Applied Psycholinguistics (2019), 1–32
doi:10.1017/S0142716419000262

CAMBRIDGE UNIVERSITY PRESS

ORIGINAL ARTICLE

The benefits of abstract word training on productive vocabulary knowledge among second language learners

Chaleece W. Sandberg[1,*], Erin Carpenter[1], Katherine Kerschen[1], Daniela Paolieri[2] and Carrie N. Jackson[1]

[1]Penn State University and [2]University of Granada
*Corresponding author. Email: cws18@psu.edu

(Received 03 July 2018; revised 14 January 2019; accepted 04 May 2019)

Abstract

This study investigates the effect of an abstract word training paradigm initially developed to treat lexical retrieval deficits in patients with aphasia on second language (L2) vocabulary acquisition. Three English–Spanish L2 learners (Experiment 1) and 10 Spanish–English L2 learners (Experiment 3) were trained on 15 abstract words within a context-category (e.g., restaurant) using a five-step training paradigm based on semantic feature analysis. In addition, 7 English–Spanish L2 learners were trained on either abstract or concrete words within a context-category (Experiment 2). Across all experiments, the majority of participants trained on abstract words showed improved production of the trained abstract words, as measured by a word generation task, as well as improvement on untrained concrete words within the same context-category (i.e., generalization). Participants trained on concrete words (Experiment 2) exhibited much smaller word production gains and no generalization to abstract words. These results parallel previous findings from aphasia research and suggest that this training paradigm can successfully be extended to L2 learning contexts, where it has the potential to be a useful tool in vocabulary instruction. We discuss the findings in terms of models of spreading activation and the underlying conceptual representations of abstract and concrete words in the L2 lexicon.

Keywords: abstract word retrieval; second language learning; vocabulary acquisition; word retrieval training

Vocabulary acquisition is a critical aspect of learning a second language (L2), as a larger vocabulary increases the ease, fluency, and efficiency of L2 communication (Barcroft, 2004; Coady & Huckin, 1997; Schmitt, 2010). However, until the early 1980s, there was relatively little research on L2 vocabulary acquisition within L2 acquisition research more generally (Meara, 1980). Fortunately, the intervening years have seen a proliferation of L2 vocabulary research, particularly with regard to the effectiveness of different instructional approaches for vocabulary learning (see Nation, 2013, for an overview).

Despite this increase in research, many challenging aspects of L2 vocabulary learning remain underinvestigated. One particular challenge is the development of productive vocabulary knowledge (the ability to produce the written or spoken form of a word to express a certain meaning; Laufer, 1998; Schmitt, 2010). Productive vocabulary knowledge typically lags behind receptive vocabulary knowledge (see Schmitt, 2014, for a review), and the instructional techniques targeting productive knowledge that have been studied to date have shown limited success (Keating, 2008; Pichette, De Serres, & Lafontaine, 2012; Webb, 2005). Another challenge is the acquisition of abstract words (e.g., *emergency*), which are learned more slowly and forgotten more quickly than concrete words (e.g., *ambulance*; de Groot & Keijzer, 2000). Despite this imbalance, few studies have tested instructional techniques aimed specifically at improving the learning of abstract words (but see Farley, Ramonda, & Liu, 2012; Pichette et al., 2012). Given these limited results in L2 acquisition research, it would be fruitful to examine other areas of language research for insights on how to improve the development of L2 productive vocabulary knowledge. One such area is treatment for language disorders, particularly aphasia. As outlined below, it is not unreasonable to assume that techniques used in aphasia therapy may translate to L2 vocabulary learning. Aphasia is a language deficit due to stroke or other acquired brain injuries. Anomia, or the inability to retrieve words from the lexicon, is a defining characteristic of aphasia. Persons with aphasia have not lost conceptual representations of words. Rather, access to or selection of the appropriate lexical label is hindered. Thus, there are parallels between persons with aphasia and L2 learners in terms of their ability to link the appropriate lexical label to a known concept. Further, aphasic syndromes in which comprehension is relatively spared, but production deficits are prominent, in many ways parallel the gap between receptive and productive vocabulary knowledge in L2 learners.

To test these intuitions, we conducted three experiments using a treatment that has been successfully implemented with persons with aphasia. This treatment for improving word finding in aphasia consists of training abstract words (e.g., *truth*) in a specific context-category (e.g., *courthouse*) by training semantic features (e.g., *generally considered positive*). Among people with aphasia, this training results not only in improvement in controlled oral production of the trained abstract words in a category association task but also in improvement in the controlled oral production of untrained concrete words in the same context-category (Kiran, Sandberg, & Abbott, 2009; Sandberg & Kiran, 2014). We adapted the protocol in Sandberg and Kiran (2014) for use with intermediate L2 learners. In Experiment 1, we trained American English L2 learners of Spanish on abstract Spanish words in two context-categories (*restaurant* and *university*), while using the context-category *soccer* as a control. In Experiment 2, we trained a second group of American English L2 learners of Spanish either on abstract or concrete words in one context-category (*restaurant*). In Experiment 3, we trained Spanish L2 learners of English on abstract words in the context-category *restaurant*, while using the context-category *soccer* as a control.

Productive versus receptive vocabulary in L2 acquisition

Both first language (L1) and L2 vocabulary knowledge can be broken down into two types: receptive and productive knowledge. Receptive knowledge is the ability to

recall the meaning of a word when its form (written or spoken) is presented; productive knowledge is the ability to produce the correct form (written or spoken) to express a given meaning (Laufer, 1998; Nation, 2013). Comprehension typically precedes production in L1 vocabulary acquisition, and even in adult L1 speakers, receptive vocabulary knowledge exceeds productive vocabulary knowledge (Clark, 1993, 2009). These patterns are mirrored in L2 acquisition, with the gap between receptive and productive knowledge being even more exaggerated, particularly at lower proficiency levels (Schmitt, 2014; Webb, 2008).

Despite the plethora of studies that have been conducted on the gap between receptive versus productive vocabulary size (see Schmitt, 2014) and a general awareness that low productive vocabulary knowledge impedes learning in the L2 classroom, there has been little research on instructional techniques to improve productive vocabulary knowledge. Some previous studies have compared output-based training tasks to receptive-based tasks. Webb (2005) contrasted a sentence-production task, in which participants had to write a sentence containing the target word, with a sentence-reading task (see also Keating, 2008; Pichette et al., 2012). Input about the target word's meaning was provided by presenting the target word with an L1 translation equivalent. De La Fuente (2002) conducted a similar comparison, but input about the target words was contextualized via L2 labels rather than L1 translation equivalents. While the learners in these studies showed greater improvement in the output-based training conditions, the cued forward recall tasks used to assess learning have notable limitations as measures of productive vocabulary knowledge. In these tasks, the learner must produce the L2 words following a prompt such as a picture of the L1 translation equivalent. This shows knowledge of the basic form–meaning link, which is a central component of vocabulary knowledge (Nation, 2013); however, the ability to produce words in context, even in a constrained context such as category association (as in the current study), likely requires a depth of knowledge that goes beyond this basic link (Meara, 2009).

These intervention studies also all focused on novel word learning. The target items were selected to be completely unfamiliar to the participants. However, when investigating productive vocabulary knowledge, it is important to consider not only the acquisition of new lexical items but also the development of productive knowledge for words that are already known receptively (Meara, 1997). To our knowledge, Lee (2003) is the only study to have tracked the progression from receptive to productive knowledge for specific items following a training intervention (Lee & Muncie, 2006, conducted a similar study but did not measure the receptive knowledge of specific items). Lee investigated the productive use of vocabulary in thematically constrained written compositions. She found that after reading a text passage containing the target vocabulary, the learners correctly produced only 13% of the target words they knew receptively. After a reading comprehension exercise plus explicit vocabulary instruction, the learners correctly produced 64% of the receptively known words in their compositions. In addition, they correctly produced 43% of the target items, which had not been known receptively at pretest. There was a small decline in their productive knowledge at a delayed posttest 3 weeks later. This study provides evidence that productive knowledge for new, as well as previously known, words can be gained and maintained following a targeted intervention, but to date this is an understudied topic in L2 vocabulary research.

The concreteness effect in vocabulary learning

When investigating any method of vocabulary instruction, it is important to account for word-level variables, as certain word types pose additional challenges to L2 learners. One such variable is concreteness, which describes the extent to which a concept can be perceived by the senses, as well as how easy it is to create a mental image of the concept in question. Concrete words (e.g., *ambulance*) score higher on concreteness and imageability than abstract words (e.g., *emergency*). Consistent differences have been observed between abstract and concrete words in L1 and L2 acquisition and processing. In the L1, concrete words are acquired earlier (Schwanenflugel, 1991; Spätgens & Schoonen, 2018) and are processed more quickly than abstract words (de Groot, 1989; Schwanenflugel & Shoben, 1983). In L2 lexical processing, concrete words are translated faster and more accurately than abstract words (de Groot, 1992), and in L2 vocabulary acquisition they are learned more quickly (de Groot & Keijzer, 2000) and retained better (van Hell & Mahn, 1997; see Altarriba & Basnight-Brown, 2012, for discussion of further distinctions between emotion words and abstract words). These advantages of concrete over abstract words are collectively termed the "concreteness effect."

Several theories have been developed to explain this effect. The dual coding theory suggests that abstract words are encoded with only verbal information, while concrete words are encoded with both verbal and visual information (Paivio, 1971). The context availability theory posits that concrete words have stronger associations to contextual information than abstract words (Schwanenflugel, Harnishfeger, & Stowe, 1988). In addition, concrete words are thought to have more readily generated semantic features than abstract words (Jones, 1985; Plaut & Shallice, 1991). While the theories cited here differ in their particulars, and this is by no means an exhaustive list of theories, they all highlight key differences between abstract and concrete words that explain persistent difficulties in the acquisition and retrieval of abstract words.

