Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Listening to disfluent speech: Robust effect at processing may not extend to learning

Emma Libersky [*], Margarita Kaushanskaya

Department of Communication Sciences and Disorders & Waisman Center, University of Wisconsin-Madison, Madison, USA

ARTICLE INFO

Keywords:
Disfluency
Word learning
Language processing
Predictive processing

ABSTRACT

Fillers, such as uh and um in English, are reliable predictors of upcoming novelty, and listeners harness these statistics to process disfluent speech efficiently. Predictive processing begets prediction error, and therefore it is possible that fillers impact word learning when they precede unexpected referents. However, the effect of fillers on word learning is unknown, especially over different time scales. Across five experiments, we sought to investigate the impact of fillers on language processing and learning, focusing on the relationship between fillers and predictive processing. Following exposure to novel words under fluent and disfluent learning conditions, we measured participants' word recognition at immediate vs. delayed testing. Participants consistently used fillers to predict upcoming novelty at exposure, but they tended to demonstrate similar retention of words taught in fluent and disfluent conditions, whether or not they made prediction errors. We frame our findings in the context of predictive processing and disfluency, concluding that disfluency shapes listeners' predictions but this effect does not carry over to word retention.

Introduction

Fillers, or filled pauses, are a common disfluency type that exists across languages in language-specific forms (Clark & Fox Tree, 2002). English-speaking adults produce the fillers "uh" and "um" roughly once per 100 words, though these statistics vary according to environmental and speaker characteristics (Bortfeld et al., 2001). Importantly, fillers are more likely to be produced ahead of low-frequency or otherwise difficult-to-retrieve words (Beattie & Butterworth, 1979; de Jong, 2016), and they are a reliable signal to upcoming delays in the speech stream (Clark & Fox Tree, 2002). Adult listeners are able to harness these statistics to make predictions for "what comes next" in a disfluent utterance, although the mechanism underlying this process remains under debate and the downstream consequences for learning are poorly understood. According to the listener-oriented view, fillers are intentionally produced as a cue to upcoming difficulty (Clark and Fox Tree, 2002). Alternatively, the disfluency-novelty link may stem purely from production demands (the speaker-oriented view, Corley & Stewart, 2008). The process underlying disfluency processing has critical implications for learning, as disfluencies and the assumptions people have about them may make people less likely to trust and learn from a disfluent speaker (White et al., 2020). In the present study, we rely on the predictive processing hypothesis to examine adults' expectations following fillers and the consequences of fillers for novel word learning.

Processing fillers

There is robust evidence that listeners assign meaning to fillers and use them to process language more efficiently (e.g., Arnold et al., 2007; Arnold et al., 2004; Arnold et al., 2003; Corley et al., 2007; Kidd et al., 2011; Collard et al., 2008; Morin-Lessard & Byers-Heinlein, 2019; Owens & Graham, 2016; Watanabe et al., 2008; Karimi et al, 2019). In one example of filler perception research, Arnold and colleagues (2007) used eye tracking and gating tasks to assess listeners' predictions following fillers. Participants looked at a screen with novel and familiar stimuli on either side and heard fluent or disfluent instructions to "click on the/thee uh" object. In the eye-tracking task, participants heard the full sentence and researchers analysed looks to target following the fluent or disfluent determiner. They found a bias toward the novel object upon hearing the disfluent determiner, but no such bias for fluent determiners. On the gating task, participants only heard a portion of the sentence: either "click on," "click on the/thee uh," or "click on the/thee uh red." None of the sentence fragments contained the target object, and participants were asked to guess which object the speaker was referring

[*] Corresponding author.
E-mail addresses: libersky@wisc.edu (E. Libersky), kaushanskaya@wisc.edu (M. Kaushanskaya).

https://doi.org/10.1016/j.jml.2025.104704
Received 17 February 2025; Received in revised form 29 September 2025; Accepted 3 October 2025
Available online 13 October 2025

to. They found that participants were more likely to choose the unfamiliar object on disfluent trials and the familiar object on fluent trials, consistent with the hypothesis that participants use disfluency as a cue to upcoming novelty. Disfluent prosody (i.e., "click on" was longer and higher pitch in the disfluent vs. fluent condition) was not sufficient to guide predictions, rather, participants needed to hear the filler in order to make reliable predictions.

Neurolinguistic work has also indicated effects of fillers on listeners, in that fillers make unpredictable words less surprising to listeners. Corley and colleagues (2007) measured adults' Event-Related Potentials (ERPs) during a listening task. One important ERP component in language processing is the N400, in which voltages change as a result of the demands associated with integrating semantic information. Corley and colleagues told participants to listen to prerecorded snippets of conversational speech for understanding, as they would in normal conversation. Some of the stimuli ended predictably (e.g., "That drink's too hot; I've just burnt my tongue"), whereas others ended unpredictably (e.g., "That drink's too hot; I've just burnt my nails"). When the stimuli were fluent, they found an N400 effect in the unpredictable condition, indicating that the referent did, indeed, surprise listeners. However, when the unpredictable word was preceded by a filler (e.g., "That drink's too hot; I've just burnt my *er* nails"), this effect was tempered.

Listeners' expectations for disfluent speech appear malleable, and expectations can be manipulated through speaker information (e.g., Arnold et al., 2007; Bosker et al., 2014; Bosker et al., 2019; Thacker et al., 2018; Barr & Seyfeddinipur, 2010; Orena & White, 2015). For example, Arnold and colleagues (2007) replicated their eye-tracking task with additional speaker information, which provided cover stories for why the speaker might be disfluent in reference to easily accessible referents. In one replication, participants were told that the speaker had object agnosia, a condition that results in difficulty retrieving even highly familiar words. In this manipulation, participants did not use fillers predictively, suggesting that, in light of an alternative explanation for disfluency, participants were able to limit their use of disfluency as a cue to upcoming novelty. In another study, Bosker and colleagues (2014) examined the impact of nonnative accent on listener predictions following fillers. They used a similar paradigm to Arnold and colleagues, except that their on-screen images were either high- or low-frequency familiar objects. Like Arnold and colleagues, they found that participants adjusted their expectations for post-disfluency referents as a function of speaker information: participants used disfluencies to anticipate low-frequency referents in the native accent condition only. Taken together, these findings indicate that listeners are highly sensitive to the factors that influence disfluency and can adapt their predictions according to available context.

Fillers, memory, and learning

Although the research on filler processing is extensive, literature describing the impact of fillers on memory and learning is limited. In one of the first studies of the impact of fillers on memory, Corley and colleagues (2007) measured memory for familiar target words embedded in fluent and disfluent sentences. Adults remembered words presented alongside disfluencies better than those presented in fluent sentences. Other researchers have tested memory for familiar words following a similar procedure, with mixed results. Diacheck and Brown-Schmidt (2022) reported a memory boost for words presented in a disfluent condition, whereas Bosker and colleagues (2014) reported no differences in recall. While these studies have focused on memory for specific items, there is some evidence that disfluency may impact memory for gist. Fraundorf and Watson (2011) had participants listen to and then recall brief passages. The passages included 14 plot points, some of which were presented disfluently. When asked to retell the passages immediately after hearing them, participants demonstrated better recall for disfluent plot points, reinforcing the possibility that disfluency may boost memory in a variety of contexts.

Investigating the impact of disfluency on learning specifically, Toftness and colleagues (2018) measured the consequences of disfluency on college students' learning by manipulating instructor disfluency in pre-recorded mock lectures. In a between-subjects manipulation, students watched lectures that were identical except in delivery. Here, disfluency was conceptualized broadly, with the instructor demonstrating a general lack of confidence and producing numerous unfilled pauses. Regardless of testing schedule, performance on the multiple-choice test was similar across conditions. Though it is impossible to isolate the consequences of disfluent pauses for students' performance from the other qualities of the disfluent instructor, Toftness and colleagues (2018) provided early evidence for the role (or lack thereof) of disfluency in learning.

The impact of disfluency on *word learning* remains unclear. To our knowledge White and colleagues (2020) conducted the first study in this area and Libersky and colleagues (2023) were the first to investigate this question in adults. White and colleagues (2020) emphasized the impact of disfluency on speaker credibility. In their first experiment, preschoolers were introduced to fluent and disfluent puppets. The puppets then labelled competing objects with the same novel label, and children were asked to endorse one of the word-object pairs. Children were more likely to endorse the pairing made by the fluent speaker. In their second experiment, they tested the impact of disfluency on retention of novel labels and found no effect. In a series of three experiments, Libersky and colleagues (2023) taught monolingual and bilingual adults novel words in a paired-associate learning task. Participants saw an illustration of a single novel fish on a screen and heard its name in a fluent or a disfluent sentence (e.g., "Thee [uh] gagek lives in warm streams in sunny areas"). After all novel fish were presented, participants advanced immediately to a three-alternative forced choice recognition task. Monolinguals performed similarly across conditions in all three experiments. Bilinguals performed similarly across conditions when only novel fish were presented but demonstrated better performance in the fluent condition when the task was padded with trials featuring familiar fish, indicating that learners with relatively less English-language experience may benefit from fluency in certain contexts. An important limitation of this prior study was that the design only allowed for measuring the effect of disfluency on retention, rather than the cascading consequences of disfluency at both processing and learning. That is, the impact of disfluency on processing was inferred rather than observed. In the present study, we address this gap by measuring the impact of disfluency on processing and learning of novel words *in the same individuals*. Thus, we first confirm whether or not learners used disfluency predictively during processing and then ask whether disfluency has consequences for learning.

Prediction error and word learning

Within the present study, we consider the impact of disfluency on word learning through the prism of prediction error, considering the possibility that predictive processing of fillers has cascading consequences for word learning. The predictive processing hypothesis posits that, in all areas of cognition, humans are constantly making and updating predictions as they interact with the world around them (e.g., Clark, 2013; Schrimpf et al., 2021). Within the realm of language processing, listeners rely on syntactic (Yano, 2018; Kamide et al., 2003), semantic (Gambi et al., 2021; Kamide et al., 2003), and talker information (Van Berkum et al., 2008; Borovsky & Creel, 2014) to anticipate upcoming information. The literature on filler processing fits neatly into this framework, as listeners harness fillers to anticipate upcoming referents, whether through statistical learning or through extracting pragmatic information (Kidd et al., 2011; Clark & Fox Tree, 2002). When given additional talker information (Arnold et al., 2007; Bosker et al., 2014) or when listening to a speaker who produces disfluencies in an atypical manner (Lowder et al., 2019), listeners update their predictions and stop utilizing fillers as a cue to novelty.

