Memory & Cognition (2023) 51:273–289
https://doi.org/10.3758/s13421-022-01348-5

Lexical inferencing as a generation effect for foreign language vocabulary learning

Steven Dessenberger[1] · **Kelly Wang**[1] · **Evan Jordan**[1] · **Mitchell Sommers**[2]

Accepted: 14 July 2022 / Published online: 27 July 2022

Abstract
Prior research suggests that second language (L2) vocabulary learning often occurs through *lexical inferencing* (translations based on context), but there has been less emphasis on how lexical inferencing compares with other methods of L2 word learning. The present study compared lexical inferencing to simply studying word lists for L2 learning. A secondary goal was to determine whether any effect of inferencing is mediated by the *generation effect of memory*, a phenomenon wherein generated information (inferencing) is remembered better than obtained information (reading). Across four experiments, participants read English sentences with embedded Swahili words and were asked either to infer the word meaning using context or were provided with translations before reading the sentence (reading condition). In contrast to our initial hypotheses, the inference condition resulted in lower rates of retention compared with the reading condition. In addition, the data suggest a number of differences between lexical inferencing and the generation effect, that argue against the proposal that lexical inferencing operates as a type of generation effect

Keywords Memory · Language acquisition · Associative learning

To comprehend a foreign language requires learning thousands of new word forms and their meanings (Nation, 2006). Even after years of study, language learners often encounter unfamiliar words when communicating with others. Without knowing the translation of unfamiliar words, the learner must rely on contextual cues to make accurate translations that can result in the acquisition of the novel vocabulary word and incorporation into the existing lexicon. Researchers refer to the process of learning vocabulary through context using various terminology, including meaning-inferencing (Mondria, 2003), contextual-word learning (Frishkoff et al., 2016), and what we will refer to in the current study as *lexical inferencing* (de la Garza & Harris, 2017; Geva et al., 2017; Shen, 2010).

Lexical inferencing has occasionally been considered an example of a phenomenon referred to as the *generation effect*, a phenomenon wherein memory for generated information is typically better than for information that is simply read (Bertsch et al., 2007; Slamecka & Graf, 1978). For example, if given the word *hot* and instructed to generate an antonym (*cold*), memory retention for the words *hot* and *cold* improves compared with simply reading the *hot–cold* word pair (Bertsch et al., 2007; Slamecka & Graf, 1978). Thus, both the generation effect and lexical inferencing require learners to generate information rather than be given information and this generation may promote learning. However, the benefits of lexical inferencing have not been compared with a read-only control condition, as is typical in studies examining the generation effect. The present study investigated the impact of lexical inferencing on L2 vocabulary learning compared with a read-only condition to determine whether lexical inferencing improves memory retention and if inferencing represents a special case of the generation effect.

Lexical inferencing

As noted, lexical inferencing occurs when individuals use semantic context to derive the meaning of new word forms. In one of the first demonstrations of lexical inferencing, Saragi

✉ Steven Dessenberger
sdessenberger@wustl.edu

[1] Psychological and Brain Sciences, Washington University in St Louis, St. Louis, MO, USA

[2] Washington University Psychology, St. Louis, MO, USA

et al. (1978) had English-speaking participants read the book *A Clockwork Orange* (Burgess, 1975) and tested their memory for the meaning of unfamiliar words. The book was written in English but contained a considerable number of Nadsat (Russian slang) words. For example, instead of describing pockets full of *money*, participants encountered sentences describing pockets full of *deng*, the Nadsat word for money. Participants were unfamiliar with Nadsat words, so they could rely only on context to derive their meaning. After participants finished reading the book, they were given a surprise multiple-choice test on the meaning of 90 Nadsat words. Participants on average correctly identified the meaning of 76% of the words. Based on these findings, Saragi et al. concluded that participants were able to learn the meaning of the words through inferencing.

Lexical inferencing has received considerable attention in second language (L2) vocabulary research, with the general finding that lexical inferencing can be an effective method for learning L2 vocabulary (Barcroft, 2002; Mondria, 2003; Rott, 1999; Webb & Chang, 2015). For example, Rott (1999) had participants read paragraphs that each contained an L2 word. Participants were exposed to 12 different words ranging from one to six times before taking a recall test. Participants never received the translations for the words and could only rely on the surrounding context to infer their meaning. On a final recall test, participants correctly translated approximately half of the target words. Other lexical inferencing studies have reported successful L2 vocabulary learning (Bordag et al., 2015; Frishkoff et al., 2016; Mondria, 2003; Vidal, 2011), but not all lexical inferencing studies demonstrate high rates of retention as Mondria (2003) found that while some vocabulary learning did occur after lexical inferencing, without additional reinforcement (providing the correct answer after inferencing), participants were able to retain fewer than 20% of the target words.

One potential method for improving the effects of lexical inferencing is providing the correct answer to language learners after an inference attempt (Mondria, 2003; Zou, 2016). Zou had participants infer the meaning of unknown words with half of the participants allowed to consult the dictionary after the inference attempt to confirm their answers as a form of corrective feedback (the other half did not). The participants that confirmed their answers remembered more of the target words on a subsequent posttest compared with the group that did not. This finding suggests that learning during lexical inferencing can be further improved through corrective feedback.

Lexical inferencing and the generation effect

Lexical inferencing shares several similarities with the generation effect. The generation effect, or a benefit to memory for generated information over obtained information, is commonly demonstrated by providing participants with a cue and a rule to guide generation such as antonym generation (Mulligan, 2004), sentence completion (Kane & Anderson, 1978), and word-stem completion tasks (McDaniel & Waddill, 1990). Memory after generation for the generated word is compared with a control condition, such as simply reading the information. Typically, generating information results in improved memory compared with the control and the difference between the conditions is referred to as the generation effect (Bertsch et al., 2007). The size of the generation effect can be moderated by making generation tasks more difficult (Tyler et al., 1979), providing the correct answer after generation attempts (Potts & Shanks, 2014), manipulating the final test format (i.e., multiple-choice test, cued recall, or free recall tests; see Gardiner, 1989), or by extending the retention interval (Bertsch et al., 2007) all of which can increase the size of the generation effect.

Lexical inferencing and the generation effect both require participants to generate target items rather than encoding presented material which suggests they may rely, at least in part, on overlapping cognitive processes. Specifically, the lexical inferencing task can be considered a generation task in which the rule is "provide a translation for the novel word based on the surrounding context." For example, in a lexical inferencing task, participants might see the sentence "The cowboy rode the *caballo*" and would then be asked to provide a translation (*horse*) for the word *caballo* based on the surrounding semantic context. According to past research on the generation effect (Bertsch et al., 2007; Mulligan, 2004), generation of a target word (*horse*) should improve memory compared with passive encoding (e.g., simply providing participants with the statement "*Caballo* means *horse*"). One goal of the present study is to provide a direct comparison between lexical inferencing and passive encoding as a method for learning L2 vocabulary.

One issue to consider in comparing the generation effect and lexical inferencing is that generation tasks typically emphasize memory for the generated target whereas vocabulary learning requires encoding both the L2 (new word form) and the L1 (semantic meaning), as the end goal is to bind them together. If lexical inferencing has similar outcomes to typical generation effect studies, memory for the cue may not always be promoted. This would be extremely problematic for L2 learning, as the goal is to acquire L2 word forms and map them onto existing semantic representations. However, there is some suggestion that the generation effect can potentially bolster memory for both cue words and generated targets (Greenwald & Johnson, 1989; McDaniel & Waddill, 1990). McDaniel and Waddill (1990) demonstrated this with a word-fragment completion task during which participants were given a cue word (e.g., strum) and were asked to generate a target item (guitar). At test, instead of providing the cue words (*strum)* and asking for the targets,

participants were given the targets (*guitar)* and asked to recall the corresponding cue word. Participants remembered the cue words better in the generation condition compared with the control (read only) condition which suggests that generating translations through context cues could potentially improve memory for both the generated L1 and the L2 cue words. However, unlike the *strum–guitar* example, the L2 cue word during lexical inferencing does not inform the generation but rather the surrounding context guides the learner to the target. In the example "The cowboy rode the *caballo*," the word *caballo* is the to-be-learned L2 word, but the generation is guided by the context of "the cowboy rode the ____." Therefore, it may be the case that the cue word *caballo* receives little to no attention during inferencing with learners focusing instead on the context. This may mean memory for the L2 could suffer as a result relative to control conditions.