Although the concreteness effect is well documented in the L1 and L2 psycholinguistic literature, few studies that focus on instructional techniques for L2 vocabulary acquisition have considered concreteness as a variable. Where this factor has been investigated, studies have focused on receptive vocabulary knowledge. Van Hell and Mahn (1997) manipulated concreteness as a factor in their comparison of rote rehearsal versus the keyword method on receptive recall in the L2. In both training conditions, translation accuracy on the recall task was significantly higher and speed of recall was faster for concrete versus abstract words, particularly at a 2-week delayed posttest. Zhao and Macaro (2016) similarly assessed learning outcomes for teacher-provided L1 translations versus L2-only definitions with concrete and abstract words. The treatment group, which received L1 translations of the target words, scored better on receptive recall of both concrete and abstract words than the comparison group, which received L2 definitions; however, as the concrete and abstract items were not matched on lexical characteristics, the authors could not compare the differential effects of the two learning conditions on item type. In addition, these studies focused on novel word learning, meaning that the participants had no previous knowledge of the target items. In contrast, the persons with aphasia tested in the studies by Sandberg and colleagues (Kiran et al., 2009; Sandberg &

Kiran, 2014) and the L2 learners in the current study had (at least partial) prior receptive knowledge of the items in the training.

To our knowledge, only Pichette et al. (2012) have specifically investigated how different training techniques can influence the productive vocabulary knowledge of concrete versus abstract words. They found that an output-based learning condition (writing sentences containing the target item) led to better productive recall of both concrete and abstract words over a receptive learning condition (reading sentences containing the target item) at immediate and 1-week delayed posttests. Recall of concrete items was higher than of abstract items across both learning conditions, particularly at immediate posttest, while recall accuracy of all items declined from immediate to delayed posttest. Though these studies show that direct training increases the short-term knowledge of abstract words, these gains typically decline posttreatment, and the learning of abstract words lagged behind that of concrete words.

Aphasia treatment as a model for vocabulary training

A defining characteristic of aphasia is anomia, or the inability to retrieve words from the mental lexicon. Treatments for anomia often focus on strengthening the semantic system to increase activation of the target concept, and thus improve target word selection. One widely used therapy is semantic feature analysis (SFA), first applied to persons with aphasia by Boyle and Coelho (1995). The idea is that training the semantic features of a concept activates the network surrounding the target via spreading activation (Collins & Loftus, 1975). This increases the strength and specificity of the activation of the target concept, which in turn improves the likelihood of choosing the correct lexical label for the target concept (Coelho, McHugh, & Boyle, 2000). While this technique is very effective with directly trained items, results are mixed regarding the transfer of benefit to untrained items, hereafter referred to as generalization (Boyle, 2010).

One promising approach to promote generalization is to combine the basic procedure of SFA with stimulus selection based on the complexity account of treatment efficacy (CATE; Thompson, Shapiro, Kiran, & Sobecks, 2003). Initially, Thompson et al. (2003) found that training more syntactically complex structures (e.g., object relative clauses) results in improvement for not only the directly trained structures but also simpler, related structures (e.g., object clefts) in persons with aphasia. Kiran and Thompson (2003) extended CATE to the semantic realm by training atypical items (e.g., *ostrich*, *artichoke*) in natural categories (e.g., *birds*, *vegetables*). In a series of studies using not only natural but also well-defined (e.g., *shapes*) and ad hoc (e.g., *things at a garage sale*) categories, Kiran found that persons with aphasia exhibited generalization to typical category exemplars when atypical exemplars were trained (see Kiran, 2008, for a review).

More pertinent to the current study, Sandberg and Kiran applied CATE to anomia treatment in aphasia using concreteness as a mode of complexity (Kiran et al., 2009; Sandberg & Kiran, 2014). Abstract words are classified as the more complex items because they are less accurate, take longer to process, have fewer semantic features, have less contextual information associated with them, and are more

difficult to conjure an image for than concrete words (Jones, 1985; Paivio, 1971; Plaut & Shallice, 1991; Schwanenflugel et al., 1988). The word-finding training protocol used by Sandberg and Kiran (Kiran et al., 2009; Sandberg & Kiran, 2014) is an adaptation of SFA, combined with the application of CATE via the use of abstract words as the more complex items. Specifically, the retrieval of abstract words (e.g., *truth*) within a context-category (e.g., *courthouse*) is trained through the selection and discussion of semantic features (e.g., *is generally considered positive*) for each target abstract word. Subsequent performance for the target abstract words is measured via a verbal fluency task within the context-category. Critically, predetermined target concrete words within the same context-category (e.g., *jury*) are not directly trained, but are still tracked to measure generalization to concrete words. In their studies, Sandberg and Kiran have found generalization to concrete words even when only abstract words are trained, but not vice versa.

The success of this training protocol in persons with aphasia merits exploration of applications to other populations, such as L2 learners. It is not uncommon to compare L2 learning and L1 relearning in aphasia (e.g., Cornelissen et al., 2004; López-Barroso & de Diego-Balaguer, 2017; Wray, 2009). For example, Wray (2009) compared the use of formulaic language across different groups and found that both people with aphasia and L2 learners—particularly those L2 learners who struggled to learn target L2 vocabulary—relied on formulaic chunks and imitation during initial stages of vocabulary learning. López-Barroso and de Diego-Balaguer (2017) identified similar possible compensatory mechanisms between L2 learning and language rehabilitation in persons with aphasia, in that both populations recruit either the dorsal or the ventral speech processing streams to support success during language-related tasks.

An important parallel between the two populations is that creating form–meaning connections is a crucial element of vocabulary learning. In persons with aphasia, the connections between previously known forms (the orthographic and acoustic representations of a word) and semantic meanings must be reestablished (Jefferies & Lambon Ralph, 2006), while in adult L2 learners, meanings from the L1 semantic network must be connected to new L2 forms (Jiang, 2000; Kroll & Stewart, 1994). While not identical processes, they share many characteristics, such as the interplay of already known and new information. A further parallel is that these connections may not be established equally at the receptive and productive levels. For L2 learners there is a gap between receptive and productive knowledge (Schmitt, 2014) that may not be fully closed even with direct training (Lee, 2003). Similarly, in several aphasic syndromes comprehension is relatively spared while lexical retrieval deficits in production are pronounced (Papathanasiou & Coppens, 2017).

The current study

The current study extends the training protocol from Sandberg and Kiran (2014) to second language acquisition. As in a lot of research on persons with aphasia, this study was carried out using a multiple-baseline, single-subject research design. In

single-subject research design, intervention is provided at the individual participant level, data are analyzed at the individual participant level, and each participant serves as her own control through the collection of baseline data prior to introducing the intervention (i.e., manipulation of the independent variable). The dependent (outcome) variable is repeatedly measured within and across different phases (e.g., baseline and intervention phases) to determine the effect of manipulating the independent variable on the dependent variable. When the manipulation of the independent variable (e.g., introduction or withdrawal of treatment) results in a change in the dependent variable, the intervention is said to have an effect. Each time the introduction (or withdrawal) of an intervention phase results in changes in the dependent variable, whether within or across participants, the effect of intervention is said to be replicated. It is this replication that establishes methodological rigor. This design is ideal for understanding the specific participant characteristics and environmental conditions under which an intervention may or may not be effective (Kratochwill et al., 2010).

This study design also addresses a typical shortcoming in L2 vocabulary acquisition studies, which is that learning outcomes are often measured only one or two times, and frequently only immediately following treatment. Consequently, any claims about longer term learning are speculative at best (Schmitt, 2010). The design of the current study circumvents this issue, as the generative naming task continues to be administered as a probe even after training in a particular context-category has ended and training in the next context-category begins. Therefore, it is possible to assess the retention of productive vocabulary knowledge across a medium-term time span after the introduction of new information via the training paradigm. This is analogous to a typical L2 classroom, in which learners ideally maintain previously acquired vocabulary while being continuously confronted with new topics and vocabulary.

In persons with aphasia, training the retrieval of abstract words (e.g., *truth*) in a specific context-category (e.g., *courthouse*) by training their semantic features (e.g., *is generally considered positive*) results in both a direct training effect, in which retrieval of the trained abstract words improves, and a generalization effect, in which the retrieval of related but untrained concrete words (e.g., *jury*) also improves. Conversely, training the retrieval of concrete words results in a direct training effect but no generalization effect to untrained abstract words in that context-category. If this training works similarly in L2 acquisition, we predict

1. Training abstract words in a context-category will improve production of the trained abstract words and generalize to the untrained concrete words within the same context-category, but will not promote improvement of words in untrained context-categories.
2. Training concrete words in a context-category will improve production of the trained concrete words, but will not show generalization to the untrained abstract words within the same context-category.