Beyond its role in language processing, prior work has indicated a strong role for prediction in word learning. Some studies tie correct predictions (i.e., efficient processing) to language growth (Fernald & Marchman, 2012), but there is also a rich literature linking prediction *error* and learning. Within the prediction error framework, better learning happens when the learner is surprised (Metcalfe, 2017), or, in a more extreme view, learning *only* happens when a prediction error occurs (Fitz & Chang, 2019). To test the prediction error mechanism in word learning, Gambi and colleagues (2021) taught participants novel words in high- or low-constraint contexts. These led to stronger or weaker predictions of familiarity, respectively. Adults tested after a 5-minute delay learned best when strong predictions were disconfirmed, but children did not show this effect. In another study with children, Reuter and colleagues (2019) found that the children who learned best were the ones who used a predict and redirect strategy at learning—i.e., they looked to the familiar object based on sentence structure then redirected to the novel object upon hearing the novel label. Taken together, these findings suggest that the ability to recover and learn from prediction errors develops over time, but, when developed, is a powerful learning mechanism.

Counterintuitively, within the context of fillers, a prediction error effect would improve word retention under *fluent* conditions. Fillers serve as a cue to novelty, and numerous processing studies have indicated that listeners are able to harness this cue (Arnold et al., 2007; Bosker et al., 2014). Consequently, listeners may expect a novel referent following a filler and expect a familiar referent in the absence of a filler. In a word learning task where both fluent and disfluent sentence contexts are present, learners are likely to make prediction errors in situations where they hear a fluent trial containing a novel word and predict a familiar rather than a novel referent. It is on these trials that we would expect listeners to be surprised when a novel referent is named within a *fluent* sentence. Fluency would, thus, only impact learning when it cooccurs with a prediction error.

We also investigate the time course of word learning by testing adults either immediately or after a delay. Though memory traces decay rapidly (Baddeley, 1975; Barrouillet et al., 2004; Barrouillet et al., 2007) it is not always the case that word retention is worse after a delay. Some learning contexts that are more difficult in the moment may yield slower rates of decay—i.e., they serve as a desirable difficulty (Schmidt & Bjork, 1992). In the present study, we compare immediate performance with performance after a six-minute delay and a 24-hour delay. The six-minute delay is consistent with the word learning paradigm used by Gambi and colleagues (2021) that we adapted for the present study, as well as that used by other researchers who work within the desirable difficulties framework (e.g., Knabe et al., 2023) and has yielded effects consistent with desirable difficulties when compared with performance at immediate delay. The 24-hour delay was intended to make the delay condition more demanding and to allow for novel words to be consolidated into the lexicon (Davis et al., 2009; Gaskell & Dumay, 2003).

Current study

In the present study, we investigated the impact of fillers on language processing and word retention over different time scales (immediate, six-minute delay, 24-hour delay) across five experiments. We had two predictions:

(1) If participants are able to harness the statistics of fillers, they will be more likely to predict a novel referent when a speaker is disfluent than when a speaker is fluent.
(2) If fillers impact word retention via prediction error, recognition of novel words will be better under fluent learning conditions, assuming that disfluencies trigger prediction of novelty, and, thus, the fluent context yields prediction errors. If prediction error serves as a desirable difficulty, memory for words learned under the fluent condition may decay more slowly than memory for words learned in the disfluent condition or even improve after a delay.

The specific aim and design of each of the five experiments is summarized in Table 1.

Data Availability.

All data and scripts associated this paper are available on the Open Science Framework (OSF) at https://osf.io/yj7pa/. Unless otherwise noted, model syntax and summaries are found in the “frequentist_analyses.rmd” file.

Experiment 1

Participants

A total of 118 adults (66 immediate condition, 52 delay condition) were included in the final sample. Our recruitment target was set to 70 participants (minimum of 50 after exclusions) per immediate and delay condition and preregistered (see OSF Registrations tab, https://osf.io/yj7pa/). This target was derived from Gambi and colleagues (2021), as we adapted their paradigm and included a similar number of parameters in our models. All participants gave informed consent through a protocol approved by the UW-Madison Minimal Risk Research IRB (protocol ID: 2013–0984, Effects of Bilingualism on Learning and Memory). Our inclusionary criteria were preregistered and required that participants were between ages 18–40 (inclusive), resided in the United States, had no hearing impairments, were monolingual speakers of English (i.e., self-rated speaking proficiency less than “adequate” in any languages besides English, as reported on the Language Experience and Proficiency Questionnaire, LEAP-Q; Marian et al., 2007), and, for participants in the delayed test condition, passed an attention check. The

Table 1
Summary of experiments.

Exp.	Participants	Aim	Item distribution	Items randomly assigned to condition?	Length of delay
1	66 immediate 52 delay	Initial experiment	Six disfluent and novel Two fluent and novel Four fluent and familiar	No	Six minutes
2	75 immediate 52 delay	Address possibility of item effects in Exp. 1	Six disfluent and novel Two fluent and novel Four fluent and familiar	Yes	Six minutes
3	68 immediate 60 delay	Extend delay (vs. Exp. 2)	Six disfluent and novel Two fluent and novel Four fluent and familiar	Yes	24 h
4	62 immediate 58 delay	Remove in-experiment association between disfluency and novelty	Four disfluent and novel Four fluent and novel Four fluent and familiar	Yes	Six minutes
5	68 immediate 63 delay	Extend delay (vs. Exp. 4)	Four disfluent and novel Four fluent and novel Four fluent and familiar	Yes	24 h

attention check was included for participants in the delayed test condition only in order to ensure that they were seated at their computer watching the video that played during the delay. An additional four participants were tested in the immediate test condition but were excluded for reporting high proficiency in a language other than English (a self-rated speaking proficiency of "adequate" or better, as reported on the LEAP-Q; Marian et al., 2007). An additional 30 participants were tested in the delayed test condition but excluded for reporting high proficiency in a language other than English ($n = 6$), failing the attention checks ($n = 23$), or both ($n = 1$). All participants were recruited via Prolific (Prolific, 2014) prolific.co and tested remotely. See Table 2 for participant characteristics.

Procedure

Participants were tested remotely, via Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Following consent, participants completed a sound check (James, 2019) to ensure that auto-play was enabled for auditory stimuli. To assess language experience and ability, participants completed the LEAP-Q (Marian et al., 2007), and English language proficiency was assessed using a speeded version of the Lexical Test for Advanced Learners of English (LexTALE; Lemhöfer & Broersma, 2012; Poort & Rodd, 2019). The LEAP-Q is a self-report measure of demographics and language background, in which participants self-rate their proficiency across their languages and describe when and how they acquired the languages. The LEAP-Q has been shown to have good internal validity and to align well to other (non-self-report) measures of language proficiency (Marian et al., 2007). The LexTALE is a lexical decision task designed to assess English-language knowledge quickly in L2 English speakers and has also been used to assess L1 proficiency (Dijkgraaf et al., 2017; Poort & Rodd, 2017). Participant characteristics, including language background, are described in Table 2. Participants completed study tasks in the following order: Consent, sound check, LEAP-Q, experimental task, and LexTALE.

Materials

Exposure trials

The design of the exposure phase was adapted from Gambi and colleagues (2021) and was the same for all participants. Participants saw two objects on the computer screen, one familiar and one novel, and heard a female native speaker of American English tell them to "Click on | the | [object]" or "Click on | thee uh/um | [object]." The sentences were broken up into three parts and presentation was self-paced, such that participants had to select an object following "click on," "the/thee uh/thee um," and "[object]" in order to hear the full sentence. When selecting an object, participants were instructed to "guess the final word of [the] sentence." Participants completed two training trials followed by twelve exposure trials (six novel-disfluent, two novel-fluent, and four familiar-fluent, described below). The procedure is described in Fig. 1.

Table 2
Experiment 1 participant characteristics.

	Immediate test	Delayed test	*p*
Sample size (Count)	66	52	
Gender (Count)[a]	–	–	.73
Female	29	24	
Male	34	26	
Nonbinary / Prefer to Self ID	3	1	
Age[b]	27.2 (5.91)	26.8 (6.11)	.72
Years Education[c]	15.1 (2.16)	15.6 (2.61)	.26
LexTALE Score	86.7 (10.4)	88.5 (7.47)	.27

Note: Values indicate mean (sd) unless otherwise indicated.

[a] One participant in the delayed test condition did not report gender.

[b] One participant in the delayed test condition did not report age. Screener responses indicated that they were between 18 and 40 years old.

[c] Two participants in the immediate test condition and one participant in the delayed test condition did not report years education. In instances where participants reported fewer than 12 years of schooling but also reported that they had finished high school, years of education was transformed as follows: high school = 12 years, some college = 14 years, college or some graduate school = 16 years (as the length of graduate school is variable), and masters = 14 years.

