Phonological versus semantic prediction in focus and repair constructions: No evidence for differential predictions

Article *in* Cognitive Psychology · May 2019

DOI: 10.1016/j.cogpsych.2019.04.001

Hossein Karimi
University of California, Davis

Trevor Brothers
University of California, Davis

Cognitive Psychology 112 (2019) 25–47

Contents lists available at ScienceDirect

Cognitive Psychology

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

Phonological versus semantic prediction in focus and repair constructions: No evidence for differential predictions

Hossein Karimi[a,*], Trevor Brothers[b], Fernanda Ferreira[c]

[a] Pennsylvania State University, United States
[b] Tufts University, United States
[c] University of California, Davis, United States

ARTICLE INFO

Keywords:
Language comprehension
Differential prediction
Disfluency
Spreading activation

ABSTRACT

Evidence suggests that the language processing system is predictive. Although past research has established prediction as a general tendency, it is not yet clear whether comprehenders can modulate their anticipatory strategies in response to cues based on sentence constructions. In two visual world eye-tracking experiments, we investigated whether focus constructions (*not the hammer but rather the …*) and repair disfluencies (*the hammer uh I mean the …*) would lead listeners to generate different patterns of predictions. In three offline tasks, we observed that participants preferred semantically related continuations (hammer – nail) following focus constructions and phonologically related continuations (*hammer – hammock*) following disfluencies. However, these offline preferences were not evident in participants' predictive eye-movements during online language processing: Semantically related (*nail*) and phonologically related words (*hammock*) received additional predictive looks regardless of whether the target word appeared in a disfluency or in a focus construction. However, significantly less semantic and phonological activation was observed in two "control" linguistic contexts in which predictive processing was discouraged. These findings suggest that although the prediction system is sensitive to sentence construction, is it not flexible enough to alter the type of prediction generated based on preceding context.

1. Introduction

A primary goal of language research is to understand how listeners process and extract meaning from the ongoing speech signal. One clear challenge for listeners is that everyday speech is filled with auditory noise and overt speech errors that can impair online processing. To deal with these errors, listeners must use a variety of cues (semantic, syntactic, prosodic) to fill in these perceptual gaps and infer a speaker's intended meaning. Predictive coding models argue that *anticipatory processing* provides a critical mechanism to achieve this goal, with higher-level conceptual predictions constantly being transmitted to lower sensory levels. These predictions can then provide a scaffolding on which to compare incoming sensory information, allowing top-down conceptual constraints to influence early stages of information processing.

One important method for investigating prediction during online language processing has been the *visual world paradigm* (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In this task, participants' eye movements are recorded while they listen to

* Corresponding author at: Department of Psychology, Pennsylvania State University, 109 Moore Building, University Park, State College, PA 16801, United States.
E-mail address: karimi@psu.edu (H. Karimi).

https://doi.org/10.1016/j.cogpsych.2019.04.001
Accepted 16 April 2019

a spoken sentence and simultaneously view related images on a visual display. Several decades of work using this method have demonstrated a tight coupling between linguistic processing and visual attention, with participants rapidly directing their gaze towards named objects in a scene (see Cooper, 1974; Henderson & Ferreira, 2004; Huettig, Rommers, & Meyer, 2011, for review). Critically, objects that are predictable in context have been shown to receive anticipatory fixations – looks to the target object before it is named, which presumably reflect an increase in activation of memory representations associated with these items (e.g., Altmann & Kamide, 1999; Huettig & McQueen, 2007; Lowder & Ferreira, 2016a; Rommers, Meyer, Praamstra, & Huettig, 2013). For example, while hearing the sentence *The boy will eat…*, participants will direct fixations to edible objects in the scene even before the onset of the critical word *cake* (Altmann & Kamide, 1999).

While it is clear that readers and listeners engage in some degree of anticipatory processing (e.g., Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Frisson, Rayner, & Pickering, 2005; Huettig, 2015; Kuperberg & Jaeger, 2016; Kutas, DeLong, & Smith, 2011; see also, Clark, 2013), two important questions about prediction during language processing have yet to be answered. The first question is whether the prediction machinery is flexible enough to predict differentially depending on different contextual cues. For instance, could the prediction system generate predictions based on semantic but not phonological features of preceding words under some conditions, and predictions based on phonological but not semantic features of preceding context under other conditions? The second question concerns the relative strength of semantic versus phonological predictions during online processing. While it is fairly well established that prediction occurs at the level of sematic representations (e.g., Brothers, Swaab, & Traxler, 2015; Federmeier & Kutas, 1999; Ito, Corley, Pickering, Martin, & Nieuwland, 2016; Kutas & Hillyard, 1984), evidence for phonological prediction is rather weak (DeLong, Urbach, & Kutas, 2005; Nieuwland et al., 2017; see also Ito, Martin & Nieuwland, 2017; and, DeLong, Urbach, & Kutas, 2017). Thus, it is important to examine the magnitude of semantic and phonological predictions, particularly in the same sentential contexts.

According to some predictive coding accounts, a major source of bottom-up processing difficulty during language comprehension is encountering material that is inconsistent with prior predictions. These “prediction errors” are costly for the comprehender, both metabolically and behaviorally (Kuperberg & Jaeger, 2016), but they also serve the important function of updating the comprehender’s pre-existing priors so that future predictions can be more accurate (see Chang, Dell, & Bock, 2006; Clark, 2013). There is ample evidence that the brain tracks and makes use of statistical probabilities to generate predictions during online language processing (see Levy, 2008a for a review). During reading comprehension, words that are predictable in context (Ehrlich & Rayner, 1981) or that have higher transitional probabilities (McDonald & Shillcock, 2003) are read more quickly. Moreover, comprehenders are able to quickly adapt to novel syntactic structures (Kaschak & Glenberg, 2004) or speech patterns (Kleinschmidt & Jaeger, 2015), and to apply these newly acquired statistics within specific environments (Kaschak & Glenberg, 2004; Kleinschmidt, Fine, & Jaeger, 2012).

In one recent theory of predictive processing, multiple mechanisms are assumed to guide how comprehenders generate online linguistic predictions. According to Huettig (2015), Type I prediction mechanisms rely on automatic activation of associated information. Spreading activation can include related “nearby” features in semantic space, or even associated syntactic or phonological information (also see Kukona, Fang, Aicher, Chen, & Magnuson, 2011). In contrast, Type II mechanisms are thought to reflect a slower, deliberative mechanism that may operate at a conscious level (Ferreira & Chantavarin, 2018). This “smarter” Type II system has been linked to flexible anticipatory mechanisms, allowing context-specific information and non-automatic associations to guide predictions. For example, in a classic semantic priming study, Neely (1977) observed fast, automatic priming for associated word pairs (body – ARM, building – SCHOOL), and delayed “strategic” priming effects for arbitrary associations that participants learned during the experiment (fruit – ROBIN).

While there is ample evidence for automatic Type I priming mechanisms during natural sentence comprehension, the evidence for flexible, Type II mechanisms is less robust. One piece of evidence for this mechanism is the finding that the global validity of predictive cues can influence the magnitude of contextual priming effects during reading (Brothers, Swaab, & Traxler, 2017). Brothers et al. observed that when lexical predictions were regularly disconfirmed (*The volleyball shot barely made it over the car*) participants no longer showed reading time benefits for contextually predictable words. While this finding suggests flexibility in the overall *probability* of predictive processing, it is unclear whether the specific *content* of a prediction (e.g. semantic vs. phonological) can also shift flexibly across contexts. We investigate this question in the present study by comparing predictive processing in disfluent speech and in focus constructions. In addition, our experiments assess the relative strength of semantic and phonological predictions. While prediction at both the semantic and phonological levels has been reported (*semantic*: e.g., Federmeier & Kutas, 1999; Kutas & Hillyard, 1984, *phonological*: e.g., DeLong et al., 2005; DeLong et al., 2017), the relative strengths of the two types of predictions in the same context is not yet understood. One possibility is that semantic anticipation will always win out over phonological prediction. This possibility is supported by the fact that evidence in favor of prediction at the semantic level is fairly strong, but evidence for pre-activation of form-based features of the expected input has thus far been weak and mixed (see Nieuwland et al., 2017). Another possibility is that no matter what type of features are pre-activated (semantic or phonological), those with the highest levels of activation will lead to prediction at their respective representational level. This idea implies that if phonological information is activated to a sufficient degree, it will win against semantic prediction.

1.1. The present study

Prediction appears to be particularly important for processing disfluent speech. Natural speech contains a variety of hesitations and revisions, with approximately 6–10% of utterances containing some kind of disfluency (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Tree, 1995). One common type of disfluency is known as a repair, illustrated in *Can you please pass the salt, uh I mean,* *the pepper*. Repair disfluencies involve an unintentionally spoken word (*salt*), which we will refer to as the "reparandum". This reparandum is sometimes followed by a signal that an error has occurred (*uh I mean*) and then the intended word (*pepper*), which we will call the "repair".

Recent studies using the visual world paradigm have suggested that disfluent speech may trigger enhanced anticipatory processing. For instance, Lowder and Ferreira (2016a) showed that when listeners encounter a reparandum such as *salt* in C*an you pass the salt, uh I mean…*, they do not passively wait for the repair to occur - instead, they actively anticipate plausible continuations using information from the prior discourse, the reparandum, and the surrounding visual scene (see also Lowder & Ferreira, 2016b). In combination with the visual world methodology, repair disfluencies offer a novel tool for investigating predictions during natural speech processing, particularly since "auto-correct" or noisy channel mechanisms have independently been argued to operate during normal language comprehension, allowing listeners to deal efficiently with speaker error (Brill & Moore, 2000; Gibson et al., 2013; Gibson, Bergen, & Piantadosi, 2013; Levy, 2008b).

An important feature of speech errors is that they may give rise to different types of lexical predictions. In some cases, a speaker might make a *semantic substitution* error in which one word is accidentally replaced with one that is semantically related (*the dog, uh I mean the cat*). In disfluencies of this type, successful prediction of the upcoming repair could be generated based on semantic overlap or contextual plausibility, similar to predictions generated in fluent sentence contexts (*add some salt and…; the web was spun by the…)*. Speakers might also make *phonological* substitution errors in which the target lexical form is replaced with a word that is phonologically related (e.g., *he was resting in the hammer, I mean the hammock…*). In cases of phonological speech errors, helpful anticipation of the repair would be based on prediction of a phonologically similar, not semantically similar, form.

Focus constructions contrast in important and potentially diagnostic ways. Consider a sentence such as *The man was looking around for not a hammer but rather a yardstick*. In these forms, the first noun phrase (henceforth, NP) is *semantically* contrasted with the second in order to establish emphasis (e.g., *not a hammer but rather a yardstick*). Some evidence suggests that interpretation of the NP and the focusing elements (in this example, the combination of *not* and *but rather*) is based on activation of an alternative set of semantically-related words generated in part based on the comprehender's discourse model (Husband & Ferreira, 2016), in anticipation of the second NP. Importantly, notice that phonological contrast seems distinctly odd in focus constructions (e.g., *not the hammer but rather the hammock*) and, as we will show, is dispreferred. Comprehenders, then, would seem to be less likely to predict phonologically related items in focus constructions than in repairs, allowing us to test our hypothesis that the language processor's sensitivity to different sentence forms allows it to make appropriately tailored predictions—predictions that are both semantic and phonological for repairs, but semantic only for focus forms.

In the present experiments, our goal was to investigate semantic and phonological anticipation during online processing of repair disfluencies and focus constructions. Specifically, we pursued two main goals: First, we wished to examine whether comprehenders shift the content of their lexical predictions (semantic or phonological) based on surrounding linguistic contextual cues. That is, to the extent that disfluencies and focus constructions are biased in favor of different types of substitutions, will listeners generate differential semantic versus phonological predictions? Second, we sought to assess the magnitude of semantic versus phonological prediction in the same linguistic contexts.

To investigate these questions, we conducted two eye-tracking experiments using the visual world paradigm. While viewing objects in a visual display, participants heard sentences such as (1), in which an initial NP was either edited to arrive at a second NP through a disfluency repair (1a), or was negated to for contrastive emphasis (1b).

(1)
(a) **Disfluency:** The man was looking around for a hammer <u>uh I mean a</u> yardstick.
(b) **Focus:** The man was looking around for not a hammer <u>but rather a</u> yardstick.

