12-17-2018

I See What You Meant to Say: Anticipating Speech Errors During Online Sentence Processing

Matthew W. Lowder
University of Richmond, mlowder@richmond.edu

Fernanda Ferreira

I See What You Meant To Say:

Anticipating Speech Errors During Online Sentence Processing

Matthew W. Lowder[1] and Fernanda Ferreira[2]

1. University of Richmond
2. University of California, Davis

Address correspondence to:

Fernanda Ferreira
Department of Psychology
Young Hall, One Shields Avenue
University of California, Davis
Davis, CA 95616

fferreira@ucdavis.edu

Abstract

Everyday speech is rife with errors and disfluencies, yet processing what we hear usually feels effortless. How does the language comprehension system accomplish such an impressive feat? The current experiment tests the hypothesis that listeners draw on relevant contextual and linguistic cues to anticipate speech errors and mentally correct them even before receiving an explicit correction from the speaker. In the current visual-world eyetracking experiment, we monitored participants' eye movements to objects in a display while they listened to utterances containing reparandum-repair speech errors (e.g., …*his cat, uh I mean his dog.*…). The contextual plausibility of the misspoken word, as well as the certainty with which the speaker uttered this word, were systematically manipulated. Results showed that listeners immediately exploited these cues to generate top-down expectations regarding the speaker's communicative intention. Crucially, listeners used these expectations to constrain the bottom-up speech input and mentally correct perceived speech errors even before the speaker initiated the correction. The results provide powerful evidence regarding the joint process of correcting speech errors that involves both the speaker and the listener.

Keywords: speech errors; repairs; prediction; eye movements

Imagine you and a colleague are writing an important email, and your colleague says, “Please hit send, uh I mean save.” Although comprehension of this utterance feels effortless, successful interpretation relies on the rapid and efficient coordination of mental processes that access and combine phonological, lexical, syntactic, and semantic information. An extra challenge in this case involves understanding that *send* was a speech error and that the listener’s initial interpretation to send the email should be disregarded and replaced by the speaker’s intended request to save the email. A central goal of psycholinguistics is to explain the incremental, moment-by-moment operations that give rise to spoken language comprehension, including those that support people’s ability to go beyond the input and recover the speaker’s intention. By far the most common methodological approach to achieving this goal is to use the so-called *visual-world paradigm*, pioneered by Cooper (1974) and popularized by Tanenhaus et al. (1995). In this paradigm, participants’ eye movements are tracked as they listen to spoken language while simultaneously viewing visual displays that contain relevant objects. The probability of fixating a particular object within a particular time window then serves as the dependent measure that informs us about the timescale on which linguistic interpretations are activated. The crucial linking hypothesis is that the bottom-up input associated with hearing the word activates the word’s representation in memory, which automatically triggers a saccade to the relevant object in the display (Allopenna et al., 1998; Tanenhaus et al., 2000). This helps explain the finding that listeners direct their gaze toward relevant objects within about 200 ms of hearing a target word, even in the absence of an explicit task.

But successful language comprehension in the real world relies on much more than simply decoding linguistic input and integrating it within a broader context. Instead, speakers and listeners must cooperate with one another in order to optimize communication (see, e.g.,

Clark & Wilkes-Gibbs, 1986; Schober & Clark, 1989). Considering speech errors from this perspective, speaker behaviors signaling a speech error or a correction may be treated by the listener as cues to begin mentally correcting the input, perhaps even before the speaker begins overtly repairing the utterance.

Indeed, more recent theoretical approaches to sentence processing have increasingly emphasized the role of prediction in guiding comprehenders' interpretations prior to receiving the input (for a review, see Kuperberg & Jaeger, 2016). One recent framework for understanding the role of predictive processing in language comprehension is the Noisy Channel model (Gibson et al., 2013), which starts with the assumption that communication involves the transmission of a linguistic signal across a noisy channel ("noise" here is conceptualized as any distortion of the input due to producer, perceiver, or environmental factors). The idea is that because noise regularly distorts the input, comprehenders take an active, forward-looking role during language processing, combining the input with relevant linguistic and contextual knowledge and mentally correcting perceived errors. Importantly, up to now these theoretical claims have largely been investigated in reading experiments. In one experiment, for example, participants who read an implausible sentence (*The mother gave the candle the daughter*) in which the insertion of a short function word would make it plausible (*The mother gave the candle [to] the daughter*) were likely to make this mental edit and adopt the more plausible interpretation (Gibson et al.).

Although it certainly seems reasonable that mental corrections of this sort occur when we read, it is likely more common for comprehenders to encounter errors in the domain of spoken language, in which speakers regularly produce filled pauses such as *uh* and *um* or begin their utterances spontaneously without planning out the full scope of the sentence and then make errors and have to backtrack. Indeed, analyses of spontaneous speech reveal that speakers tend to

produce between six and 10 disfluencies for every 100 words (Bortfeld et al., 2001; Fox Tree, 1995). Although the majority of these disfluencies tend to be filled pauses or repetitions, speakers also reliably produce errors that they then go back and repair (see, e.g., Bortfeld et al., Lau & Ferreira, 2005; Shriberg, 1996). Moreover, in contrast to the examples of outright ill-formedness that have been the target of research within the Noisy Channel framework[1], the study of disfluencies holds promise as a more ecologically valid approach to understanding how listeners deal with input that is noisy—"noisy" in the general sense that some of the content might fail to reflect the speaker's communicative intention and therefore will need to be reinterpreted and reconstructed by the listener.