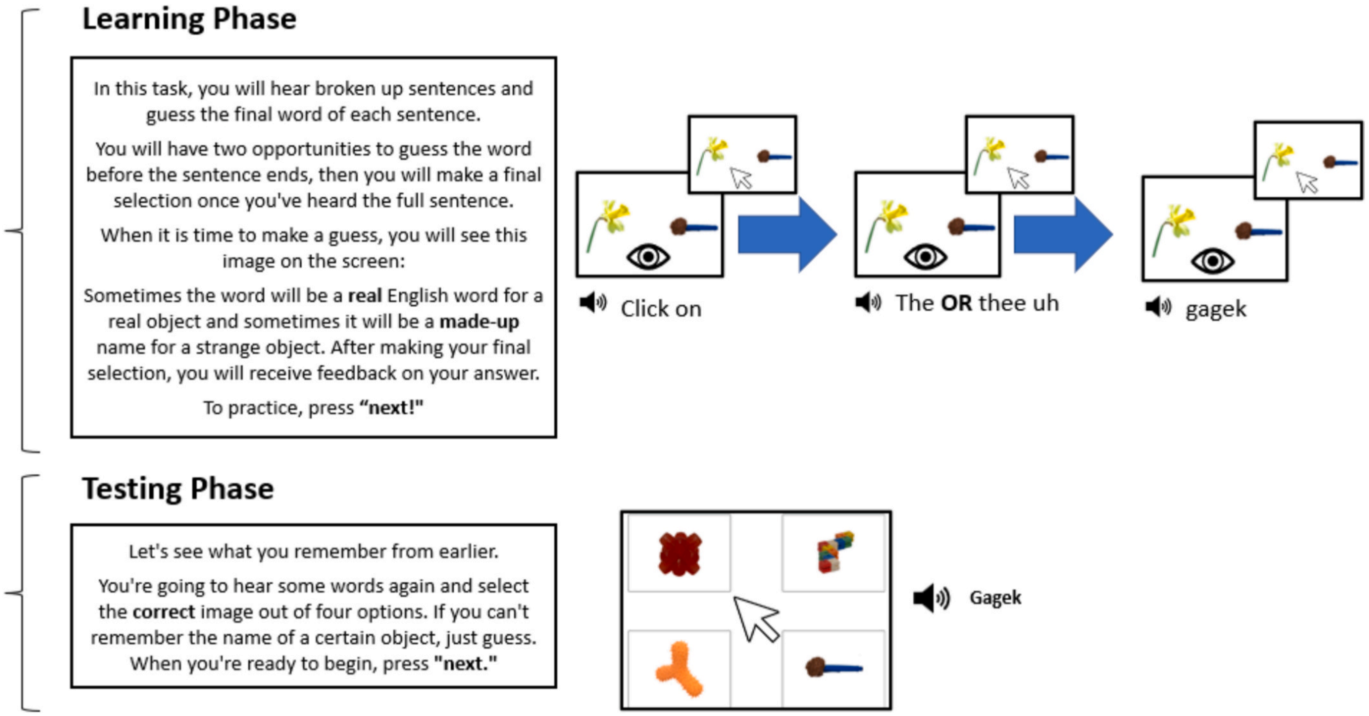
The two training trials were both fluent; one ended with a novel stimulus and one ended with a familiar stimulus. Participants were told at the end of each training trial if their final selection was accurate. Of the twelve exposure trials, eight ended in a novel label for the novel image and four were filler trials and ended in a familiar label that referred to the familiar object. The twelve trials were presented in a random order. Novel words were all English-like, two-syllable words selected from Gupta and colleagues (2004). They were associated with novel objects selected from the Novel Object and Unusual Name (NOUN) Database (Horst and Hout, 2016). Familiar objects and their labels were all selected from the Bank of Standardized Stimuli (BOSS; Brodeur et al., 2010), with the requirement that they had a two-syllable modal name. Because familiar and novel objects appeared on the computer screen side by side during exposure, we matched novel and familiar images on visual complexity (proxied by JPG file size; Bates et al., 2003) using Match (van Casteren & Davis, 2007).

All filler trials, on which the familiar object was referenced, were presented in a fluent sentence frame, i.e., "Click on the [object]." In contrast, novel objects were preceded by a disfluency ("thee uh" or "thee um") on six trials and presented in a fluent sentence frame on two trials. We verified that novel stimuli in the fluent and disfluent conditions did not meaningfully differ (see the Discussion for more on limitations of the design). These ratios were adapted from Gambi and colleagues (2021) and reflected the predictive nature of disfluencies in spontaneous speech. As in spontaneous speech, disfluencies were a cue to upcoming novelty. The most natural-sounding disfluencies were selected from a larger pool of recordings based on consensus from lab members. Recordings were spliced in Audacity (Audacity-Team, 2018) and normalized to 70 dB in Praat (version 6.1.16; Boersma & Weenink, 2020). All participants learned the same six objects in the disfluent condition, and item effects were modelled in our analyses.

Test trials

Participants were tested on the eight novel objects in a field of four recognition task. Participants heard the name of the novel object, from a new female speaker of American English. The recordings at test were from Gupta and colleagues' (2004) database and thus assessed participants' ability to generalize learning to a novel speaker. On each trial, the target object was shown alongside three alternatives, and participants clicked on the image that the speaker was referring to. The alternatives were all novel words that participants had learned a label for, meaning that familiar stimuli were not included at test. Participants were tested on their recognition of each novel word three times, with position on the screen and foils varying across trials. Test trials were pseudorandomized such that the entire list was tested once before any words were tested a second time, and each object was used as a foil six times. Feedback was not provided at test. Words were tested multiple times to increase the number of observations at test and to minimize the possibility that participants were merely guessing the correct answer. Post-hoc analyses indicated that performance improved with each subsequent testing block, but testing cycle did not interact with any of the effects of interest. Moreover, the pattern of results was the same when we reran our analyses on the first testing block only. Thus, we report the planned analyses that included all three testing blocks.

The assignment to testing schedule (immediate vs. delayed) varied between subjects and was randomized. Approximately half of the participants advanced to the test trials immediately after the exposure phase whereas the other half had a short break between exposure and test trials. Participants in the delay condition watched a short cartoon that had no spoken language (Mole as a Gardener; Miler, 1969) and then responded to two comprehension questions. The comprehension



In this task, you will hear broken up sentences and guess the final word of each sentence.

You will have two opportunities to guess the word before the sentence ends, then you will make a final selection once you've heard the full sentence.

When it is time to make a guess, you will see this image on the screen:

Sometimes the word will be a **real** English word for a real object and sometimes it will be a **made-up** name for a strange object. After making your final selection, you will receive feedback on your answer.

To practice, press **"next!"**

Let's see what you remember from earlier.

You're going to hear some words again and select the **correct** image out of four options. If you can't remember the name of a certain object, just guess. When you're ready to begin, press **"next."**

Fig. 1. Graphical depiction of experimental protocol.

questions referenced important details from the video and were multiple choice. These questions were included to verify that participants were watching the video during the delay, rather than doing another activity that might influence their test performance. Including the instructions, video, and multiple-choice questions, the delay lasted six minutes. Participants in the delay condition were excluded from the study if they responded to either question incorrectly.

Results

Data analysis

We analysed participants' object selections on learning trials following the fluent or disfluent determiner (intermediate-selections; the second selection point) and following the full sentence (final-selections), as a function of fluency. We analysed their accuracy on retention trials as a function of fluency, delay (immediate vs. six-minute delay), and prediction accuracy (whether upon hearing the determiner they incorrectly selected the familiar object or correctly selected the novel object).

Item-level data were used in all analyses. Logistic mixed effects models were used to analyse data in R (R Core Team, 2020) with the LME4 package (Bates et al., 2015). Data were cleaned based on reaction times (RTs); observations were excluded if RTs were below 150 ms, above 5000 ms, or beyond 2.5SDs from the individual participants mean RT. Based on these criteria, 313 observations were excluded from the exposure data, across all selection points (out of 4956), and 177 observations were excluded from the recognition data (out of 2832). Maximal random effects structures were retained except when the maximal model had convergence or singularity issues (Barr et al., 2013), and the buildmer package (Voeten, 2023) was used to identify the maximum feasible random effects structure as necessary. Full model specifications (including estimates for random effects) are available on OSF.

Prediction of target at exposure

At exposure, participants' final selections (selections upon hearing the full object name) on filler familiar trials were highly accurate, $M = 0.98$, $SD = 0.13$. To predict selections on novel trials, we regressed accuracy on condition, coded as −.5 for fluent and.5 for disfluent. Participants (correctly) selected the novel object at similar rates across disfluent, $M = 0.86$, $SD = 0.35$, and fluent, $M = 0.79$, $SD = 0.41$, conditions, $B = 0.74$, $SE = 0.51$, $X^2(df = 1) = 2.15$, $p = .143$.

Upon hearing the determiner (the second selection point), participants selected the familiar object in filler trials at above chance levels, $M = 0.61$, $SD = 0.49$; $p < .001$. To predict selections on novel trials, we regressed accuracy on condition (fluent and disfluent, coded as −.5 and.5). Upon hearing the determiner, participants were more likely to select the novel object in the disfluent condition, $M = 0.51$, $SD = 0.50$, than in the fluent condition, $M = 0.39$, $SD = 0.49$, $B = 0.54$, $SE = 0.18$, $X^2(df = 1) = 9.03$, $p = .003$.

Retention

Overall, participants learned novel words to above chance levels ($M = 0.44$, $SD = 0.50$, $p < .001$; chance $= 0.25$). To predict recognition task accuracy, we regressed accuracy at testing on condition (fluent vs. disfluent), delay (immediate vs. delay, coded as −.5 and.5), prediction accuracy (prediction error vs. correct prediction, coded as −.5 and.5), and all interactions. We observed a significant condition by delay interaction, $B = -0.42$, $SE = 0.20$, $X^2(df = 1) = 3.89$, $p = .048$. Simple effects indicated that participants in the immediate test group performed better on disfluent items, $B = 0.11$, $SE = 0.28$, $X^2(df = 1) = 0.07$, $p = .787$, while participants in the delay test group performed better on fluent items, $B = -0.31$, $SE = 0.29$, $X^2(df = 1) = 0.07$, $p = .787$. No other variables were significant predictors of accuracy. Post-hoc analyses in which we removed items that participants failed to map (i.e., items where participants made an incorrect final selection at exposure) indicated a significant group by condition interaction, $B = -0.66$, $SE = 0.27$, $X^2(df = 1) = 4.70$, $p = .030$. Again, simple effects indicated that participants in the immediate test group performed better on disfluent items, $B = 0.22$, $SE = 0.34$, $X^2(df = 1) = 0.019$, $p = .890$, while participants in the delay test group performed better on fluent items, $B = -0.44$, $SE = 0.35$, $X^2(df = 1) = 0.019$, $p = .890$. No other variables were significant predictors of accuracy.

Discussion

We found that adult listeners were more likely to predict upcoming novelty upon hearing a filler than upon hearing a fluent determiner, consistent with prior findings (Arnold et al., 2007; Bosker et al., 2014; Kidd et al., 2011). While analyses indicated different retention patterns for disfluent and fluent items, and no impact of prediction error, we hesitate to overinterpret these findings due to design limitations.

Specifically, our findings may reflect item effects. To assuage concerns that the results of Experiment 1 were muddied by memorability of the specific items across the fluent and the disfluent conditions, we ran additional analyses comparing the novel items assigned to the fluent vs. disfluent condition on several factors that could impact learning including characteristics of the novel word (mean biphone probability of the novel word), the novel object (visual complexity, familiarity, nameability, colour saliency, texture saliency), and the familiar object with which the novel object was paired (word frequency). These analyses yielded no significant differences between conditions, although we acknowledge that these analyses are likely underpowered (see OSF for these analyses). Moreover, because the same items were always assigned to the same condition, the potential for item effects was high despite careful matching of item characteristics across the conditions. To address this design limitation, we conducted a second experiment in which items were randomly assigned to condition on a participant-by-participant basis.