Each display contained an object corresponding to the first-mentioned NP (NP1, *hammer*) and another object corresponding to a second-mentioned, intended NP (NP2, *yardstick*). Critically, the display also included an image that was related to NP1 either phonologically (*hammock*) or semantically (*nail*). By measuring anticipatory eye-movements prior to the onset of the second NP (during the underlined sections in 1), we can assess the relative preference for semantic versus phonological predictions across the two sentence structures.

In addition to our main eye-tracking experiments, we conducted a judgment study (Experiment 1A), and two completion studies (Experiments 1B and 1C) to assess participants' preferences for a phonologically related NP2 in the two constructions. Then, in Experiments 2 and 3, we examined whether this knowledge would influence online processing and prediction in the visual world task. Finally, to evaluate to what extent our effects are attributable to Type I versus Type II prediction mechanisms (Huettig, 2015), we conducted a third visual world experiment (Experiment 4) using contexts that should discourage active prediction.

2. Experiment 1A: Judgment task

For this task, we created 44 quadruplets of sentences, crossing Sentence Type (Disfluency vs. Focus) and Repair Type (Semantic vs. Phonological), as illustrated below:

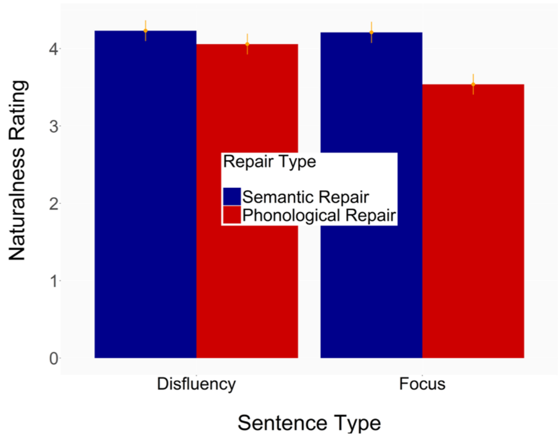

Fig. 1. Naturalness rating results by Sentence and Repair Type. The error bars represent the standard error of the mean.

(2)
(a) **Disfluency/Semantic Repair:** The man was looking around for a hammer, uh I mean a nail.
(b) **Disfluency/Phonological Repair:** The man was looking around for a hammer, uh I mean a hammock.
(c) **Focus/Semantic Repair:** The man was looking around for not a hammer but rather a nail.
(d) **Focus/Phonological Repair:** The man was looking around for not a hammer but rather a hammock.

Note that although there is technically no “reparandum” or “repair” in a focus construction, for consistency we will refer to NP2 in both the disfluency and focus constructions as the “repair” and to NP1 in both constructions as the “reparandum”. These 44 critical sentences were distributed across four experimental lists in a within-subjects design such that each participant received only one version of each critical sentence.

The experiment was carried out in Qualtrics Survey Software, which was interfaced with the UC Davis’s Research Participation Website (SONA). Participants were told that each sentence was a transcript of a spoken sentence, and they were asked to read and rate each sentence for how natural it sounded using a 7-point scale, with 1 denoting very unnatural and 7 very natural. In addition, each participant saw 80 filler sentences, including 60 simple active sentences and 20 anomalous catch trials. Sixty undergraduate students who were native speakers of American English took part in this judgment experiment in exchange for course credit.

To analyze the data, we performed linear mixed effects regression models (Baayen, Davidson, & Bates, 2008; Jaeger, 2008) using the lme4 package in R (version 3.3.4). The model included Sentence Type (Disfluency vs. Focus), Repair Type (Semantic vs. Phonological) and their interaction as the predictors, with Naturalness Rating (1–7) as the dependent variable. Following Barr, Levy, Scheepers, and Tily (2013), we used a maximal random effects structure for this model—that is, we added random intercepts for both subjects and items as well as by-subjects and by-items random slopes for Sentence Type, Repair Type and their interaction. We also centered both predictors to minimize multicollinearity. We used the normal approximation technique to calculate p-values.

2.1. Results

Fig. 1 displays the findings and Table 1 reports the results of our regression model. First, we found that focus constructions were rated less natural overall, as were phonological repairs in either construction. Importantly, and in line with our predictions, we also found a significant interaction between Sentence Type (Disfluency vs. Focus) and Repair Type (Semantic vs Phonological): Whereas semantic and phonological continuations were judged to be equally natural in disfluency contexts, phonologically related continuations were judged significantly less natural in the focus construction (*not a hammer but rather a hammock*).

Table 1
Results for the plausibility norming study. “*Sem*” and “*Phon*” denote “Semantic” and “Phonological” repairs, respectively.

Predictor	β	*SE*	*t*	*p*
Intercept	4.01	0.14	28.63	< .001
Sentence Type	−0.27	0.09	−2.71	.006
Repair Type	0.42	0.12	3.45	< .001
Interaction	0.49	0.13	3.59	< .001
Simple effects				
Intercept	3.53	0.16	21.72	< .001
Sem vs. *Phon* within Focus	0.67	0.13	4.95	< .001
Intercept	4.05	0.17	23.60	< .001
Sem vs. *Phon* within Disfluency	0.17	0.14	1.18	.23

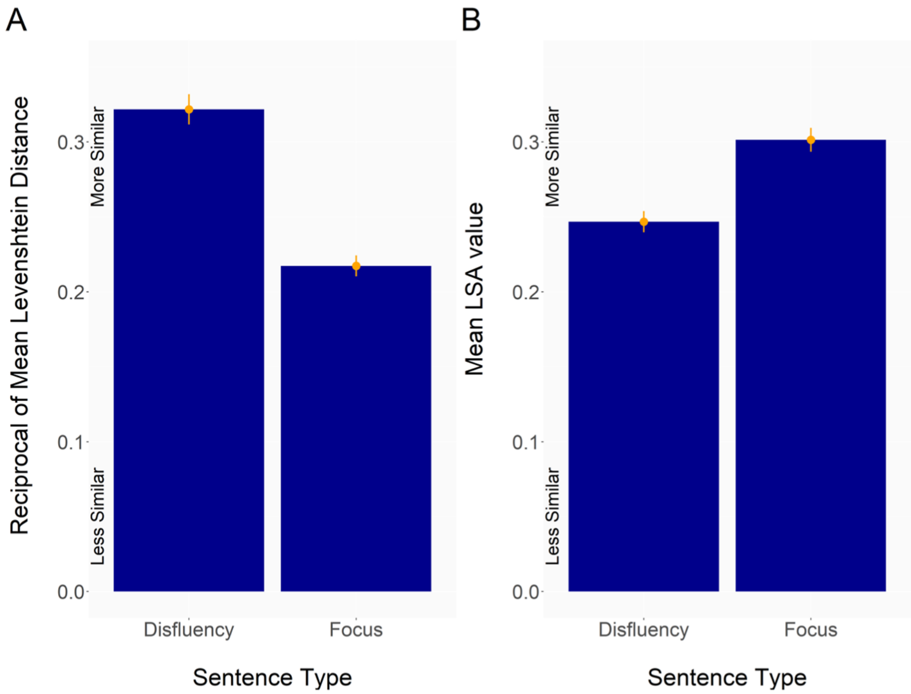

Fig. 2. (A) Mean Reciprocal of Levenshtein distance for each Sentence Type. (B) Latent Semantic values for each Sentence Type. In both panels, the error bars represent the standard error of the mean.

These judgment data show that readers are sensitive to the relative likelihood of a phonological relationship between NP1 and NP2 in disfluency and focus constructions: Specifically, they view phonological repairs as more natural than phonological contrasts.

3. Experiment 1B: Traditional cloze task

To further investigate people's judgments of phonological and semantic substitutions in Focus versus Disfluency contexts, we conducted a sentence completion study using the same 44 materials, as illustrated in (3). The experimental sentences were intermixed with 30 fillers.

(3)
(a) **Disfluency:** The man was looking around for a hammer, uh I mean a ….
(b) **Focus:** The man was looking around for not a hammer but rather a ….

Sixty participants were recruited from Amazon's Mechanical Turk and were presented with the critical sentence frames, one at a time, and were asked to provide the first continuation that came to mind. They were instructed to type only one word to complete each sentence. Qualifications were set to restrict the participants to native speakers of English who resided in the US, had a college education, and with "hit" approval rates of equal or greater than 95%. The task took approximately 20 min to complete and the participants were paid $1 for their time.

We performed two types of analyses of the data: First, we calculated the Levenshtein distance between the *pronunciations* of the noun given post-verbally in the sentence (i.e., NP1s) and the responses (i.e., NP2s) using the Carnegie-Mellon Pronouncing Dictionary, version 0.7b (http://www.speech.cs.cmu.edu/cgi-bin/cmudict) and the *adist* function of R. To assess semantic similarity, we obtained Latent Semantic Analysis cosine values between each response and the preceding NP2 (http://lsa.colorado.edu/). Consistent with Experiment 1A, participants produced more phonologically related continuations (i.e. responses with *smaller*

Table 2
Results for the cloze norming study.

Analysis	Predictor	β	*SE*	*t*	*p*
Phonological similarity	Intercept	5.80	0.33	17.43	< .001
	Sentence Type	1.30	0.30	4.30	< .001
Semantic similarity	Intercept	0.24	0.01	12.84	< .001
	Sentence Type	0.05	0.01	3.28	= .001

Levenshtein distances) in disfluency contexts (Fig. 2A) and more semantically related continuations in focus contexts (Fig. 2B). Note that because smaller values of Levenshtein distance indicate *greater* phonological similarity, we graphed the *reciprocal* of Levenshtein distance in Fig. 2A so that the y-axis reads more naturally (i.e., moving up in the y-axis indicates greater phonological similarity). The results of the statistical models are reported in Table 2.

4. Experiment 1C: Cloze task in the visual world

Because our goal was eventually to compare the results of these norming studies with those from the Visual World eyetracking experiments, we conducted an additional norming experiment in which participants were presented with the same visual displays (see Fig. 4 below) and the same critical sentences they would see in those eyetracking experiments. The stimuli were altered in only two ways: First, the repairs were deleted as illustrated in (3) above, and second, the sentences were presented in the written rather than the auditory modality. This norming experiment ensured that results from the previous two norming studies would generalize to the Visual World paradigm where the prediction choices are restricted to the number of pictures on the display (in our case four).[1] The 44 critical sentences (illustrated in 3) were intermixed with 50 fillers. Forty-eight native speakers of American English from the participant pool of University of California, Davis took part in the experiment, which was programmed in PsychoPy (v1.82.01, Peirce et al., 2019). The visual displays were presented to the participants together with the corresponding sentences which appeared at the bottom of each display. To maximize comparability between this norming study and the main eye-tracking experiments, there was a 2500 ms delay between the appearance of the display and the appearance of the corresponding sentence in each trial (see below). The participants' task was to read the sentences and click on the picture that best completed each sentence. We then analyzed the probability of choosing a semantic versus a phonological competitor (a binomial measure) as a function of Sentence Type (Focus vs. Disfluency).

Fig. 3 illustrates the probability of choosing the semantic competitor out of semantic plus phonological responses. We used a generalized mixed-effects regression with random intercepts for both subjects and items, as well as by-subjects and by-items random slopes for the effect of Sentenced Type (Barr et al., 2013) to analyze the data. The results revealed that the probability of choosing the semantic competitor was reliably greater given Focus constructions relative to Disfluency constructions ($\beta = 1.40$, $SE = 0.20$, $z = 6.82$, $p < .001$). The flip side of this result is that the probability of choosing the phonological competitor was significantly greater following Disfluency constructions compared to Focus constructions. Note that, obviously, participants also chose the repair and even the reparandum as their response, but such responses were removed from the statistical analysis (and also from calculating the proportions shown in Fig. 3) as they are not relevant to assessing our study's hypotheses. However, we report the raw frequency and the mean of all the responses in Appendix A.

Thus, across these two studies, we observed the expected interaction between sentence type (Disfluency vs. Focus) and Repair Type (Semantic vs Phonological): While semantic and phonological continuations were judged to be equally natural in disfluent forms, phonologically related continuations were viewed as significantly less natural in the focus construction. Moreover, disfluency contexts led to the *production* of more phonologically-related and less semantically-related substitutions relative to the focus context, indicating that the two sentence types are associated with differential expectations for the upcoming NP, at least in a completion task. Finally, given visual displays containing both semantic and phonological competitors, the probability of choosing the semantic (vs. the phonological) competitor was greater following Focus relative to disfluency constructions, whereas the probability of choosing the phonological competitor was greater following disfluency constructions relative to focus constructions.