A useful theoretical framework for comparing the generation effect and lexical inferencing is the Type of Processing Resource Allocation (TOPRA) model of L2 vocabulary learning (Barcroft, 2002; Barcroft & Sommers, 2005). According to the TOPRA model, two types of processing occur during L2 vocabulary learning, *form processing* of the novel word form and *semantic processing* of the meaning of the word. The model stipulates that the allocation of cognitive resources to these two types of processing during study impacts memory outcomes for form and meaning, respectively. Specifically, according to the model, there is a limit to the cognitive resources available to an individual (see Kahneman, 1973), and if one of these tasks (e.g., processing semantic meaning) necessitates additional cognitive resources, then resources will be directed toward that task and away from other tasks (e.g., learning the L2 word form). In the context of lexical inferencing, generating the meaning of a vocabulary word focuses almost entirely on processing semantic meaning, and therefore resources available for encoding the word form and mapping the new word form onto semantic representations may be reduced during lexical inferencing compared with simple reading. Effectively, the TOPRA model suggests lexical inferencing may result in an overall deficit to vocabulary learning due to disproportionate emphasis on processing the semantic meaning compared with the word form.

The present study

The present study sought to answer two questions: (1) Does lexical inferencing produce better L2 vocabulary learning than a read-only control, and (2) is the learning that occurs during lexical inferencing mediated at least in part by the same mechanisms as the generation effect?

To answer the first question, we conducted three experiments that compared memory retention for unfamiliar L2 (Swahili) words following either lexical inferencing or a read-only control. The inferencing task provided participants with native language (L1) sentences with one word replaced by its L2 equivalent and had participants infer its meaning based on context (e.g., "He fed the hay to the *farasi.*"). The read-only control provided the translation before revealing the sentence, eliminating the need for participants to infer the meaning. Two of the three experiments (Experiments 1A and 1B, Experiment 3) sought to determine the influence of lexical inferencing relative to a control condition on memory for the association between L2 word forms and their semantic meanings. Experiment 2 sought to examine lexical inferencing from the perspective of the TOPRA model and determine how different types of memory are influenced by lexical inferencing relative to a control condition. In Experiment 2, we will compare memory for the L1 and L2 items separately followed by a test of memory for their association or ability to translate from L1 to L2.

To answer the second question (is any observed benefit of lexical inferencing relative to a read only condition due to a generation effect?), the present study included several manipulations known to modulate the size of the generation effect, including manipulation of generation difficulty (Experiment 1-3), inclusion of corrective feedback (Experiment 1A-1B), and altering the delay before final test (Experiment 3; for a meta-analysis of these factors, see Bertsch et al., 2007). If lexical inferencing represents a type of generation effect, factors that are known to modulate the generation effect should exhibit the same pattern of effects on lexical inferencing.

The first moderator that we will examine is the influence of task difficulty. Increasing generation-task difficulty is thought to increase the cognitive effort required during initial encoding (i.e., a type of levels of processing effect; Craik & Lockhart, 1972), which predicts that deeper processing of information results in improved memory retention Tyler et al. (1979) compared memory retention for low-difficulty and high-difficulty word stem completion tasks. Word stems solved in the high-difficulty condition, which required more cognitive effort, were remembered better at final test compared with words generated in the low-difficulty condition. In the present experiment, we applied a difficulty manipulation by including sentences in all experiments that were high in predictability (easier generation) and sentences low in predictability (harder generation) for the target word to determine whether task difficulty played a role in memory retention after lexical inferencing similar to what has been observed in the generation task.

The second moderator of the generation effect that we examined in the current study was corrective feedback. One issue with using generation-based learning is that outcomes

are reliant on generation accuracy. An incorrect generation typically improves memory retention for incorrect information. Corrective feedback or giving the correct answer after a generation can improve retention for the correct answer regardless of generation accuracy (Bertsch et al., 2007; Kane & Anderson, 1978; Potts & Shanks, 2014). In Experiments 1A and 1B, we manipulated the presence of feedback for half of the lexical inferencing trials to determine whether inferencing would show a similar benefit from corrective feedback.

The third moderator of the magnitude of the generation effect that we examined is the manipulation of retention interval or the time between the generation attempt and the final test. When the final test is given within 24 hours of the generation attempt, the generation effect is smaller ($d < .42$) while retention intervals of 24 hours or longer have generally been found to produce larger generation effect ($d = .64$; Bertsch et al., 2007). In Experiment 3, we manipulated the retention interval by giving the final test after a delay of either 5 minutes,12 hours, or 24 hours.

Experiment 1A

Method

Participants

For Experiment 1, An a priori power analysis was conducted with the intent to focus resources to identify the presence of a potential generation effect using the effect size reported by Bertsch et al. (2007) for within-subject generation effect research designs, or $d = .50$. That being the case, the analysis determined that 35 participants would provide .8 power to detect a difference a main effect of lexical inferencing over the read-only control if a generation effect was present at the alpha = .05 level. Fifty-one participants were recruited from a private U.S. research university (33 females, $M_{age} = 19.1$ years, $SD_{age} = 1.4$ years), and of those 51, 12 participants did not follow instructions during training (see Experiment 1A Results) and were therefore excluded from the analyses, resulting in a total of 39 participants. All participants were fluent speakers of English and reported no prior knowledge of Swahili.

Materials

Sixty English (L1) sentences were used during the experiment (see Appendix). Within each sentence, a target word was replaced with the Swahili translation surrounded by asterisks (e.g., He fed the hay to the *farasi*). Sentences were categorized by predictability using cloze values obtained from a sample of 33 English-speaking participants recruited from Amazon's MTurk web platform and who did not participate in the inferencing experiment. Participants were shown the sentences with the target word omitted entirely and were asked to guess the meaning of the missing word. A cloze value of 1 means every participant correctly guessed the meaning of the missing word while a value of 0 would mean that no participant was able to guess the missing word. Half of the sentences (high context) had relatively high cloze values ($M_{cloze} = .92$, $SD_{cloze} = .03$), meaning that the target word was guessed frequently based on context. The remaining sentences (low context) had relatively low cloze values ($M_{cloze} = .38$, $SD_{cloze} = .09$).

All procedures were conducted in the laboratory using desktop computers with PsychoPy software (Peirce et al., 2019) installed to run the experiment

Design

A 3 × 2 within-subjects design investigated the effects of training (read vs. lexical inference with feedback vs. lexical inference without feedback) and sentence context (high vs. low) on memory retention for Swahili–English word pairs. In the lexical inferencing conditions, participants were given an English sentence with a single embedded Swahili word and were instructed to infer its meaning based on the sentence context (example: "Please translate the word in asterisks: cowboys often ride *farasi*"). In the inference with feedback condition, participants were given the correct answer (*farasi–horses*) at the end of the trial while the correct answer was not provided for the no-feedback condition. In the read-only condition, participants were given the word pair prior to viewing the sentence to prevent an inference attempt. To manipulate the difficulty of inferencing, in each training condition, half of the sentences were low-context sentences and half were high-context sentences (context conditions were evenly distributed across training conditions). Memory retention was measured after a 2-minute delay using a cued recall test (Swahili cues with English targets).

Procedure

Participants completed three experimental blocks in a randomized order, one block for each training condition. Each block contained 20 unique sentences, such that there were 10 high-context sentences and 10 low-context sentences in each block and sentence assignments were counterbalanced between blocks. Each block consisted of four distinct phases: (1) an instruction phase, (2) a study phase, (3) a delay period, and (4) a test phase.

At the start of a block, participants were given a set of instructions. For the read condition, the instructions indicated that participants would be given the Swahili–English word pair to read on its own followed by a sample English

sentence that replaced the target English word with the Swahili equivalent surrounded by asterisks. For both of the lexical inference conditions, instructions indicated that participants would first be shown the sentence, and their task was to guess the meaning of the Swahili word based on the sentence context. The lexical inference condition that included feedback also notified participants that they would be shown the correct answer at the end of the trial. The no-feedback condition notified participants that they would not be shown the correct answer. All participants were made aware that they would be tested on the words at the end of the block.

After receiving instructions, participants completed 20 trials in a randomized order. For a schematic of trial timing, see Fig. 1. Trials in all conditions lasted a total of 10s to control for exposure time. For the read-only condition, the correct Swahili–English translation appeared at the top of the screen at the trial start and the English sentence was presented after a 2-s delay. Both the word pair and sentence remained on-screen until the end of the trial. For the feedback condition, sentences appeared immediately at trial start and the correct Swahili–English word pair appeared at the top of the screen after an 8-s delay. The no-feedback condition also started with the sentence appearing on-screen, but the correct answer was never shown.

After the training, participants played Tetris for 2 minutes followed by a final test. The final test consisted of a cued recall test of the words seen during the most recent training block, one at a time. The order of the words was the same as they appeared in the block to ensure the total time between study and test was approximately equal for all words. Participants were asked to translate Swahili cue words to English via typing and had up to 10 s to submit an answer. No feedback was provided. Once participants had completed all 20 final test trials for a block, they proceeded onto the next training condition or if they had completed all other blocks, the experiment ended. Total duration of the experiment was approximately 30 minutes.