Experiment 2

To address the possibility of item effects in Experiment 1, we conducted a follow-up in which novel items were randomly assigned to the disfluent or fluent condition. The design of the experiment was otherwise the same, and participants met the same inclusionary criteria as in Experiment 1.

Participants

A total of 127 adults (75 immediate condition, 52 delay condition) were included in the final sample. As in Experiment 1, our recruitment target was set to 70 participants (minimum of 50 after exclusions); some people who reported technical issues opening the task were sent new links, resulting in 75 participants in the immediate condition. All participants gave informed consent through a protocol approved by the UW-Madison Minimal Risk Research IRB (protocol ID: 2013–0984, Effects of Bilingualism on Learning and Memory). Our inclusionary criteria were the same as in Experiment 1. An additional two participants were tested in the immediate test condition but excluded for reporting high proficiency in a language other than English ($n = 1$) or age outside of the targeted range ($n = 1$). An additional 42 participants were tested in the delayed test condition but were excluded for reporting high proficiency in a language other than English ($n = 5$), reporting a history of hearing impairment ($n = 1$), failing the attention check ($n = 34$), failing the attention check and reporting high proficiency in a second language ($n = 1$), or failing the attention check and reporting a history of hearing impairment ($n = 1$). All participants were recruited via Prolific (Prolific, 2014) and tested remotely. See Table 3 for participant characteristics.

Procedure

As in Experiment 1.

Materials

Exposure trials

The same experimental protocol and stimuli developed for Experiment 1 were used in Experiment 2, but we randomized the assignment of trials to condition. Sentence frames were randomized such that any novel item could be presented in either the disfluent or fluent condition.

Table 3
Experiment 2 participant characteristics.

	Immediate test	Delayed test	*p*
Sample size (Count)	75	52	
Gender (Count)	–	–	.91
Female	41	27	
Male	32	24	
Nonbinary / Prefer to Self ID	2	1	
Age[a]	30.8 (5.44)	30.9 (5.16)	.88
Years Education[b]	14.9 (2.01)	15.4 (2.47)	.27
LexTALE Score	88.1 (10.1)	88.8 (8.81)	.67

Note: Values indicate mean (sd) unless otherwise indicated.

[a] One participant in the delayed test condition did not report age. Screener responses indicated that they were between 18 and 40 years old.

[b] One participant in the immediate test condition did not report years education. In instances where participants reported fewer than 12 years of schooling but also reported that they had finished high school, years of education was transformed as follows: high school = 12 years, some college = 14 years, college or some graduate school = 16 years (as the length of graduate school is variable), and masters = 14 years.

Test trials

As in Experiment 1, recognition for each item was tested three times. Each item was tested once before they were cycled through a second and then third time.

Results

Data analysis

Our analytical approach was the same as in Experiment 1. Based on RT criteria, 190 observations were excluded from the exposure data, across all selection points (out of 4572), and 256 observations were excluded from the recognition data (out of 3048). Full model specifications (including estimates for random effects) are available OSF.

Prediction of target at exposure

At exposure, participants' final selections (selections upon hearing the full object name) on filler familiar trials were highly accurate, $M = 0.99$, $SD = 0.09$. To predict selections on novel trials, we regressed accuracy on condition, fluent and disfluent, coded as −.5 and.5. Participants (correctly) selected the novel object across disfluent, $M = 0.89$, $SD = 0.31$, and fluent, $M = 0.83$, $SD = 0.38$, conditions, but they were more likely to select the correct object in the disfluent condition, $B = 0.73$, $SE = 0.24$, $X^2(df = 1) = 9.10$, $p = .003$.

Upon hearing the determiner, participants selected the familiar object in filler trials at above chance levels, $M = 0.76$, $SD = 0.43$; $p < .001$. To predict selections on novel trials, we regressed accuracy on condition (fluent and disfluent, coded as −.5 and.5). Upon hearing the determiner, participants were more likely to select the novel object in the disfluent condition, $M = 0.61$, $SD = 0.49$, than in the fluent condition, $M = 0.50$, $SD = 0.50$, $B = 1.03$, $SE = 0.16$, $X^2(df = 1) = 39.44$, $p < .001$.

Retention

Overall, participants learned novel words to above chance levels ($M = 0.43$, $SD = 0.49$, $p < .001$; chance = 0.25). To predict recognition task accuracy, we regressed accuracy at testing on condition (fluent vs. disfluent), delay (immediate vs. delay, coded as −.5 and.5), prediction accuracy (prediction error vs. correct prediction, coded as −.5 and.5), and all interactions. Condition, delay, prediction error and their interactions were not significant predictors of accuracy. Post-hoc analyses indicated a similar pattern of results after we removed items that participants failed to map (i.e., items where participants made an incorrect final selection at exposure; analyses included on OSF with other data and scripts).

Discussion

In Experiment 2, we randomly assigned items to fluent and disfluent conditions, whereas in Experiment 1 the items were not randomly assigned. The impact of disfluency on processing was consistent across experiments: in both experiments, participants used disfluency to anticipate the novel object. However, the impact of disfluency on retention was different. In Experiment 1, we observed better retention in the disfluent condition at immediate testing, and better retention in the fluent condition in the delayed condition, although neither effect was strong enough to yield simple effects. Here, we observed similar retention across conditions, for participants in both the immediate and delayed test conditions. In both experiments, the effect of prediction error on retention was not significant. In interpreting these results, we considered the possibility that the absence of robust disfluency effects in Experiments 1 and 2 was due to the relatively brief delay of six minutes. In Experiment 3, we ran a partial replication of Experiments 1 and 2, in which we increased the length of the testing delay from six-minutes to 24-hours.

Experiment 3

Although six-minutes should be sufficient to clear working memory based on prior work (Angwin et al., 2019; Knabe et al., 2023), it did not yield detectable differences in retention in Experiments 1 and 2. In Experiment 3, we tested the impact of disfluency on word retention over time by increasing the delay between teaching and testing to 24-hours. Critically, this delay should allow time for participants to consolidate novel words into their lexicons.

Participants

A total of 128 adults (68 immediate condition, 60 delay condition) were included in the final sample. As before, our recruitment target was set to 70 participants (minimum of 50 after exclusions). All participants gave informed consent through a protocol approved by UW-Madison Minimal Risk Research IRB (protocol ID: 2013–0984, Effects of Bilingualism on Learning and Memory). Our inclusionary criteria were the same as in previous experiments. Two additional participants in the immediate test condition and three additional participants in the delay condition were tested but excluded for reporting high proficiency in a language other than English. An additional 18 participants were tested in the delayed test condition but did not return on day two. All participants were recruited via Prolific (Prolific, 2014) and tested remotely. See Table 4 for participant characteristics.

Procedure

The procedure for the immediate test group was the same as in Experiments 1 and 2, but the procedure for the delay group was modified to include a longer delay. On day one, participants were recruited via Prolific (Prolific, 2014) and made aware that the experiment would be completed over two days. After consenting to participate, participants completed the exposure phase only. On day two, participants were tested for retention of the novel items before completing the LEAP-Q and LexTALE. Participants were recruited during a two-hour window on day one, and the task was reopened for a four-hour window on day two. The middle of the day-two window was exactly 24-hours after the middle of the day one window. On day two, participants were messaged when the experiment reopened and before it closed. Participants were allowed to participate after the day two window had closed if they messaged on the same day.

Materials

As in Experiment 2, novel items were randomly assigned to the fluent or disfluent condition.

Table 4
Experiment 3 participant characteristics.

	Immediate test	Delayed test	*p*
Sample size (Count)	68	60	
Gender (Count)	–	–	1
Female	35	31	
Male	33	29	
Age	31.3 (5.33)	31.2 (5.67)	0.911
Years Education[a]	14.7 (2.57)	15.4 (2.85)	0.139
LexTALE Score	87.9 (10.4)	87.4 (9.18)	0.796

Note: Values indicate mean (sd) unless otherwise indicated.

[a] Three participants in the immediate test condition did not report years education. In instances where participants reported fewer than 12 years of schooling but also reported that they had finished high school, years of education was transformed as follows: high school = 12 years, some college = 14 years, college or some graduate school = 16 years (as the length of graduate school is variable), and masters = 14 years.

Results

Data analysis

Our analytical approach was the same as in Experiment 1. Based on RT criteria, 212 observations were excluded from the exposure data, across all selection points (out of 4608), and 184 observations were excluded from the recognition data (out of 3072). Full model specifications (including estimates for random effects) are available on OSF.

Prediction of target at exposure

At exposure, participants' final selections (selections upon hearing the full object name) on familiar filler trials were highly accurate, $M = 0.99$, $SD = 0.11$. To predict selections on novel trials, we regressed accuracy on condition, fluent and disfluent, coded as −.5 and.5. Participants were more likely to select the novel object in the disfluent condition, $M = 0.88$, $SD = 0.32$, than in the fluent, $M = 0.83$, $SD = 0.38$, conditions, $B = 0.56$, $SE = 0.23$, $X^2(df = 1) = 5.70$, $p = .0.17$.

Upon hearing the determiner, participants selected the familiar object in filler trials at above chance levels, $M = 0.65$, $SD = 0.48$; $p < .001$. To predict selections on novel trials, we regressed accuracy on condition (fluent and disfluent, coded as −.5 and.5). Upon hearing the determiner, participants were more likely to select the novel object in the disfluent condition, $M = 0.57$, $SD = 0.50$, than in the fluent condition, $M = 0.35$, $SD = 0.48$, $B = 0.98$, $SE = 0.20$, $X^2(df = 1) = 25.18$, $p < .001$.

Retention

Overall, participants learned novel words to above chance levels ($M = 0.40$, $SD = 0.49$, $p < .001$; chance = 0.25). To predict recognition task accuracy, we regressed accuracy at testing on condition (fluent vs. disfluent), delay (immediate vs. delay, coded as −.5 and.5), prediction accuracy (prediction error vs. correct prediction, coded as −.5 and.5), and all interactions. Delay was the only significant predictor of accuracy, with participants in the 24hr delay condition performing worse than participants in the immediate test condition, $B = -0.47$, $SE = 0.14$, $X^2(df = 1) = 12.22$, $p < .001$. Post-hoc analyses indicated a similar pattern of results after we removed items that participants failed to map (i.e., items where participants made an incorrect final selection at exposure; analyses included on OSF with other data and scripts).