Investigations of disfluency demonstrate that listeners do not process disfluencies passively, but instead use them as information to help reconstruct the speaker's communicative intention. For example, listeners attempt to assess the causes of speaker's filled pauses, treating them as evidence the speaker is about to refer to a new entity in the discourse (Arnold et al., 2004, 2007), and as a cue to speaker truthfulness and reliability (Loy et al., 2017). The Noisy Channel framework suggests that listeners will also take an active role in the process of error correction, zeroing in on cues that signal a likely speech error and attempting to uncover the speaker's intended meaning. At the same time, speakers also monitor their output for coherence and will often explicitly signal that an error has occurred. From this perspective, the listener and speaker work together to optimize communication, with the speaker explicitly repairing speech errors and the listener exploiting any cues that suggest a speech error has occurred and mentally correcting it, perhaps even before the speaker initiates the correction.

[1] A notable exception is the field of computational linguistics, which has successfully applied this framework to the processing of spoken language (see, e.g., Honal & Schultz, 2003; Johnson & Charniak, 2004; Zwarts et al., 2010).

To make this more concrete, consider another example utterance: *Bill likes to throw the Frisbee in the park with his uhh cat, uh I mean his dog.…* The speaker here notices he has made an error and initiates the process of explicitly replacing the reparandum (*cat*) with the repair (*dog*). However, if the listener is actively modeling the speaker's communicative intentions, she may mark *cat* as a potential error and mentally repair it with the more plausible *dog*. This example contains two potentially relevant cues to the listener. First, listeners may rely on their real-world knowledge that people typically play in the park with dogs rather than cats to rapidly judge *cat* as implausible, anticipate that this is an error, and correct it with the more plausible *dog*. It is well known that semantic information about the utterance is used to predict upcoming words (e.g., Altmann & Kamide, 1999). In contrast, the crucial question we ask here is whether the listener's higher-level model of what the speaker is trying to communicate can constrain and potentially override the bottom-up input when that input is perceived as a speech error. Second, the presence of the filled pause (*uhh*) immediately before *cat* may serve as an additional cue to the listener. Specifically, we predicted that listeners would interpret this signal as evidence that the speaker was having trouble producing the intended word, which may then make listeners more likely to mark it as a potential error, regardless of the semantic context. This prediction follows from previous findings demonstrating that speakers tend to use filled pauses when uncertain about an upcoming word (Brennan & Williams, 1995; Smith & Clark, 1993). Further, Fraundorf and Watson (2011) have proposed that filled pauses are effective attention-orienting devices, based on their evidence that memory for complex discourses tends to be enhanced when the discourse contains fillers. In fact, Fraundorf and Watson suggested that one reason for this effect might be that the presence of filled pauses leads listeners to infer that the speaker is not

confident in the utterance, which may then prompt the listener to devote extra effort to comprehension.

Although we predicted that the filled pause would lead the listener to be less confident in the speaker's utterance, regardless of the semantic context, other hypotheses are also possible. For example, Corley et al. (2007) provided evidence using event-related potentials that the presence of a filled pause may lead listeners to reduce the extent to which they use real-world knowledge to predict an upcoming word. On this account, one might predict an interaction between the plausibility of the reparandum and the presence of the filled pause, such that listeners might be more likely to predict an unexpected utterance when the speaker becomes disfluent.

At a broader theoretical level, these ideas align with the Good-Enough language-processing framework (Ferreira et al., 2002; Ferreira & Lowder, 2016). According to this approach, comprehension mechanisms will sometimes fail to deliver a faithful interpretation of a sentence if such an interpretation would violate the comprehender's syntactic or semantic expectations. Up to now, most work conducted within this framework has emphasized distortions and normalizations that are potentially problematic, including misinterpretations based on syntactic errors (Christianson et al., 2001; Ferreira, 2003), and shallow interpretations that fail to distinguish between closely related but distinct concepts (Barton & Sanford, 1993). As Ferreira (2003) argued, distortions clearly have the potential to undermine effective communication, but these normalizations may sometimes be helpful given that speakers make errors and often either fail to correct them, or correct them only after some delay. It therefore might sometimes be adaptive for the comprehension system to normalize input that is unexpected.

In the current experiment, we investigated the extent to which listeners are sensitive to cues of plausibility and speaker certainty during online sentence processing, and whether they can use these cues to rapidly anticipate and mentally correct errors. Participants listened to sentences containing reparandum-repair disfluencies (…*cat uh I mean dog*…) in which we systematically manipulated the plausibility of the reparandum and whether or not it was preceded by a filled pause. Listeners heard these sentences while viewing four-image displays that included images corresponding to the reparandum, the repair, and two distractor images. By analyzing fixation patterns to the two critical images across the time course of the utterance, we can examine how changes in the plausibility of the reparandum and the speaker's relative certainty about the utterance influence listeners' commitment to the reparandum versus the repair and how these commitments change as the utterance unfolds in real time. Importantly, this approach allows us to directly compare listeners' relative weighting of the bottom-up input that comes from the speech signal versus their top-down expectations regarding the communicative intentions of the speaker, by examining how these sources of information affect where listeners direct their fixations. If listeners rapidly exploit cues about plausibility and speaker certainty to create a strong prediction about what the speaker intends to communicate, then we should observe that listeners show a decreased tendency to fixate the reparandum picture (e.g., cat) when it is implausible in the context, or is uttered with uncertainty, or both, and instead show an increased tendency to fixate the repair (e.g., dog) even before the speaker starts to articulate an explicit correction. This pattern would provide evidence that listeners use their top-down expectations to constrain the bottom-up speech input and mentally correct the perceived speech error even before receiving any overt signal from the speaker that an error has occurred.