5. Experiment 2: Visual world eyetracking with semantic and phonological competitors

Having established that an NP2 that is phonologically related to NP1 is preferred in disfluent constructions, and that an NP2 that is semantically related to NP1 is preferred in focus constructions, we can now ask whether this knowledge is rapidly applied during language processing to support different online predictions. Based on the results thus far, we expected an overall greater tendency to predict NP2s that were semantically related to NP1s regardless of construction (Experiment 1A), and we also expected to observe differential predictions, with more predictive looks to the semantic competitor in focus constructions and more predictive looks to the phonological competitor following disfluent sentences. To test these possibilities, we presented participants in this experiment with visual displays such as the one shown in Fig. 3 and recorded their eye movements while they listened to sentences such as (1), repeated here for convenience.

(1)
(a) **Disfluency:** The man was looking around for a hammer <u>uh I mean a</u> yardstick.
(b) **Focus:** The man was looking around for not a hammer <u>but rather a</u> yardstick.

5.1. Participants

Forty-eight undergraduate students from the participant pool of the University of California, Davis took part in the experiment in

[1] We thank an anonymous reviewer for suggesting this experiment.

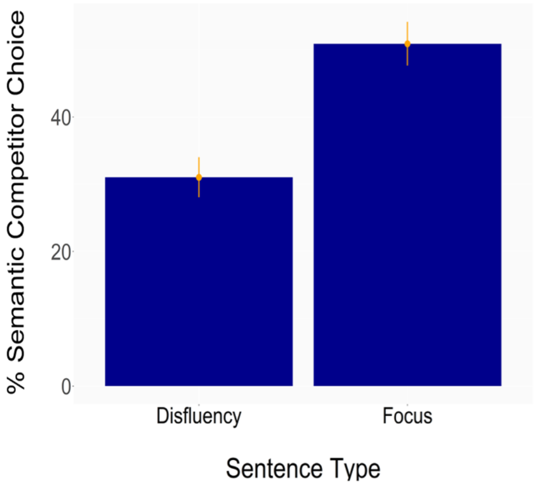

Fig. 3. Cloze task results by Sentence and Repair Type. Experiment 1C. The error bars represent the standard error of the mean. Note that the bars represent the probability of choosing the semantic competitor out of the sum of semantic and phonological competitors.

Fig. 4. Sample visual display for Experiment 2.

exchange for course credit. The data from one of the participants were removed from analysis due to poor tracking quality. All subjects were native speakers of American English and reported normal or corrected-to-normal vision and hearing.

5.2. Materials

Forty-four critical sentences such as (1) were created (see Appendix B for a list all the critical sentences used in this and the following experiments). Each sentence contained two NPs: a first-mentioned NP (NP1, *hammer*), and a second-mentioned, intended NP (NP2, *yardstick*). Two versions of each sentence were created to establish the Disfluency and Focus sentence conditions.

Each sentence was paired with a visual display that contained the images for NP1 and NP2 (*hammer* and *yardstick*) as well as two critical competitors: a semantic competitor (*nail*) and a phonological competitor (*hammock*) (see Fig. 3). The competitors were selected based on their semantic relatedness to NP1, which was determined using Latent Semantic Analysis (accessible at: http://lsa.colorado.edu/). Crucially, semantic competitors were semantically most related to NP1 (Mean = 0.40) and phonological competitors were semantically least related to NP1 (Mean = 0.08). NP2s always fell in-between with regards to their semantic relationship to NP1 (Mean = 0.14). Simple *t*-tests revealed that semantic competitors were significantly more semantically related to NP1s than were the phonological competitors ($t(86) = 10.02$, $p < .001$). NP2s were also more semantically related to NP1s than phonological competitors ($t(86) = 3.09$, $p = .002$). Moreover, semantic competitors were significantly more semantically related to NP1s than were the NP2s ($t(86) = 7.76$, $p < .001$). There were no statistically reliable differences between semantic and phonological competitors in

SUBTLEX-US word frequency ($t(86) = 1.16$, $p = .24$).

The prediction window (underlined in 1)—the window in which we will measure anticipatory looks to the displayed pictures—was defined separately for each condition: In the Disfluency condition, this window started at the onset of the "editing term" (i.e., "uh I mean") and ended at the onset of NP2. In the Focus condition, the predictive window started at the onset of *but* in the sequence *but rather* and ended at the onset of NP2. These windows were chosen because the linguistic signal indicating that NP1 will be contrasted with an upcoming NP becomes available with the onset of "uh I mean" in the Disfluency condition, but at the offset of NP1 in the Focus condition (note that the negation of NP1 already makes it clear that another NP is imminent). Thus, predictive saccades should start at slightly different time points in the two conditions. The mean length of the predictive window was 1200 ms and was matched across the two conditions.

The 44 critical sentences were distributed across four lists and were intermixed with 108 filler sentences of various types. Eighty-four of the filler sentences were spoken fluently. Of the 24 remaining fillers, 12 contained a repair disfluency and 12 were focus constructions similar to the critical items. To equate the utility of phonological and semantic prediction overall, 12 fillers resolved to phonological and 12 to semantic competitors. The sentences were recorded by a native speaker of English (the second author). The position of images on the visual displays was randomized for each trial and the trials were randomized for each participant.

5.3. Procedure

The experiment was programmed with the SR Research Experiment Builder software and eye movements were recorded using an SR Research EyeLink 1000 Plus eyetracker sampling at 1000 Hz. Participants were seated in front of a computer screen and were instructed to listen to and follow the sentences on the visual display. Auditory stimuli were presented binaurally using headphones. After calibration of the eyetracker, the participants listened to three practice trials to become acquainted with the experiment. At the beginning of each trial, a fixation dot appeared at the center of the screen. Participants were asked to fixate this dot and press the spacebar for the trial to begin. After pressing the spacebar, the visual display appeared on the screen, and, after a 2500 ms delay, the sentence associated with the display was presented. The trial automatically ended at sentence offset. Participants could take a break whenever they wished. If the subject elected not to take a break, the experimenter paused the experimental session halfway through for a 5-min break. The experiment took approximately 60 min to complete.

5.4. Statistical analyses

We first removed any fixations shorter than 80 ms to ensure that all looks were indeed intended and not random. Then, we fitted logistic mixed effects regression (Baayen et al., 2008; Jaeger, 2008) with Image Type (NP1 vs. NP2 vs. Semantic Competitor vs. Phonological Competitor) and Sentence Type (Disfluency vs. Focus) as the fixed effects, using the *glmer* function from the lme4 package in the statistical software R (version 3.3.4) to model our data. The dependent variable was binomial and measured whether a critical image received a fixation or not. Specifically, we created a "reference" file containing all conditions that the participants were exposed to. We then looked at the actual eye-tracking data within the predictive window and recoded whether there was a fixation to any of the 4 critical images, in a specific condition, for a specific subject and for specific item. Following Barr et al. (2013), we fitted models with the full random effects structure, that is, with random intercepts for both subjects and items as well as by-subjects and by-items random slopes for the two predictors and their interaction. We achieved convergence for all the models reported for this experiment.

It is important to mention that although logistic regression reduces continuous eye-tracking data to a binomial measure and might consequently result in loss of some detail, there are two important benefits to this approach over those which use continuous measures, such as the 'empirical logit' of fixation proportions (Barr, 2008): First, unlike the empirical logit approach, logistic regression does not require collapsing the data over one random variable at a time (by-subjects and by-items analyses) and therefore affords simultaneous generalization over both subjects and items. Second, and more importantly, recent simulations have shown that empirical logits can lead to spurious interactions (Donnelly & Verkuilen, 2017). Given that the critical test of our hypothesis is whether we observe an interaction between context type and repair type, we decided to use logistic regression rather than empirical logits to analyze our data, although we also analyzed the data from all of our experiments using the empirical logit of fixation proportions (Barr, 2008), with very similar results.

Since we were interested only in looks to the semantic competitor, the phonological competitor, and NP2, we excluded looks to NP1 when performing our statistical analyses.[2] We performed two independent analyses to capture the effects of interest. In the "Differential Prediction Analysis", we looked at the potential 2 × 2 interaction between Sentence Type (Focus vs. Disfluency) and Image Type (Semantic vs. Phonological). The dependent variable in this analysis was binomial and reflected whether there was a look to either the semantic or the phonological competitor within the predictive window (looks to the repair were removed from this analysis). Importantly, the Differential Analysis would reveal whether looks to the semantic and phonological competitors reliably varied as a function of Sentence Type. Both Sentence Type and Image Type were centered in the Differential Prediction Analysis to reduce multicollinearity. In the "Image Type Analysis", we examined the probability of looking to the images associated with the semantic and phonological competitors relative to NP2 in the predictive window, pooling over Sentence Type. Again, the dependent

[2] However, we included NP1s when calculating the fixation proportions within the predictive window (see the bar graphs below), as well as when plotting how the fixations unfolded over time (see the line graphs below).

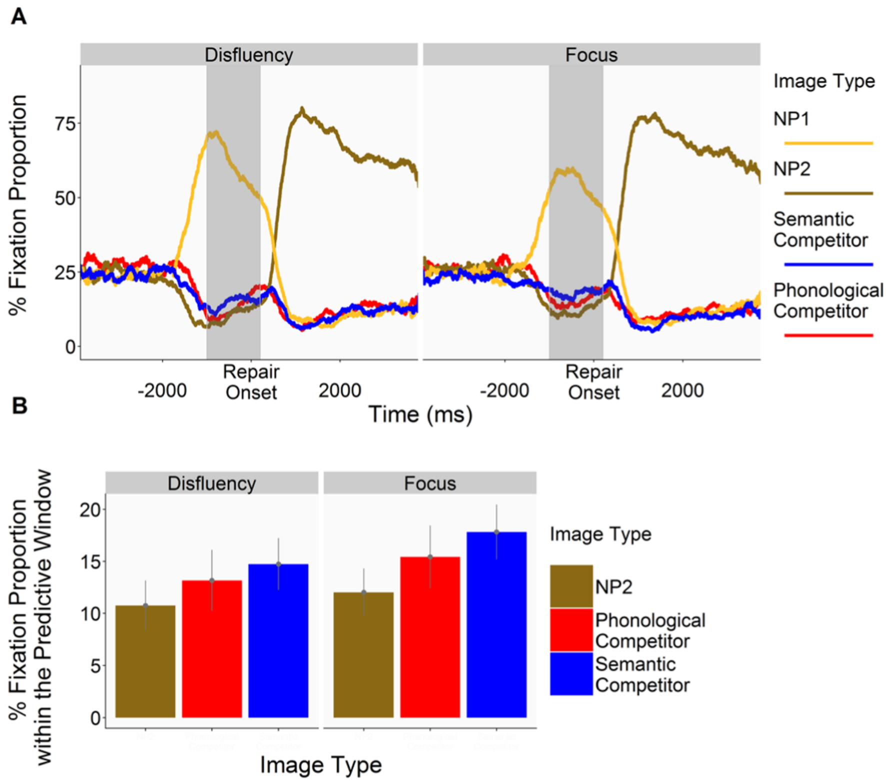

Fig. 5. Panel A: Proportion of looks to all images for each Sentence Type over time in Experiment 2. Panel B: Proportion of looks to the distractor, NP2, the phonological and the semantic competitors within the predictive window.

variable was binomial and reflected whether there was a look to one of the three images or not. For this analysis, Image Type was entered in the regression model as a three-level predictor (semantic competitor, phonological competitor and NP2). Then, we used treatment coding to assess the relative number of looks to the three images. The logic is that the Differential Analysis answers our primary research question concerning whether comprehenders predict semantic versus phonologically related words based on the preceding sentence construction (Focus vs. Disfluency), and the Image Type Analysis answers our second research question concerning the overall strength of phonological relative to semantic predictions (i.e., regardless of sentence construction).

