effective. However, the effectiveness of this sort of deliberate vocabulary learning is premised on the assumption that the learner is exposed to sufficient input so that

the word, once it is learned, is likely to be encountered again before it is forgotten. When input is closely monitored, as it can be in the design of pedagogical materials, careful attention should thus be given to recycling of words. Finally, the current results should not be construed to categorically deny the potential benefits of learning words in collocations. When learners have some familiarity with the target word, paired associate learning of the word in a common collocation may lead to deeper and more stable knowledge of the word. To establish empirically grounded pedagogical recommendations, future research should investigate how the various components of word knowledge are acquired over time and the potential role of deliberate study of a word's high-frequency collocations within this process.

References

Bein, O., Livneh, N., Reggev, N., Gilead, M., & Goshen-Gottsein, Y. (2015). Delineating the effect of semantic congruency on episodic memory: The role of integration and relatedness. *PLoS ONE, 10*(2). doi:10.1371/journal.pone.0115624

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning and Technology, 11*(3), 38-63.

Cortese, M. J., & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, 36*(3), 384-387.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, UK: Cambridge University.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671-684.

Davies, M. (2008-). The Corpus of Contemporary American English: 450 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

De Groot, A. M. B. (1992). Determinants of word translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(5), 1001-1018.

Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology, 79*(2), 162-170.

Dóczi, B., & Kormos, J. (2015). *Longitudinal developments in vocabulary knowledge and lexical organization.* New York, NY: Oxford University.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning, 43*(4), 559-617.

Eysenck, M. W. (1978). Levels of processing: A critique. *British Journal of Psychology, 69*(2), 157-169.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice.* New York, NY: Cambridge University Press.

Griffin, G. F. (1992). *Aspects of the psychology of second language vocabulary list learning.* (PhD), University of Warwick.

Kanayama, K., & Kasahara, K. (2016). How many cues are effective in known-and-unknown combination vocabulary learning? *Journal of Hokkaido University of Education, 66*(2), 17-27.

Kasahara, K. (2011). The effect of known-and-unknown word combinations on intentional vocabulary learning. *System, 39*(4), 491-499.

Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal, 28*(1), 89-108.

McQuillan, J., & Krashen, S. D. (2008). Commentary: Can free reading take you all the way? A response to Cobb (2007). *Language Learning and Technology, 12*(1), 104-109.

Nakanishi, T. (2015). A meta-analysis of extensive reading research. *TESOL*

Quarterly, 37(1), 6-37.

Nation, P., & Webb, S. (2011). *Researching and analyzing vocabulary.* Boston, MA: Heinle.

Ouyang, L., Boroditsky, L., & Frank, M. C. (2017). Semantic coherence facilitates distributional learning. *Cognitive Science, 41*(S4), 855-884.

Peters, E. (2012). The differential effects of two vocabulary instruction methods on EFL word learning: A study into task effectiveness. *International Review of Applied Linguistics in Language Teaching, 50*(3), 213-238.

Prince, P. (1996). Second language vocabulary learning: The role of context versus translations as a function of proficiency. *The Modern Language Journal, 80*(4), 478-493.

Seibert, L. C. (1930). An experiment on the relative efficiency of studying French vocabulary in associated pairs versus studying French vocabulary in context. *Journal of Educational Psychology, 21*(4), 297-314. doi:http://dx.doi.org/10.1037/h0070517

Steinel, M. P., Hulstijn, J. H., & Steinel, W. (2007). Second language idiom learning in a paired-associate paradigm: Effects of direction of learning, direction of testing, idiom imageability, and idiom transparency. *Studies in Second Language Acquisition, 29*(3), 449-484.

Thomson, D. M., & Tulving, E. (1970). Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology, 86*(2), 255-262.

Tinkham, T. (1993). The effect of semantic clustering on the learning of second language vocabulary. *System, 21*(3), 371-380.

Waring, R. (1997). The negative effects of learning words in semantic sets: A replication. *System, 25*(2), 261-274.

Webb, S. (2007a). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46-65.

Webb, S. (2007b). Learning word pairs and glossed sentences: The effects of a

single context on vocabulary knowledge. *Language Teaching Research, 11*(1), 63-81.

Webb, S., & Nation, P. (2017). *How vocabulary is learned.* Oxford, UK: Oxford University.

Appendix A: Target Words in Exp. 1 and Exp. 2

Note: Depending on the condition, words were learned in isolation (only the bolded word) or in collocations (the collocation containing the bolded form).

High Imageability Words		Low Imageability Words	
fresh **bait**	leech **bites**	**alms** for the poor	a temporary **lull** in the fighting
a plastic **bib**	a body in the **morgue**	the problem of urban **blight**	a romantic **ode**
pilot a **blimp**	butterflies and **moths**	a **boon** to mankind	the country's economic **plight**
the handle of a **broom**	tighten a **noose**	the **brunt** of the blame	a marketing **ploy**
a cat's **claw**	a dog's **paw**	the **chasm** between rich and poor	a clever **pun**
a yearling **colt**	a wooden **pier**	political **clout**	have moral **qualms**
a wooden shipping **crate**	a violent **pimp**	on the **cusp** of adulthood	a complete **rout** of the enemy
a baby's **crib**	a snow **plow**	a **dearth** of research	the **scourge** of terrorism
cake **crumbs**	roasted **quail**	an embarrassing **farce**	a doctor's **scrawl**
pie **crust**	a leaf **rake**	an amazing **feat**	a metallic **sheen**
cookie **dough**	a healed **scab**	a **gale** warning	a funny **spoof**
poisonous **fangs**	a flowering **shrub**	a world oil **glut**	a large **swath** of land
green **ferns**	throw a **spear**	a cruel **hoax**	diplomatic **tact**
a **flock** of sheep	a ruptured **spleen**	raised public **ire**	a huge **throng** of reporters
biting **gnats**	a **spool** of wool	a weekend **jaunt**	a **tinge** of sadness
a stack of **hay**	soap **suds**	a **knack** for spotting opportunities	agree to a **truce**
a garden **hoe**	a muddy **swamp**	a temporary **lapse** in judgment	on a sudden **whim**
fat **hogs**	a sharp **thorn**	according to local **lore**	religious **zeal**