Method

Participants

Thirty-two students at the University of California, Davis participated in this experiment in exchange for course credit. They were all native English speakers and reported normal or corrected-to-normal vision. All participants provided informed consent, and all procedures were approved by the Institutional Review Board at UC Davis.

Materials

Forty sets of experimental materials were created, as in Example (1). Each sentence contained an introductory preamble, a reparandum (e.g., *cat*), an edit interval (*uh I mean*), a repair (e.g., *dog*), and then a continuation of the sentence. The reparandum-repair pairs were strong semantic associates of one another (e.g., *cat-dog*, *salt-pepper*, *cake-pie*) and thus seemed that they might reasonably elicit semantic substitution errors from a speaker. Plausibility was manipulated with respect to the fit of the reparandum with the preceding sentence context: in the Plausible condition, the reparandum fit naturally with what preceded it, whereas in the Implausible condition, the reparandum constituted a semantically odd continuation of the sentence. In contrast, the repair fit plausibly in all contexts. A filled pause immediately preceded the reparandum in the Uncertain condition, but not in the Certain condition. Each set of experimental items was associated with a visual display that consisted of four color images (see Figure 1) representing two critical entities—the reparandum (e.g., a cat) and the repair (e.g., a dog)—as well as two unrelated distractors (e.g., a plant and a dishtowel)[2]. The images used in

[2] Semantic similarity between the reparandum and repair was assessed using latent semantic analysis (Landauer & Dumais, 1997). Mean semantic similarity between reparandum-repair pairs (e.g., cat-dog) was 0.45. Similarity between reparandum-repair pairs was significantly higher than semantic similarity between reparanda and the unrelated distractors (0.13), $t = 9.25$, $p < .001$, as well as semantic similarity between repairs and the unrelated distractors (0.16), $t = 8.19$, $p < .001$.

the experiment came from a combination of images we had collected and used in previous visual-world experiments, as well as Google image searches. The two distractor images for each trial were chosen mostly randomly. In some cases, one or both of the distractor images were consistent with contextual elements of the sentence, but the distractor images were never mentioned in any of the utterances, nor were they meant to be predicted based on the utterance. For example, one item introduced the context of a wedding with the critical utterance being "…the groom, uh I mean the bride…" In addition to the reparandum and repair pictures for this trial (i.e., a groom and a bride), one of the distractor images was a church, whereas the other was a chair. Despite the thematic similarity of some distractor images to the context of the utterance, participants very rarely fixated the distractor images during the critical portion of the utterance, as we will report in the Results section.

1a. *Every Saturday, Bill likes to grab a book and sit on the couch with his cat, uh I mean his dog, where they spend the afternoon.* (Plausible-Certain)

1b. *Every Saturday, Bill likes to grab a Frisbee and go to the park with his cat, uh I mean his dog, where they spend the afternoon.* (Implausible-Certain)

1c. *Every Saturday, Bill likes to grab a book and sit on the couch with his uhh cat, uh I mean his dog, where they spend the afternoon.* (Plausible-Uncertain)

1d. *Every Saturday, Bill likes to grab a Frisbee and go to the park with his uhh cat, uh I mean his dog, where they spend the afternoon.* (Implausible-Uncertain)

To ensure the validity of our plausibility manipulation, we presented written versions of the Plausible-Certain and Implausible-Certain conditions up to and including the reparandum to 24 participants drawn from the same population, none of whom participated in the eyetracking portion of this study. Items were counterbalanced across two lists and randomized. Participants judged how likely they believed the events described by the sentence were on a scale from 1 (highly unlikely) to 7 (highly likely). The Plausible condition (5.80) was rated as significantly more plausible than the Implausible condition (2.80), $t(39) = 15.71$, $p < .001$.

The materials also included a set of 54 filler items representing a variety of different sentence types, each with a corresponding visual array of four images. None of the filler sentences contained speech errors or disfluencies. All sentences were recorded by a female native speaker of American English who was naïve to the purposes of the recordings. Sentences were recorded separately for each condition so as to make them sound as natural as possible. We analyzed the maximum fundamental frequency (F0 max) of the reparandum across conditions and found that F0 max was significantly higher in the Uncertain condition (304 Hz) than the Certain condition (236 Hz), $F(1,39) = 27.73$, $p < .001$. This is consistent with the idea that the speaker was communicating uncertainty about the word that was articulated immediately after the filled paused and tended to produce it with a higher pitch (e.g., *uhh cat*) (see, e.g., Brennan & Williams, 1995; Smith & Clark, 1993). Thus, it is not only the presence or absence of the filled pause that signaled uncertainty to the listener, but also the prosodic features of the reparandum. Crucially, however, there was no effect of Plausibility on F0 max, nor was there an interaction between Plausibility and Certainty (Fs < 2.1, ps > .15). Further, there were no systematic differences in the timing of the critical windows across any conditions (described in detail below).

The items were counterbalanced across four lists. Written versions of all experimental and filler sentences, along with their accompanying images, are available as supplemental online material.

Procedure

Eye movements were recorded with an EyeLink 1000 Plus system (SR Research) at a sampling rate of 1,000 Hz. The tracker was calibrated at the beginning of each session and throughout the session as needed. Participants were told they would view images on the screen

while also listening to sentences. They were explicitly told, “Some of the sentences might contain errors,” but were instructed to try to understand each sentence; there was no explicit task other than to listen to the sentences for meaning. At the start of each trial, a fixation point was presented in the center of the screen. When gaze was steady on this point, the experimenter pressed a button that presented the visual display. After a 3,000 ms delay, the corresponding sentence was presented via headphones. After the sentence finished playing, the images disappeared and the fixation point for the next trial appeared.