Results

Despite the instructions, 12 participants did not type out their generated translation in the lexical inference conditions and we could not verify their inference accuracy. Instead of excluding these participants, we conducted two separate analyses, one including the 39 participants that typed out their guess and the other that included all 51 participants. We found the same pattern in both analyses and thus the following analysis is based on the 39 participants who typed their lexical inference responses.

Training

As a check of the inference difficulty manipulation, we conducted a paired-sample *t* test comparing inference accuracy in the high- versus low-predictability contexts during the initial training period. As can be seen in Fig. 2, high-context sentences resulted in higher accuracy rates compared with low-context sentences, $t(38) = 13.05$, $p < .001$, $d = 4.18$.

Final test

A multilevel logistic regression model analyzed the influence of training (read vs. lexical inference with feedback vs. lexical inference without feedback) and sentence context (high context vs. low context) on memory retention for Swahili–English word pairs as measured by a cued recall

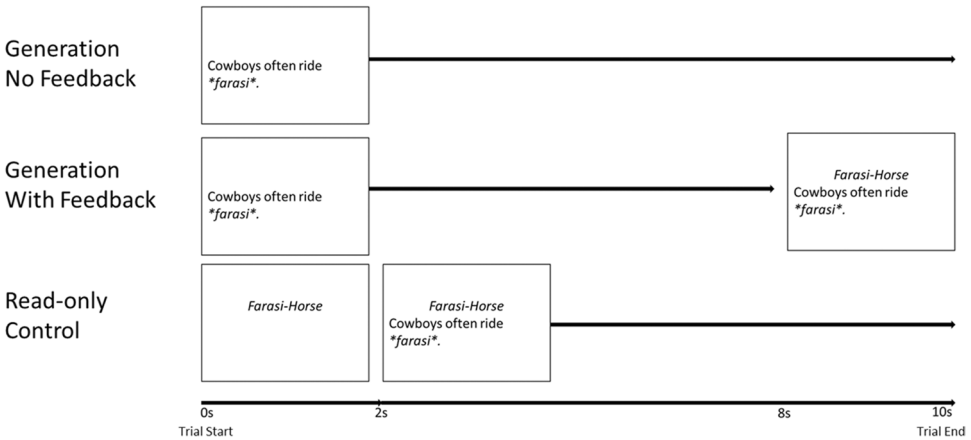

Fig. 1 Schematic of the trial structure for all three conditions in Experiments 1A. Trials lasted 10 seconds in all conditions. For generation trials, the cue sentence would appear on-screen, and participants were instructed to guess the meaning of the word surrounded by asterisks. If the trial was assigned to the feedback condition, during the last 2 s of the trial, the correct answer would appear on-screen. For the read-only control, the correct answer appeared on-screen at the start of the trial, and the sentence did not appear until 2 s after the trial started. Both items would remain on-screen until the trial ended

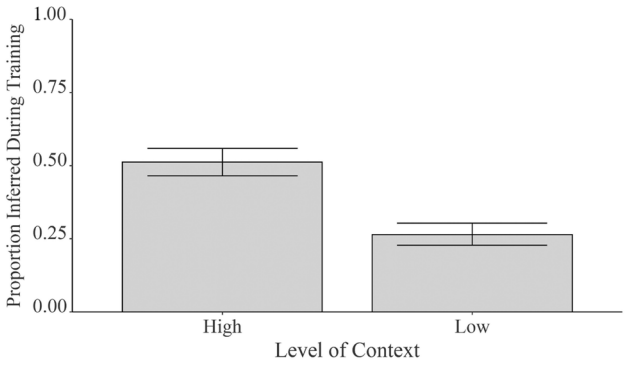

Fig. 2 In Experiments 1A, high-context sentences lead to the most accurate generations compared with the low-context sentences. Error bars are 95% confidence intervals

test. Final test scores can be seen in Fig. 3 (top) and were calculated as mean participant performance for all words within a training block, separated by level of context. The individual participant was set as the Level 2 error term to account for differences across participants. The full model was determined a priori and included both training-type and context-level as predictors as well as their interaction. We applied a step-wise procedure to determine model fit, starting with a base model that contained no predictors, adding a single predictor or interaction term at each step. Predictors were dummy-coded such that the read-only control was the reference group for the training variable and the low context sentences was the reference group for the context variable. Each model was compared with the previous model using chi-squared goodness of fit tests. Models were created using the R statistical analysis software with the *lme4* package (Bates et al., 2015; R Core Team, 2018), and post hoc analyses were conducted using the *multcomp* package (Horthorn et al., 2008). Model formula, description, and code are all available in the supplemental materials.

The coefficients for the full model can be found in Table 1. Adding the training type predictor to the model significantly improved model fit compared with the base model, $\chi(2) = 61.21$, $p < .001$. Adding the context predictor also significantly improved model fit, $\chi(1) = 19.86$, $p < .001$. Adding the interaction term did not significantly improve fit, $\chi(2)$ 4.75, $p = .09$.

As show by Fig. 3 (top), post hoc linear comparison of the full model indicated there was a significant main effect of training such that the read condition produced greater memory retention at final test compared with the generation with feedback condition ($z = 5.44$, $p < .001$) and the generation without feedback condition ($z = 7.49$, $p < .001$). There was no significant difference between the two generation conditions ($z = 2.32$, $p = .08$). There was also a main effect of context such that high-context sentences resulted in greater accuracy compared with low-context sentences ($z = 4.6$, $p < .001$). All post hoc p values were corrected using the Bonferroni method.

Conditional analysis

To address concerns as to whether the negative effects of lexical inferencing relative to the read-only control were attributable to inference accuracy, exploratory analysis investigated the influence of inference accuracy on final

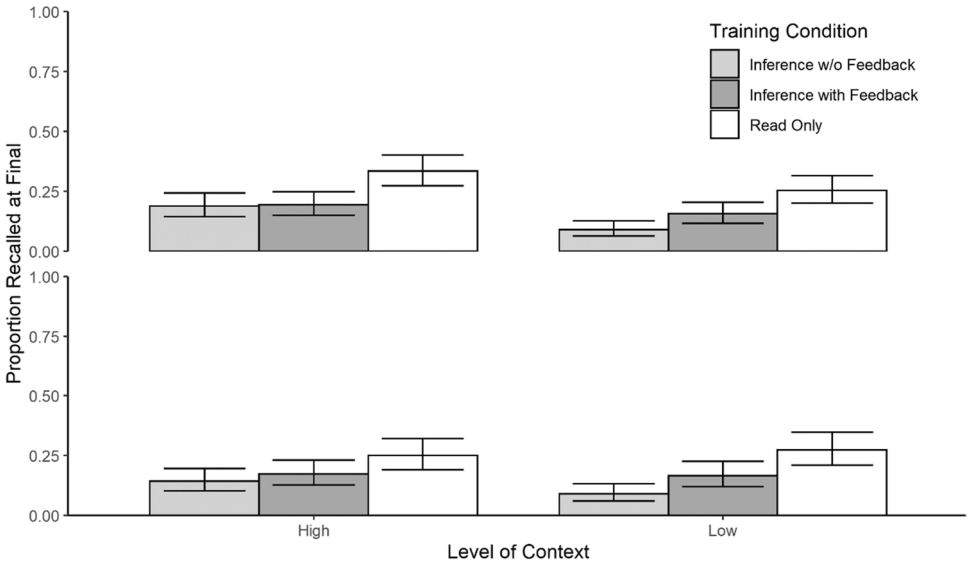

Fig. 3 In Experiments 1A (top), a negative generation effect was found with both inference with feedback and inference without feedback conditions leading to lower performance compared with the read-only control condition, regardless of the level of context. The finding was replicated in Experiments 1B (bottom). Error bars are 95% confidence intervals

Table 1 Experiment 1A and 1B multilevel logistic regression model fixed effects output for the final cued recall test scores (Intercept is the read-only control with high context)

Fixed Effects	β	*SE*	*z*	*p*
Experiment 1A				
Intercept	−0.69	0.15	4.66	<.001
Inference No Feedback	−0.77	0.17	−4.54	<.001
Inference with Feedback	−0.73	0.17	−4.36	<.001
Low Context	−0.39	0.16	−2.42	0.02
Inference No Feedback × Low Context	−0.46	0.27	−1.74	0.08
Inference with Feedback × Low Context	0.13	0.25	0.51	0.60
Experiment 1B				
Intercept	−1.10	0.18	6.23	<.001
Inference No Feedback	−0.70	0.18	3.81	<.001
Inference with Feedback	−0.47	0.18	2.63	0.01
Low Context	0.11	0.17	0.69	0.49
Inference No Feedback × Low Context	−0.64	0.28	2.25	0.02
Inference with Feedback × Low Context	−0.17	0.26	0.64	0.52

β and standard error are presented in logit units

test performance relative to the control condition. We categorized the two inference variables based on inference accuracy creating four new conditional variables (accurate inference with and without feedback as well as inaccurate inference with and without feedback). Variables were inserted into the model and compared with the read-only control condition, effectively changing the original 3 × 2 model to a 5 × 2 within-subjects model. As the primary motivation for this analysis was the unforeseen main effect of read-only control over inferencing, linear comparisons of the full 5 × 2 model only compared the conditional variables with the control condition to preserve statistical power, and all reported *p* values were corrected using the Bonferroni method. The comparisons indicated that each of the conditional inference variables resulted in a decrease in final test performance relative to the read-only control: accurate inferencing with feedback ($z = 3.34$, $p < .01$) and inaccurate inferencing with feedback ($z = 4.73$, $p < .001$), accurate inferencing without feedback ($z = 3.18$, $p = .01$), and inaccurate inferencing without feedback ($z = 7.48$, $p < .001$). This exploratory analysis suggests that the negative effect of lexical inferencing on memory for the word pairs was not attributable to poor performance during inferencing alone.