Discussion

In Experiment 3, we increased the delay between teaching and testing from six-minutes to 24-hours while otherwise replicating Experiment 2. Again, we found that participants used disfluency to anticipate the novel object. However, the delay—not the fluency condition and not prediction error—was the only manipulation to impact retention. Across the three experiments, the effect of disfluencies and prediction error on retention was largely null, but one common design feature across the experiments may have influenced this pattern. Specifically, the limited number of items in the fluent condition ($n = 2$) may have contributed to the null results of disfluency and prediction error on retention, since it is well known that low power increases the risk of spurious findings (Ioannidis, 2005; Wacholder et al., 2004).To address this limitation, we conducted a follow-up experiment in which both the fluent and disfluent conditions featured four items.

Experiment 4

To address the possibility the findings in Experiments 1–3 were due a small number of items in the novel-fluent condition, we conducted another partial replication. In Experiment 4, we increased the number of novel-fluent trials and randomly assigned novel items to the fluent and disfluent conditions on a participant-by-participant basis. This approach yielded a larger sample of fluent items on which to base our conclusions and minimized the possibility of item effects. The trade-off is that the within-experiment association between disfluency and novelty is weakened relative to Experiments 1–3. Consequently, any effect of disfluency would be reflective of participants day-to-day experiences rather than within-experiment experiences with disfluency.

Participants

A total of 121 adults (62 immediate condition, 58 delay condition) were included in the final sample. As before, our recruitment target was set to 70 participants (minimum of 50 after exclusions). All participants gave informed consent through a protocol approved by the UW-Madison Minimal Risk Research IRB (protocol ID: 2013–0984, Effects of Bilingualism on Learning and Memory). Our inclusionary criteria were the same as in prior experiments. One additional participant was tested in the immediate test condition but excluded for reporting high proficiency in a language other than English. An additional 33 participants were tested in the delayed test condition but were excluded for reporting high proficiency in a language other than English (n = 3) or failing the attention checks (n = 30). All participants were recruited via Prolific (Prolific, 2014) and tested remotely. See Table 5 for participant characteristics.

Results

Data analysis

Our analytical approach was the same as in Experiment 1. Based on RT criteria, 174 observations were excluded from the exposure data, across all selection points (out of 4320), and 184 observations were excluded from the recognition data (out of 2880). Full model specifications (including estimates for random effects) are available OSF.

Prediction of target at exposure

At exposure, participants' final selections (selections upon hearing the full object name) on fluent filler trials were highly accurate, M = 0.99, SD = 0.11. To predict selections on novel trials, we regressed accuracy on condition, fluent and disfluent, coded as −.5 and.5. Participants (correctly) selected the novel object at similar rates across disfluent, M = 0.88, SD = 0.32, and fluent, M = 0.86, SD = 0.35, conditions, B = 0.22, SE = 0.23, $X^2(df = 1) = 0.93$, p = .33.

Upon hearing the determiner, participants selected the familiar object in filler trials at above chance levels, M = 0.58, SD = 0.49; $p < .001$. To predict selections on novel trials, we regressed accuracy on condition (fluent and disfluent, coded as −.5 and.5). Upon hearing the determiner, participants were more likely to select the novel object in the disfluent condition, M = 0.53, SD = 0.50, than in the fluent condition, M = 0.39, SD = 0.49, B = 0.65, SE = 0.20, $X^2(df = 1) = 11.12$, $p < .001$.

Table 5
Experiment 4 participant characteristics.

	Immediate test	Delayed test	*p*
Sample size (Count)	62	58	
Gender (Count)	–	–	0.002
Female	33	19	
Male	24	39	
Nonbinary / Prefer to Self ID	5	0	
Age	28.8 (5.29)	31.0 (5.12)	0.02
Years Education[a]	14.9 (2.57)	15.1 (2.59)	0.75
LexTALE Score	87.2 (10.7)	87.1 (11.9)	0.96

Note: Values indicate mean (sd) unless otherwise indicated.

[a] In instances where participants reported fewer than 12 years of schooling but also reported that they had finished high school, years of education was transformed as follows: high school = 12 years, some college = 14 years, college or some graduate school = 16 years (as the length of graduate school is variable), and masters = 14 years.

Retention

Overall, participants learned novel words to above chance levels (M = 0.40, SD = 0.49, $p < .001$; chance = 0.25). To predict recognition task accuracy, we regressed accuracy at testing on condition (fluent vs. disfluent), delay (immediate vs. delay, coded as −.5 and.5), prediction accuracy (prediction error vs. correct prediction, coded as −.5 and.5), and all interactions. Condition predicted accuracy, with participants performing better in the fluent condition, B = -0.22, SE = 0.09, $X^2(df = 1) = 5.65$, p = .017. We also observed a significant three-way interaction among the three predictor variables – condition, delay, and prediction accuracy, B = −1.09, SE = 0.48, $X^2(df = 1) = 5.26$, p = .021. To interpret this interaction, we recentered the group and condition variables and obtained the simple effects. Participants in the immediate test group benefitted from prediction error only in the fluent condition, whereas participants in the delay test condition benefitted from prediction error only in the disfluent condition; see Fig. S2 for visualization and OSF for simple effects. Post-hoc analyses after we removed items that participants failed to map (i.e., items where participants made an incorrect final selection at exposure), also yielded a significant three-way interaction and a marginal condition effect, B = −0.20, SE = 0.11, $X^2(df = 1)$ = 3.60, p = .058; analyses included on OSF with other data and scripts. Because the immediate and delay groups differed on age and gender, we reran both models with age and gender as covariates. The model with all observations yielded a significant effect of condition but the three-way interaction became marginal; the model fitted on the filtered data set yielded a significant effect of condition and a significant three-way interaction. Because the pattern of findings was similar across analyses, we report the models without covariates, but the analyses containing covariates are available on OSF.

Discussion

In Experiment 4, we addressed another limitation of the prior experiments: the small number of items within the fluent novel condition. In Experiments 1–3, we chose to have six items in the disfluent-novel condition and two items in the fluent-novel condition based on a study by Gambi and colleagues (2021). In this way, Experiments 1–3 presented disfluencies in an ecologically valid manner, since in spontaneous speech disfluencies are more likely to precede discourse-novel information (Beattie & Butterworth, 1979; de Jong, 2016). Therefore, in Experiments 1–3, the effect of disfluency (if observed) would have been a reflection of both the statistics learned within the experiments and the cumulative experience with disfluent language. Another issue with Experiments 1–3 was the limited number of items in the fluent-novel condition, which could have made the effect of disfluency on word retention unstable. In Experiment 4, we presented participants with 4 items in both the fluent and the disfluent condition, increasing power, and at the same time, eliminating the within-experiment statistical distribution that would cue participants to the disfluency as signalling novelty. Critically, in Experiment 4, we once again replicated the effect of disfluency on processing, despite boosting the number of fluent-novel items in the task. This suggests that participants were making their predictions based on their cumulative experience with disfluency rather than the relationship between disfluency and novelty within the experiment itself. That said, the effect of disfluency on retention was starkly different from those observed in Experiments 1–3. Here, we found that participants performed better in the fluent condition, consistent with the literature on prediction error in word learning. We also observed an interaction among condition, test delay, and prediction error, such that the two-way interaction was different for the immediate and delay groups. The immediate group benefitted from prediction error in the fluent condition whereas the delay group benefitted from prediction error in disfluent condition. In interpreting this interaction, we exercise caution as the effect was weak, but it may suggest a desirable difficulty wherein conditions that are challenging during learning yield slower decay over time. To test whether this effect would remain with a longer delay, we ran a final experiment that replicated Experiment 4 while extending the delay from six-minutes to 24-hours.

Experiment 5

In Experiment 5, we tested the impact of disfluency on word learning over time by including a 24-hour delay between teaching and testing. As in Experiment 4, novel items were equally likely to occur in the fluent or disfluent learning conditions.

Participants

A total of 131 adults (68 immediate condition, 63 delay condition) were included in the final sample. As in Experiment 1, our recruitment target was set to 70 participants (minimum of 50 after exclusions). All participants gave informed consent through a protocol approved by the UW-Madison Minimal Risk Research IRB (protocol ID: 2013–0984, Effects of Bilingualism on Learning and Memory). Our inclusionary criteria were the same as in Experiment 1. An additional three participants were tested in the immediate test condition but excluded for reporting high proficiency in a language other than English. An additional 21 participants were tested in the delayed test condition but did not return on day two (n = 18) or were excluded for reporting high proficiency in a language other than English (n = 1). All participants were recruited via Prolific (Prolific, 2014) and tested remotely. See Table 6 for participant characteristics.

Table 6
Experiment 5 participant characteristics.

	Immediate test	Delayed test	*p*
Sample size (Count)	68	63	
Gender (Count)	–	–	.54
Female	32	26	
Male	32	35	
Nonbinary / Prefer to Self ID	4	2	
Age[a]	30.5 (5.80)	30.3 (5.23)	0.826
Years Education[b]	15.3 (2.99)	15.6 (2.22)	0.448
LexTALE Score	90.2 (8.48)	91.1 (7.51)	0.499

Note: Values indicate mean (sd) unless otherwise indicated.

[a] One participant in the delayed test condition did not report age. Screener responses indicated that they were between 18 and 40 years old.

[b] One participant in the immediate test condition did not report years education. In instances where participants reported fewer than 12 years of schooling but also reported that they had finished high school, years of education was transformed as follows: high school = 12 years, some college = 14 years, college or some graduate school = 16 years (as the length of graduate school is variable), and masters = 14 years.

Procedure

The procedure for the immediate test group was the same as in prior experiments, and the procedure for the delay group was the same as in Experiment 3, which also included a 24-hour delay. On day one, participants were recruited via Prolific (Prolific, 2014) and made aware that the experiment would be completed over two days. After consenting to participate, participants completed the exposure phase only. On day two, participants were tested for retention of the novel items before completing the LEAP-Q and LexTALE. Participants were recruited during a two-hour window on day one, and the task was reopened for a four-hour window on day two. The middle of the day-two window was exactly 24-hours after the middle of the day-one window. On day two, participants were messaged when the experiment reopened and before it closed. Participants were allowed to participate after the day two window had closed if they messaged on the same day.