Participants were first presented with four of the filler sentences. After this warm-up block, the remaining sentences were presented pseudorandomly under the constraint that no more than two trials from the same condition could be presented consecutively. The locations of the four images within a visual display were randomized on each trial.

Analysis

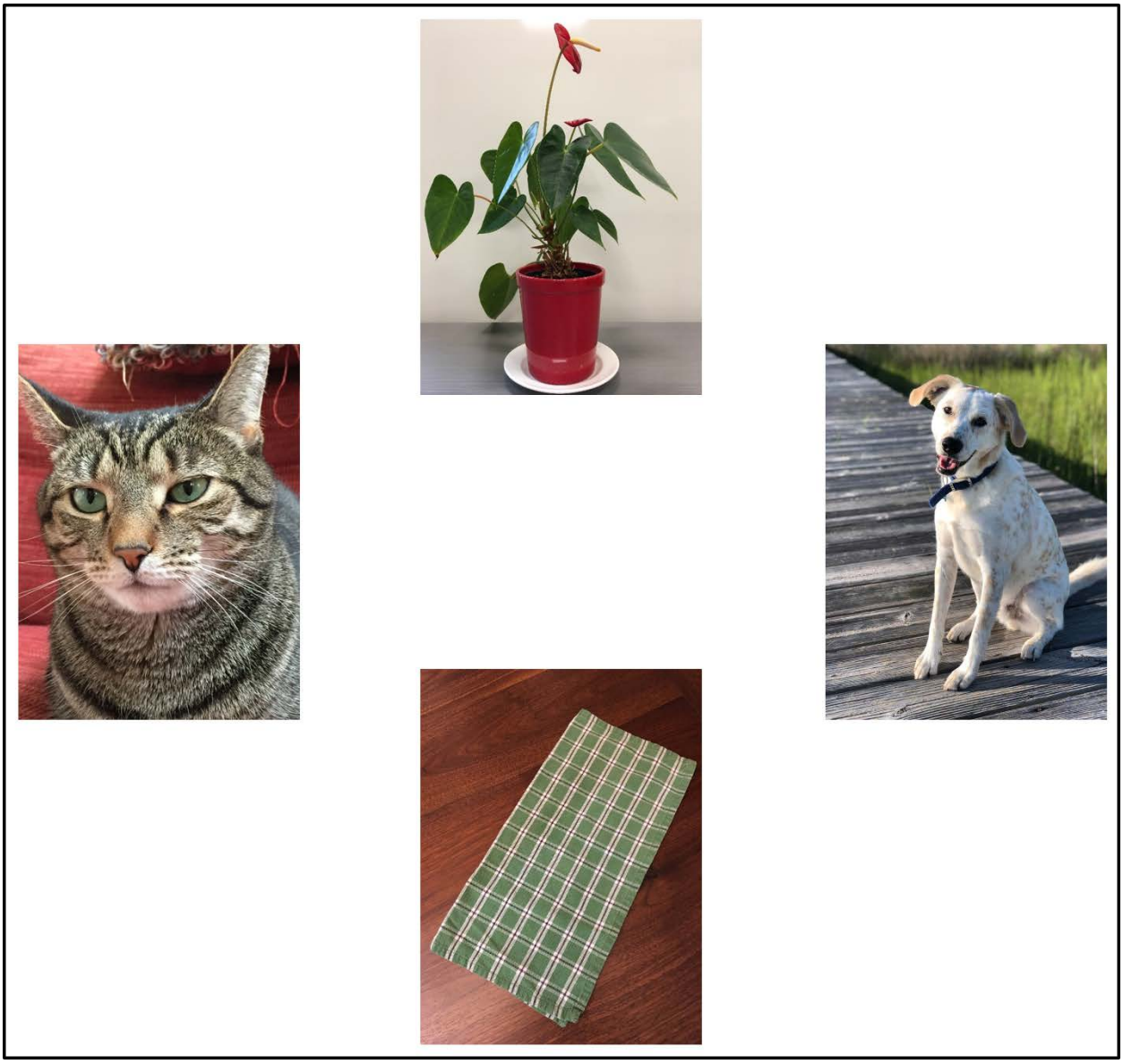
Statistical analyses were performed using the lme4 package in R. The dependent variable was the proportion of fixations per participant per trial to the pictures of interest[3]. We fit linear mixed-effects regression models that included Plausibility, Certainty, and their interaction as fixed effects as well as subjects and items as crossed random effects. Fixed effects were sum-coded. The random-effects structure included the maximally appropriate random intercepts as well as by-subject and by-item random slopes for the factors Plausibility, Certainty, and their interaction. In cases where the maximal model failed to converge, the random-effects structure was sequentially simplified. All *p*-values were obtained using the lmerTest package in R (Kuznetsova et al., 2016).

[3] Analyses were conducted on proportion of fixations so as to align more closely with the fixation plots. Logistic regression analyses using the probability of a fixation within a given time window produced an identical pattern of results.