Discussion

Counter to our hypothesis, Experiment 1A found that lexical inferencing resulted in a decreased rate of memory retention compared with the control condition, regardless of task difficulty and the presence of feedback. Additionally, we had predicted based on prior research (see Bertsch et al. 2007), that low-context sentences would produce better memory retention compared with high-context sentences in the inference tasks due to the increased cognitive effort required by the task, but our prediction was not supported, as high-context sentences resulted in better memory retention for the Swahili–English word pairs compared with the low-context sentences.

Neither feedback nor difficulty during inferencing moderated the rate of memory retention in a fashion consistent with past research on the generation effect (see Bertsch et al., 2007). Given that these findings were unexpected and due to concerns of power to detect the novel effect found in Experiment 1A, we sought to replicate them in Experiment 1B. We also made two changes to the procedure to address some potential concerns: First, written directions better highlighted that participants should type their inferences during the training phase, and participants were verbally instructed by the researcher to type their inferences during training. Second, we wanted to determine whether the results of Experiment 1A replicated after randomizing the order of the questions in the final test.

Experiment 1B

Methods

Participants

Forty participants were recruited from a private U.S. research university (30 females, $M_{age} = 18.8$ years, $SD_{age} = .98$ years). Sample size was determined a priori in effort to maintain consistency with Experiment 1A for the replication process. All participants were fluent English speakers, reported no prior knowledge of the target language (Swahili), and none had participated in Experiment 1A.

Design, materials, procedure

The design, materials, and procedure were identical to those of Experiment 1A, with two exceptions: (1) in addition to the on-screen instructions, the participants were explicitly told by the researcher that they should type their guesses during the lexical inference phases, and (2) the order of the final test trials was randomized within each block.

Results

Final test

A multilevel logistic regression model was used to establish the influence of training (read vs. lexical inferencing with feedback vs. lexical inferencing without feedback) and

sentence context (high context vs. low context) on memory retention for Swahili–English word pairs. The coefficients for the full model can be found in Table 1. Model fit was determined using the same step-wise procedure described in Experiment 1A. Adding the training type predictor to the model significantly improved model fit compared with the base model, $\chi(2) = 51.85$, $p < .001$. Adding the context predictor did not significantly improve model fit, $\chi(1) = .72$, $p = .39$, and adding the interaction term did not significantly improve fit, $\chi(2) = 5.1$, $p = .08$.

Similar to Experiment 1A, post hoc linear comparisons of the full model indicated there was a significant main effect of training type such that the read condition produced greater memory retention at final test compared with the generation with feedback condition ($z = 4.25$, $p < .001$) and the generation without feedback condition ($z = 7.15$, $p < .001$). There was also a significant difference between the two generation conditions ($z = 3.14$, $p = .005$) such that generation with feedback resulted in greater memory retention compared with generation without feedback.

Conditional analysis

Exploratory analysis investigated the interaction of initial inference accuracy on final test performance relative to the control condition. We used the same model described in Experiment 1A and all reported p-values were corrected using the Bonferroni method. Linear comparisons of the full model indicated that each of the conditional inference variables resulted in a decrease in final test performance relative to the read-only control: accurate inferences with feedback ($z = 2.61$, $p = .04$), inaccurate inferences with feedback ($z = 4.36$, $p < .001$), accurate inferences without feedback ($z = 4.43$, $p < .001$), and inaccurate inferences without feedback ($z = 6.70$, $p < .001$). This exploratory analysis suggests that the negative effect of lexical inferencing on memory for the word pairs was not attributable to poor performance during inferencing.

Power analyses of Experiments 1A and 1B

An a priori power analysis assuming a generation effect based on previous findings by Bertsch et al. (2007) indicated that 35 participants would be necessary to achieve a power of .80. Given the unexpected results of Experiment 1A, we replicated the Experiment with the same sample size goal in Experiment 1B and found a similar effect. To determine the achieved level of power for detecting this novel effect during Experiments 1A and 1B, we conducted three sets of power analyses using Monte Carlo simulations. We first sampled with replacement 39 participants from Experiment 1A 1,000 times and determined the proportion of significant findings ($p < .05$) in the aforementioned step-wise comparisons and the calculated *t*-distribution outputs shown in Table 1 to determine power achieved ($1 - \beta$). For the second set of power analyses, we did the same thing with Experiment 1B (instead sampling 40 times). The achieved power in the step-wise analysis was as follows (as a reminder, Step 1 compared the training variable with the base model, Step 2 introduced the context variable, and Step 3 introduced the interaction). In Experiment 1A: Step 1 ($1 - \beta = 1.00$), Step 2: ($1 - \beta = .76$), Step 3: ($1 - \beta = .19$), in Experiment 1B: Step1 ($1 - \beta = 1.00$), Step 2: ($1 - \beta = .79$), Step 3: ($1 - \beta = .18$). For the *t* distributions extracted from the model, the power achieved for detecting differences between lexical inferencing with feedback and the control in Experiment 1A was $1 - \beta = .87$ and Experiment 1B was $1 - \beta = .89$ and the difference between the control and the no-feedback condition in Experiment 1A was $1 - \beta = .94$ and Experiment 1B = .94. For the main effects of context and the interactions, achieved power was below the .8 threshold in both experiments.

For the third power analysis, we investigated the sample size required to reliably detect a negative effect of lexical inferencing relative to the control condition by sampling with replacement participants from both Experiment 1A and Experiment 1B 1,000 times, starting with 10 participants and increasing in increments of five until we reached 80 participants. According to this power analysis, to reliably detect the effect of Step 1 necessitates 10 participants, Step 2 would require 25 participants, and Step 3 could not be reliably detected with fewer than 80 participants. In the extracted *t* distributions of the multilevel models, analysis indicated approximately 40 participants in the lexical inferencing with feedback condition and 25 participants in the lexical inferencing without feedback condition were required to achieve a power of .8. For the main effect of context and interactions, power did not exceed .8 during the analysis. This suggests that the studies were adequately powered to detect the detriment of lexical inferencing with or without feedback relative to the control condition.

Discussion

Experiment 1B replicated the key findings of Experiment 1A: Memory retention after inferencing decreased in comparison to the read-only control condition regardless of changes in task difficulty and the presence of feedback (Fig. 3). We initially hypothesized that lexical inferencing would improve learning relative to simply reading word pairs with sample sentences, given the similarities between lexical inferencing and the generation effect, a robust method for improving memory retention. Our hypothesis was not supported by Experiments 1A and 1B.

One possible explanation for these unexpected findings is that a generation effect might be occurring during lexical inferencing, but the cued recall format used in Experiments

1A and 1B was not sensitive to these changes. According to the TOPRA model (Barcroft, 2002), the type of processing that occurs during study, whether it is form processing or semantic processing, impacts what is later recalled. Lexical inferencing emphasizes processing of the meaning but provides little emphasis on form. This lack of form processing could result in the word form being forgotten at a higher rate, which in turn would explain why Experiment 1A and Experiment 1B showed decreased final test performance on a cued recall task that required knowledge of the word form and its semantic meaning. To address this issue, Experiment 2 included a variety of test formats that would be sensitive to changes in memory for the English words and Swahili word forms separately, as well as sensitive to changes in memory for their association.