To test the effectiveness of the training paradigm from Sandberg and Kiran (2014) with L2 learners, in Experiment 1 we trained American English L2 learners of

Table 1. Biographical and language background information for all participants

	Experiment 1		Experiment 2		Experiment 3	
	English L2 Spanish		English L2 Spanish		Spanish L2 English	
	M (*SD*)	*Range*	*M* (*SD*)	*Range*	*M* (*SD*)	*Range*
Current age (years)	19.7 (0.6)	19–20	19.8 (1.0)	18–21	24.8 (3.0)	22–31
Use of L2 (years)	8.3 (1.2)	7–9	7.4 (3.0)	5–13	8.6 (2.9)	6–15
Self-rated L2 proficiency (max 7)						
Listening	5.0 (0.0)	5–5	4.7 (1.1)	3–6	3.4 (1.3)	1–5
Speaking	4.7 (0.6)	4–5	4.3 (1.0)	3–6	3.2 (1.0)	2–5
Reading	4.7 (0.6)	4–5	5.4 (1.1)	4–7	4.5 (0.8)	3–6
Writing	5.0 (0.0)	5–5	5.0 (1.3)	3–7	3.5 (1.0)	2–5

Spanish on abstract Spanish words in two context-categories (*restaurant* and *university*) and tracked their productive vocabulary knowledge for abstract and concrete words in these two categories prior to, during, and after training. In Experiment 2, we trained a new group of American English L2 learners of Spanish either on abstract words or concrete words in the context-category *restaurant* to replicate the findings from Experiment 1 and to investigate whether training concrete words rather than abstract words leads to generalization (cf. Kiran et al., 2009). In Experiment 3, we trained Spanish L2 learners of English on abstract words in the context-category *restaurant*, to further replicate findings from Experiment 1 with a larger population with a different L2; we also included a translation recognition task prior to training to systematically investigate how preexisting receptive L2 vocabulary knowledge impacts the effectiveness of this training paradigm.

Experiment 1

Method

Design

This experiment used a single-subject design, with a baseline phase, two training phases, a posttesting phase, a delayed posttest, and a control condition. The order of training phases was counterbalanced across participants. Each participant served as his own control, with the opportunity for replication within and across participants.

Participants

Three American English L2 learners of Spanish (all female) were recruited from intermediate-level Spanish content courses at Penn State University. All participants completed a language background questionnaire (Li, Zhang, Tsai, & Puls, 2014). All participants started learning Spanish after age 7 at school. See Table 1 for complete biographical information.

Materials

Context-category selection. Context-categories and the vocabulary for each category were selected from survey data previously collected from L1 Spanish speakers. Each L1 Spanish speaker generated as many abstract and concrete words as possible within eight categories that represent different contexts in which a variety of abstract and concrete words are thematically related. These speakers also provided a rating of the cultural relevance of each category. For this study, we selected context-categories (a) with a high cultural relevance rating, (b) for which there were a large number of abstract and concrete words produced by at least two survey respondents, and (c) that are normally included in Spanish language learning curricula. We selected *restaurant* and *university* as trained context-categories and *soccer* as the control context-category.

Vocabulary selection. For each context-category (*restaurant*, *university*, and *soccer*), we selected 15 abstract and 15 concrete target words (see Appendix A, Table A.1 for a full list of target words in each context-category). Target words were selected from the L1 Spanish survey data and from Spanish textbooks, in collaboration with L1 Spanish speakers from the Spanish Department at the same university from which the participants were recruited. Abstract words were defined as existing only in the mind or those that are theoretical. Concrete words were defined as existing in reality, or concepts that can be experienced by the senses. Abstract words had a concreteness rating of 400 or lower and concrete words had a concreteness rating of 500 or higher. Within each context-category, abstract and concrete words significantly differed in imageability and concreteness ($ps < .001$) but were balanced on frequency and familiarity ($ps > .1$). In addition, each context-category was balanced on concreteness, imageability, familiarity, and frequency ($ps > .2$). All norms were obtained in English from the MRC Psycholinguistic database (Coltheart, 1981), except frequency, which was obtained from the SUBTLEX databases for American English (Brysbaert, New, & Keuleers, 2012) and Spanish (Cuetos, Glez-Nosti, Barbón, & Brysbaert, 2012). There were no differences between the English and Spanish frequencies within each context-category ($ps > .3$).

Feature selection. A total of 45 semantic features were used in training for each participant. The majority of these features were developed by Sandberg and Kiran (2014). They created 17 general features using dictionary definitions of abstract and concrete (e.g., *exists outside the mind*; *is an object*) and perceptual characteristics (e.g., *can be touched*; *can be tasted*). Fifteen distractor features (e.g., *is poisonous*) were taken from a previous study (Kiran, 2008). Alongside these 32 standard features, 13 individualized, category-specific features were generated by each participant during the first training session for each context-category.

Procedure

Prior to training, participants completed three baseline vocabulary probe sessions (approximately 15 min in length), at least 1 day apart, but no more than 2 weeks apart. After baseline, participants completed six training sessions for each trained context-category (each approximately 30- to 45-min long), at least 2 days apart, but no more than 1 week apart.[1] At each training session, participants completed five

different steps, outlined in greater detail below. During training in the first context-category, the second context-category continued to be baselined. Participants switched to the second trained context-category after they had completed six training sessions in the first context-category. During training of the second context-category, the first context-category continued to be posttested. Vocabulary probes were administered at the beginning of each training session, as well as at delayed posttest, which was approximately 2 weeks after the conclusion of all training for the first two participants and approximately 4 weeks after all training for the third participant.

Vocabulary probes. Productive vocabulary in the three context-categories (*restaurant*, *university*, and *soccer*) was tested through a generative naming task. Participants were instructed to name as many concrete and abstract words as they could within each context-category in 2 min. Responses were divided into four categories: (a) target concrete words (e.g., *camarero* "waiter," *mantel* "tablecloth") (b) target abstract words (e.g., *satisfecho* "satisfied," *calidad* "quality") (c) other concrete words (e.g., *vela* "candle," *propina* "tip"), and (d) other abstract words (e.g., *orden* "order," *felicidad* "happiness"). Responses were counted as correct if they were clear and intelligible productions of the target word or a semantic variant of the target word (e.g., *satisfacción* "satisfaction" for *satisfecho* "satisfied"). Other abstract and other concrete words were words that were appropriate members of the particular context-category but were not one of the target items. Any responses that did not fit into the four categories listed above were not counted. All sessions were audio recorded for reliability.

Training. Each training session consisted of five steps: category sorting, feature selection, yes/no feature questions, synonym generation and concreteness judgment, and free generative naming. These steps were carried out using Qualtrics survey software.

Category sorting. Each participant was presented with 30 (15 abstract and 15 concrete) words from the trained context-category (*restaurant* or *university*) and the control context-category (*soccer*) on the left-hand side of the computer screen in a random order. The participant then sorted the words into the two context-categories by using a mouse to drag the word from the list to a box under the appropriate category heading. If a word was sorted incorrectly, the examiner brought it to the attention of the participant, who self-corrected the error. Feedback was not needed after the first session.

Feature selection. For each trained word, the participant was presented with 45 features (17 general features, 13 category-specific features generated by the participant in the first training session, and 15 unrelated distractor features) in random order at the left-hand side of the computer screen. The participant went down the list and selected the first 6 most descriptive features for each word by using a mouse to drag the feature to a box located under the target word on the right-hand side of the screen.

Yes/no feature questions. The experimenter asked each participant 15 questions about each abstract word using the 45 features. Five questions were taken from each

of the feature categories (e.g., *¿Esto está asociado con un precio más alto?* "Is this associated with a higher cost?"). Care was taken to present each feature an approximately equal number of times for each word across sessions.

Synonym generation and concreteness judgment. Each participant was prompted to recall each word, provide a synonym for the word, and judge whether the word was abstract or concrete.

Free generative naming. Each participant was prompted to name as many words as she could in the trained context-category, with no time limit. During this step, participants were prompted to try to recall the words they had trained that day, as well as any other words they could think of.

Training effect size calculation. Effect sizes (ES) for both trained and untrained (generalized) items were calculated for each participant using a variation of Cohen's *d* statistic as proposed by Beeson and Robey (2006). The mean of the baseline probe scores was subtracted from the mean of the posttreatment probe scores, and then divided by the standard deviation (*SD*) of the baseline probe scores. In cases where the *SD* was 0, the smallest *SD* for that participant was used. The resulting ES were interpreted based on guidelines by Beeson and Robey (2008) for trained (small > 6.5, medium > 8, large > 9.5) and untrained (small > 2, medium > 5, large > 8) items. Although subject-level ES is the primary measure in single-subject research design, the average ES across participants was also calculated to estimate overall efficacy.

Reliability. Interrater reliability was performed on at least 25% of the probes. In Experiment 1, the exact percentage was 37%. A research assistant who did not provide the training, and who spoke both English and Spanish, listened to the audio recordings and scored the probes. A third party then compared the original probe scores with the reliability probe scores and calculated percent agreement. In Experiment 1, the percent agreement was 95%.