5.5. Results

Fig. 5, Panel A shows fixation proportions for the semantic competitor, the phonological competitor, and NP2, time-locked to the onset of NP2 (NP1 is included for completeness, even though it is not analyzed). The gray area shows the predictive window, which spanned from 1000 ms before until 200 ms after NP2 onset. This 200 ms shift in the predictive window was to account for saccade programming time, which takes approximately 200 ms (e.g., Saslow, 1967). Fig. 5, Panel B, displays in bar graph form the proportion of fixations to NP2, the phonological and the semantic competitors within the predictive window.

As can be seen in Fig. 5, the proportion of looks to both the semantic and the phonological competitors is greater than the proportion of looks to NP2. In addition, there were also generally fewer looks to NP1 in the focus condition (see Panel A), presumably

Table 3
Logistic regression results for Experiment 2. "*Sem*" and "*Phon*" denote "Semantic" and "Phonological" competitors, respectively.

Analysis	Predictor/Contrast	β	*SE*	*Z*	*p*
Differential prediction	Intercept	−1.43	0.13	−10.68	< .001
	Sentence Type	0.18	0.09	1.96	.04
	Image Type	0.51	0.16	3.14	=.001
	Interaction	0.01	0.20	0.06	.94
Image type	Intercept	−1.72	0.15	−11.02	< .001
	Sem vs. NP2	0.55	0.12	4.29	< .001
	Phon vs. NP2	0.03	0.13	0.26	.79
	Intercept	−1.68	0.18	−9.26	< .001
	Sem vs. *Phon*	0.52	0.16	3.22	=.001

because NP1 was explicitly negated (*not the hammer but*…). As a result, there was a corresponding increase in looks to the other three images (as can be seen in panels A and B). However, the crucial comparison is looks to the semantic versus the phonological competitors *relative* to NP2; although looks to all three images generally increased in the Focus condition, this increase may or may not be equal for the two competitors relative to NP2.

The results of our models are reported in Table 3. The Differential Prediction analysis showed a significant Sentence Type effect, with more looks to both the semantic and the phonological competitor in the Focus compared to the Disfluency condition. There was also a main effect of Image Type, with reliably more looks to the semantic competitor than to the phonological competitor. Crucially, there was no interaction between Image Type and Sentence Type.

The Image Type analysis revealed that the probability of looking at the semantic distractor was significantly greater than the probability of looking at NP2, but the probabilities of looking at the phonological competitor and NP2 were statistically the same. Finally, the probability of looks to the semantic competitor was reliably greater than the probability of looks to the phonological competitor.

5.6. Discussion

The results of this first eyetracking experiment suggest that when there is a signal in speech indicating that the currently spoken word was not intended and is going to be edited to a new word, the language comprehension system does not simply wait for the intended word, but rather starts to predict it, which is consistent with Lowder and Ferreira (2016a)'s findings. However, the lack of interaction between the type of sentence heard (disfluency vs. focus construction) and the type of word predicted (semantic vs. phonological) suggests that prediction during the processing of repair disfluencies is not flexible enough to vary based on the contextual information that we established in Experiments 1A, 1B, and 1C influences naturalness judgments, free completion preferences, as well as completion choices in the Visual World contexts. Since differential preferences for semantic versus phonological substitutions were observed in all three norming experiments, insensitivity to such cues runs counter to psycholinguistic theories maintaining that subtle statistical regularities and co-occurrences in the linguistic input should be readily available for generating predictions during online language processing (Chang et al., 2006; Clark, 2013; Kleinschmidt et al., 2012; Levy, 2008a; McDonald & Shillcock, 2003).

Moreover, we observed that, overall, the probability of looks to the semantic competitor was significantly greater than the probability of looks to the phonological competitor, suggesting that prediction during the comprehension of focus constructions and repair disfluencies is primarily driven by semantic rather than phonological relationships. Finally, the main effect of Sentence Type showed that the average probability of looks to all the images was greater in the Focus condition relative to the Disfluency condition, which is most likely due to the presence of the negation operator in the focus construction, which depressed looks to NP1 and thus made room for more looks to all three of the other images in the visual display. Note that this result is not relevant to evaluating our hypotheses because we were interested in the *difference* between probability of looks to semantic versus phonological competitors as a function of Sentence Type.

6. Experiment 3: Visual world eyetracking with semantic or phonological competitors

The results of the second experiment suggest that although listeners anticipate mention of an upcoming NP, they are insensitive to distributional differences concerning the likelihood of semantic versus phonologically associated words for focus versus repair disfluencies. The results also provided no evidence for pre-activation of phonological information during the processing of disfluencies. However, it is possible that these results were obtained because the semantic and the phonological competitor were simultaneously present in the visual display, forcing listeners to choose between them. As such, any tendency to predict the phonological competitors could have been suppressed by the presence of the more salient semantic competitor in the display. In addition, contextual cues might not have modulated prediction preferences because the processor might be incapable of simultaneously activating the memory representations associated with both competitors *and* the respective sentential contexts in which those competitors occur more frequently. To eliminate the need for simultaneous prediction of semantic and phonological information, we designed our visual displays so that each one contained only one type of competitor: either phonological or semantic.

In addition, one potential limitation of Experiment 2 was that the repair (NP2) was used as a baseline for comparing anticipatory looks to the semantic and phonological competitors. Considering the strong influence of semantic relatedness on anticipatory eye movements, the greater semantic association between NP2 and NP1 (the reparandum, Mean = 0.14) relative to the phonological competitor and NP1 (Mean = 0.08, $t(86) = 3.09$, $p = .002$) may have prevented us from observing a clear phonological prediction effect. To provide a more neutral baseline condition, we included a random distractor image that was both semantically and phonologically unrelated to the NP1s that were used in this experiment.

6.1. Participants

Sixty undergraduate students from the participant pool of the University of California at Davis took part in the experiment in exchange for course credit. All the participants were native speakers of American English and reported normal or corrected to normal vision and hearing.

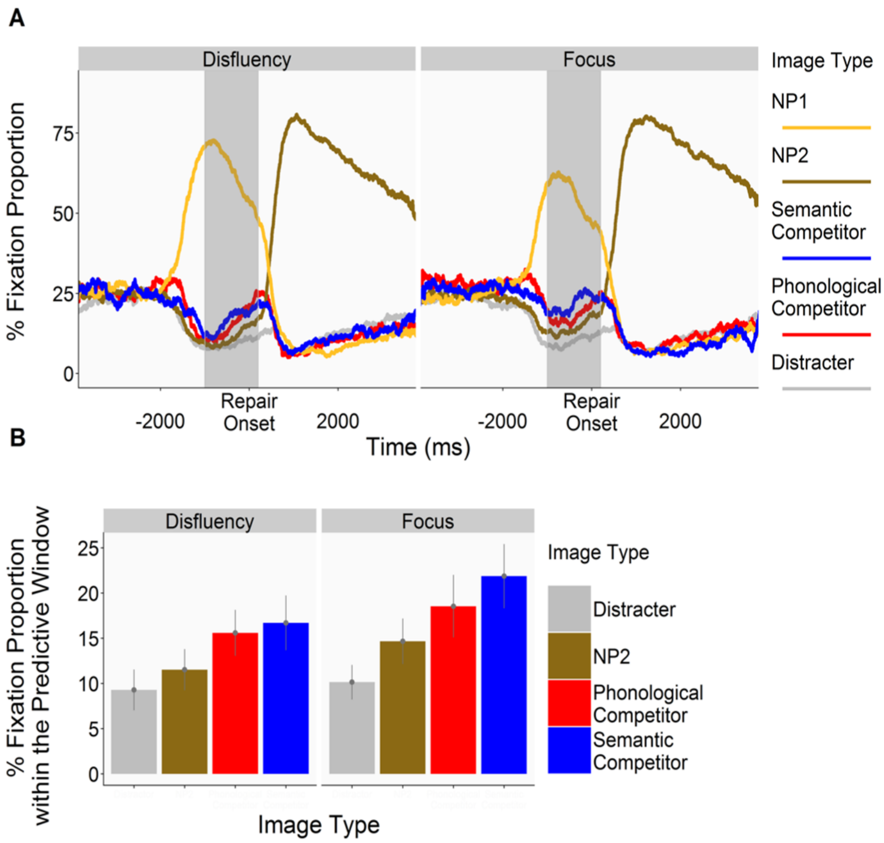

Fig. 6. Panel A: Proportion of looks to all images for each Sentence Type over time in Experiment 3. Panel B: Proportion of looks to the distractor, NP2, the phonological and the semantic competitors within the predictive window.

6.2. Materials

We used the same linguistic materials as in Experiment 2. The visual displays were changed so each display contained either the semantic or the phonological competitor, but not both. For this experiment, the fourth image was a random distractor—an image associated with a semantic or phonological competitor from other critical items. These random distractors were reliably less semantically related to NP1 than were the semantic competitors ($t(86) = 9.93$, $p < .001$), and NP2s ($t(86) = 2.90$, $p = .004$), but the phonological competitors and the random distractors were equally semantically related to NP1 ($t(86) = -0.27$, $p = .78$).

6.3. Procedure and statistical analyses

The procedure was identical to that in Experiment 2. The statistical analyses were also identical to those in Experiment 2, except that for the Image Type Analysis, in two separate analyses we compared looks to the two critical competitors with looks to both the distractor as well as to NP2. Models including the full random-effects structure always converged in this experiment.

6.4. Results

Fig. 6, Panel A shows the proportion of fixations to all images for each sentence type. The gray area highlights the analysis widow spanning from 1000 ms before to 200 ms after NP2 onset. (As mentioned above, the 200 ms shift in the predictive window was to accommodate the time needed for saccade programing; Saslow, 1967.) We collapsed the five critical images in the same graph for convenience and to facilitate comparison of these results with those of Experiment 2. Fig. 6, Panel B displays the proportion of looks to the distractor, NP2, the phonological competitor, and the semantic competitor within the predictive window.

Table 4 reports the results of logistic mixed-effects regression models on the data. In the Differential Prediction analysis, the effect of Sentence Type was significant, and the effect of Image Type was marginally significant. Crucially, and consistent with Experiment 2, we again found no evidence of an interaction between Sentence Type and Image Type.

As before, the Image Type analysis revealed significantly more looks to the semantic competitor than to the random distractor. In addition, we found significantly more looks to the phonological competitor than to the random distractor. Moreover, the probability of looking at the semantic competitor was significantly greater than the probability of looking at NP2. Interestingly, and unlike

Table 4
Logistic regression results for Experiment 3. "*Sem*", "*Phon*" and "*Dist*" denote "Semantic competitor", "Phonological competitor" and "Distracter", respectively.

Analysis	Effect/Contrast	β	*SE*	*Z*	*p*
Differential prediction	Intercept	−1.13	0.10	−10.45	< .001
	Sentence Type	0.22	0.11	1.94	.05
	Image Type	0.22	0.11	1.83	.06
	Interaction	0.22	0.23	0.96	.33
Image Type: Distractor as baseline	Intercept	−1.81	0.13	−13.20	< .001
	Sem vs. Dist	0.79	0.13	6.08	< .001
	Phon vs. Dist	0.59	0.13	4.46	< .001
Image Type: NP2 as baseline	Intercept	−1.53	0.11	−13.13	< .001
	Sem vs. NP2	0.51	0.11	4.51	< .001
	Phon vs. NP2	0.31	0.12	2.48	=.01

Experiment 2, the probability of looks to the phonological competitor was greater than the probability of looks to NP2.

6.5. Discussion

Similar to what we found in the previous Visual World eyetracking experiment, in Experiment 3 we observed that semantic and phonological competitors were predicted to the same degree in disfluency and semantic focus constructions, suggesting that the prediction system cannot vary the content of predictions based on these subtle cues during online language processing. Considering that participants exhibited differential preferences for semantic and phonological continuations in our judgment and completion tasks, the obvious question is why these preferences did not influence online predictions. One possibility is that exploiting subtle contextual cues to make differential predictions might be too demanding given the time pressures of online processing (Mani & Huettig, 2012). Before exploring this possibility in greater detail, we conducted an analysis combining the data from both eyetracking experiments to confirm that there were no differential effects of sentence context.