Materials

As in Experiment 4, of the twelve target trials at exposure, four were novel-disfluent, four were novel-fluent, and four were familiar-fluent.

Results

Data analysis

Our analytical approach was the same as in Experiment 1. Based on RT criteria, 230 observations were excluded from the exposure data, across all selection points (out of 4716), and 212 observations were excluded from the recognition data (out of 3144). Full model specifications (including estimates for random effects) are available on OSF.

Prediction of target at exposure

At exposure, participants' final selections (selections upon hearing the full object name) on fluent filler trials were highly accurate, M = 0.99, SD = 0.08. To predict selections on novel trials, we regressed accuracy on condition, fluent and disfluent, coded as −.5 and.5. Participants (correctly) selected the novel object at similar rates across disfluent, M = 0.87, SD = 0.33, and fluent, M = 0.88, SD = 0.33, conditions, B = −0.08, SE = 0.23, $X^2(df = 1)$ = 0.12, p = .732.

Upon hearing the determiner, participants selected the familiar object in filler trials at above chance levels, M = 0.63, SD = 0.48, $p < .001$. To predict selections on novel trials, we regressed accuracy on condition (fluent and disfluent, coded as −.5 and.5). Random slopes were removed due to convergence and singularity issues. Upon hearing the determiner, participants were more likely to select the novel object in the disfluent condition, M = 0.57, SD = 0.50, than in the fluent condition, M = 0.36, SD = 0.48, B = 1.05, SE = 0.21, $X^2(df = 1)$ = 24.31, $p < .001$.

Retention

Overall, participants learned novel words to above chance levels (M = 0.40, SD = 0.49, $p < .001$; chance = 0.25). To predict recognition task accuracy, we regressed accuracy at testing on condition (fluent vs. disfluent), delay (immediate vs. delay, coded as −.5 and.5), prediction accuracy (prediction error vs. correct prediction, coded as −.5 and.5), and all interactions. Delay was a significant predictor of word retention, B = -0.46, SE = 0.14, $X^2(df = 1)$ = 10.55, p = .001. Participants in the immediate test condition performed better than participants in the 24-hour delay test condition. Prediction accuracy was also a significant predictor, B = -0.23, SE = 0.10, $X^2(df = 1)$ = 6.26p = .012: items where participants made a prediction error were recognized more accurately at test, independent of condition or delay. No other predictors were significant. Post-hoc analyses indicated a similar pattern of results after we removed items that participants failed to map (i.e., items where participants made an incorrect final selection at exposure; analyses included on OSF with other data and scripts).

Discussion

In our final experiment, participants received an equal number of novel-disfluent, novel-fluent, and familiar-fluent trials at exposure. We again found that participants used disfluency to anticipate upcoming novelty, but this did not impact retention. As in Experiment 3, the delay between teaching and testing impacted word retention, but the facilitative effect of fluency observed in Experiment 4 was not replicated in Experiment 5. Unlike in Experiment 4, where prediction error exerted a subtle effect on retention in tandem with condition and delay (i.e., participants in the immediate condition benefitted from prediction error on fluent items and participants in the delay condition benefitted from prediction error on disfluent items), in Experiment 5, prediction error influenced retention more broadly, with prediction errors yielding better retention independent of condition and delay.

Post hoc analysis of disfluency on word retention

Across experiments, the impact of disfluency on retention was largely null. Results are summarized in Table 7. To strengthen our conclusions and to alleviate concerns that the negligible effect of disfluency reflected low statistical power, we ran two sets of post hoc analyses recommended by reviewers. First, we pooled data across experiments (excluding Experiment 1, as items were not randomly assigned to condition) and ran the same retention models reported for individual experiments. As before, we regressed accuracy at testing on condition (fluent vs. disfluent), delay (immediate vs. delay, coded as −.5 and.5), prediction error (correct prediction vs. prediction error) and all interactions. Across 10,241 observations and 504 participants, there was no effect of condition, $B = 0.01$, $SE = 0.05$, $X^2(df = 1) = 0.008$, $p = .977$. A similar pattern of results was observed when we removed items that participants failed to map: across 8968 observations and 502 participants, there was no effect of condition ($B = -0.019$, $SE = 0.06$, $X^2(df = 1) = 0.274$, $p = .601$. Analyses are available on OSF.

We also reanalysed our retention data via Bayesian multilevel analysis. While frequentist analyses serve to "fail to reject the null hypothesis," Bayesian analyses enable us to estimate the amount of evidence for null and alternative models. Bayesian mixed effects models were used to analyse data in R (R Core Team, 2020) with the brms package (Bürkner, 2017). We derived Bayes inclusion factors for the condition effect. Inclusion factors above 1 indicate that data are more likely to occur under the alternative hypothesis (i.e., models with the predictor) and inclusion factors below 1 indicate that data are more likely under the null hypothesis. Again, we included the full dataset as well as the filtered dataset excluding items that participants failed to map at exposure. In the full dataset, Bayes inclusion factors for condition ranged from 0.009 in Experiment 2 to 0.153 in Experiment 5. In the filtered dataset, Bayes inclusion factors ranged from 0.008 (Experiment 4) to 0.966 (Experiment 1). Thus, all inclusion factors fell below 1, indicating that the data were more likely to occur under the null hypothesis (i.e., models without condition). These results reinforce the conclusion from our frequentist models, that disfluency does not impact word retention. See OSF for full statistical output ("combined_bayes_followup.rmd").

General Discussion

The impact of fillers on word learning has received relatively little attention in comparison to the impact of fillers on online processing, and this study is one of the first to investigate the effects of fillers on learning. Critically, there is a growing body of literature on the importance of predictive processing, and the role of prediction error in word learning. In the current work, we linked the occurrence of fillers to predictive processing and examined whether prediction errors stemming from a mismatch between disfluency and novelty (i.e., disfluency leads listeners to predict novelty) would impact learning. We found that adult listeners were more likely to predict upcoming novelty upon hearing a filler than upon hearing a fluent determiner, consistent with prior findings (Arnold et al., 2007; Kidd et al., 2011; Bosker et al., 2014). Moreover, this effect persisted across five experiments, and was observed even in Experiments 4 and 5, where an increase in the number of items in the fluent-novel object condition resulted in a weaker association between disfluency and novelty than in Experiments 1—3. Critically, although the impact of disfluency on processing was consistently observed across all experiments, the impact of disfluency on learning was inconsistent and largely null. Effects of prediction error were also week and unstable: prediction error was involved in a three-way interaction in Experiment 4 and yielded a benefit for word retention in Experiment 5 but otherwise had no impact on retention. Effects of delayed testing suggest that a longer delay posed more difficulty for recall than a shorter delay, in line with theories emphasizing decay (Baddeley, 1975; Barrouillet et al., 2004, 2007).

Processing

During the processing task, we found that adults were more likely to select the novel object following a disfluent "thee uh/um" than after a fluent "the." However, even at the final selection point, after participants had heard the novel word, selection of the novel object was not at

Table 7
Results summary for all experiments (for effects of interest).

	Experiment				
	1	2	3	4	5
Performance at Exposure	–	–	–	–	–
Final item selection	Null	More likely to select novel item in disfluent vs. fluent condition	More likely to select novel item in disfluent vs. fluent condition	Null	Null
Item selection after filler	More likely to select novel item in disfluent vs. fluent condition	As in Exp. 1	As in Exp. 1	As in Exp. 1	As in Exp. 1
Performance at Retention	–	–	–	–	–
All items	Condition by delay interaction (refer to text for explanation)	Null	Better performance by participants in immediate vs. delay condition	Better performance in fluent vs. disfluent condition; Three-way interaction (refer to text for interpretation)	Better performance by participants in immediate vs. delay condition; Benefit of prediction error
Items mapped correctly at exposure	Condition by delay interaction (refer to text for explanation)	Null	Better performance by participants in immediate vs. delay condition	Marginally better performance in fluent vs. disfluent condition; Three-way interaction (refer to text for interpretation)	Better performance by participants in immediate vs. delay condition; Benefit of prediction error

ceiling. This seems to reflect a bias to choose the familiar object, as participants were more likely to choose the familiar object even at the first selection point (when all they had heard was “click on”). Thus, we conclude that the presence of a filler makes adults significantly more likely to predict a novel referent, although a familiar object bias persists. Our findings—in tandem with recent web-based disfluency research by Karimi and colleagues (2019) and Diachek and Brown-Schmidt (2022)—demonstrate that effects of disfluency are observable even under remote testing conditions that are relatively less sensitive than the in-lab eye-tracking and EEG methodologies used in earlier studies. Because the sentences are segmented, the task is somewhat less naturalistic than tasks that present sentences in full. Here, participants are forced to make a selection at two points in the sentence, which may not reflect the type of prediction making that adults do when listening to a sentence without interruptions. That said, because the self-paced nature of our task limits the typically present disfluent prosody that accompanies fillers (Soderstrom & Morgan, 2007) and the length of the segment with the determiner was identical across conditions (i.e., participants were not simply responding to a delay), the present study also strongly points to filler *form* as the nexus of this effect—although further research will be necessary to confirm this hypothesis.

Learning

Although the findings regarding processing were consistent across experiments, the impact of disfluency and prediction error on word retention was weak or null. In Experiment 1, we found that participants in the immediate and delay groups were impacted by fluency differently, with participants tested immediately demonstrating better performance on fluent items and participants tested after a delay showing better performance on disfluent items. In Experiment 4, participants performed better in the fluent condition independent of which group they were in; but we also observed a complex three-way interaction, where the impact of fluency and prediction error differed across immediate (benefit of prediction error on fluent items) and delayed (benefit of prediction error on disfluent items) test conditions. In Experiment 5, we found that learners benefitted from prediction error independent of fluency condition or group. Given the inconsistency of these findings and the null effect of disfluency on word retention in our post-hoc analyses, we are inclined to view these results as a failure to replicate rather than meaningful differences that can be interpreted.