Experiment 2

Experiment 2 sought to investigate why lexical inferencing resulted in poorer memory outcomes in comparison to the control condition despite its similarity to typical generation effect paradigms in Experiments 1A and 1B. According to the TOPRA model, it is possible that lexical inferencing is a process that improves memory for the semantic meaning of a vocabulary word but not the L2 word form itself. The cued recall test used in Experiments 1A and 1B indicated that the lexical inferencing condition resulted in poorer memory for the L2 words. However, the cued recall test format requires memory for both word form and semantic meaning as participants are required to not only recall the word (*farasi)* but also have to be able to trace it back to its meaning (*farasi* means *horse*). Therefore, any declines in cued recall performance could be attributed to either poor memory for the L2 word form after inferencing or its associated meaning or both. To determine whether the poorer performance for inferencing relative to reading found in Experiments 1A and 1B could be attributed to poor memory for the Swahili word or for its association to the English meaning (or both), Experiment 2 included three different memory measures that would allow us to asses changes in memory for novel word forms and semantic meanings separately: Specifically, in addition to the final recognition test of the association between L2 and L1, we also included a free recall test of the L1 meanings, a free recall test of the L2 word forms, and a multiple-choice test that measured memory for the association between the L1 meaning and L2 word form. Using these three tests will help identify which types of memory were improved by lexical inferencing and which resulted in a memory deficit relative to control conditions.

We predicted a positive benefit of inferencing for the English (L1) words compared with a read-only condition, similar to past generation effect research (Bertsch et al., 2007; Slamecka & Graf, 1978). As for the L2 word forms, given the negative effect on memory for L2 vocabulary seen in Experiments 1A and 1B, we predicted poorer performance for the Swahili words after inferencing compared with the control. For the multiple-choice test, we predicted that the pattern seen in Experiment 1A and Experiment 1B would be replicated, such that the read-only condition would lead to greater test performance compared with the inference condition. If confirmed, these hypotheses would suggest that poor L2 word form memory after inferencing is responsible for the negative generation effect seen in Experiments 1A and 1B. To increase statistical power, we eliminated the no-feedback condition used in Experiments 1A and 1B, as results indicated no differences between the feedback and no-feedback conditions.

Methods

Participants

Forty participants were recruited from a private U.S. research university (29 females, M_{age} = 19.4 years, SD_{age} = 1.1 years). Sample size was again determined a priori based on estimated effect size taken from Bertsch et al. (2007) and is supported by the post hoc power analysis of Experiments 1A and 1B. All participants reported no prior training/knowledge of the target language (Swahili) and none had participated in Experiments 1A or 1B.

Materials

We used the same materials and lab equipment described in Experiments 1A and 1B, and the data were also collected in the laboratory using PsychoPy Software

Design

A 2 × 2 within-subjects design was used to investigate the effects of training (read vs. lexical inference with feedback) and sentence context (high context vs. low context) on memory retention for Swahili–English word pairs. Memory retention was measured using two free recall tests, one requesting participants recall all Swahili words learned and the other all English translations learned during the most recent block of training and the order of the tests was counterbalanced. Participants then completed a multiple-choice test, which unlike the previous cued recall tests, provided both L1 meaning and L2 word form to the participants and necessitates only that participants recall the correct association between the cue word and the presented lures. The multiple-choice test was comprised of English cue words and the Swahili targets accompanied by three Swahili lures (an English–Swahili multiple-choice test was deemed to be more challenging

and less likely to produce ceiling effects compared with a Swahili–English multiple-choice test). All lures were taken from the same block of training as the target word. The multiple-choice test was always given after the free recall tests to avoid additional exposures to the Swahili words.

As in Experiments 1A and 1B, the training condition was manipulated by adjusting when participants saw the correct Swahili–English word pair in relation to the sample sentence during a training trial. For the lexical inference condition, the correct pairing was shown 8 s after the sentence appeared to allow participants to make inferences about the meaning of the novel L2 word. For the read condition, the word pair appeared 2 s before the sentence and stayed on-screen to prevent any need for inference. Half of the items in each block were high-context sentences and half were low-context sentences.

Procedure

Experiment 2 repeated the procedure described in Experiment 1B but had 30 trials per block instead of 20 (15 high-context and 15 low-context sentences per block). Additionally, there was no cued recall test. Instead, there were two untimed free recall tests, one requesting all Swahili words learned during the previous block and the other requesting all English translations learned during the previous block. The order of these free recall tests was randomized for each participant. Once both free recall tests were completed, participants completed a multiple-choice test that consisted of an English cue word and asked participants to select the correct translation from four possible Swahili options.

Results

All full models for the three final tests (free recall English, free recall Swahili, and multiple-choice English–Swahili test) were determined a priori and the same step-wise procedure used in Experiments 1A and 1B was applied to determine model fit. Coefficients for all three full models can be found in Table 2. All post hoc p values were corrected using the Bonferroni method.

Free recall—English

A multilevel logistic regression model analyzed the influence of training and sentence context on memory retention for the English translations, measured by a free recall test. Adding the training type predictor significantly improved model fit compared with the base model, $\chi(1) = 31.98$, $p < .001$. Adding the context predictor significantly improved model fit, $\chi(1) = 6.31$, $p = .01$, and adding the interaction term also significantly improve fit, $\chi(1) = 4.87$, $p = .03$. As show by Fig. 4 (top), post hoc linear comparisons indicated that there was a significant main effect of training type such that the lexical inferencing condition produced greater memory retention at final test compared with the read condition ($z = 5.46$, $p < .001$). While the effect of context was significant in the step-wise comparisons, in the post hoc analysis of the full model, there was no main effect of context after correcting for family-wise error ($z = 2.09$, $p = .15$). As for the interaction, the effect of context was significant for the lexical inference condition ($z = 3.34$, $p = .003$) such that the low-context target words were remembered at a higher rate than the high-context words, but context was not significant in the read condition ($z = 0.09$, $p = 1.00$).

Free recall—Swahili

As shown by Fig. 4 (bottom) the free recall scores for the L2 word form in all conditions were near floor and the following analysis should be interpreted with caution. A multilevel logistic regression model analyzed the influence of training and sentence context on memory retention for the Swahili translations, measured by a free recall test. Adding the training type predictor to the model significantly improved model fit compared with the base model, $\chi(1) = 21.82$, $p < .001$. Adding the context predictor did not significantly improve model fit, $\chi(1) = 1.39$, $p = .24$. Adding the interaction term did not significantly improve fit, $\chi(1) = .91$, $p = .34$. As show by Fig. 4 (bottom), linear comparisons of the full model indicated there was a significant main effect of training type such that the read condition produced greater memory retention at final test compared with the generation

Table 2 Experiment 2 multilevel logistic regression models fixed effects output (Intercept is the read-only control with high context)

Fixed Effects	β	*SE*	*z*	*p*
English Free Recall				
Intercept	1.89	0.15	12.92	<.001
Inference	0.36	0.16	2.24	0.02
Low Context	−0.01	0.17	0.08	0.93
Inference × Low Context	0.49	0.22	2.22	0.03
Swahili Free Recall				
Intercept	−2.88	0.20	13.98	<.001
Inference	−0.86	0.30	2.84	.004
Low Context	−0.13	0.25	0.51	0.61
Inference × Low Context	−0.47	0.48	0.96	0.38
Multiple-Choice Final				
Intercept	0.72	0.18	3.98	<.001
Inference	0.39	0.13	2.99	0.003
Low Context	0.17	0.13	1.32	0.19
Inference × Low Context	0.10	0.18	0.55	0.58

β and standard error are presented in logit units

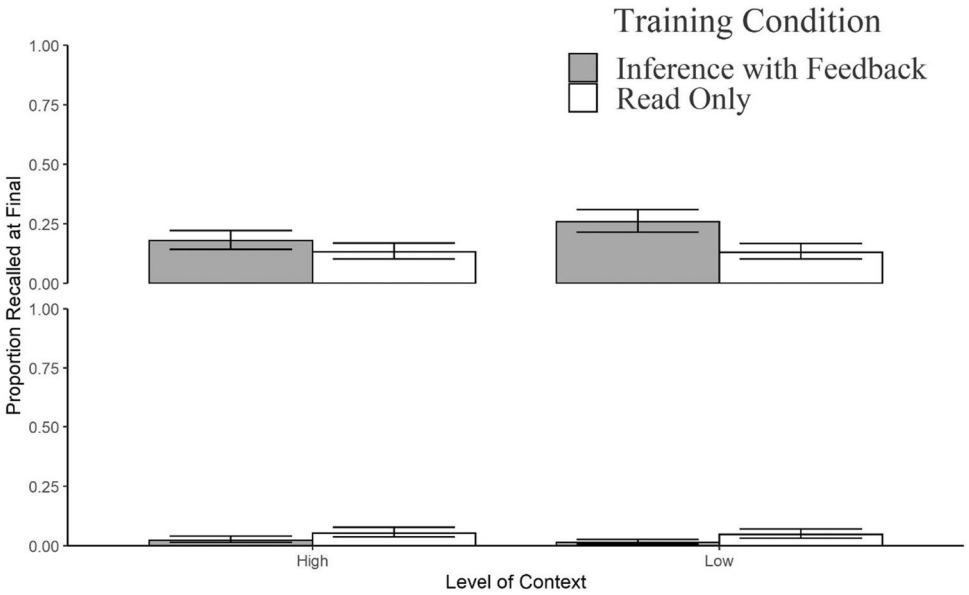

Fig. 4 In Experiment 2, we found a positive generation effect during the English free recall task (Top). When participants were asked to recall all the English words learned during the previous training block, they remembered more in the generation condition compared with the generation condition. (Bottom) In the Swahili free recall task, scores are near floor, but the negative generation effect found during Experiment 1A and 1B was replicated. Error bars are 95% confidence intervals

with feedback condition ($z = 4.38$, $p < .001$). There were no other significant differences found.