Results

The results for each context-category for each participant are presented in Figure 1 and Table 2 provides all ES for each participant. As seen in Figure 1, participants generally produced more concrete words (light gray lines) than abstract words (dark gray lines) across all three context-categories in the baseline phase. Critically, across both trained context-categories (i.e., *restaurant* and *university*) there was little sustained improvement in either concrete or abstract word production during the baseline phase, prior to training. After the onset of training, there is a significant increase in target abstract word production across all three participants in both context-categories, with smaller increases in concrete word production. Of the three participants, two (P1 and P2) showed large ES for the trained abstract words in both trained context-categories (*restaurant*, *university*), and one (P3) showed a small ES in the trained context-category *university*. Of the two participants who showed training effects in the context-category *restaurant*, one (P2) showed a small generalization effect to untrained concrete words. All three participants showed training

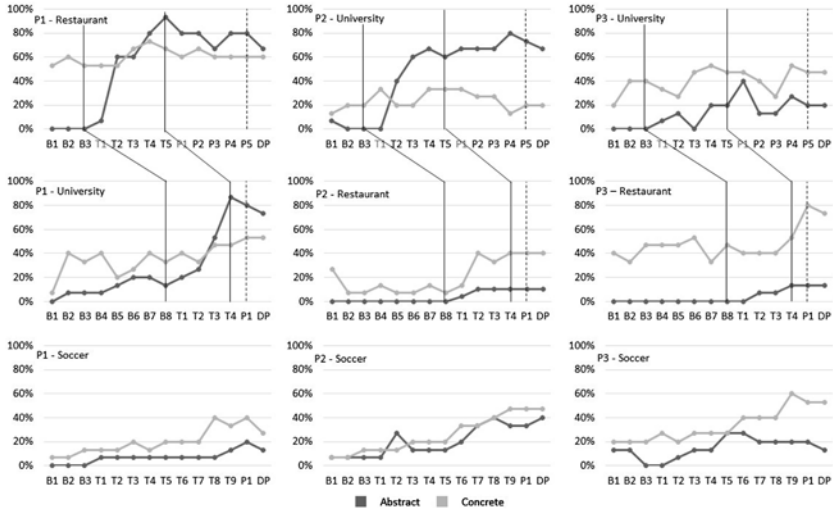

Figure 1. Graphs of participant performance across phases for each context-category in Experiment 1. Each column represents data from a single participant. The top graph is the context-category that was trained first, the middle graph is the context-category that was trained second, and the bottom graph is the control context-category. Solid lines indicate the start and end of training. The dotted line separates the final posttest from the delayed posttest. Performance on abstract words is shown in dark gray, and performance on concrete words is shown in light gray. B = baseline phase; T = training phase; P = post-testing phase; DP = delayed posttest.

effects in the context-category *university*, and one (P1) showed generalization to untrained concrete words. Across participants, there were large average ES for the trained abstract words in both trained context-categories (13.6 and 10.2) and a small generalization ES for concrete words in *restaurant* (3.6). Of interest, both P1 and P2 showed improvement for abstract words in the untrained context-category *soccer*, and all three participants showed improvement for concrete words in the context-category *soccer*. Across participants, there was a small average generalization ES for abstract words in *soccer* (4.7) and a medium generalization ES for concrete words in *soccer* (7.5). All three participants also showed maintenance of the training effects 2–4 weeks after training ended, with only a 0%–13% drop in accuracy across participants.

Discussion

The majority of participants showed the predicted pattern of improved production of both the trained abstract words and the untrained concrete words within the trained context-category. However, a few unexpected results also emerged. First, not all three participants exhibited the expected outcome. P3 did not show a direct training effect for abstract words in the context-category *restaurant*, and although she showed a direct training effect for abstract words in the context-category *university*, it was small and she did not show generalization. This may be due to some inconsistency in the training schedule of P3: she had approximately 3 weeks between the first two sessions of the context-category *restaurant*. Oddly, P3 showed

Table 2. *Effect sizes for target items in each context-category for each participant in Experiment 1*

	Trained				Untrained	
	Restaurant		University		Soccer	
Participant	Abstract	Concrete	Abstract	Concrete	Abstract	Concrete
P1	20.1***	1.5	9.8***	2.0*	5.2**	8.1***
P2	18.7***	4.1*	17.8***	1.6	7.5**	9.8***
P3	1.9	5.1**	3.2*	0.8	1.4	4.7*

Note: *Small effect size. **Medium effect size. ***Large effect size.

improvement for concrete words in the context-category *restaurant*. This pattern is not generalization in the strictest sense, as it was not accompanied by a parallel increase in the production of trained abstract words in the context-category *restaurant*, but may indicate some general enhancement that occurred during training. In addition, while P1 showed generalization in the context-category *university*, she did not show generalization in the context-category *restaurant*, and the reverse is true for P2. However, they were both within .5 of the small ES cutoff for significant generalization effects in each of these respective categories. This points to the possibility that generalization could have occurred across the board—in both context categories for both participants—had additional training taken place.

Second, all three participants improved in the control context-category *soccer*. One possible explanation for this result is that performance may have improved due to testing effects from completing the generative naming task multiple times throughout treatment. *Soccer* is a frequent topic in Spanish language textbooks, and the participants may have had high baseline knowledge of words in this context-category, which was then revealed as they became better at the generative naming task over time. However, *university* and *restaurant* are also common topics in the L2 Spanish classroom, and while *university* showed a similar improvement as *soccer* prior to training, *restaurant* did not. An alternative explanation is that improvement in the context-category *soccer* was due to either learning or activation of existing knowledge via exposure in the category sorting step of training. During the training periods for *restaurant* and *university*, participants saw the words in the context-category *soccer* and had to assign them to the correct category. Activation of the words during this task could have led to an increased ability to retrieve and produce them in the correct category during the generative naming task. However, this hypothesis cannot be disentangled from the possible testing effects discussed above. Thus, in Experiment 2, we implemented two control categories: *university* and *soccer*. *University* was the unexposed control context-category, which appeared only in testing probes, while *soccer* was the exposed control context-category, which was included in the category sorting step during training. This allowed us to isolate the effect of exposure from the effects of task practice.

Another open question from Experiment 1 is that only abstract words were trained, making it difficult to determine whether generalization effects were due to the training protocol in general, or specifically the training of abstract words. In Kiran et al. (2009), the aphasic participants did not show generalization to

abstract words when concrete words were trained using this paradigm, but this effect has not yet been tested in L2 learners. Thus, in Experiment 2, we trained half of the participants using concrete words and half using abstract words. Finally, given that Experiment 1 only involved three participants, Experiment 2 also serves more generally as a replication, to see if the pattern of findings from Experiment 1 can be replicated with a second group of American English L2 learners of Spanish.

Experiment 2

Method

Design

As in Experiment 1, this Experiment 2 used a single-subject design. In Experiment 2, there was a baseline phase, a training phase, and a posttesting phase. The type of training (abstract vs. concrete) was counterbalanced across participants (see Procedure).

Participants

Seven American English L2 learners of Spanish (six females, one male) were recruited from the same intermediate-level Spanish content courses at Penn State University as those participants from Experiment 1. All participants started learning Spanish after age 7 at school. See Table 1 for complete biographical information.

Materials

The stimuli and materials were identical to those in Experiment 1.

Procedure

The procedures and testing measures were identical to those in Experiment 1, except that four of the seven participants were only trained on abstract words in the context-category *restaurant*, and three participants were only trained on concrete words in the context-category *restaurant*. In addition, while both *university* and *soccer* served as control categories, only words from the context-category *soccer* were included in the category sorting step during training, while participants were not exposed to the context-category *university* at all during training. Interrater reliability was performed on 29% of the data with 95% agreement.

Results

The results for the context-category *restaurant* for each participant are presented in Figure 2 and Table 3 provides all ES for each participant. As in Experiment 1, participants generally produced more concrete words (light gray lines) than abstract words (dark gray lines) in the baseline phase, and there was little sustained improvement in either concrete or abstract word production during the baseline phase, prior to training. After the onset of training, there is a significant increase in target word production.

All four participants who were trained on abstract words in the context-category *restaurant* (P1, P3, P4, P5) showed medium to large direct training ES. Two participants (P3, P4) also exhibited small generalization effects for concrete words in the context-category *restaurant*. Across participants, the average ES for the trained

Table 3. *Effect sizes for target items in each context-category for each participant in Experiment 2*

	Trained		Untrained			
	Restaurant		University (Unexposed)		Soccer (Exposed)	
Participant	Abstract	Concrete	Abstract	Concrete	Abstract	Concrete
Trained on abstract						
P1	22.4***	1.2	3.4*	1.0	11.5***	1.7
P3	8.7**	2.0*	0.6	2.8*	1.7	3.0*
P4	10.4***	4.0*	0.0	0.6	−2.3	1.2
P5	20.8***	0.0	0.4	3.0*	1.3	2.0*
Trained on concrete						
P2	0.0	4.6	0.0	−2.0	0.0	4.6*
P6	0.0	9.2**	1.2	0.6	0.0	2.3*
P7	0.0	3.7	1.7	0.6	1.7	0.6

Note: *Small effect size. **Medium effect size. ***Large effect size.

Figure 2. Graphs of participant performance across phases for the trained context-category *restaurant* in Experiment 2. Graphs on the left are from the participants who were trained on concrete words; graphs on the right are from the participants who were trained on abstract words. Solid lines indicate the start and end of training. Performance on abstract words is shown in dark gray, and performance on concrete words is shown in light gray. B = baseline phase; T = training phase; P = posttesting phase.

abstract words was large (15.6), while the average generalization ES for concrete words in the trained category approached the small threshold (1.8). In contrast, although all three participants who were trained on concrete words in the context-category *restaurant* showed some improvement, only one participant (P6) achieved a medium direct training ES. None of these participants showed generalization to abstract words in the trained context-category. The average ES across participants did not reach the small threshold for trained concrete words (5.9) and was zero for untrained abstract words.

Turning to the control categories, P1, who improved on trained abstract words in the context-category *restaurant*, also showed a small effect for untrained abstract words in the unexposed control context-category *university* and a large effect for untrained abstract words in the exposed control context-category *soccer*. P3 and P5, who also improved on trained abstract words in the context-category *restaurant*, showed small effects for untrained concrete words in both the unexposed and the exposed control categories. P2, who did not improve on trained concrete words in the context-category *restaurant*, and P6, who did, both showed small effects on untrained concrete words in the exposed control context-category, *soccer* (see Appendix A, Figures A.1 and A.2 for the corresponding figures of the untrained context-categories *university* and *soccer* for each participant). Across all participants, average ES did not reach the small threshold for either abstract (1.0) or concrete (0.9) words in *university*, and just reached the small threshold for both abstract (2.0) and concrete (2.2) words in *soccer*.