Within each sound file, we marked the onsets of three critical words: the reparandum, the onset of the edit interval (i.e., *uh I mean*), and the repair. Four windows were constructed around these onsets. Window 1 encompassed the 1,000 ms before the onset of the reparandum. Window 2 measured from the onset of the reparandum to the onset of the edit interval[4]. Window 3 measured from the onset of the edit interval to the onset of the repair. Window 4 measured from the onset of the repair until 800 ms had elapsed. There were no main effects or interactions of Plausibility and Certainty on the duration of these windows across experimental items (Fs < 2.6, ps > .11). All windows were shifted forward 200 ms to account for the time required to launch a signal-based saccade.

[4] The mean duration of Window 2 was 1,075 ms; however, the mean duration of the spoken reparandum was 510 ms. Thus, the speaker tended to utter the reparandum, insert a pause (mean duration of 565 ms), and then initiate the repair.

Figure 1
Example visual display for a trial. The locations of the four images were randomized on each trial.



Results

Fixation plots are presented in Figure 2, and results of the statistical analyses are presented in Table 1. In addition, Figure 3 presents mean proportion of fixations to the two critical pictures (i.e., the reparandum and repair pictures) in each of the four time windows. Visual inspection of Window 1 suggests that in the Plausible condition, listeners tended to consider the two critical pictures as equally likely continuations of the sentence, whereas in the Implausible condition, listeners showed a strong tendency to consider the picture corresponding to the repair. Consistent with this observation, statistical analyses of fixation proportions during this time window revealed significant main effects of Plausibility in looks to both pictures: listeners were more likely to consider the reparandum picture in the Plausible condition versus the Implausible condition, whereas they were more likely to consider the repair picture in the Implausible condition versus the Plausible condition. In addition, there was a significant main effect of Certainty in looks to the reparandum picture with listeners being more likely to look at the reparandum in the Uncertain condition versus the Certain condition. Although inspection of Figure 2 suggests that this main effect of Certainty may have been driven by the Implausible-Uncertain condition, the test of the interaction was nowhere close to being significant ($t = 0.64$). This is consistent with the idea that the listener interpreted the filled pause in the Uncertain condition as a signal that the speaker was not entirely sure what she would say next, and so the listener tended to fixate multiple candidate objects, inflating fixation proportions on the cat.

Within Window 2, there were robust effects of Plausibility and Certainty in looks to both the reparandum and repair pictures. Listeners were more likely to consider the reparandum picture and less likely to consider the repair picture in the Plausible condition versus the

Implausible condition. Listeners were also more likely to consider the reparandum picture and less likely to consider the repair picture in the Certain condition versus the Uncertain condition.

Within Window 3, there was again a significant effect of Certainty in looks to both the reparandum and repair pictures, with listeners being more likely to consider the reparandum and less likely to consider the repair in the Certain condition versus the Uncertain condition. No significant effects emerged in Window 4.

The transition from Window 1 to Window 2 (i.e., before and then after the onset of the reparandum) represents a key test of our hypothesis that listeners would use the presence of a filled pause to put less weight on the speaker's utterance and perform a mental correction of the input. That is, whereas the plausibility manipulation occurred early in the utterance, affecting fixation patterns even before Window 1, the certainty manipulation (i.e., presence versus absence of a filled pause) occurred toward the end of Window 1. Thus, one might predict that upon hearing the onset of the reparandum in Window 2, listeners should show a reduced tendency to look at the reparandum picture in this window in the Uncertain condition versus the Certain condition.

This hypothesis was tested by constructing linear mixed-effects regression models as described above, but including Window as a fixed effect (i.e., Window 1 versus Window 2), along with the fixed effects of Plausibility, Certainty, and their interactions. Regarding proportion of looks to the reparandum picture, there was a significant Certainty-by-Window interaction (estimate = 0.03, SE = 0.01, t = 4.04, $p < .001$) such that there was a smaller increase in looks to the reparandum from Window 1 to Window 2 when the reparandum was preceded by a filled pause compared to when it was spoken with certainty (see Figure 3). A similar Certainty-by-Window interaction was observed in proportion of looks to the repair picture (estimate = -

0.02, *SE* = 0.01, t = -2.72, *p* < .01) such that listeners in the Uncertain condition tended to stay on the repair picture from Window 1 to Window 2 (despite hearing the speaker utter the reparandum), whereas listeners in the Certain condition showed a large decrease in looks to the repair picture. For these analyses, no other two-way or three-way interactions were significant (*t*s < 1.5, *p*s > .14).

This pattern from Window 1 to Window 2 supports the idea that listeners used the filled pause to update their mental representation of what the speaker was trying to communicate, thereby placing less weight on the bottom-up input and more weight on the top-down prediction of what the speaker had likely intended to say. Further, proportions of fixations to the distractor images during Window 2 were very low (i.e., mean of 0.06, which did not vary by condition), suggesting that listeners very rarely made a mental correction from the reparandum to one of the other entities. These results suggest that listeners make their mental corrections strategically, drawing on cues such as plausibility and speaker certainty to derive a confident internal representation of what the speaker intended to communicate, which can override the actual bottom-up input when what is said conflicts with the listener's prediction.