Multiple choice

A multilevel logistic regression model analyzed the influence of training and sentence context on memory retention for the Swahili–English word pairs, measured by a multiple-choice test. Adding the training type predictor to the model significantly improved model fit compared with the base model, $\chi(1) = 13.72$, $p < .001$. Adding the context predictor did not improve model fit, $\chi(1) = 1.78$, $p = .18$, nor did adding the interaction term improve fit, $\chi(1) = .29$, $p = .59$. As show by Fig. 5, post hoc linear comparisons of the full model indicated there was a significant main effect of training type such that the read condition produced greater memory retention at final test compared with the generation condition ($z = 3.71$, $p < .001$). There were no other significant differences found.

Conditional analyses

Exploratory analysis investigated the interaction of initial inference accuracy on final test performance relative to the control condition. We used the same procedure described in Experiment 1A to divide the lexical inferencing condition into accurate and inaccurate trials and compared each to the read-only control in a 3×2 multilevel model. We again focused analysis on the main effect of interest to preserve statistical power and reported p values were corrected using the Bonferroni method (corrections were done separately for each of the three tests). Linear comparisons of the full model in the English free recall test indicated that the inference condition resulted in an increase in final test performance relative to the read-only control for both accurate ($z = 4.43$, $p < .001$) and inaccurate ($z = 4.71$, $p < .001$) trials. For the Swahili free recall, the linear comparisons identified decreases in test perform for the lexical inferencing condition relative to the control for both accurate ($z = 3.16$, $p < .01$) and inaccurate ($z = 3.72$, $p < .001$) trials, with a similar finding for the multiple-choice test as both accurate ($z = 2.34$, $p = .04$) and inaccurate ($z = 3.74$, $p < .001$) trials resulted in decreases in memory after inferencing relative to the control. These findings suggest that regardless of accuracy, inferencing provides a benefit for memory of the L1 targets compared with the read-only control, but memory for the L2 word forms and the L1–L2 association are negatively impacted compared with the control condition.

Discussion

As predicted, in the free recall tests we observed a positive generation effect for English targets and we observed a negative generation effect for the Swahili words. This suggests that a generation effect is produced during lexical inferencing only for the English words. According to the TOPRA model, the Swahili words are likely being forgotten after inferencing because the available cognitive resources are

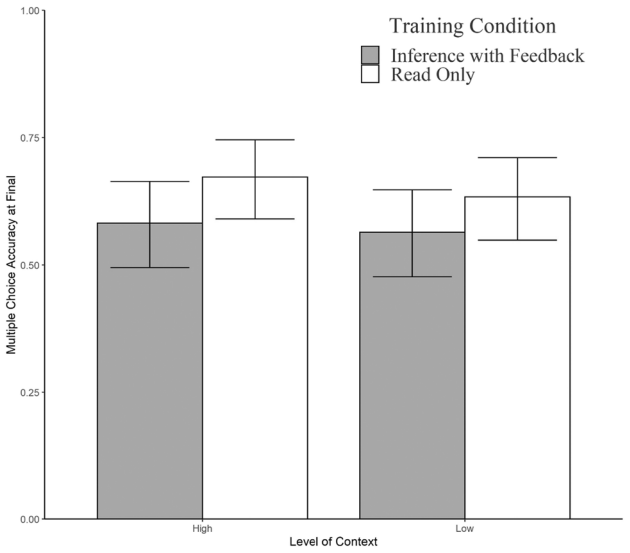

Fig. 5 In Experiment 2, similar to the negative generation effect found in Experiment 1A and Experiment 1B, a negative generation effect was also found during a multiple-choice test. Error bars are 95% confidence intervals

being devoted to processing the semantic meaning and not to encoding the word form. As for the multiple-choice test which measured the association between the Swahili and English words, the negative effect of inferencing relative to the control found in Experiments 1A and 1B persisted which suggests there is no benefit of generation towards translation ability.

Experiment 3

One limitation of the Experiment 2 is the potential testing effect introduced by the free recall tests that always preceded the multiple-choice test. The testing effect, or benefit to memory retention derived from memory retrieval, is a robust memory phenomenon that could have affected the outcome of the multiple-choice test (Roediger & Karpicke, 2006; Rowland, 2014). To address this limitation, we included only the multiple-choice test in Experiment 3, as it was determined that, of the three tests in Experiment 2, the multiple-choice was best representative of translation ability (the skill of interest in language learning).

In addition to controlling for potential testing effects, Experiment 3 sought to determine whether the influence of retention interval after lexical inferencing mimicked its effects on generation-based learning (Bertsch et al., 2007). Prior research suggests the generation effect increases in magnitude as the retention-interval extends beyond 24 hours (Bertsch et al., 2007). It is possible that generation during inferencing could produce a positive benefit to memory retention compared with the control given a lengthy retention interval. In Experiment 3, we compared memory retention on a delayed test relative to an immediate test to determine if memories after inferencing were retained longer compared with the read-only control. We chose three different delays (5 min, 12 h and 24 h) to test for linear trends. These points were chosen to avoid floor and ceiling effects based on the findings of Experiments 1A, 1B, and 2. We predicted that memory retention after lexical inferencing would decline at a slower rate over time compared with the control condition.

Method

Participants

Seventy-seven participants were recruited from a private U.S. research university (58 females, M_{age} = 19.6 years, SD_{age}= 1.33 years). Target sample size was doubled compared with previous experiments due to the inclusion of a between-subjects manipulation allowing for two groups, one of 38 and one of 39. All participants reported no prior training/knowledge of Swahili, and none had participated in any of the previous experiments.

Materials

We used the same materials described in Experiment 2, with the exception that Experiment 3 was conducted remotely using the Qualtrics web platform to facilitate the use of delay periods.

Design

A 3 × 2 × 2 mixed design was used to investigate the effects of delay (5 min vs. 12 h vs. 24 h), training (read vs. lexical inference with feedback) and sentence context (high context vs. low context) on memory retention for Swahili–English word pairs. Memory retention was measured using a multiple-choice test with an English cue word and presented four Swahili choices (the target word and three lures). All lures came from the same block of training as the target word.

The delay condition was manipulated between subjects. All participants were trained on 30 words and took a test after a 5-minute delay. All participants were then trained on the remaining 30 words and randomly assigned to either the 12-hour (N = 38) and or 24-hour delay (N= 39) condition. After their respective delays, participants were given a test on the second set of 30 words. The training condition was manipulated within-subject by adjusting when participants saw the correct Swahili–English word pair in relation to the sample sentence during a training trial. As in previous experiments, for the read condition, the word pair appeared 2 s before the sentence to encourage participants to

simply read the words and not infer the translation. For the lexical inference condition, the correct pairing was shown 8 s after the sentence appeared to encourage participants to infer the translation followed by 2 s of feedback. To preserve the within-subject nature of the task, the blocks were of a mixed-list design, meaning that half the trials within each block were inference trials while the other half were reading trials.

Procedure

Participants completed all procedures online using the Qualtrics web platform. The procedure mimicked previous experiments, as participants were first provided with instructions, completed one block of study, had a delay period followed by a final test. All participants first studied one block of words that included both inferencing and the read-only control condition, followed by a 5-minute delay and then a multiple-choice test. Participants then studied the second block and were dismissed for either 12 hours or 24 hours when they would then be asked to complete another multiple-choice test covering material from the second block.

Results

A multilevel logistic regression model analyzed the influence of training (read vs. lexical inference with feedback), sentence context (high context vs. low context), and delay (5 min vs. 12 h vs. 24 h) on memory retention for the Swahili English word-pairs, measured by a multiple-choice test. The full model for the final test was determined a priori and the same step-wise procedure used in Experiments 1A, 1B, and 2 was applied to determine model fit. The final multiple-choice test score served as the dependent variable. Coefficients for the full model can be found in Table 3.