The results of this third experiment also demonstrated that, in the absence of a semantic competitor, the phonological competitor is predicted as the likely referent during the processing of repair disfluencies. Interestingly, the phonological competitor was not only predicted more often than the random distractor, it was also predicted more than the actually heard NP2, even though NP2 was significantly more semantically related to NP1. This implies that prediction during language processing can involve pre-activation of not only meaning-based but also form-based information, and that strongly activated phonological features can influence online predictions more than weakly activated semantic features can. In other words, this pattern of results suggests that the prediction system employs the most available information to make predictions regardless of whether that information stems from semantic or phonological levels of linguistic representations (Levy, 2008a; MacDonald, 1994; MacDonald, Pearlmutter, & Seidenberg, 1994; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002).

As in Experiment 2, semantic prediction effects were larger in magnitude than the effects of phonological prediction, within both disfluency and focus constructions. This suggests that overlapping semantics may provide the primary source of information for generating predictions about upcoming linguistic input (also see, Ito et al., 2016; Pickering & Garrod, 2007, 2013). However, if phonological information is strongly and directly activated by the preceding noun phrase (i.e., the reparandum), anticipatory looks to the phonological competitor will occur, rather than looks to a weakly associated semantic competitor such as NP2. In other words, this finding suggests that when phonological features reach a strong enough activation level such that a phonologically related word is more accessible than one that might be activated based on semantic features, that activation can drive prediction to a phonological competitor over a semantic competitor. This result runs counter to theories based on which phonological information can only be accessed after semantic information (Ito et al., 2016; Pickering & Garrod, 2007, 2013) and is consistent with theories of prediction maintaining that the language processing system uses whatever information is most available to guide prediction—whether semantic or phonological (e.g., Clark, 2013; Levy, 2008a).

7. Experiments 2 and 3 combined

Because our primary question was the potential flexibility of predictive processing across different language contexts, to maximize statistical power we combined the results of Experiment 2 and 3 into a single dataset. Note that although the visual displays were not identical in the two experiments, the semantic and phonological competitors (*nail* and *hammock*) as well as NP2 (*yardstick*) were present in both (within a single trial in Experiment 2, and across trials in Experiment 3). As such, it is possible to directly compare the proportion of looks to the two critical competitors relative to NP2 across the two datasets.

7.1. Statistical analyses

Since we wished to focus only on the probability of looks to the two competitors as a function of Sentence Type, we ran only the Differential Analysis (which was identical to those in Experiments 2 and 3) and not the Image Type analyses. The full random structure model converged for this model (Barr et al., 2013). In addition, to assess the strength of evidence for or against the null

Table 5
Logistic regression results for Experiments 2 and 3 combined.

Analysis	Predictor/Contrast	β	SE	Z	p
Differential prediction	Intercept	−1.25	0.08	−14.47	< .001
	Sentence Type	0.19	0.07	2.64	=.008
	Image Type	0.36	0.11	3.31	< .001
	Interaction	0.09	0.14	0.67	.49

hypothesis, we conducted a Bayesian paired sample *t*-test on the critical interaction term (e.g., Gallistel, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009). That is, we calculated the difference in the average of looking probabilities to the semantic vs. the phonological competitor for each subject for each Sentence Type. We then conducted a Bayesian *t*-test predicting the calculated difference score as a function of hearing a disfluency versus a focus construction. This analysis reveals the level of statistical confidence we could have on the null results for the critical interaction between Sentence Type and Image Type. To verify the robustness of the Bayes Factor, we used three different Cauchy scaling factors, 0.5, 0.707, 1.0, which correspond to alternative hypotheses of small, medium, or large effect sizes, respectively. We ran this test with the JASP software, Version 0.8.1.4 (https://jasp-stats.org/).

7.2. Results and discussion

Table 5 reports the results of generalized linear mixed-effects regression models on the combined data. As can be seen in this table, even in the combined dataset, there was no interaction between Sentence Type and Image Type, suggesting that the prediction system is not flexible enough to make differential predictions based on contextual cues related to the sentence construction—in this case, repair disfluencies versus focus constructions (cf., Clark, 2013; Levy, 2008a). Rather, the prediction system seems to operate primarily based on semantic and phonological likelihood of the predicted words given the preceding information (Huettig, 2015).[3] In contrast, in three norming studies, we did observe a differential preference for semantic versus phonological competitors depending on whether the construction involved focus or a disfluency, suggesting that although listeners may possess the knowledge of the different lexical alternates associated with focus and disfluency constructions, they are unable to rapidly access that knowledge quickly enough to guide online predictions.

Our Bayesian *t*-test revealed Bayes Factors of 3.57, 4.88 and 6.76 in favor of the null hypothesis for the three Cauchy scaling factors, respectively (see above). Given that Bayes Factors between 3 and 10 are considered as moderate evidence in favor of one hypothesis over the other (Jeffreys, 1961; Wetzels et al., 2011), the results appear to be most consistent with the null hypothesis, and therefore with the results from our mixed-effects models, indicating that sentence contexts do not influence the type of prediction made—to a semantic or phonologically related item—during online processing. Thus, unlike in Experiments 1A, 1B, and 1C in which people exhibited differential expectations based on preceding sentence type, in real-time comprehension participants do not generate a differential pattern of looks to the semantic and phonological competitors in light of the preceding context.

While the interaction term was not significant in this combined analysis, we did observe a main effect of Sentence Type, with more looks to both the semantic and the phonological competitor in the Focus condition. The main effect of Image Type was also significant, with more overall looks to the semantic than to the phonological competitor.

Across both experiments, then, the semantic competitor received more looks than the phonological competitor, suggesting that the prediction system is primarily driven by semantic rather than phonological overlap.

8. Experiment 4: To what extent are looks to semantic and phonological competitors driven by lexical priming from the reparandum?

The results of both eyetracking experiments suggest that comprehenders do not, or cannot, use their knowledge of sentence context to make differential semantic versus phonological predictions online—they make robust predictions, but the content of those predictions is not influenced by whether the sentence is a focus structure or a reparandum-repair disfluency. What might account for this discrepancy between these eyetracking results and the findings from the judgment and completion tasks (Experiments 1A, 1B and 1C)? One possibility is that listener's looks in the predictive window are automatically directed to any object with semantic and phonological overlap with the reparandum (NP1) through lexical priming. In other words, looks to the semantic and phonological competitors might have been driven not so much by active, *context*-based prediction, but rather by *word*-based feature overlap. In fact, previous visual world studies have shown that eye movements are generated to both semantically and phonologically associated words in a visual scene (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Huettig & Altmann, 2005; Mirman & Graziano, 2012; Yee & Sedivy, 2006). A second possibility, however, is that looks to objects in the visual display that were made before the repair was spoken reflect prediction, but the content of such predictions is not sensitive to sentential context (e.g., Kukona et al., 2011).

For our eyetracking experiments, then, we wished to assess the amount of this lexical priming from NP1 to NP2, to determine

[3] We also ran this same analysis on a subset of items where the difference in naturalness ratings between semantic and phonological substitutions was maximal in our judgment task (Experiment 1A). The results of this analysis were nearly identical to those reported above, with no interaction between Sentence Type and Image Type.

whether the effects we observe in the focus and disfluency constructions are over and above this baseline spreading activation based on a single word. For this purpose, we conducted a final eyetracking experiment that included two novel sentence versions: a Coordination condition (…*the hammer and also the*…), as well as a Silence condition in which there was no NP2, but instead, 1200 ms of silence (… *the hammer* < *silence* > .). Neither of these conditions should motivate predictive looks because there is no signal in the linguistic input indicating that NP1 is either spoken in error or is to be contrasted with another NP. The silence condition simply includes the preamble to NP1, along with NP1 itself. In the case of coordination structures, Lowder and Ferreira (2016a) demonstrated across multiple experiments that listeners predict far less in coordination contexts than in either repair or focus constructions. Consequently, by examining looks to semantic and phonological competitors in the visual display, we can assess the baseline tendency to anticipate some type of NP2 independent of our manipulation of sentence type: focus construction versus disfluency. We accordingly used the visual displays from Experiment 3 (which contained only one competitor type, semantic or phonological) to evaluate this tendency.

8.1. Participants, procedure, materials and statistical analyses

Fifty-two participants, all native speakers of American English, were recruited from the participant pool of the University of California, Davis. One participant's data were removed because the session was not fully recorded. The procedure and the statistical analyses were identical to those of Experiment 3, except that we performed only the Image Type analysis. However, and as mentioned above, in this experiment the NP2s were either conjoined with the coordinating phrase *and also*, or the repair was removed altogether and replaced with silence, producing a Coordination conditions and a Silence condition, respectively, as illustrated in (4):

(4)
(a) **Coordination:** The man was looking around for a hammer and also a yardstick.
(b) **Silence:** The man was looking around for a hammer < 1200 ms silence > .

The stimuli for the Silence condition were recorded independently by the same individual who recorded the stimuli for the previous eye-tracking experiments (i.e., the second author); they were not created by truncating the stimuli made for the Coordination condition. Moreover, trials in the Silence condition automatically ended at the offset of the 1200 ms silence period, which triggered the removal of the visual display.

8.2. Results

Figs. 7 and 8 show the proportion of looks in the Coordination and the Silence conditions, respectively. In the Coordination condition (Fig. 7), the gray area marks the time window from 1000 ms before NP2 onset up to 200 ms after the NP2 onset. In the

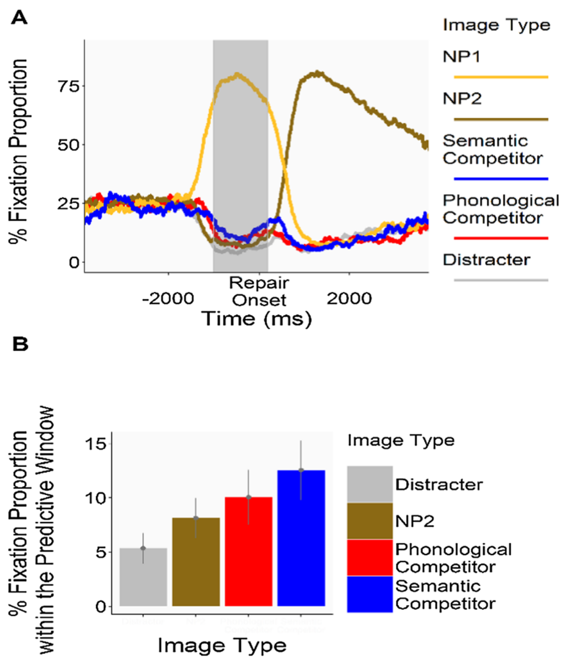

Fig. 7. Proportion of looks in the Coordination condition in Experiment 4. Panel A shows the proportion of looks to all images over time. Panel B displays the proportion of looks to the Distracter, NP2, the phonological and the semantic competitors within the critical window. competitors within the predictive window.

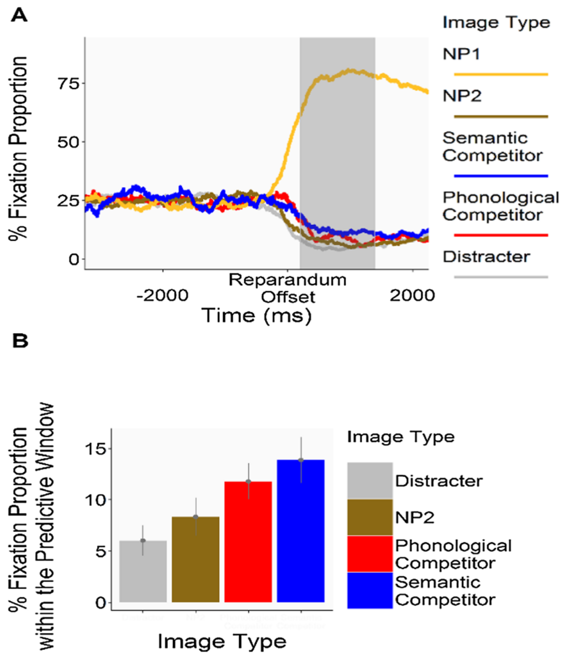

Fig. 8. Proportion of looks in the Silence condition in Experiment 4. Panel A shows the proportion of looks to all images over time. Panel B displays the proportion of looks to the Distracter, NP2, the phonological and the semantic competitors within the critical window. competitors within the predictive window.

Table 6
Logistic regression results of Experiment 4. "*Sem*" and "*Phon*" denote "Semantic" and "Phonological" competitors, respectively.