Why might effects of disfluency on learning fail to replicate while effects on processing are so consistent? For one, processing effects seem to be much more robust. Numerous studies have found that adults and children use disfluencies to anticipate upcoming referents, and this has held across a variety of paradigms (e.g., visual world eye tracking tasks: Arnold et al., 2007, EEG tasks: Corley et al., 2007, remote behavioural tasks: Karimi et al., 2019; Diachek and Brown-Schmidt, 2022). In contrast, the impact of disfluency on learning and memory remains poorly understood and inconsistent. Some studies have found a disfluency advantage (Corley et al., 2007; Diachek & Brown-Schmidt, 2022; Fraundorf & Watson, 2011), others have found a fluency advantage (Libersky et al., 2023), and others have found no difference (Bosker et al., 2014; Toftness et al., 2018). Notably, Libersky and colleagues (2023) found a fluency advantage for bilinguals only, whereas monolinguals showed similar performance across conditions. This suggests that the impact of disfluency may depend on prior language experience (i.e., learners with sufficient exposure to the target language learn similarly well from disfluent and fluent input, although what constitutes sufficient exposure remains to be explored). Additionally, there are competing processes at play when learning from disfluent speakers. Although disfluency may orient learners to upcoming novelty, disfluent speakers are considered less credible (Toftness et al., 2018). Future work might unpack how much speaker knowledge is necessary for learners to view a disfluent speaker as a qualified instructor and enable learners to effectively use disfluency as a cue to novelty; if the impact of disfluency on perception of speakers is alleviated, processing effects might cascade to learning. Another point is that testing word retention via recognition may not be the most sensitive approach to understanding the impact of fluency on learning; future work could assess free recall of novel words or competition with known words (i.e., consolidation) in order to build a fuller picture of disfluent word learning.

Furthermore, learning studies are necessarily less well powered than processing studies. In one-shot learning tasks like ours, participants can only be taught so many items before retention dips below chance. Compounding this, assuming any effect of disfluency on learning has a smaller effect size than the effect of disfluency on processing, the number of necessary items to observe an effect is actually *higher* than the items needed to observe an effect on processing. In practice, studies of disfluency on processing have had a greater number of items (Arnold et al., 2007; Morin-Lessard & Byers-Heinlein, 2019) than studies of disfluency on learning (Libersky et al., 2023; White et al., 2020). Because underpowered studies may yield spurious findings (Ioannidis, 2005; Wacholder et al., 2004), it is not entirely surprising that the effect of disfluency on learning does not replicate, either across studies or within a single set of studies like ours. Westfall and colleagues (2014) provide some guidance on power in situations where the number of items is limited by what participants can feasibly respond to (i.e., more participants are required), but the number of participants required for studies with a limited number of items is not feasible in many cases. For example, Brysbaert and Stevens (2018; within the context of reaction time experiments) recommend 1,600 observations per condition through some combination of participants and items (e.g., 40 participants each respond to 40 items per condition). Applying this guideline to a learning task in which participants can only feasibly learn ten words per condition, 160 participants would be required *per condition.* That said, post hoc frequentist analyses pooling data across Experiments 2–5 (yielding 10,241 observations in the full dataset and 8968 observations in the dataset filtered of incorrect mappings) indicated a null effect of condition (fluent vs. disfluent) on retention, as did post hoc Bayesian analyses of each experiment. However, the omnibus analyses only partially address concerns about power. Although the number of participants is increased, the number of stimuli remains a limiting factor. Per Westfall and colleagues (2014), a counterbalanced design with 60 participants and 8 stimuli has 0.80 power to detect a minimum effect size of 0.79; increasing the number of participants to 500 while retaining the same number of items only lowers the minimum effect size to 0.76. This is a common problem in word learning research, as presenting too many items would result in below-chance retention. Therefore, although post-hoc findings substantiate our original results, they were unplanned and conducted at the recommendation of reviewers and do not address limits on the number of items presented.

Previously, researchers have investigated the role of disfluency on memory for gist information or events (e.g., what happened in a story, Fraundorf & Watson, 2011; which familiar words appeared in a list, Corley et al., 2007; what was taught in a lecture, Toftness et al., 2018) and have typically shown either a disfluency advantage or no effect. This contrasts with the direction of the effect observed here, perhaps due to the role of speaker credibility in learning or the nature of novel word learning. Prior work has indicated that listeners associate disfluencies with lack of topic knowledge (Brennan & Williams, 1995), and people prefer to learn from fluent instructors (Toftness et al., 2018; White et al., 2020). Consequently, learners may deprioritize items that are preceded by disfluency (with the caveat that work on disfluency and credibility has used between- rather than within-speaker manipulations). Moreover, word learning is a unique memory task in that it involves encoding a unique, precise phonological sequence and mapping it to a referent. Prior work has indicated that memory for gist (e.g., what happened in a story) is better retained than memory for items (Zeng et al., 2021); in the case of item memory, familiar items (i.e., homophones, Storkel & Maekawa, 2005, or words made up of familiar sounds or characters, Reder et al., 2016) are learned better than unfamiliar items. Given the relative difficulty of novel word learning (vs. prior tasks used in disfluency research), it is possible that adults treat disfluency differently during novel word learning vs. other processing and memory tasks, leading to a different pattern of results.

Regarding retention over time, performance in the 24-hour delay was worse than performance measured immediately. Reduced performance after a delay over 12 h has been shown in many word learning studies (e.g., Kaushanskaya, 2012; Schneider et al., 2002). However, brief delays have also been shown to impact word retention (Angwin et al., 2019; Knabe et al., 2023) as they should capture the initial transfer of items from short- to long-term memory (Tehan & Tolan, 2007). Yet, in our study, we did not observe reduced performance after the six-minute delay. Instead, we observed an interaction between condition and delay in one experiment (Experiment 1) and only after filtering out the items that participants failed to map at exposure. The 24-hour delay is qualitatively different from the six-minute delay in that it involves sleep, and new words are consolidated into the lexicon during sleep (Davis et al., 2009; Gaskell & Dumay, 2003). Although we did not specifically query if and how long participants slept in the 24-hour window between sessions, and this would need to be addressed in any follow up studies. Future work could explore the consequences of disfluency for consolidation of novel words and integration into the lexicon via longer delays (e.g., one-week) between exposure and testing (see Gaskell & Dumay, 2003), although our preliminary findings suggest a minimal impact of disfluency on long-term retention.

Conclusion

Although disfluency robustly shapes listeners' predictions of word novelty, little prior work has tested the effect of fillers on word learning. In the current study, we first confirmed that participants were sensitive to disfluency during processing and then tested the cascading consequences of disfluency and prediction error for learning. We also included a testing delay manipulation, investigating the possibility that fillers may impact word learning differently over time. Across five experiments, we consistently observed that adult listeners use disfluencies to anticipate upcoming novelty, but the impact of disfluency and prediction error on word retention was largely null.

Theoretically, it will be important to further investigate the mechanisms underlying disfluency effects in both online processing and learning. Analysing the impact of disfluency across different tasks would be a fruitful next step in understanding how and when disfluencies affect task performance. Manipulating listener characteristics (e.g., age and language proficiency) and speaker characteristics (e.g., reliability and credibility), may further advance our understanding of how and when disfluency affects listeners. Such investigations will clarify the processes underlying effects of disfluency on language processing as well as identify possible circumstances under which disfluencies have downstream effects on learning.

CRediT authorship contribution statement

Emma Libersky: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Margarita Kaushanskaya:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by a National Institutes of Health (NIH) grants (R01 DC016015 and R01 DC020447) to M. K., an Oros Bascom Professorship to M. K., an NIH grant (F31 DC021386) to E.L., an Emma Allen Award from the University of Wisconsin-Madison to E.L., a Gorilla Grant from Cauldron Science to E. L., and a center grant (U54 HD090256) to the Waisman Center. We thank our participants and member of the Language Acquisition and Bilingualism Lab for making this research possible. We also thank Audra Sterling and Haley Vlach for providing literature suggestions and giving feedback on the manuscript, and Daniel Bolt for statistical consulting.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2025.104704.

Data availability

All data and scripts associated this paper are available on the Open Science Framework (OSF) at https://osf.io/yj7pa/.

References

Angwin, A. J., Wilson, W. J., Ripollés, P., Rodriguez-Fornells, A., Arnott, W. L., Barry, R. J., Cheng, B. B. Y., Garden, K., & Copland, D. A. (2019). White noise facilitates new-word learning from context. *Brain and Language, 199*(September), Article 104699. https://doi.org/10.1016/j.bandl.2019.104699

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2020). Gorilla in our Midst: An online behavioral experiment builder. *Behavior Research Methods, 52*, 388–407. https://doi.org/10.3758/s13428-019-01237-x

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research, 32*(1), 25–36. https://doi.org/10.1023/A:1021980931292

Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning Memory and Cognition, 33*(5), 914–930. https://doi.org/10.1037/0278-7393.33.5.914

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new disfluency and reference resolution. *Psychological Science, 15*(9), 578–582. https://doi.org/10.1111/j.0956-7976.2004.00723.x

Audacity-Team. (2018). *Audacity(R): Free Audio Editor and Recorder* (2.3.3).

Baddeley, A. D. (1975). Word Length and the Structure of Short-Term Memory. *Journal of Verbal Learning and Verbal Behavior, 14*, 575–589.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language.* https://doi.org/10.1016/j.jml.2012.11.001

Barr, D. J., & Seyfeddinipur, M. (2010). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes, 25*(4), 441–456. https://doi.org/10.1080/01690960903047122

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time Constraints and Resource Sharing in Adults' Working Memory Spans. *Journal of Experimental Psychology: General, 133*(1), 83–100. https://doi.org/10.1037/0096-3445.133.1.83

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 570–585. https://doi.org/10.1037/0278-7393.33.3.570

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

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., Herron, D., Lu, C. C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, K., Gerdjikova, I., Gutierrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., & Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin and Review, 10*(2), 344–380. https://doi.org/10.3758/BF03196494

Beattie, G. W., & Butterworth, B. L. (1979). Determinants of pauses and errors in spontaneous speech. *Language and Speech, 22*(3), 201–211.

Boersma, P., & Weenink, D. (2020). *Praat: Doing phonetics by computer* (6.1.16). http://www.praat.org.

Borovsky, A., & Creel, S. (2014). Children and adults integrate talker and verb information in online processing. *Developmental Psychology*, *50*(5), 1600–1613. doi: doi:10.1037/a0035591.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech, 44*(2), 123–147.