Discussion

As in Experiment 1, participants who were trained on abstract words showed generalization to concrete words in the same context-category. Three of the four participants who improved on the trained abstract words (P1, P3, P4) also showed some improvement for untrained concrete words, although only two (P3, P4) achieved small ES. This replicates the training and generalization effects from Experiment 1, supporting the notion that training abstract words results in not only the improvement of trained abstract words but also a transferrable benefit to concrete words in the same context-category.

In stark contrast, of the three participants who were trained on concrete words, although all three showed some improvement, only one (P6) showed a medium ES for trained concrete words. This is most likely due to the lower variability of P6 at baseline (*SD*s: P6 = 0.58 vs. P2 = 1.15, P7 = 1.53), as the raw difference in gains from pretest to posttest is similar across these participants (Δs: P6 = 5.33 vs. P2 = 5.33, P7 = 5.67). More importantly, none of these three participants showed any improvement on the untrained abstract words in the same context-category. Thus, paralleling previous findings among persons with aphasia (Kiran et al., 2009), American English L2 learners of Spanish show generalization from abstract to concrete words, but not vice versa, as predicted by the CATE (Thompson et al., 2003).

Turning to the question of activation via exposure, although three participants showed small effects in the unexposed control context-category *university* (P1, P3, P5), five participants showed small to large effects in the exposed control context-category *soccer* (P1, P2, P3, P5, P6). This bias toward improvement in the exposed context-category suggests that it is the exposure to items during training that is driving improvement in untrained categories, rather than simply repeated practice with the testing probes over the course of the experiment.

One open question from Experiments 1 and 2 is whether the positive effect of this training paradigm is population specific, as both experiments were completed with American English L2 learners of Spanish. Thus, Experiment 3 was performed using the same approach as Experiment 2 (except that all participants were trained on

abstract words) with Spanish L2 learners of English in Granada, Spain. Another outstanding question is the impact of preexisting receptive L2 vocabulary knowledge on the effectiveness of this training paradigm and, as a corollary, whether receptive L2 vocabulary changes as a result of training. Previous research has shown that receptive vocabulary knowledge precedes productive knowledge, but also that there is a correlation between receptive vocabulary size (knowledge) and productive vocabulary knowledge (Laufer, 1998; Webb, 2008). Therefore, it seems likely that higher preexisting receptive knowledge of the target words would lead to greater improvement in productive knowledge. It is also possible that receptive knowledge will increase as a result of the training (see, e.g., Laufer & Goldstein, 2004). Thus, in Experiment 3, we implemented a translation recognition task that participants performed before and after the training to test this hypothesis.

Experiment 3

Methods

Design

As in Experiment 2, Experiment 3 used a single-subject design, in which there was a baseline phase, a training phase, and a posttesting phase. In Experiment 3, only one training condition was used across participants.

Participants

Ten Spanish L2 learners of English (8 females, 2 males) participated in this study. All participants were native Spanish speakers living in Granada, Spain, who started learning English after age 7. See Table 1 for complete biographical information.

Materials

The stimuli and materials were identical to those in Experiments 1 and 2, except that the target words were in participants' L2 English. In addition, as a measure of receptive vocabulary knowledge, participants completed a translation recognition task (Sunderman & Kroll, 2006) prior to and after training.[2] This task contained the 15 abstract and 15 concrete L2 English words belonging to the context-category *restaurant*, all of which were paired with their correct L1 Spanish translation. In addition, the task contained 47 concrete L2 English filler items with correct L1 Spanish translations, 87 concrete L2 English filler items with incorrect L1 Spanish translations, 43 abstract L2 English filler items with correct L1 Spanish translations, and 33 abstract L2 English filler items with incorrect L1 Spanish translations, based on items from Sunderman and Kroll (2006). Across the entire task, participants saw an equal number of correct and incorrect translation pairs.

Procedure

The procedures were identical to those in Experiment 2, except that all participants were trained on abstract words in the context-category *restaurant*, and all training was conducted in English. Participants also completed the translation recognition task immediately following the final baseline probe and after the completion of the

Table 4. *Results of Wilcoxon signed-rank tests*

	Baseline				Posttest			
	M(*SD*)	Bootstrapped 95% CI	*Mdn*	IQR	*M*(*SD*)	Bootstrapped 95% CI	*Mdn*	IQR
Concrete	82.7 (11.0)	[76.0, 89.3]	80.0	16.7	97.3 (3.4)	[95.3, 99.3]	100.0	6.7
Abstract	76.7 (17.7)	[66.7, 86.0]	76.7	30	95.3 (6.3)	[91.3, 98.7]	96.7	6.7

Note: IQR, interquartile rank. CI, confidence interval.

final posttest probe. In the translation recognition task, each trial began with a fixation point, which remained on the screen until participants pressed the space bar. After an ISI of 100 ms, an L2 English word appeared in the center of the screen for 400 ms. After an additional ISI of 100 ms, a Spanish word appeared on the screen. The Spanish word remained on the screen until participants pressed a "yes" or "no" button on the keyboard to indicate whether the Spanish word was a correct or incorrect translation of the English word. Interrater reliability was performed on 32% of the probe data with 98% agreement.

Results

Translation recognition

Participants' accuracy on concrete and abstract *restaurant* words at baseline and posttest are presented in Table 4. Wilcoxon signed-rank tests revealed no significant difference in participants' recognition of abstract versus concrete words at either baseline or posttest (baseline: $Z = -0.99$, $p = .324$, $d = 0.41$; posttest: $Z = -0.63$, $p = .527$, $d = 0.40$). However, they exhibited significant improvement in their recognition of both abstract and concrete words from baseline to posttest (concrete: $Z = -2.54$, $p = .011$, $d = 1.79$; abstract: $Z = -2.20$, $p = .028$, $d = 1.40$).

Training outcomes

The results for the context-category *restaurant* for each participant are presented in Figure 3, and Table 5 provides all ES for each participant. As seen in Figure 3, participants generally produced more concrete words (light gray lines) than abstract words (dark gray lines) in the baseline phase and there was little sustained improvement in either concrete or abstract word production during the baseline phase, prior to training. After the onset of training, there was a significant increase in word production in the trained context-category (*restaurant*) for the majority of the participants, with an average direct training ES of 9.4 (bordering on large), and an average generalization ES of 8.5 (large). Specifically, 7 of the 10 participants (P1, P3, P4, P5, P6, P7, P10) showed medium to large ES for the trained abstract words. All 7 participants who exhibited direct training effects also exhibited small to large generalization effects. The 3 participants who did not show improvement on the trained abstract words (P2, P8, P9) nonetheless showed improvement for the untrained

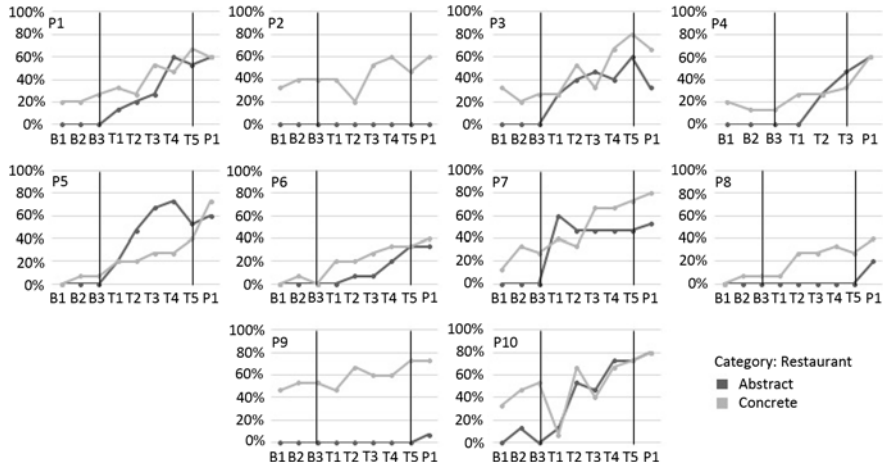

Figure 3. Graphs of participant performance across phases for the trained context-category *restaurant* in Experiment 3. Solid lines indicate the start and end of training. Performance on abstract words is shown in dark gray, and performance on concrete words is shown in light gray. B = baseline phase; T = training phase; P = posttesting phase.

concrete words in the context-category *restaurant*. While this pattern is not generalization per se, as it is not accompanied by a direct training effect with abstract words, it may indicate some general enhancement that occurred during training.

As seen in Table 5, only two participants (P3, P9), one of whom showed no direct training effects (P9), showed small ES in the unexposed control context-category *university* (see Figure A.3 in Appendix A for the corresponding figures for each participant). Across participants, the average ES for abstract and concrete words in the unexposed control context-category were 0.0 and 0.5, respectively. In comparison, all 10 participants showed improvement for concrete words in the exposed control context-category *soccer*, with an average ES of 6.0 (medium). Seven participants (P1, P3, P4, P5, P7, P9, P10), one of whom showed no direct training effects (P9), showed improvement for abstract words in the exposed control context-category as well, and the average ES across all 10 participants was 5.6 (medium; see Figure A.4 in Appendix A for the corresponding figures for each participant). At the same time, the mean slopes of improvement for abstract and concrete words in the exposed control context-category *soccer* (abstract slope: $M = 0.65$, $SD = 0.49$; concrete slope: $M = 0.75$, $SD = 0.39$) were descriptively lower than those of the trained context-category *restaurant* (abstract slope: $M = 0.93$, $SD = 0.98$; concrete slope: $M = 1.14$, $SD = 0.41$). The steeper slope in the trained context-category suggests not only a larger improvement (as also evidenced by the larger ES) but also a more immediate effect.