Predictor/Contrast	β	*SE*	*t*	*p*
Intercept	−2.35	0.13	−17.76	< .001
Sem vs. Distracter	0.85	0.13	6.16	< .001
Phon vs. Distracter	0.63	0.13	4.79	< .001
Intercept	−1.93	0.12	−15.28	< .001
Sem vs. NP2	0.45	0.13	3.21	< .001
Phon vs. NP2	0.21	0.13	1.60	.10

Silence condition (Fig. 8), the gray area demarcates the time window spanning from 200 ms up to 1400 ms after NP1 offset. As before, the critical window was shifted forward by 200 ms to account for saccade programming time (e.g., Saslow, 1967). In both figures, Panel B displays the proportion of fixations to all images within the critical window (i.e., the gray area).

As can be seen in these figures, the proportion of looks to both the semantic and the phonological competitors is greater than looks to the distracter and to NP2. Because we did not find any differences between the Coordination and Silence conditions, we collapsed them and examined the probability of looking at the semantic and phonological competitors in relation to the distracter and to NP2. The results of our statistical analyses are reported in Table 6. As is clear from this table, the probability of looking to either the semantic or the phonological competitor was significantly greater than the probability of looking to the distracter. Moreover, looking probability was reliably greater for the semantic competitor relative to NP2. Finally, there was a trend towards more looks to the phonological competitor than to NP2.

In order to investigate the differences between Experiment 3 where prediction was encouraged by disfluency or focus constructions and Experiment 4 in which there was no apparent motivation for prediction, we also compared the overall proportion of looks to all images on the visual display (i.e., looks to the semantic competitor + looks to the phonological competitor + looks to NP2 + looks to the distracter), as well as looks to each image individually (i.e., the semantic competitor, the phonological competitor, NP2, and the distracter) across these two experiments. The results of these analyses are reported in Table 7. As can be seen in this table, the probability of looking at all images was reliably greater in Experiment 3 compared to Experiment 4, both in the aggregate as well as for each individual image. This can be seen by comparing the y-axis range for Panel Bs of Fig. 5 with those of Figs. 6 and 7).[4]

[4] When comparing looks to the phonological competitor and the distracter, the full models did not converge and we simplified the random-effects to a "slopes-only" structure (Barr et al., 2013).

Table 7
The results of comparisons between looks to potential repairs across Experiments 3 and 4. "*Sem*" and "*Phon*" denote "Semantic" and "Phonological" competitors, respectively.

Effect/Contrast	β	*SE*	*t*	*p*
Intercept	−1.43	0.10	−14.32	< .001
All potential repairs	−0.47	0.14	−3.31	< .001
Intercept	−1.00	0.11	−8.65	< .001
Sem	−0.48	0.17	−2.85	=.004
Intercept	−1.19	0.12	−9.91	< .001
Phon	−0.50	0.17	−2.89	=.003
Intercept	−1.51	0.10	−13.75	< .001
NP2	−0.42	0.15	−2.64	=.008
Intercept	−1.79	0.13	−13.67	< .001
Distracter	−0.51	0.17	−2.92	=.003

8.3. Discussion

The *absolute* proportion of looks to all potential substitutions was significantly lower in Experiment 4, in which prediction was based on lexical priming only, compared to Experiment 3, in which prediction was encouraged via disfluency and focus contexts. This pattern of results shows that looks to the potential substitutions in Experiment 3 were driven by more than just spreading activation from NP1 (the reparandum), and likely involved active, context-based prediction as well. Note that because this argument is based on a between-subjects and across-experiments comparison, the conclusion that there was more than just lexical priming going on in Experiments 2 and 3 might seem rather weak. However, this argument receives support from multiple fronts: First, Lowder and Ferreira (2016a) have already shown, in two separate *within*-subjects experiments, that disfluencies and focus constructions result in significantly more looks to likely upcoming NPs than do coordination structures. Our current between-subjects results are fully consistent with theirs. Second, because NP1 was identical across Experiments 3 and 4, the greater number of looks to the potential repairs in Experiment 3 could not have possibly been caused by "more lexical priming" from NP1 in that experiment. Third, if one were to include that priming is the sole mechanism responsible for the effects observed in Experiments 2 and 3, then the results of these experiments should be interpreted as showing that the prediction system does not generate any predictions when it has the necessary information and time to do so (1200 ms), which runs against a whole host of research findings that have shown evidence for prediction during language comprehension (e.g., Altmann & Kamide, 1999; Brothers et al., 2015, 2017; Huettig & McQueen, 2007; Ito et al., 2016; Lowder & Ferreira, 2016a; Rommers et al., 2013; also see Levy, 2008a).

However, and importantly, the *relative* differences between looks to the potential substitutions remained the same across Experiments 3 and 4. This means that active and context-based prediction (as observed in Experiment 3) might sometimes essentially "ride on" lexical priming (as observed in Experiment 4). In other words, these results suggest that the underlying mechanism for both lexical priming and active prediction might sometimes be the same: spreading activation. The difference between lexical priming and active, context-based prediction would then be the source of this spreading activation: In lexical priming (Experiment 4), the source is a single word, whereas during active prediction, spreading activation might arise from a combination of words and/or contextual cues such as the presence of a disfluency. Thus, collectively, the results from Experiments 3 and 4 suggest that looks to the semantic and phonological competitors were triggered by both lexical priming as well as active prediction. However, as discussed above, because semantic and phonological competitors were predicted to the same degree within disfluency and semantic focus constructions in Experiment 3 (and 2), the prediction system seems not to be able to flexibly vary the content of predictions based on specific contextual cues, as reflected in the offline difference between disfluency and focus constructions that we observed in Experiment 1.

Our observation that anticipatory looks during online language processing are driven by both local priming from preceding words as well as active prediction is also entirely consistent with the findings of Kukona et al. (2011) who demonstrated that following a constraining verb such as *arrest*, people fixate on the image associated with a potential agent (*policeman*) as well as a potential patient (*crook*), even if the agent role is already filled by another NP (e.g., *Toby*). These results clearly suggest that local priming from the verb triggers looks to all associated items in a visual scene independent of preceding syntactic constraints. However, the authors also reported an advantage for looks to the patients over the agents, indicating that active prediction was also at work (see also Chow, Smith, Lau, & Phillips, 2016).

9. General discussion

In this study, we examined the processing of fluent and disfluent speech as a means to investigate two psycholinguistic questions. Our primary question was whether comprehenders make semantic and phonological predictions differentially depending on surrounding contextual cues, and our second question concerned the extent to which the predictive system relies on semantic versus phonological information to generate predictions. Below, we separately discuss the results of our experiments with regards to these two questions.

9.1. Differential prediction

The results of our norming studies established that comprehenders possess the tacit knowledge that speakers can generate both semantic and phonological repairs following a disfluency (*the hammer, uh I mean…*) and that semantic repairs are more plausible than phonological repairs following a focus construction (*not the hammer but the…*). However, during online language processing, the participants failed to access and/or engage this knowledge to flexibly predict likely repairs. Instead, anticipatory looks to semantic and phonological competitors were highly similar for both repair disfluencies and focus constructions. This pattern of results suggests that although the brain records and maintains the statistical regularities of the linguistic input, it may not be able to use those regularities in real time to guide prediction (cf. Clark, 2013; Kleinschmidt et al., 2012; Levy, 2008a; McDonald & Shillcock, 2003).

As mentioned above, Huettig (2015) distinguishes between "dumb" (Type I) and "smart" (Type II) routes to prediction (see also Kahneman, 2011). "Dumb" predictions are generated based on automatic spreading activation from the preceding context, whereas "smart" predictions are driven by "active reasoning", inference generation, or the combination of multiple contextual cues. This distinction might correspond to two ends of a continuum of prediction, with predictions falling in a range extending from spreading activation based on a single word, spreading activation based on multiple words, a combination of spreading activation and context-dependent prediction, and finally, active reasoning and inference-based prediction based on multiple linguistic and/or non-linguistic cues. Viewed in this way, the absence of differential predictions within disfluency contexts (Experiments 2 and 3), as well as the greater probability of overall anticipatory looks to the semantic and phonological competitors within disfluency contexts (Experiment 3) relative to coordination and silence contexts (Experiment 4) are explained simultaneously: The prediction system is smart enough to take disfluency and focus constructions into account to generate more predictions under these contexts compared to coordination and silence contexts, but not smart enough to generate during online processing differential semantic vs. phonological predictions based on disfluency vs. focus constructions.

Recall that the probability of looking to all potential substitutions (i.e., NP2, the phonological competitor, the semantic competitor and the distracter) was reliably greater in Experiment 3, which encouraged prediction via disfluency and focus contexts, relative to Experiment 4, in which the linguistic signal did not encourage prediction. Compared to the non-predictive contexts employed in Experiment 4, the presence of disfluency signals (such as "uh I mean") and focus elements ("not the X but rather the Y") triggered participants to actively entertain upcoming competitors, which suggest that the prediction system is "smart" enough to take the context into account. Despite this *overall* increase in predictive looks, however, the *relative* difference between looks to the potential substitutions remained very similar, suggesting that the prediction machinery is not "smart" enough to vary the content of prediction based on the specific type of sentence construction it encounters during online language processing.

In the current experiments, we argued that predictive looks to semantic and phonological competitors are likely driven by both spreading activation from the reparandum as well as the use of contextual information. This raises the question as to how much of previous findings supporting context-based prediction might be reduced to spreading activation from one or multiple words. In some instances, such as in Altmann and Kamide (1999)'s study, predictive looks to the image of a *cake* following *The boy will eat…* might also have arisen from semantic associations between the target word (*cake*) and the preceding verb (*eat*). Indeed, when examining the stimuli employed by Altmann and Kamide (1999), we observed higher LSA semantic similarity between the verbs and predicted targets (*eat – cake*, Mean = 0.38) than for those verbs and the average of all the distracters (*eat – train*, Mean = 0.10, $t(15) = 4.02$, $p < .001$). It is an open question, then, whether a Type I priming-based mechanism can account for some of the findings in the prediction literature.

Our results concerning differential predictions also revealed a dissociation between offline and online use of stored memory representations. Specifically, across two offline norming tasks, participants exhibited differential preferences for semantic and phonological repairs depending on whether they appeared in a disfluency or in a focus construction. However, this pattern failed to emerge during online prediction in the visual world task. These results may suggest that the representations associated with the statistical co-occurrences of word substitutions and the contexts in which they occur might be stored in memory *passively*, such that they may be accessed during offline processing, but not during online processing. There is some precedence for observing dissociations between offline and online measures of prediction. In a series of studies, Chow and colleagues examined the processing of "canonical" sentences and their role-reversed counterparts ("*…which customer the waitress served*" vs. "*…which waitress the customer served*"). While role-reversed sentences produced lower cloze probabilities in an offline completion task, the authors observed *no* differences in the amplitude of the N400 during online reading (Chow, Smith, Lau, Phillips, 2015). Critically though, when readers were given additional time to access the critical argument-role information, N400 effects did emerge (Chow, Lau, Wang, & Phillips, 2018). In future studies it would be valuable to examine whether pragmatic information regarding disfluencies may also be available following a delay. Similar discrepancies between offline and online processing have been reported by studies examining aspects of language processing other than prediction. For instance, using offline measures, Davies and Katsos (2013) and Engelhardt, Bailey, and Ferreira (2006) found that comprehenders are sensitive to violations of the Gricean maxims of cooperative language use. However, Fukumura and van Gompel (2017) found no support for sensitivity to Gricean maxims during online reading.

Our results delimit the current theories of prediction during sentence processing according to which pre-stored statistical probabilities should be readily available during online processing to guide prediction (e.g., Clark, 2013; Jaeger, 2010; Levy, 2008a). Note that the failure to use such pre-stored information is sometimes attributed to the "noise" in the language processing system (the noisy-channel model of language processing; e.g., Jaeger & Levy, 2007). Such noise might stem from lapses in the attentional and/or perceptual systems, or from the environment, among other sources. Relevant to our results, such noise might have hindered the availability of pre-stored, statistically-learned information for prediction in real time. Although such a scenario is a logical possibility, it is very difficult to falsify, and we therefore argue that the prediction system might not always be able to access the memory representations associated with statistical regularities of the linguistic input. However, it is important to note that the current study is, to the best of our knowledge, the first study that investigates differential phonological vs. semantic predictions based on disfluency cues. As such, we acknowledge that our results are limited to the two contextual cues that we employed in our experiments (i.e., repair disfluencies and focus constructions). More research is needed to investigate whether other contextual cues could lead to differential predictions during online language processing.