Table 1
Results of Mixed Effects Analyses

Model parameters	Reparandum Estimate (95% CI)	*SE*	*t*	*p*	Repair Estimate (95% CI)	*SE*	*t*	*p*
Window 1								
Intercept	.26 (.23, .30)	.02	14.42	<.001	.39 (.35, .43)	.02	17.16	<.001
Plausibility	-.06 (-.09, -.03)	.01	-4.35	<.001	.07 (.04, .11)	.02	4.22	<.001
Certainty	.03 (.01, .05)	.01	2.61	.013	.01 (-.02, .03)	.01	.57	>.250
Plausibility*Certainty	.01 (-.01, .03)	.01	.64	>.250	-.00 (-.03, .02)	.01	-.44	>.250
Window 2								
Intercept	.43 (.38, .47)	.02	18.14	<.001	.35 (.31, .39)	.02	15.38	<.001
Plausibility	-.04 (-.07, -.01)	.01	-2.89	.006	.05 (.02, .08)	.01	3.70	<.001
Certainty	-.04 (-.06, -.02)	.01	-3.16	.004	.05 (.02, .08)	.01	3.34	.002
Plausibility*Certainty	.00 (-.02, .02)	.00	.00	>.250	-.00 (-.03, .02)	.01	-.40	>.250
Window 3								
Intercept	.25 (.22, .29)	.02	12.93	<.001	.53 (.47, .59)	.03	17.26	<.001
Plausibility	-.02 (-.05, .00)	.01	-1.63	.112	.01 (-.02, .04)	.01	.83	>.250
Certainty	-.05 (-.08, -.02)	.02	-3.04	.004	.06 (.02, .09)	.02	3.38	.002
Plausibility*Certainty	.01 (-.01, .03)	.01	1.15	>.250	-.01 (-.03, .02)	.01	-.42	>.250
Window 4								
Intercept	.12 (.09, .15)	.01	7.81	<.001	.64 (.57, .71)	.03	18.37	<.001
Plausibility	-.01 (-.02, .01)	.01	-.70	>.250	-.00 (-.02, .02)	.01	-.02	>.250
Certainty	-.01 (-.03, .01)	.01	-.82	>.250	.00 (-.02, .03)	.01	.23	>.250
Plausibility*Certainty	.01 (-.00, .02)	.01	1.44	.149	.01 (-.02, .03)	.01	.49	>.250

Note. CI = confidence interval; SE = standard error

Figure 2

Proportion of fixations to each picture type across the four conditions. All fixation plots were anchored to the onset of the reparandum (vertical line at time zero). The second vertical line represents the mean onset of the edit interval (i.e.,"uh I mean"). The third vertical line represents the mean onset of the repair. These lines represent the real-time onsets of these words, whereas all time windows were shifted forward 200 ms for the statistical analyses to account for the time required to launch a signal-based saccade. Error bands represent 95% confidence intervals.