Discussion

The majority of participants showed the predicted pattern of improved production of both the trained abstract words and the untrained concrete words within the

Table 5. *Effect sizes for target items in each context-category for each participant in experiment*

	Trained		Untrained			
	Restaurant		University (Unexposed)		Soccer (Exposed)	
Participant	Abstract	Concrete	Abstract	Concrete	Abstract	Concrete
P1	15.6***	9.8***	0.0	−1.7	5.2**	8.0***
P2	0.0	5.8**	0.0	0.0	1.2	3.0*
P3	8.7**	6.0**	0.0	2.3*	8.7***	2.3*
P4	14.4***	11.5***	−0.6	−0.2	2.9*	8.1***
P5	15.6***	17.9***	0.0	0.0	10.4***	11.2***
P6	8.7**	9.8***	0.0	0.6	1.7	4.0*
P7	13.9***	5.5**	0.6	0.9	10.4***	7.5**
P8	5.2	9.2***	0.0	−1.0	0.0	4.5*
P9	1.7	5.8**	0.0	2.9*	3.5*	4.1*
P10	9.8***	3.5*	0.0	1.2	12.1***	7.5**

Note: *Small effect size. **Medium effect size. ***Large effect size.

trained context-category *restaurant*. This shows that this training is effective not only for American English L2 learners of Spanish but also for Spanish L2 learners of English. This finding is underscored by the fact that receptive knowledge of the abstract and concrete words did not significantly differ prior to training, as seen in the results of the translation-recognition task.[3] Therefore, we can conclude that the observed improvements in productive knowledge on both abstract and concrete words was not due to higher preexisting receptive knowledge of one word type versus the other. In addition, training led to an increase in receptive knowledge of both abstract and concrete words in the context-category *restaurant* from baseline to posttest. This parallels previous findings that training productive use can lead to improvement in receptive knowledge (Laufer & Goldstein, 2004; Mondria & Wiersma, 2004), and extends this notion to encompass the transfer of generalization effects in production to receptive knowledge.

The effectiveness of the training is also shown by the lack of improvement in the untrained, unexposed control context-category *university* across all testing times. However, the control context-category *soccer*, which appeared in the category sorting step of each training session, showed improvement for both abstract and concrete words. Both *university* and *soccer* are highly familiar categories for these participants and had the same testing schedule, suggesting that this improvement in the untrained exposed context-category *soccer* stems not from practice effects, but rather repeated exposure through the category sorting step. This exposure may help improve productive vocabulary through activation of existing knowledge, which then becomes more accessible during the productive generative naming task. Critically, the average ES and the slopes of improvement for both abstract and concrete words in the exposed context-category *soccer* were lower than those in the

trained context-category *restaurant*, underscoring the value of the full training procedures over simple exposure via the category sorting task.

General discussion

In all three experiments, the majority of participants showed the predicted pattern of improved production of the trained abstract words within the trained context-category and at least small generalization effects to the untrained concrete words in the same context-category. This effect emerged in both American English L2 learners of Spanish and Spanish L2 learners of English, demonstrating that these effects are not unique to a specific population. In Experiment 2, when concrete words were trained, generalization to abstract words was nonexistent. In Experiment 3, results from a translation-recognition test revealed that participants had receptive knowledge of the trained items prior to training, and that receptive vocabulary for both abstract and concrete words improved in posttraining measures. Each of these findings will be addressed in detail below.

Paralleling studies of word-finding therapy in aphasia (Kiran et al., 2009; Sandberg & Kiran, 2014), across all three experiments, training abstract words led to improved production of the trained abstract words and this generalized to improved production of untrained concrete words. Those participants directly trained on concete words in Experiment 2 exhibited no generalization effects to the subsequent production of abstract words, also paralleling previous research on persons with aphasia (Kiran et al., 2009). Further, it is worth noting that while all three participants who were trained on concrete words in Experiment 2 made some improvement, only one participant achieved a significant direct training ES for concrete words, suggesting that even direct training effects are smaller when participants are trained on concrete versus abstract words. Together, these results are in line with CATE (Thompson et al., 2003), in that training more complex items (i.e., abstract words) results in generalization to less complex, related items (i.e., concrete words), but not vice versa.

The application of the CATE model to abstract word learning relies on the assumption that concepts are stored within a semantic network with weighted connections to other, related concepts, and that when one concept in the network is activated during language comprehension or production, related items are also activated through spreading activation (Collins & Loftus, 1975). Within the semantic network, abstract words generally have fewer semantic features (Plaut & Shallice, 1991), and are more semantically diverse (i.e., are found in more contexts and have more variability in meaning) than concrete words (Hoffman, Lambon Ralph, & Rogers, 2013). This semantic diversity means that abstract concepts are weakly connected to a large number of other concepts within the network, with more spreading activation, while concrete concepts have strong and specific representations, with less spreading activation. These stronger and more specific representations for concrete concepts lead to faster and more accurate retrieval of lexical forms relative to abstract concepts during language production, even among nonclinical populations (Newton & Barry, 1997). However, in terms of generalization, the specificity of concrete words, which limits spreading activation, may also limit generalization. While generalization has been

observed after concrete word training, it appears to be limited to concrete words that share semantic features with the trained items (e.g., Kiran, 2008).

Though the underlying lexical–semantic network likely differs between the L1 in persons with aphasia and the L2 in adult language learners, there may be similarities that lead to parallel findings between L1 aphasia research and the results of the current study. In aphasia, concepts are not lost, but retrieving the accurate lexical form during language production is difficult due to brain injury. In L2 learners, established L1 concepts have acquired a new L2 lexical mappings that are often less stable, especially at lower proficiency levels (Kroll & Stewart, 1994). In both cases, regardless of potential differences in the underlying organizational structure, the connections between lexical forms and their corresponding concepts are weak. The training paradigm developed by Kiran and Sandberg (Kiran et al., 2009; Sandberg & Kiran, 2014) targets these lexical–conceptual connections. By training semantic features, the conceptual representation is strengthened, allowing for faster and more accurate retrieval of the lexical form (Newton & Barry, 1997). Because abstract concepts are weaker and more diverse to begin with, this training method is particularly effective for boosting productive knowledge of abstract words. An added benefit of this type of training is that because of the broad associative network for abstract concepts, the connections between the trained abstract concepts and other concepts in the semantic network are similarly strengthened through spreading activation, leading to improvement on retrieval of the untrained, but related words in the same context-category (see Kiran et al., 2009; Sandberg & Kiran, 2014).

In Experiment 3, we included a receptive translation-recognition vocabulary task, the results of which showed that participants had significant receptive knowledge of the trained items prior to beginning the training intervention, even though their performance in the timed generative naming task was low at baseline, particularly for abstract words. This is not unexpected, as receptive vocabulary size is typically larger than productive vocabulary size (Nation, 2013; Webb, 2008). This also suggests that the participants may have already had some knowledge of the forms and meanings of the L2 target words, but did not yet have the ability to actively recall and produce them in a controlled context, namely, associations within a context-category (Laufer & Nation, 1999; Nation, 2013). We also found that receptive knowledge of the target words improved as a result of the abstract word training, suggesting that among L2 learners, the benefits of this training extend to receptive vocabulary knowledge. However, it must be noted that the task we used (translation recognition) represents the most basic form of receptive knowledge (passive recognition, according to the taxonomy outlined in Laufer & Goldstein, 2004). It is uncertain whether performance at pretest—and subsequent improvement at posttest—would be similarly high had we used a different measure of receptive knowledge that required a higher level of skill.

The fact that the participants in Experiment 3 exhibited a high level of receptive vocabulary knowledge at baseline also raises the question of whether this training would be effective for learners at lower L2 proficiency levels, who have an overall smaller receptive vocabulary size and who are less likely to have any preexisting knowledge of L2 words in a particular context-category. Preliminary results from a follow-up study in an L2 classroom setting suggest that this training paradigm can also lead to increased productive knowledge of abstract and

concrete vocabulary among less-proficient L2 learners (Kerschen, Sandberg, Carpenter, & Jackson, 2018).

Unexpectedly, participants in all three experiments also showed improvement outside of the trained context-category, mainly in the exposed control context-category *soccer*. This is most likely due to the inclusion of this context-category during the category sorting step of training. In this task the participants were exposed to the target abstract and concrete words in *soccer* and had to process them for meaning in order to place them into the appropriate category. They also received feedback on their responses until all words had been sorted correctly. Therefore, this task provided some training of the form and meaning of the target words in *soccer*, which could have led to the acquisition of and/or activation of existing knowledge for these words. Evidence to support this interpretation comes from Experiments 2 and 3, in which little improvement was found on an unexposed control context-category (*university*), where participants were not exposed to any target words at any point during training. Of importance here, both *university* and *soccer* were tested equally often during the baseline and testing probes and both categories likely had a similar level of baseline familiarity the learners, suggesting that the improved performance in the context-category *soccer* was not simply the result of repeated testing. This improvement was also surprisingly robust, with multiple participants showing large ES, suggesting that the category sorting task played a key role in the observed training effects. Nonetheless, this small amount of training for *soccer* differed in quantity and quality from the full training for the context-category *restaurant*, and results from all three experiments showed that improvement on the generative naming task was greater in the directly trained context-category. However, these findings suggest that future research should test individual training components separately, and more accurately measure receptive knowledge of target words in the control categories prior to training, to help identify how each task contributes to the development of productive knowledge of target vocabulary.