Given that statistical distributions of words and contexts have been shown to influence online language processing (e.g., Kaschak & Glenberg, 2004; Kleinschmidt et al., 2012; McDonald & Shillcock, 2003), an obvious question is why would such statistical probabilities not be employed during the processing of disfluent speech or focus constructions? We argue that this could be because focus constructions and disfluencies likely require cognitive control processes such as suppression, or updating of the contents of working memory (Miyake & Friedman, 2012), which may engage working memory resources (Engelhardt, Corley, Nigg, & Ferreira, 2010; Engelhardt, Nigg, & Ferreira, 2013; Husband & Ferreira, 2016; Osaka, Nishizaki, Komori, & Osaka, 2002). Limited working memory resources might in turn weaken predictive processing. Specifically, during the processing of repair disfluencies and focus constructions, the memory representation associated with the reparandum must either be actively suppressed, or the contents of working memory should be updated (such that the reparandum is no longer part of the current discourse representation). Implementing such cognitive control processes and predicting upcoming information at the same time might strain working memory, limiting its capacity to predict. The possibility that working memory limitations might be the reason why differential predictions did not occur in our study is consistent with research showing that individuals with higher verbal memory capacities are better predictors than those with smaller working memory capacities (Huettig & Janse, 2016; Mani & Huettig, 2012), and also with research showing that older adults are less capable of predicting upcoming information than are younger adults (Federmeier, Kutas, & Schul, 2010; Lau, Holcomb, & Kuperberg, 2013), presumably because cognitive control processes decline with age (e.g., Braver & Barch, 2002; Braver et al., 2001). In addition to working memory span, other individual differences such as processing speed (Kukona et al., 2016) and vocabulary knowledge (Borovsky, Elman, & Fernald, 2012) have also been shown to influence prediction. These cognitive factors are relevant to our results because they might interact with working memory capacity in influencing the ability to predict. For example, individuals with higher processing speeds and/or broader vocabulary knowledge might not need as large a working memory to generate predictions.

A potential concern about our results is that because the NP2s (i.e., the repairs) were slightly semantically related to NP1, and because few sentences actually resolved to phonological substitutions (six fillers in total, see above), the lack of differential expectations might have been caused by the global experimental environment, meaning that even if participants started out with expectations for phonological substitutions, the overall few number of phonological substitutions could have weakened any initial tendency over the course of the experiment. However, such an explanation of our findings is unlikely for two main reasons. First, if prediction of phonological competitors were undermined in our experiments, participants should have always looked at NP2s (i.e., repairs) more than phonological competitors. However, although repairs were reliably more semantically related to NP1s, participants still looked significantly more at the phonological competitors than the repairs (see the results of Experiment 3). Second, in Experiments 1A and 1C, we did not have any fillers that resolved to phonological substitutions and nevertheless we observed the critical interaction between Sentence Type and Image Type. Given these features of our results, we believe expectations for phonological competitors were not unduly dampened in our experiments. Of course, this is an issue to be explored more definitively in future work.

To the best of our knowledge, our study is the first to systematically assess differential semantic versus phonological predictions. Despite the clear and consistent results we obtained, clearly more research from different perspectives is needed, including investigation of other linguistic and non-linguistic contexts, and exploration of these issues in languages other than English. Moreover, because previous research has shown that the emergence of effects that are as well-established as lexical frequency depends on the method employed in the study (e.g., Kretzschmar, Schlesewsky, & Staub, 2015), future research might also examine differential predictions using a different methodology (e.g., self-paced/natural reading, electroencephalography/ERPs) to assess the robustness of the current set of results. Our hope is that our findings provide a clear path forward for future work on prediction generally and on the content of predictions made at different representational levels in the language system.

9.2. The magnitude of semantic vs. phonological predictions

Our results also have implications for the relative strength of meaning-based and form-based predictions. In the combined data from Experiments 2 and 3, we observed that phonological competitors received more anticipatory looks than did random distractors (Experiment 2), or repairs (Experiments 2 and 3 combined), suggesting that prediction during language comprehension could involve pre-activation of phonological features of the predicted words (DeLong et al., 2005; Laszlo & Federmeier, 2009). At the same time, we also observed reliably more looks to the semantic distractor than to the phonological distractor, suggesting that prediction is more robust at the semantic level than at the phonological level (Ito et al., 2016; Pickering & Garrod, 2007, 2013). The fact that we observed more anticipatory looks to the phonological competitor than to the repair (i.e., NP2), but fewer anticipatory looks to the phonological competitor than to the semantic competitor, provides support for an efficient language processing system where plausible candidates are predicted based on whatever information is currently most available (Fine & Jaeger, 2013; Jaeger & Tily, 2011; Jaeger, 2010; Kurumada & Jaeger, 2015). Interestingly, these results are compatible with a continuous range of prediction mechanisms: When there is strong phonological overlap between the reparandum and a potential repair, the phonological features of the potential repair are sufficiently activated to cause predictive saccades to a phonologically related competitor. In other words, prediction of phonologically related words is possible if the cue for such prediction is sufficiently strong, as is the case of a reparandum-repair pair.

Given the relatively mixed literature on phonological prediction, we acknowledge that our results may not be generalizable to all situations, and that generating predictions based on the phonology of preceding words might depend on various factors such as the idiosyncrasies of the specific task at hand, the availability of phonological information and the utility of such predictions. Therefore, although our results provide support for form-based prediction during online language processing, such predictions might be limited to situations where form-based information is highly and directly available (such as through a preceding reparandum).

Moreover, the phonological prediction effects observed in the current study differed in some ways from the way phonological prediction has been previously assessed in the literature (DeLong et al., 2005; Drake & Corley, 2015). While previous studies have used semantic context to constrain the form of a specific article (DeLong et al., 2005) or noun (Drake & Corley, 2015; Ito et al., 2016), we used phonological overlap between a reparandum and repair (*hammer* - *hammock*) to induce form-based predictions. Based on the current experiments, we believe it is still an open question whether, in the absence of explicit phonological priming, semantic constraints routinely result in the pre-activation of form level features (see Brothers et al., 2015; Nieuwland et al., 2017 for a discussion). Nonetheless, from the current results, we can conclude that semantic and (to a lesser extent) phonological overlap can guide predictive eye movements in a visual scene.

9.3. Conclusion

Taken together, our results show that comprehenders have implicit knowledge of the types of words that typically follow speech disfluencies, but that listeners cannot readily access these memory representations to make differential semantic vs. phonological predictions during online processing. This finding imposes important limits on the theories of predictive processing regarding the statistical regularities readily available for generating predictions during online spoken language comprehension. We also found the prediction during sentence processing is primarily driven by semantics. However, if phonological information is sufficiently activated due to the nature of the linguistic construction, anticipation of form can override prediction of meaning.

Appendix A. The frequency and percentages of choosing each picture in Experiment 1C.

Sentence Type	Image	Frequency	Percentage
Disfluency	Reparandum	20	1.9%
Disfluency	Repair	108	10.2%
Disfluency	Phon Competitor	640	60.6%
Disfluency	Sem competitor	288	27.3%
Focus	Reparandum	12	1.2%
Focus	Repair	134	12.7%
Focus	Phon Competitor	447	42.3%
Focus	Sem competitor	463	43.8%

Appendix B. The critical stimuli for experiments 2, 3 and 4. The images on the visual display for each item are listed beneath each item and include the reparandum (NP1), the phonological competitor, the semantic competitor, the repair (N2) and the distracters, in that order. The distracters were only used in experiments 3 and 4 (one of the two was present in a display). The stimuli for the Coordination condition of Experiment 4 can be constructed by replacing “uh I mean” with “and also” in the disfluency condition. The stimuli for the Silence condition of Experiment 4 can be constructed by removing all the text after the reparandum (NP1) in the disfluency condition.

Item	Disfluency	Focus
1	Off in the distance we spied a towel uh I mean a cooler lying on the beach.	Off in the distance we spied not a towel but rather a cooler lying on the beach.
Images	towel/tower/bath/cooler/(bullets-gum)	
2	Because he was so careless, the boy broke the table uh I mean the clock when he was running.	Because he was so careless, the boy broke not the table but rather the clock when he was running.
Images	table/tablet/chair/clock/(horseshoe-sandal)	
3	The craftsman was making a saddle uh I mean a bench in his workshop.	The craftsman was making not a saddle but rather a bench in his workshop.
Images	saddle/sandal/horse shoe/bench/(pump-reel)	
4	The children playing in the park walked toward the paddle uh I mean the bat over in the grass.	The children playing in the park walked toward not the paddle but rather the bat over in the grass.
Images	paddle/puddle/kayak/bat/(chicken-snake)	
5	Her husband left the papers beside the sofa uh I mean the fridge.	Her husband left the papers beside not the sofa but rather the fridge.
Images	sofa/soda/bed/fridge/(yarn-knob)	
6	The tourist took a picture of a sailor uh I mean a runner using his new camera.	The tourist took a picture of not a sailor but rather a runner using his new camera.
Images	sailor/tailor/fisherman/runner/(chair-tablet)	
7	The little girl hated the skirt uh I mean the headband because it was so ugly.	The little girl hated not the skirt but rather the headband because it was so ugly.