Bosker, H. R., Quené, H., Sanders, T., & de Jong, N. H. (2014). Native 'um's elicit prediction of low-frequency referents, but non-native 'um's do not. *Journal of Memory and Language, 75*, 104–116. https://doi.org/10.1016/j.jml.2014.05.004

Bosker, H. R., van Os, M., Does, R., & van Bergen, G. (2019). Counting 'uhm's: How tracking the distribution of native and non-native disfluencies influences online language comprehension. *Journal of Memory and Language, 106*, 189–202. https://doi.org/10.1016/j.jml.2019.02.006

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS ONE, 5*(5), Article e10773. https://doi.org/10.1371/journal.pone.0010773

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition, 1*(1), 9. https://doi.org/10.5334/joc.10

Bürkner, P.-C. (2017). **brms**: An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal of Statistical Software, 80*(1). https://doi.org/10.18637/jss.v080.i01

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition, 84*(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Collard, P., Corley, M., MacGregor, L. J., & Donaldson, D. I. (2008). Attention orienting effects of hesitations in speech: Evidence from ERPs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(3), 696–702. https://doi.org/10.1037/0278-7393.34.3.696

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition, 105*(3), 658–668. https://doi.org/10.1016/j.cognition.2006.10.010

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Linguistics and Language Compass, 2*(4), 589–602. https://doi.org/10.1111/j.1749-818X.2008.00068.x

Davis, M. H., Di Betta, Macdonald, M. J. E., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience, 21*(4), 803–820. https://doi.org/10.1162/jocn.2009.21059

de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *IRAL - International Review of Applied Linguistics in Language Teaching, 54*(2), 113–132. https://doi.org/10.1515/iral-2016-9993

Diachek, E., & Brown-Schmidt, S. (2022). The effect of disfluency on memory for what was said. *Journal of Experimental Psychology: Learning Memory and Cognition.*. https://doi.org/10.1037/xlm0001156

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2017). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition, 20*(5), 917–930. doi: doi:10.1017/S1366728916000547.

Fernald, A., & Marchman, V. A. (2012). Individual Differences in Lexical Processing at 18 Months Predict Vocabulary Growth in Typically Developing and Late-Talking Toddlers. *Child Development, 83*(1), 203–222. https://doi.org/10.1111/j.1467-8624.2011.01692.x

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology, 111*, 15–52. https://doi.org/10.1016/j.cogpsych.2019.03.002

Fraundorf, S. H., & Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language, 65*(2), 161–175. https://doi.org/10.1016/j.jml.2011.03.004

Gambi, C., Pickering, M. J., & Rabagliati, H. (2021). Prediction error boosts retention of novel words in adults but not in children. *Cognition, 211*(February), Article 104650. https://doi.org/10.1016/j.cognition.2021.104650

Gaskell, M. G., & Dumay, N. (2003). *Lexical competition and the acquisition of novel words., 89*, 105–132. https://doi.org/10.1016/S0010-0277(03)00070-2

Gupta, P., Lipinski, J., Abbs, B., Lin, P., Aktunc, E., Ludden, D., Martin, N., & Newman, R. (2004). Space aliens and nonwords: Stimuli for investigating the learning of novel word-meaning pairs. *Behavior Research Methods, Instruments, & Computers, 36*(4), 599–603. https://doi.org/10.3758/BF03206540

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods, 48*(4), 1393–1409. https://doi.org/10.3758/s13428-015-0647-3

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine, 2*(8), e124.

James, E. (2019). *Pre-experiment sound check*. Gorilla Experiment Builder.

Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research, 32*(1), 37–55. https://doi.org/10.1023/A:1021933015362

Karimi, H., Brothers, T., & Ferreira, F. (2019). Phonological versus semantic prediction in focus and repair constructions: No evidence for differential predictions. *Cognitive Psychology, 112*(May), 25–47. https://doi.org/10.1016/j.cogpsych.2019.04.001

Kaushanskaya, M. (2012). Cognitive mechanisms of word learning in bilingual and monolingual adults: The role of phonological memory. *Bilingualism: Language and Cognition, 15*(3), 470–489. https://doi.org/10.1017/S1366728911000472

Kidd, C., White, K. S., & Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science, 14*(4), 925–934. https://doi.org/10.1111/j.1467-7687.2011.01049.x

Knabe, M. L., Schonberg, C. C., & Vlach, H. A. (2023). When time shifts the boundaries: Isolating the role of forgetting in children's changing category representations. *Journal of Memory and Language, 132*(August 2022), 104447. doi: 10.1016/j.jml.2023.104447.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods, 44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Libersky, E., Neveu, A., & Kaushanskaya, M. (2023). One fish, uh, two fish: Effects of fluency and bilingualism on adults' novel word learning. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-022-02189-8

Lowder, M. W., Maxfield, N. D., & Ferreira, F. (2019). Processing of self-repairs in stuttered and non-stuttered speech. *Language, Cognition and Neuroscience, 35*(1), 93–105. https://doi.org/10.1080/23273798.2019.1628284

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and Multilinguals. Journal of Speech Language and Hearing Research, 50(4), 940. https://doi. *Journal of Speech Language and Hearing Research, 50*(4), 940. http://jslhr.pubs.asha.org/article.aspx?doi=10.1044/1092-4388(2007/067).

Metcalfe, J. (2017). Learning from Errors. *Annual Review of Psychology, 68*(1), 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Miler, Z. (1969). *Mole as a Gardener*.

Morin-Lessard, E., & Byers-Heinlein, K. (2019). Uh and euh signal novelty for monolinguals and bilinguals: Evidence from children and adults. *Journal of Child Language, 46*(3), 522–545. https://doi.org/10.1017/S0305000918000612

Orena, A. J., & White, K. S. (2015). I Forget what that's called! Children's online processing of disfluencies depends on speaker knowledge. *Child Development, 86*(6), 1701–1709. https://doi.org/10.1111/cdev.12421

Owens, S. J., & Graham, S. A. (2016). Thee, uhh disfluency effect in preschoolers: A cue to discourse status. *British Journal of Developmental Psychology, 34*, 388–401. https://doi.org/10.1111/bjdp.12137

Poort, E. D., & Rodd, J. M. (2019). Towards a distributed connectionist account of cognates and interlingual homographs: Evidence from semantic relatedness tasks. *PeerJ, 7*, e6725.

Poort, E. D., & Rodd, J. M. (2017). The cognate facilitation effect in bilingual lexical decision is influenced by stimulus list composition. *Acta Psychologica, 180* (September), 52–63. https://doi.org/10.1016/j.actpsy.2017.08.008

Prolific (2014). *Prolific* (December 2021). Prolific. www.prolific.co.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.r-project.org/.

Reder, L. M., Liu, X. L., Keinath, A., & Popov, V. (2016). Building knowledge requires bricks, not sand: The critical role of familiar constituents in learning. *Psychonomic Bulletin and Review, 23*(1), 271–277. https://doi.org/10.3758/s13423-015-0889-1

Reuter, T., Borovsky, A., & Lew-Williams, C. (2019). Predict and redirect: Prediction errors support children's word learning. *Developmental Psychology, 55*(8), 1656–1665. https://doi.org/10.1037/dev0000754

Schmidt, R. A., & Bjork, R. A. (1992). New Conceptualizations of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training. *Psychological Science, 3*(4), 207–218. https://doi.org/10.1111/j.1467-9280.1992.tb00029.x

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What Is Learned under Difficult Conditions Is Hard to Forget: Contextual Interference Effects in Foreign Vocabulary Acquisition, Retention, and Transfer. *Journal of Memory and Language, 46*(2), 419–440. https://doi.org/10.1006/jmla.2001.2813

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., … Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences of the United States of America, 118*(45). https://doi.org/10.1073/pnas.2105646118

Soderstrom, M., & Morgan, J. L. (2007). Twenty-two-month-olds discriminate fluent from disfluent adult-directed speech. *Developmental Science, 10*(5), 641–653. https://doi.org/10.1111/j.1467-7687.2006.00605.x

Storkel, H. L., & Maekawa, J. (2005). A comparison of homonym and novel word learning: The role of phonotactic probability and word frequency. *Journal of Child Language, 32*(4), 827–853. https://doi.org/10.1017/S0305000905007099

Tehan, G., & Tolan, G. A. (2007). Word length effects in long-term memory. *Journal of Memory and Language, 56*(1), 35–48. https://doi.org/10.1016/j.jml.2006.08.015

Thacker, J. M., Chambers, C. G., & Graham, S. A. (2018). When it is apt to adapt: Flexible reasoning guides children's use of talker identity and disfluency cues. *Journal of Experimental Child Psychology, 167*, 314–327. https://doi.org/10.1016/j.jecp.2017.11.008

Toftness, A. R., Carpenter, S. K., Geller, J., Lauber, S., Johnson, M., & Armstrong, P. I. (2018). Instructor fluency leads to higher confidence in learning, but not better learning. *Metacognition and Learning, 13*(1), 1–14. https://doi.org/10.1007/s11409-017-9175-0

Van Berkum, Van Den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience, 20*(4), 580–591. https://doi.org/10.1162/jocn.2008.20054

van Casteren, M., & Davis, M. H. (2007). Match: A program to assist in matching the conditions of factorial experiments. *Behavior Research Methods, 39*(4), 973–978. https://doi.org/10.3758/BF03192992

Voeten, C. C. (2023). *buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression (Version 2.11)*. [Computer software].

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., & Rothman, N. (2004). Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies. *JNCI: Journal of the National Cancer Institute, 96*(6), 434–442. https://doi.org/10.1093/jnci/djh075

Watanabe, M., Hirose, K., Den, Y., & Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication, 50*(2), 81–94. https://doi.org/10.1016/j.specom.2007.06.002

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal* *of Experimental Psychology: General, 143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

White, K. S., Nilsen, E. S., Deglint, T., & Silva, J. (2020). That's thee, uuh blicket! How does disfluency affect children's word learning? *First Language, 40*(1), 3–20. https://doi.org/10.1177/0142723719873499

Yano, M. (2018). Predictive processing of syntactic information: Evidence from event-related brain potentials. *Language, Cognition and Neuroscience, 33*(8), 1017–1031. https://doi.org/10.1080/23273798.2018.1444185

Zeng, T., Tompary, A., Schapiro, A. C., & Thompson-Schill, S. L. (2021). Tracking the relation between gist and item memory over the course of long-term memory consolidation. *eLife, 10*, Article e65588. https://doi.org/10.7554/eLife.65588