An alternative explanation for the training effects observed in this study is that prior to training, participants had the productive knowledge, but failed to produce target words because they were not considered category exemplars and that training increased the scope of the category. While plausible, one argument against this explanation is that target words were chosen based on consensus and typical vocabulary for L2 learners, and the participants were specifically prompted to name both concrete and abstract words in the instructions for the vocabulary probes both prior to and after training. In addition, in Experiment 2, participants who were trained on concrete words saw the abstract words during category sorting and were therefore cued that these words belong in the category *restaurant*. However, they did not end up producing any of the target abstract words by the end of training, which one would have predicted to occur if the training effects were simply the result of participants learning to increase the scope of possible category exemplars over time.

Conclusion

This L2 vocabulary training study applied a treatment protocol originally developed to target lexical retrieval deficits in patients with aphasia (Sandberg & Kiran, 2014)

to increase the production of abstract words by L2 learners of English and Spanish. Paralleling findings from aphasia research, the L2 learners showed improvement in productive knowledge of the trained abstract words as well as untrained concrete words from the same context-category (Sandberg & Kiran, 2014). Given these results, this training paradigm is a potentially useful tool in L2 vocabulary instruction, as it not only is a successful direct training method that promotes the development of productive knowledge of difficult-to-learn abstract words in the L2, but also shows promise for facilitating the ability to produce untrained concrete words via generalization (Boyle, 2010; Sandberg & Kiran, 2014). In addition, it can be broadly applied to learners of different languages with different L1 backgrounds. By taking a training paradigm already in use among people with aphasia and adapting it for use with L2 learners, the results from the present study also highlight the value of increased talk between disciplines. By identifying theories and methodological approaches that can be fruitfully applied across fields, such interdisciplinary work has the potential to push language research in exciting new directions.

Acknowledgments. This work was supported in part by National Science Foundation Office of International Science and Engineer Grant OISE-0968369 (to Judith F. Kroll, PI; and Paola Dussias and Janet van Hell, co-PIs). The authors would also like to acknowledge the research assistants who performed inter-rater reliability: Lynn Ables, Carina Lindrooth, Alexandra Santos, and Sierra Clemetson.

Notes

1. For one participant there were two sessions that occurred three weeks apart because of a holiday break.
2. In addition, participants completed a semantic priming task immediately following the final baseline probe and after the completion of the final posttest probe. This task included the critical abstract and concrete words in the category *restaurant*. However, this task will not be discussed further.
3. The lack of a concreteness effect seen in the accuracy of the TR task is not unsurprising, because concreteness effects in healthy adults are most often observed in reaction time data due to ceiling effects in accuracy on basic tasks such as this.

References

Altarriba, J., & Basnight-Brown, D. (2012). The acquisition of concrete, abstract, and emotion words in a second language. *International Journal of Bilingualism*, **16**, 446–452.

Barcroft, J. (2004). Second language vocabulary acquisition: A lexical input processing approach. *Foreign Language Annals*, **37**, 200–208.

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychology Review*, **16**, 161–169.

Beeson, P. M., & Robey, R. R. (2008). *Meta-analysis of aphasia treatment outcomes: Examining the evidence. Paper presented at the Clinical Aphasiology Conference*, Jackson Hole, WY.

Boyle, M. (2010). Semantic feature analysis treatment for aphasic word retrieval impairments: What's in a name? *Topics in Stroke Rhabilitation*, **17**, 411–422.

Boyle, M., & Coelho, C. A. (1995). Application of semantic feature analysis as a treatment for aphasic dysnomia. *American Journal of Speech Language Pathology*, **4**, 94–98.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, **44**, 991–997. doi:10.3758/s13428-012-0190-4

Clark, E. V. (1993). *The lexicon in acquisition*. Cambridge: Cambridge University Press.

Clark, E. V. (2009). *First language acquisition* (2nd ed.). Cambridge: Cambridge University Press.

Coady, J., & Huckin, T. N. (1997). *Second language vocabulary acquisition: A rationale for pedagogy*. New York: Cambridge University Press.

Coelho, C. A., McHugh, R. E., & Boyle, M. (2000). Semantic feature analysis as a treatment for aphasic dysnomia: A replication. *Aphasiology*, **14**, 133–142.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, **82**, 407–428.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, **33**, 497–505.

Cornelissen, K., Laine, M., Renvall, K., Saarinen, T., Martin, N., & Salmelin, R. (2004). Learning new names for new objects: Cortical effects as measured by magnetoencephalography. *Brain and Language*, **89**, 617–622.

Cuetos, F., & Glez-Nosti, M., & Barbón, A., & Brysbaert, M. (2012). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, **33**, 133–143.

de Groot, A. M. (1989). Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**, 824–845.

de Groot, A. M. (1992). Bilingual lexical representation: A closer look at conceptual representations. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 389–412). Amsterdam: Elsevier Science.

de Groot, A. M., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, **50**, 1–56.

De La Fuente, M. J. (2002). Negotiation and oral acquisition of L2 vocabulary: The roles of input and output in the receptive and productive acquisition of words. *Studies in Second Language Acquisition*, **24**, 81–112.

Farley, A. P., Ramonda, K., & Liu, X. (2012). The concreteness effect and the bilingual lexicon: The impact of visual stimuli attachment on meaning recall of abstract L2 words. *Language Teaching Research*, **16**, 449–466.

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, **45**, 718–730.

Jefferies, E., & Lambon Ralph, M. A. (2006). Semantic impairment in stroke aphasia versus semantic dementia: A case-series comparison. *Brain*, **129**, 2132–2147.

Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, **21**, 47–77.

Jones, G. V. (1985). Deep dyslexia, imageability, and ease of predication. *Brain and Language*, **24**, 1–19.

Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research*, **12**, 365–386.

Kerschen, K., Sandberg, C. W., Carpenter, E., & Jackson, C. N. (2018). *The effect of abstract word training on productive L2 vocabulary knowledge: A classroom-based study*. Paper presentation at the Germanic Linguistics Annual Conference (GLAC), University Park, PA, 11–12 May 2018.

Kiran, S. (2008). Typicality of inanimate category exemplars in aphasia treatment: Further evidence for semantic complexity. *Journal of Speech Language and Hearing Research*, **51**, 1550–1568.

Kiran, S., Sandberg, C., & Abbott, K. (2009). Treatment for lexical retrieval using abstract and concrete words in persons with aphasia: Effect of complexity. *Aphasiology*, **23**, 835–853.

Kiran, S., & Thompson, C. K. (2003). The role of semantic complexity in treatment of naming deficits: Training semantic categories in fluent aphasia by controlling exemplar typicality. *Journal of Speech Language and Hearing Research*, **46**, 608–622.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. What Works Clearinghouse. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, **33**, 149–174.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, **19**, 255–271.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, **54**, 399–436.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, **16**, 33–51.

Lee, S. H. (2003). ESL learners' vocabulary use in writing and the effects of explicit vocabulary instruction. *System*, **31**, 537–561.

Lee, S. H., & Muncie, J. (2006). From receptive to productive: Improving ESL learners' use of vocabulary in a postreading composition task. *TESOL Quarterly*, **40**, 295–320.

Li, P., Zhang, F., Tsai, E., & Puls, B. (2014). Language history questionnaire (LHQ 2.0): A new dynamic web-based research tool. *Bilingualism: Language and Cognition*, **17**, 673–680.

López-Barroso, D., & de Diego-Balaguer, R. (2017). Language learning variability within the dorsal and ventral streams as a cue for compensatory mechanisms in aphasia recovery. *Frontiers in Human Neuroscience*, **11**(Article 476), 1–7.

Meara, P. (1980). Vocabulary acquisition: A neglected aspect of language learning. *Language Teaching and Linguistics: Abstracts*, **13**, 221–246.

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In N. Schmitt & McCarthy, M. (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 109–121). Cambridge: Cambridge University Press.

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition*. Amsterdam: Benjamins.

Mondria, J. A., & Wiersma, B. (2004). Receptive, productive, and receptive+productive L2 vocabulary learning: What difference does it make? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 79–100). Amsterdam: Benjamins.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.

Newton, P. K., & Barry, C. (1997). Concreteness effects in word production but not word comprehension in deep dyslexia. *Cognitive Neuropsychology*, **14**, 481–509.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart, and Winston.

Papathanasiou, I., & Coppens, P. (2017). *Aphasia and related neurogenic communication disorders* (2nd ed.). Burlington, MA: Jones & Bartlett Learning.

Pichette, F., De Serres, L., & Lafontaine, M. (2012). Sentence reading and writing for second language vocabulary acquisition. *Applied Linguistics*, **33**, 66–82.

Plaut, D. C., & Shallice, T. (1991). *Effects of word abstractness in a connectionist model of deep dyslexia*. Paper presented at the 13th Annual Conference of the Cognitive Science Society, Chicago, IL.

Sandberg, C., & Kiran, S. (2014). How justice can affect jury: Training abstract words promotes generalisation to concrete words in patients with aphasia. *Neuropsychological Rehabilitation*, **24**, 738–769.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, **64**, 913–951.

Schwanenflugel, P. J. (1991). Why are abstract concepts hard to understand? In P. J. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 223–250). Hillsdale, NJ: Erlbaum.

Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, **27**, 499–520.

Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **9**, 82–102.

Spätgens, T., & Schoonen, R. (2018). The semantic network, lexical access, and reading comprehension in monolingual and bilingual children: An individual differences study. *Applied Psycholinguistics*, **39**, 225–256.

Sunderman, G., & Kroll, J. F. (2006). First language activation during second language lexical processing: An investigation of lexical form, meaning, and grammatical class. *Studies in Second Language Acquisition*, **28**, 387–422.