Images	skirt/skate/socks/headband/(palmtree-bleach)	
8	The well-known Frenchman made the best soap uh I mean candles which smelled amazing.	The well-known Frenchman made not the best soap but rather the best candles which smelled amazing.
Images	soap/soup/shampoo/candle/(chariot-hook)	
9	In the storage shed, the man tripped over a wheel uh I mean a toolbox and broke his arm.	In the storage shed, the man tripped over not a wheel but rather a toolbox and broke his arm.
Images	wheel/reel/pump/toolbox/(eggplant-tornado)	
10	The worker grabbed the map uh I mean the cell phone to find his way.	The worker grabbed not the map but rather the cell phone to find his way.
Images	map/mop/compass/cell phone/(pants-bell)	
11	In the other room, the man was working with his roller uh I mean his drill while humming.	In the other room, the man was working not with his roller but rather with his drill while humming.
Images	roller/ruler/paint/drill/(pole-fist)	
12	The woman picked up the chain uh I mean the stone while working in the garden.	The woman picked up, not the chain but rather the stone while working in the garden.
Images	chain/cane/lock/stone/(shampoo-soup)	
13	The house fly landed on the plum uh I mean the toast that was on the counter.	The house fly landed on not the plum but rather the toast that was on the counter.
Images	plum/plug/grapes/toast/(arrow-bowl)	
14	The blind man couldn't find the needle uh I mean the iron.	The blind man couldn't find not the needle but rather the iron
Images	needle/noodle/pin/iron/(river-brick)	
15	The housewife found the meat uh I mean the kiwi that she had bought from the market.	The housewife found not the meat but rather the kiwi that she had bought from the market.
Images	meat/mitt/potato/kiwi/(bulb-ramp)	
16	Jack pulled on the knot uh I mean the leash while walking his dog.	Jack pulled on not the knot but rather the leash while walking his dog.
Images	knot/knob/yarn/leash/(grapes-plug)	
17	The teacher used the large cube uh I mean globe in her demonstration.	The teacher used not the cube but rather the globe in her demonstration.
Images	cube/tube/rectangle/globe/(socks-skate)	
18	The little boy was searching around for the kite uh I mean the ball that he had lost.	The little boy was searching around for not the kite but rather the ball that he had lost.
Images	kite/coat/airplane/ball/(bed-soda)	
19	The billionaire bought a still-life painting of a harp uh I mean a bracelet.	The billionaire bought a still-life painting of not a harp but rather a bracelet.
Images	harp/heart/piano/bracelet/(beans-mitt)	
20	The man tried to clean the mud off the hoof uh I mean the shovel.	The man tried to clean the mud off not the hoof but rather the shovel.
Images	hoof/hook/chariot/shovel/(sail-bolt)	
21	The seamstress was sewing the puppet uh I mean the hat that had a tear in it.	The seamstress was sewing not the puppet but rather the hat that had a tear in it.
Images	puppet/pocket/doll/hat/(slippers-rope)	
22	Aeron said the beach uh I mean the store was closed for the afternoon.	Aeron said not the beach, but rather the store was closed for the afternoon.
Images	beach/bleach/palm tree/store/(kayak-puddle)	
23	My father said that the beer uh I mean the hotdog was on sale.	My father said that not the beer but rather the hotdog was on sale.
Images	beer/bear/wine/hotdog/(piano-heart)	
24	My brother wanted to show us the belt uh I mean the glasses that he bought.	My brother wanted to show us not the belt but rather the glasses that he bought.
Images	belt/bell/pants/glasses/(starfish-shelf)	
25	Bill couldn't find the boat uh I mean the car that was in a wreck.	Bill couldn't find not the boat but rather the car that was in a wreck.
Images	boat/bolt/sail/car/(lock-cane)	
26	The expert was carving a bow uh I mean a shield out of wood.	The expert was carving, not a bow, but rather a shield out of wood.
Images	bow/bowl/arrow/shield/(chair-deck)	
27	The engineer knew that the bridge uh I mean the roof was very unstable.	The engineer knew that, not the bridge, but rather the roof was very unstable.
Images	bridge/brick/river/roof/(nail-hammock)	
28	The man was looking around for a hammer I mean uh a yardstick.	The man was looking around not for a hammer but rather a yardstick.
Images	hammer/hammock/nail/yardstick/(rectangle-tube)	
29	There was something wrong with the lamp uh I mean the stapler at the office.	There was something wrong not with the lamp but rather with the stapler at the office.
Images	lamp/ramp/bulb/stapler/(plate-train)	
30	He was examining the letter, I mean uh the book that was on the counter.	He was examining, not the letter, but rather the book that was on the counter.
Images	letter/ladder/stamp/book/(fisherman-tailor)	
31	We took a picture of the sheep uh I mean the fence at the edge of the field.	We took a picture of not the sheep but rather the fence at the edge of the field.
Images	sheep/sheet/cow/fence/(purse-walnut)	
32	The boys are near the cannon uh I mean the hydrant over on the left.	The boys are near not the cannon but rather the hydrant over on the left.
Images	cannon/canyon/rifle/hydrant/(stream-rake)	
33	The can of oil splashed on the door uh I mean the rug when he dropped it.	The can of oil splashed not on the door but rather on the rug when he dropped it.
Images	door/drawer/window/rug/(bus-trunk)	
34	My grandmother saw the tomato uh I mean the roses and was very impressed.	My grandmother saw not the tomato but rather the roses and was very impressed.
Images	tomato/tornado/eggplant/roses/(bath-tower)	
35	We found the sweater in the truck uh I mean the backpack where Linda had left it.	We found the sweater not in the truck but rather in the backpack where Linda had left it.
Images	truck/trunk/bus/backpack/(rifle-canyon)	
36	He fell down by the lake uh I mean the tent when he had a stroke.	He fell down, not by the lake, but rather by the tent when he had a stroke.
Images	lake/rake/stream/tent/(window-drawer)	

37	His younger cousin was staring at the tray uh I mean the turkey that was being carved.	His younger cousin was staring, not at the tray but rather at the turkey that was being carved.
Images	tray/train/plate/turkey/(pin-noodle)	
38	They had to repair the desk uh I mean the keyboard that was broken.	They had to repair, not the desk but rather the keyboard that was broken.
Images	desk/deck/chair/keyboard/(airplane-coat)	
39	The woman was talking about the steak uh I mean the martini because it was delicious.	The woman was talking, not about the steak, but rather about the martini because it was delicious.
Images	steak/snake/chicken/martini/(doll-pocket)	
40	We noticed the man's fish uh I mean his salad that looked pretty tasty.	We noticed not the man's fish but rather his salad that looked pretty tasty.
Images	fish/fist/pole/salad/(paint-ruler)	
41	The art student was sketching a wallet uh. I mean a scarf using a black pen.	The art student was sketching, not a wallet, but rather a scarf using a black pen.
Images	wallet/walnut/purse/scarf/(cow-sheet)	
42	Now my sister is inspecting the shell uh I mean the branch that we found on our walk.	Now my sister is inspecting, not the shell, but rather the branch that we found on our walk.
Images	shell/shelf/starfish/branch/(wine-bear)	
43	Johnnie liked the color of the gun uh I mean the keys that the man was carrying.	Johnnie liked the color of not the gun but rather the keys that the man was carrying.
Images	gun/gum/bullets/keys/(stamp-ladder)	
44	At the back of the closet he found a robe uh I mean a safe near the wall.	At the back of the closet he found, not a robe, but rather a safe near the wall.
Images	robe/rope/slippers/safe/(compass-mop)	

Appendix C. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cogpsych.2019.04.001.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language, 38*(4), 419–439.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247–264.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*(4), 457–474.

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, 68*(3), 255–278.

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology, 112*(4), 417–436.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech, 44*(2), 123–147.

Braver, T. S., & Barch, D. M. (2002). A theory of cognitive control, aging cognition, and neuromodulation. *Neuroscience & Biobehavioral Reviews, 26*(7), 809–817.

Braver, T. S., Barch, D. M., Keys, B. A., Carter, C. S., Cohen, J. D., Kaye, J. A., ... Jagust, W. J. (2001). Context processing in older adults: Evidence for a theory relating cognitive control to neurobiology in healthy aging. *Journal of Experimental Psychology: General, 130*(4), 746.

Brill, E., & Moore, R. C. (2000, October). An improved error model for noisy channel spelling correction. In Proceedings of the 38th annual meeting on association for computational linguistics (pp. 286–293). Association for Computational Linguistics.

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition, 136*, 135–149.

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language, 93*, 203–216.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*(2), 234.

Chow, W. Y., Smith, C., Lau, E., & Phillips, C. (2016). A "bag-of-arguments" mechanism for initial verb predictions. *Language, Cognition and Neuroscience, 31*(5), 577–596.

Chow, W. Y., Lau, E., Wang, S., & Phillips, C. (2018). Wait a second! Delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience,* 1–26.

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.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6*(1), 84–107.

Davies, C., & Katsos, N. (2013). Are s“peakers and listeners 'only moderately Gricean'? An empirical response to Engelhardt et al. (2006). *Journal of Pragmatics, 49*(1), 78–106.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2017). Is there a replication crisis? Perhaps. Is this an example? No: A commentary on Ito, Martin & Nieuwland (2016). *Language, Cognition, and Neuroscience*. https://doi.org/10.1080/23273798.2017.1279339.

Donnelly, S., & Verkuilen, J. (2017). Empirical logit analysis is not logistic regression. *Journal of Memory and Language, 94*, 28–42.

Drake, E., & Corley, M. (2015). Effects in production of word pre-activation during listening: Are listener-generated predictions specified at a speech-sound level? *Memory & Cognition, 43*(1), 111–120.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*(6), 641–655.

Engelhardt, P. E., Bailey, K. G., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language, 54*(4), 554–573.

Engelhardt, P. E., Corley, M., Nigg, J. T., & Ferreira, F. (2010). The role of inhibition in the production of disfluencies. *Memory & Cognition, 38*(5), 617–628.

Engelhardt, P. E., Nigg, J. T., & Ferreira, F. (2013). Is the fluency of language outputs related to individual differences in intelligence and executive function? *Acta*

Psychologica, 144(2), 424–432.
Federmeier, K. D., & Kutas, M. (1999). Right words and left words: Electrophysiological evidence for hemispheric differences in meaning processing. *Cognitive Brain Research, 8*(3), 373–392.
Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language, 115*(3), 149–161.
Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research, 1146*, 75–84.
Ferreira, F., & Chantavarin, S. (2018). Integration and prediction in language processing: A synthesis of old and new. *Current Directions in Psychological Science, 27*(6), 443–448.
Fine, A. B., & Jaeger, F. T. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science, 37*(3), 578–591.
Frisson, S., Rayner, K., & Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 862.
Fukumura, K., & van Gompel, R. P. (2017). How do violations of Gricean maxims affect reading? *Journal of Memory and Language, 95*, 1–18.
Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116*(2), 439.
Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences* 201216438.
Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science, 24*(7), 1079–1088.
Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson, & F. Ferreira (Eds.). *The integration of language, vision, and action: Eye movements and the visual world* (pp. 1–58). New York: Psychology Press.
Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*(2), 151–171.
Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research, 1626*, 118–135.
Huettig, F., & Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition, 96*(1), B23–B32.
Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience, 31*(1), 80–93.
Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language, 57*(4), 460–482.
Husband, E. M., & Ferreira, F. (2016). The role of selection in the comprehension of focus alternatives. *Language, Cognition and Neuroscience, 31*(2), 217–235.
Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language, 86*, 157–171.
Ito, A., Martin, A. E., & Nieuwland, M. S. (2017). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience, 32*(8), 954–965.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446.
Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology, 61*(1), 23–62.
Jaeger, T. F., & Levy, R. P. (2007). Speakers optimize information density through syntactic reduction. In Advances in neural information processing systems (pp. 849–856).
Jaeger, T. F., & Tily, H. (2011). On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(3), 323–335.
Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press, Clarendon Press.
Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.
Kaschak, M. P., & Glenberg, A. M. (2004). This construction needs learned. *Journal of Experimental Psychology: General, 133*(3), 450.
Kretzschmar, F., Schlesewsky, M., & Staub, A. (2015). Dissociating word frequency and predictability effects in reading: Evidence from coregistration of eye movements and EEG. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(6), 1648.
Kukona, A., Braze, D., Johns, C. L., Mencl, W. E., Van Dyke, J. A., Magnuson, J. S., ... Tabor, W. (2016). The real-time prediction and inhibition of linguistic outcomes: Effects of language and literacy skill. *Acta Psychologica, 171*, 72–84.
Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148.
Kleinschmidt, D.F., Fine, A.B., Jaeger, T.F., 2012. A belief-updating model of adaptation and cue combination in syntactic comprehension. In: *Proceedings of the annual meeting of the cognitive science society* (Vol. 34, No. 34).
Kukona, A., Fang, S. Y., Aicher, K. A., Chen, H., & Magnuson, J. S. (2011). The time course of anticipatory constraint integration. *Cognition, 119*(1), 23–42.
Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32–59.
Kurumada, C., & Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language, 83*, 152–178.
Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association.
Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. *Predictions in the Brain: Using our Past to Generate a Future,* 190–207.
Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language, 61*(3), 326–338.
Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience, 25*(3), 484–502.
Levy, R. (2008a). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177.
Levy, R. (2008, October). A noisy-channel model of rational human sentence comprehension under uncertain input. In Proceedings of the conference on empirical methods in natural language processing (pp. 234–243). Association for Computational Linguistics.
Lowder, M. W., & Ferreira, F. (2016a). Prediction in the processing of repair disfluencies: Evidence from the visual-world paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(9), 1400.
Lowder, M. W., & Ferreira, F. (2016b). Prediction in the processing of repair disfluencies. *Language, Cognition and Neuroscience, 31*(1), 73–79.
MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes, 9*(2), 157–201.
MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676.
Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake – But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance, 38*(4), 843–847.
McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science, 14*(6), 648–652.
Mirman, D., & Graziano, K. M. (2012). Individual differences in the strength of taxonomic versus thematic relations. *Journal of Experimental Psychology: General, 141*(4), 601.
Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions four general conclusions. *Current Directions in Psychological Science, 21*(1), 8–14.
Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of*

Experimental Psychology: General, 106(3), 226.
Nieuwland, M., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... Mézière, D. (2017). Limits on prediction in language comprehension: A multi-lab failure to replicate evidence for probabilistic pre-activation of phonology. *bioRxiv,* 111807.
Osaka, M., Nishizaki, Y., Komori, M., & Osaka, N. (2002). Effect of focus on verbal working memory: Critical role of the focus word in reading. *Memory & Cognition, 30*(4), 562–571.
Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods,* 1–9.
Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11*(3), 105–110.
Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(04), 329–347.
Rommers, J., Meyer, A. S., Praamstra, P., & Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia, 51*(3), 437–447.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237.
Saslow, M. G. (1967). Latency for saccadic eye movement. *JOSA, 57*(8), 1030–1033.
Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology, 45*(4), 447–481.
Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632.
Tree, J. E. F. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language, 34*(6), 709.
Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*(3), 291–298.
Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(1), 1.