Adding the training type predictor to the base model significantly improved model fit, $\chi(1) = 12.89$, $p < .001$. Adding the context predictor did not significantly improve model fit, $\chi(1) = 2.61$, $p = .11$, and adding their interaction term also did not significantly improve fit, $\chi(1) = .08$, $p = .78$. Adding the delay as a predictor significantly improved model fit, $\chi(2) = 68.98$, $p < .001$, but adding the two-interaction terms did not significantly improve model fit, $\chi(4) = .38$, $p = .82$, and adding the three-way interaction term also did not significantly improve model fit, $\chi(2) = .78$, $p = .38$.

As show by Fig. 6, post hoc linear comparisons of the full model indicated there was a significant main effect of training type such that the read condition produced greater memory retention at final test compared with the generation condition ($z = 3.50$, $p = .002$). Additionally, there was a significant difference found between the 5-minute delay condition and the 12-h delay condition ($z = 6.59$, $p < .001$) with the 5-minute delay condition resulting in improved memory retention relative to the 12-h delay condition. A significant difference was also found between the 5-mintue delay condition and the 24-h delay condition ($z = 6.52$, $p < .001$) such that the 5-minute delay condition resulted in better retention compared with the 24-h delay condition. There we no significant difference between the 12-h and the 24-h delay condition ($z = 0.42$, $p = 1.00$). All post hoc p values were corrected using the Bonferroni method.

Conditional analyses

To again address concerns as to whether the effects of lexical inferencing relative to the read-only control were attributable to performance during the lexical inferencing task, exploratory analysis investigated the interaction of initial inference accuracy on final test performance relative to the control condition. We used the same procedure described in Experiment 1A to divide the lexical inferencing condition into accurate and inaccurate trials and compared each to the read-only control in a 3 × 3 × 2 multilevel model. Given the lack of interactions in the step-wise comparisons of the original model, linear comparisons again focused on the main effects of training to preserve statistical power and all reported p values were corrected using the Bonferroni method. As in the previous experiments, lexical inferencing resulted in reduced final test scores when initial inferences were accurate ($z = 2.85$, $p = .01$) and inaccurate ($z = 2.39$, $p = .03$).

Table 3 Experiment 3 multilevel logistic regression model fixed effects output for multiple-choice final test scores (Intercept is the read-only control with high context with a 5min delay)

Fixed Effects	β	Std. Error	z	p
Intercept	0.17	0.10	1.80	0.07
Inference	−0.15	0.12	1.25	0.21
Low Context	−0.01	0.12	0.06	0.96
12-h Delay	−0.38	0.15	2.54	0.01
24-h Delay	−0.43	0.15	2.83	.005
Inference × Low Context	−0.10	0.17	0.60	0.55
Inference × 12-h Delay	−0.20	0.21	0.93	0.35
Inference × 24-h Delay	−0.15	0.21	0.71	0.48
Low Context × 12-h Delay	−0.27	0.21	1.27	0.20
Low Context × 12-h Delay	−0.17	0.21	0.80	0.42
Inference × Low context × 12-h Delay	0.33	0.30	1.11	0.27
Inference × Low context × 24-h Delay	0.21	0.30	0.70	0.48

β and standard error are presented in logit units

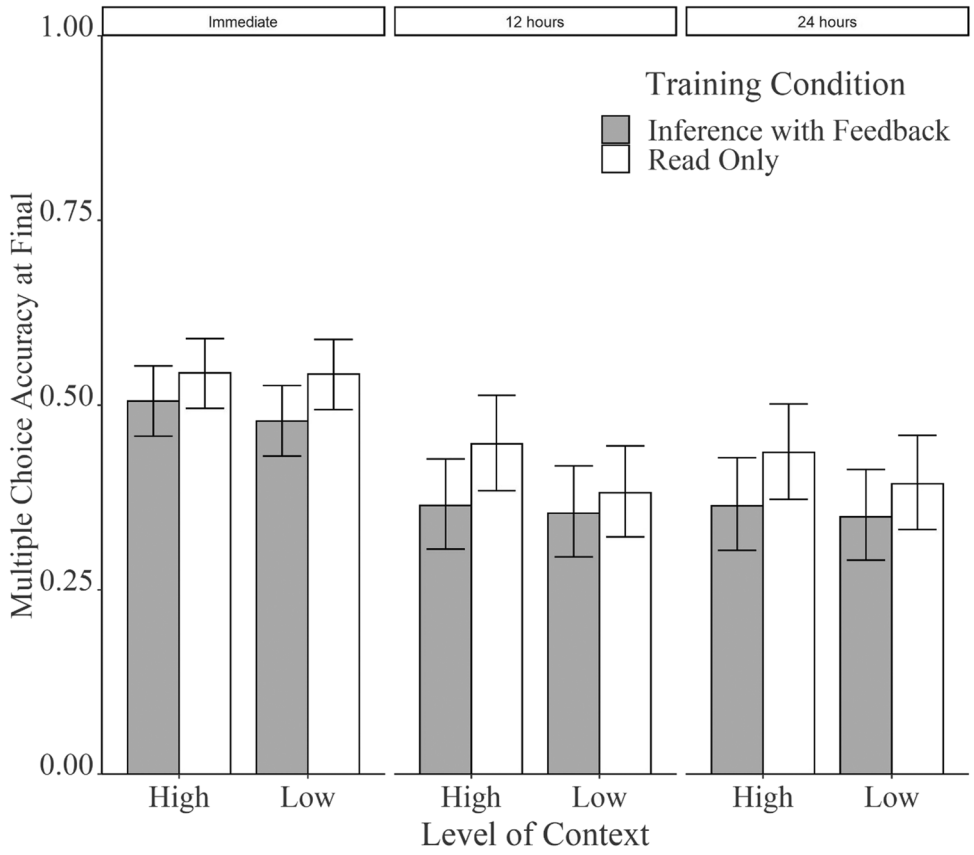

Fig. 6 In Experiment 3, the negative generation effect pattern was replicated in both the 12-hour and 24-hour delay intervals. Error bars are 95% confidence intervals

Discussion

Past research suggests that extending retention intervals to 24 h or greater increases the magnitude of the generation effect. As shown by Fig. 6, our multiple-choice test results indicated that there was no positive generation effect at any of the time points. This finding suggests that even after lengthier delays, lexical inferencing does not result in a benefit relative to a read-only condition, at least not for the ability to translate between the languages.

General discussion

The present study had two aims: (1) to determine whether lexical inferencing produced a memory benefit for foreign language (L2) vocabulary learning relative to a read-only control and (2) to determine whether learning during lexical inferencing could be attributed to the generation effect of memory. For the first aim, across all four experiments, the inference condition resulted in lower memory retention for the L1–L2 word pairs compared with the control condition. As for the second aim, the experiments applied various moderators that have been found to influence the magnitude of the generation effect to our lexical inferencing task, including manipulations of difficulty, the inclusion of corrective feedback, and delayed final tests. All of these factors failed to improve the benefits of inferencing which, in addition to the general negative impact of lexical inferencing relative to the read-only control on memory, suggests a generation effect for L2 vocabulary does not occur during lexical inferencing.

Past research suggests that learning during lexical inferencing may be linked to the generation effect (Barcroft, 2015; Frishkoff et al., 2016; Joe, 1998) because the demands of an inference task closely mimic the methodology used to assess generation effects. In both paradigms, the learner must self-generate information using a provided context or generation rule. The findings of the present study suggest that while inference tasks closely mimic generation effect paradigms, inferencing appears to operate through a different mechanism. In all experiments, when the final test format relied on the retrieval of the cue-target relationship (i.e., cued recall and multiple-choice tests), the read-only control resulted in higher accuracy compared with lexical inferencing. The free recall tests that did not rely on the cue–target

relationship suggest that generation improves memory for the L1 meaning but not memory for the L2 word form.

Further evidence of the disparity between typical generation effect findings and the present study can be found in the analysis of the three moderators included in the present study, specifically manipulations of task difficulty (Experiments 1–3), the presence of feedback (Experiments 1A and 1B), and manipulations of retention interval (Experiment 3). Increasing task difficulty, providing feedback and increasing the retention interval were predicted to increase memory retention after lexical inferencing compared with the read-only control relative to easier tasks, no feedback, and shorter retention intervals, respectively. All three moderators did not result in the expected outcome with the exception of task difficulty which improved memory, but only for free recall of the L1 during Experiment 2. These findings when considered together suggest that lexical inferencing should not be considered a special case of the generation effect.

One explanation for the reduced memory for the L2 word form after lexical inferencing in comparison to the read-only control is that during lexical inferencing, the L2 word form itself does not guide the generation, but rather the surrounding context informs the meaning. This may result in fewer cognitive resources allocated to processing the word form itself. According to the TOPRA model (Barcroft, 2002), the emphasis on meaning over form processing likely promotes memory for the meaning at the expense of memory for the word form. To address the lack of word form processing, we suggest that the attention of the learners be orientated to the word form during inferencing to promote better encoding. For example, sentences could be constructed to contain fewer animate or emotional words that potentially drew attention away from the target L2 word form.