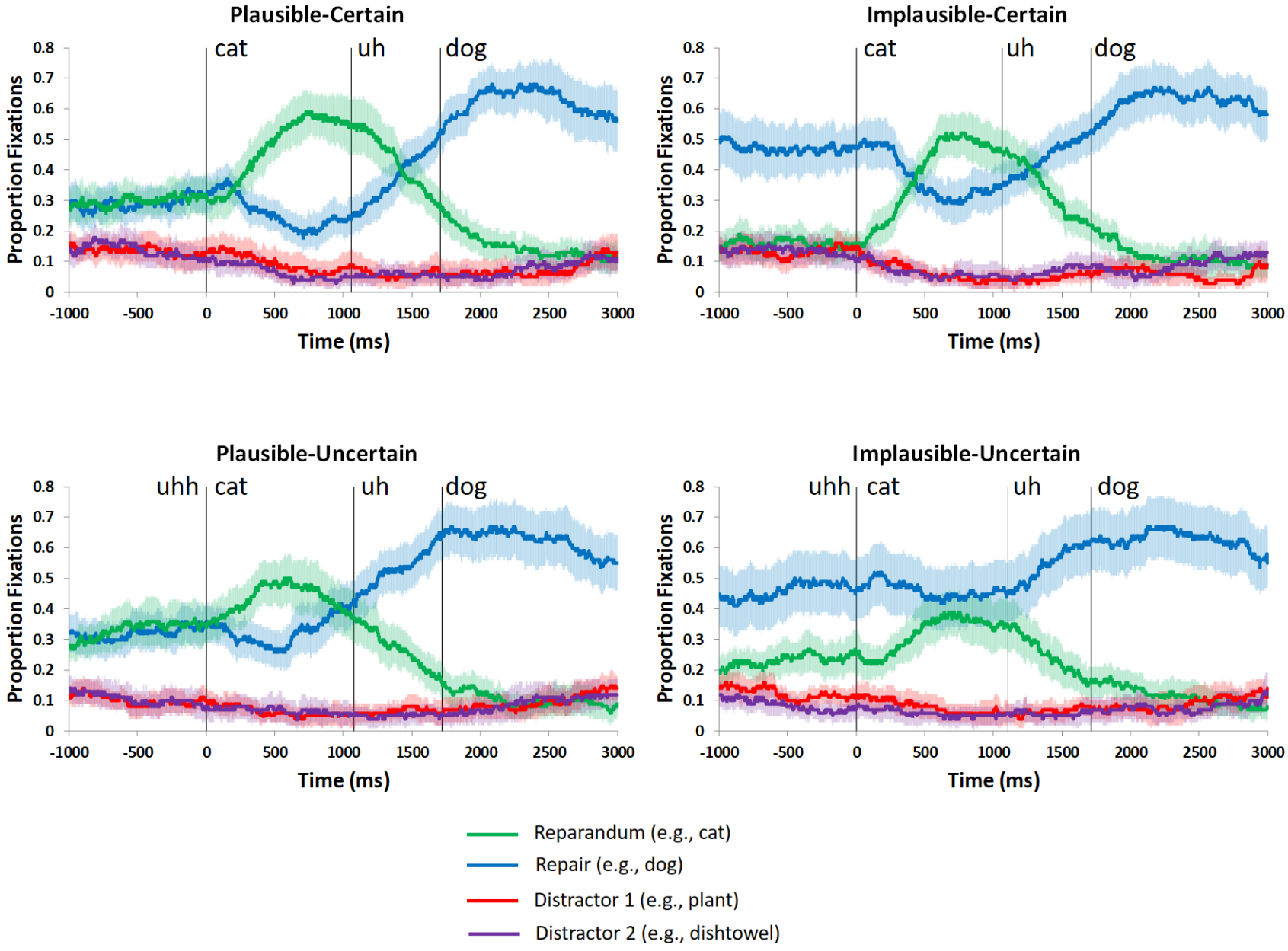
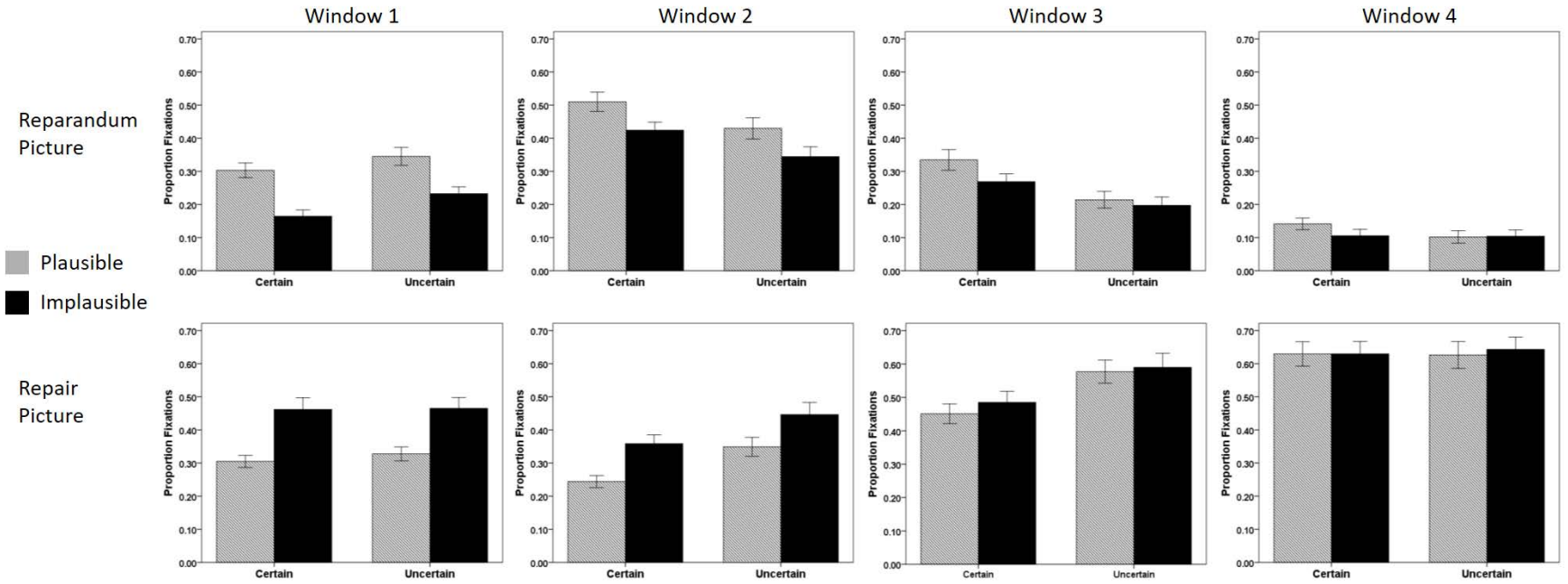

Plausible-Certain
Implausible-Certain
Plausible-Uncertain
Implausible-Uncertain
cat
uh
dog
uhh
Proportion Fixations
Time (ms)
Reparandum (e.g., cat)
Repair (e.g., dog)
Distractor 1 (e.g., plant)
Distractor 2 (e.g., dishtowel)

Figure 3

Mean proportion of fixations to the reparandum picture (top row) and repair picture (bottom row) as a function of certainty and reparandum plausibility, for each of the four time windows. Error bars represent 95% confidence intervals.



Discussion

How does language comprehension proceed so rapidly even when the speech signal is rife with errors and disfluencies? The work presented here suggests that one way we accomplish this feat is by anticipating and correcting speech errors even before receiving an overt correction from the speaker. The experiment demonstrated that listeners draw on top-down cues of contextual plausibility and speaker certainty to constrain their interpretations of the bottom-up input. Before the speaker uttered the reparandum, there was a robust effect of plausibility that guided listeners' expectations about what the speaker would say: listeners considered the cat and dog equally when either would form a natural continuation of the sentence, but there was a strong preference to consider the dog over the cat when the dog was the only item that fit the context. When the speaker uttered the reparandum (e.g., *cat*), listeners' degree of commitment to the cat versus the dog depended largely on how plausibly *cat* fit the preceding context, as well as the certainty with which the speaker produced the utterance. That is, if *cat* was implausible or was preceded by a filled pause, listeners were more likely to switch to the dog. The combined effect of an implausible reparandum that was uttered with uncertainty (see Implausible-Uncertain condition in Figure 2) led listeners to commit to the repair early and to some extent discount the reparandum. When the speaker initiated the correction, all conditions showed a shift from the reparandum to the repair, but the effect of certainty persisted. The results provide powerful evidence regarding the joint process of correcting speech errors that involves both the speaker and the listener. That is, listeners do not wait passively for speakers to explicitly correct their errors. Instead, listeners actively model the communicative intentions of speakers, rapidly integrating relevant cues to anticipate and mentally correct perceived errors.