Thompson, C., Shapiro, L., Kiran, S., & Sobecks, J. (2003). The role of syntactic complexity in treatment of sentence deficits in agrammatic aphasia: the complexity account of treatment efficacy (CATE). *Journal of Speech Language and Hearing Research*, **46**, 591–607.

van Hell, J. G., & Mahn, A. C. (1997). Keyword mnemonics versus rote rehearsal: Learning concrete and abstract foreign words by experienced and inexperienced learners. *Language Learning*, **47**, 507–546.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, **27**, 33–52.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, **30**, 79–95.

Wray, A. (2009). Formulaic language in learners and native speakers. *Language Teaching*, **32**, 213–231.

Zhao, T., & Macaro, E. (2016). What works better for the learning of concrete and abstract words: Teachers' L1 use or L2-only explanations? *International Journal of Applied Linguistics*, **26**, 75–98.

Appendix A

Table A.1. *Target abstract and concrete words for each context-category*

Restaurant	Spanish	English
abstract	satisfecho	satisfied
abstract	rico	rich
abstract	gusto	taste
abstract	picante	spicy
abstract	hambre	hunger
abstract	gratis	free
abstract	calidad	quality
abstract	caro	expensive
abstract	sabrosa	delicious
abstract	ruidoso	noisy
abstract	barato	cheap
abstract	olores	smells
abstract	crudo	raw
abstract	preferencia	preference
abstract	asco	gross
concrete	bebidas	drinks
concrete	tarjeta de crédito	credit card
concrete	servilleta	napkin
concrete	tenedor	fork
concrete	cuchara	spoon
concrete	cuchillo	knife
concrete	comida	food
concrete	cena	dinner
concrete	almuerzo	lunch
concrete	desayuno	breakfast
concrete	cuenta	check
concrete	utensilios	utensils
concrete	camarero	waiter
concrete	mantel	tablecloth
concrete	menú	menu
University	Spanish	English

(*Continued*)

Table A.1. (*Continued*)

Restaurant	Spanish	English
abstract	aprendizaje	learning
abstract	conocimiento	knowledge
abstract	curso	course
abstract	éxito	success
abstract	nota	grade
abstract	materia	subject
abstract	discurso	speech
abstract	fácil	easy
abstract	perezoso	lazy
abstract	lectura	reading
abstract	aburido	boring
abstract	especialidad	major
abstract	motivación	motivation
abstract	estrés	stress
abstract	difícil	difficult
concrete	profesor	professor
concrete	alumno	student
concrete	biblioteca	library
concrete	pizarrón	blackboard
concrete	escritorio	desk
concrete	lápiz	pencil
concrete	pluma	pen
concrete	tarea	homework
concrete	compañero	classmate
concrete	laboratorio	laboratory
concrete	dormitorio	dormitory
concrete	apuntes	notes
concrete	computadora	computer
concrete	mochila	backpack
Soccer	Spanish	English
abstract	emocionante	exciting
abstract	ganar	win
abstract	perder	lose
abstract	tirar	throw

(*Continued*)

Table A.1. (*Continued*)

Restaurant	Spanish	English
abstract	partido	game
abstract	patear	kick
abstract	competición	competition
abstract	habilidad	ability
abstract	posición	position
abstract	inválido	foul
abstract	determinación	determination
abstract	oponente	opponent
abstract	agresividad	aggression
abstract	ejercicio	exercise
abstract	gritos	yells
concrete	aficionado	fan
concrete	árbitro	referee
concrete	balón	ball
concrete	camiseta	jersey
concrete	red	net
concrete	entrenador	coach
concrete	estadio	stadium
concrete	campo	field
concrete	césped	grass
concrete	banquillo	bench
concrete	jugador	player
concrete	equipo	team
concrete	gradería	stands
concrete	marcador	scoreboard
concrete	espectador	spectator

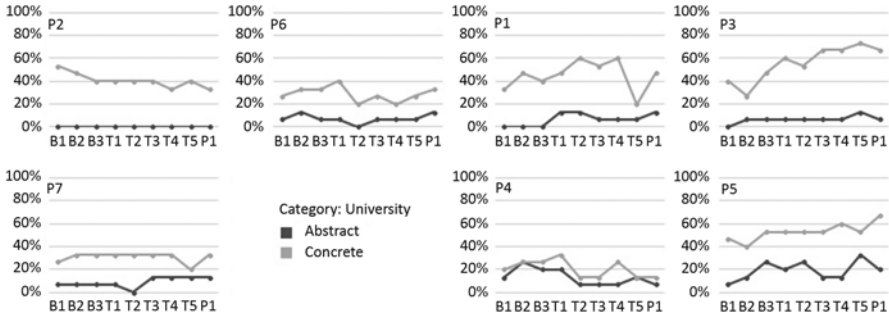

Figure A.1. Graphs of participant performance across phases for the unexposed control context-category *university* in Experiment 2. Graphs on the left are from the participants who were trained on concrete words; graphs on the right are from the participants who were trained on abstract words. Solid lines indicate the start and end of training. Performance on abstract words is shown in dark gray, and performance on concrete words is shown in light gray. B = baseline phase; T = training phase; P = posttesting phase.

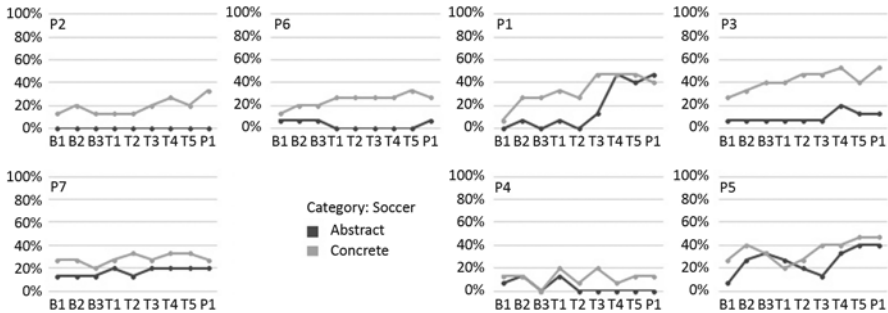

Figure A.2. Graphs of participant performance across phases for the exposed control context-category *soccer* in Experiment 2. Graphs on the left are from the participants who were trained on concrete words; graphs on the right are from the participants who were trained on abstract words. Solid lines indicate the start and end of training. Performance on abstract words is shown in dark gray, and performance on concrete words is shown in light gray. B = baseline phase; T = training phase; P = posttesting phase.

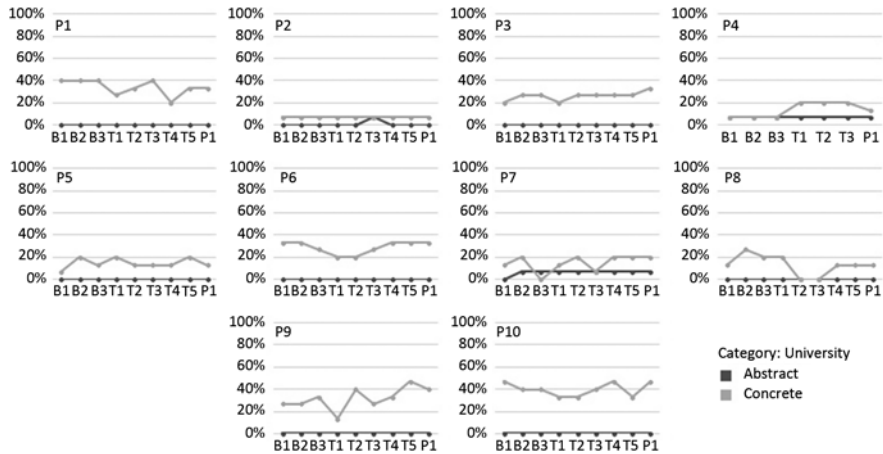

Figure A.3. Graphs of participant performance across phases for the unexposed control context-category *university* in Experiment 3. Performance on abstract words is shown in dark gray, and performance on concrete words is shown in light gray. B = baseline phase; T = training phase; P = posttesting phase.

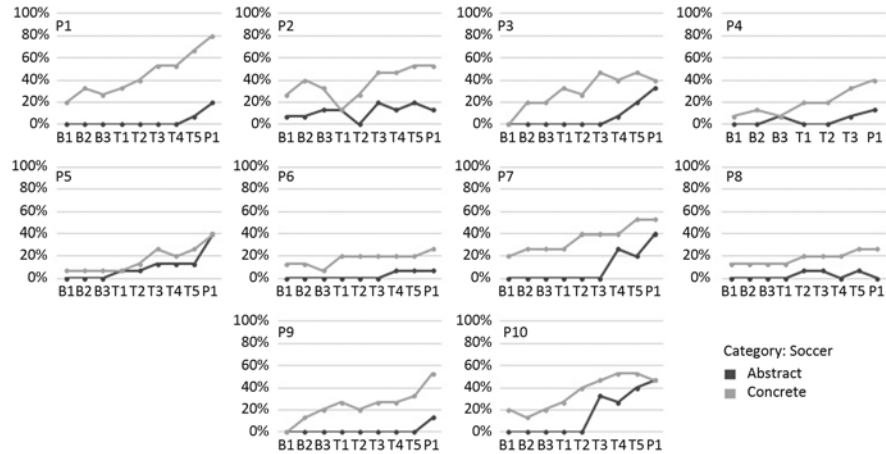

Figure A.4. Graphs of participant performance across phases for the exposed control context-category *soccer* in Experiment 3. Performance on abstract words is shown in dark gray, and performance on concrete words is shown in light gray. B = baseline phase; T = training phase; P = posttesting phase.

Cite this article: Sandberg C.W., Carpenter E., Kerschen K., Paolieri D., and Jackson C.N. The benefits of abstract word training on productive vocabulary knowledge among second language learners. *Applied Psycholinguistics*. https://doi.org/10.1017/S0142716419000262