The present study deviated from previous research on lexical inferencing in both methodology and results. Prior research on lexical inferencing, including the *Clockwork Orange* study (Saragi et al., 1978) and the study by Rott (1999), resulted in higher rates of retention compared with the present series of experiments. One key difference between those studies and the present study is the earlier studies used multiple exposures to the same target words in a variety of contexts rather than a single exposure. Studies with methodologies more similar to the ones used in the current study (relying on only a single exposure) had similar learning outcomes to the present study with fewer than 20% of the words retained during the final test (Mondria, 2003).

Future research needs to focus on improving memory for the word form of the L2 word during lexical inferencing. During lexical inferencing, there is a memory benefit for the L1 word so if lexical inferencing can be combined with another method meant to bolster L2 form learning, it is possible that inferencing could improve L2 vocabulary learning more than control conditions. One possible method would be to provide additional exposures to the target words in a variety of situations (using the words in different sentences), a method that has been shown to improve learning during lexical inferencing but has not been compared with a read-only control (see Rott, 1999). With multiple exposures, learners may be able to recognize word forms in each subsequent exposure which could prompt a covert retrieval of the memory of a previous exposure and in turn improve memory retention (Roediger & Karpicke, 2006). Episodic memory retrieval has a long history of improving memory for foreign language vocabulary through a phenomenon known as the *testing effect* (Rowland, 2014). As learners see recurring word forms, they may initiate a retrieval of prior experiences which then promotes retention, effectively making the task easier and potentially freeing up cognitive resources to be allocated to encoding the novel word form.

Interpretation of the results of the present study should be limited to people with no knowledge of the target language and may not generalize to advanced learners. Advanced language learners may be able to encode L2 word forms more easily due to familiarity with recurring phonemic patterns or prior knowledge of word structure (e.g., the meaning of affixes or roots). For example, advanced learners maybe be able to tie novel L2 words to familiar L2 words with similar meanings and build on preexisting associations between L2 words and their L1 semantic meanings, similar to how an advanced L1 learner acquires new L1 vocabulary

For pedagogy, our findings suggest that lexical inferencing may not be an efficient means of learning L2 vocabulary for novice learners. The failure to surpass the read-only condition used in the present study is most concerning given that reading alone is known to be a poor method of study for word pairs when compared with the more effective methods such as keyword mnemonics and retrieval practice routinely produce higher rates of memory retention (Barcroft, 2015; Miyatsu & McDaniel, 2019). There is potential for inferencing to eventually surpass a read-only control if future research determines how memory for the L2 forms can be enhanced during inferencing. Until that time, the findings presented here suggest that inferencing should not be the primary method used for learning new vocabulary, at least for novice learners.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13421-022-01348-5.

Code availability Code for PsychoPy and Qualtrics program as well as R analysis script are available online (https://osf.io/grxmq/).

Funding Steven J. Dessenberger was funded by National Science Foundation Graduate Research Fellowship Program (DGE-1745038).

Data availability Materials and data are available online (https://osf.io/grxmq/).

Declarations

Conflicts of interest The authors declare they have no financial interests.

Ethics approval This project was approved by the Washington University in St. Louis Institutional Review Board. The project was classified as exempt from review as it presented no greater than minimal risk to participants, and participants were all consenting adults over the age of 18 and were considered to be not part of a protected population.

Consent to participate All participants were provided with an information sheet detailing the purposes of the experiment as well as the general nature of the tasks and time requirements as well as the compensation (course credit). Participants were informed the experiment was voluntary and that they would be provided with an alternative means of acquiring course credit should they so choose. Consent to participate was collected at the time of participation.

Consent for publication All participants were provided with an information sheet detailing that the experimental data would be published, but that any and all personally identifiable information would be removed prior to publication.

References

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

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

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition, 27*(03). https://doi.org/10.1017/S0272263105050175

Barton, K. (2020). *MuMIn: Multi-Model Inference* (R Package Version 1.43.17) [Computer software]. https://CRAN.R-project.org/package=MuMIn

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*(2), 201–210. https://doi.org/10.3758/BF03193441

Bordag, D., Kirschenbaum, A., Tschirner, E., & Opitz, A. (2015). Incidental acquisition of new words during reading in L2: Inference of meaning and its integration in the L2 mental lexicon. *Bilingualism: Language and Cognition, 18*(3), 372–390. https://doi.org/10.1017/S1366728914000078

Burgess, A. (1975). *A clockwork orange (Reprinted)*. Heinemann.

R Core Team. (2018). *Fitting linear mixed-effects models using lme4*. R Foundation for Statistical Computing. https://www.R-project.org/

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. https://doi.org/10.1016/S0022-5371(72)80001-X

de la Garza, B., & Harris, R. J. (2017). Acquiring foreign language vocabulary through meaningful linguistic context: Where is the limit to vocabulary learning? *Journal of Psycholinguistic Research, 46*(2), 395–413. https://doi.org/10.1007/s10936-016-9444-0

Frishkoff, G. A., Collins-Thompson, K., Hodges, L., & Crossley, S. (2016). Accuracy feedback improves word learning from context: Evidence from a meaning-generation task. *Reading and Writing, 29*(4), 609–632. https://doi.org/10.1007/s11145-015-9615-7

Gardiner, J. M. (1989). A generation effect in memory without awareness. *British Journal of Psychology, 80*(2), 163–168. https://doi.org/10.1111/j.2044-8295.1989.tb02310.x

Geva, E., Galili, K., Katzir, T., & Shany, M. (2017). Learning novel words by ear or by eye? An advantage for lexical inferencing in listening versus reading narratives in fourth grade. *Reading and Writing, 30*(9), 1917–1944. https://doi.org/10.1007/s11145-017-9759-8

Greenwald, A. G., & Johnson, M. M. S. (1989). The generation effect extended: Memory enhancement for generation cues. *Memory & Cognition, 17*(6), 673–681. https://doi.org/10.3758/BF03202628

Horthorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal, 50*(3), 346–363 http://multcomp.r-forge.r-project.org/

Joe, A. (1998). What effects do text-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics, 19*(3), 357–377. https://doi.org/10.1093/applin/19.3.357

Johns, E. E., & Swanson, L. G. (1988). The generation effect with nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(1), 180–190.

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.

Kane, J. H., & Anderson, R. C. (1978). Depth of processing and interference effects in the learning and remembering of sentences. *Journal of Educational Psychology, 70*(4), 626–635. https://doi.org/10.1037/0022-0663.70.4.626

McDaniel, M. A., & Waddill, P. J. (1990). Generation effects for context words: Implications for item-specific and multifactor theories. *Journal of Memory and Language, 29*(2), 201–211. https://doi.org/10.1016/0749-596X(90)90072-8

Miyatsu, T., & McDaniel, M. A. (2019). Adding the keyword mnemonic to retrieval practice: A potent combination for foreign language vocabulary learning? *Memory & Cognition*. https://doi.org/10.3758/s13421-019-00936-2

Mondria, J.-A. (2003). The effects of inferring, verifying, and memorizing on the retention of l2 word meanings: An experimental comparison of the "meaning-inferred method" and the "meaning-given method". *Studies in Second Language Acquisition, 25*(4), 473–499. https://doi.org/10.1017/S0272263103000202

Mulligan, N. W. (2004). Generation and memory for contextual detail. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(4), 838–855. https://doi.org/10.1037/0278-7393.30.4.838

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General, 143*(2), 644–667. https://doi.org/10.1037/a0033194

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition, 21*(4), 589–619. https://doi.org/10.1017/S0272263199004039

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System, 6*(2), 72–78. https://doi.org/10.1016/0346-251X(78)90027-1

Shen, M. (2010). Effects of perceptual learning style preferences on L2 lexical inferencing. *System, 38*(4), 539–547. https://doi.org/10.1016/j.system.2010.09.016

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory, 5*(6), 607–617. https://doi.org/10.1037/0278-7393.5.6.607

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition: Incidental vocabulary acquisition. *Language Learning, 61*(1), 219–258. https://doi.org/10.1111/j.1467-9922.2010.00593.x

Webb, S., & Chang, A. C.-S. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research, 19*(6), 667–686. https://doi.org/10.1177/1362168814559800

Zou, D. (2016). Comparing Dictionary-induced Vocabulary Learning and Inferencing in the Context of Reading. *Lexikos, 26*(1). https://doi.org/10.5788/26-1-1345

Open Practices Statement

Data and materials for all experiments are available online (https://osf.io/grxmq/), and none of the experiments was preregistered.