The fixation patterns in Window 2 (i.e., at the onset of the reparandum) are particularly relevant, as this was the point in the utterance when the speaker made the error but before she initiated the explicit correction. Even though the speaker had not yet initiated the repair, listeners were less likely to fixate the reparandum if it was implausible or if the speaker signaled some degree of uncertainty, and instead, listeners were more likely to direct their fixations toward the repair. The additive nature of these effects shows that our manipulation of speaker certainty led listeners to put less confidence in the speaker's utterance regardless of its plausibility. These results are important because the extant literature concerning how listeners might use the filled pause to inform their predictions suggests that filled pauses primarily signal an unexpected upcoming word or expression (see Corley et al., 2007). Our results show that a filled pause also leads listeners to place less confidence in the speaker's utterance, regardless of the context, which is an important contribution to the literature on how speaker disfluencies inform the inferences listeners may draw concerning the upcoming input and the speaker. It should be noted, however, that our experiment was designed so that the filled pause always preceded an upcoming speech error and never signaled that an unexpected word or phrase might follow it. Thus, an important goal of future work will be to more carefully explore the factors that signal to listeners when a filled pause communicates that the speaker is about to say something unexpected as opposed to when the speaker does not have a high degree of confidence in the utterance, and how these expectations might interact.

The notion that the filled pause signaled uncertainty to listeners in the current experiment is consistent with previous work by Arnold et al. (2004, 2007) demonstrating that listeners can use a filled pause to inform their predictions about what the speaker will say next. For example, Arnold et al. (2004) showed that when participants heard an instruction to move one object and

then another instruction that contained a filled pause (e.g., *Put the grapes above the candle. Now put thee uh.…*), they rapidly predicted that the speaker would refer to something new to the discourse (as shown by eye-movement patterns). This previous work and the current experiment are similar in that they demonstrate listeners' sensitivity to filled pauses as an informative cue signaling that the speaker is having a problem with language production, which prompts listeners to actively anticipate what might come next. The current experiment goes beyond these previous findings in demonstrating that listeners can use these cues to generate a confident top-down prediction of what the speaker is trying to communicate, which can override the bottom-up input when it does not match the listener's prediction.

Overall, the results are consistent with the predictions of Noisy Channel models of language comprehension but expand this approach in important ways. According to this framework, the noise inherent in everyday language necessitates a comprehension system in which the perceived input is weighed against assumptions about the communicative intent. The current results are consistent with this framework in demonstrating that the linguistic context in which the error is encountered, as well as the speaker's certainty about the utterance, are two crucial sources of information that listeners exploit to detect errors. Our work also highlights the importance of studying the error correction mechanisms that are inherent to the Noisy Channel model in contexts beyond the written domain. As we have noted, the cooperative nature of spoken language comprehension makes this a particularly fruitful area in which to investigate the comprehender's relative weighting of bottom-up linguistic input versus top-down expectations about the overarching message, as speakers and listeners work together to optimize communication. Indeed, our work demonstrates that listeners can mentally repair speech errors even before the speaker has time to explicitly initiate a correction.

In addition, we have recently argued (Ferreira & Lowder, 2016; Lowder & Ferreira, 2016a, 2016b) that the process of mentally correcting speech errors may be akin to the process of generating sets of contrastive alternates in semantic focus constructions. Speakers often use focus constructions to contrast one entity with another, as illustrated in *The woman went to the animal store and brought home not a dog but rather a…* Our work provides support for the hypothesis that the focused item in a focus construction and the reparandum in a disfluent utterance (e.g., *…a dog, uh I mean a…*) both function to activate a set of related concepts or semantic features that then allow the listener to anticipate the rest of the sentence. This view is consistent with formal approaches to dialogue that treat disfluencies as one of many clarification devices available to interlocutors (Ginzburg et al., 2014) to facilitate mutual understanding, as well as an effective method of orienting the listener's attention (Fraundorf & Watson, 2011). A key area of future study thus involves uncovering the similarities and differences between the predictive mechanisms that underlie the processing of speech errors and the processing of focus more generally. Further, a major goal for future investigation will be generalizing these claims using experimental approaches beyond the visual-world paradigm, given the inherent limitations associated with this paradigm (e.g., presenting participants with a limited array of images).

Regarding models of language processing, this work extends the Good-Enough framework in a new direction. Although this approach has been very influential in the sentence-processing literature, it has tended to focus exclusively on the ways in which comprehenders misinterpret linguistic input such that they often derive interpretations that are underspecified, incomplete, or inaccurate. In contrast, the current work shows that there are also cases in which normalizing the input can be quite adaptive to the comprehension system, serving to make

communication between speaker and listener more efficient by allowing the comprehender to understand the message the speaker intended to convey.

This finding of adaptive normalization has important broader implications. The current work shows that the listener's internal model of the speaker's production system strongly constrains how the input is weighted. That is, if the listener is able to derive a confident top-down representation of the speaker's intended message, the bottom-up signal will be adjusted in favor of the representation derived from the listener's expectations. If listeners' ability to normalize linguistic input is based on their model of speakers' communicative behavior and intentions, then we would predict that the more a listener knows about a speaker, the more successful the normalization will be, opening up a line of research into the links between strength of social relationship and communicative success (see Grodner & Sedivy, 2007, for evidence that speaker reliability can affect the inferences listeners draw during language comprehension). Additionally, speaker modeling might rely on mechanisms linked to theory of mind (Pickering & Garrod, 2014), which are known to undergo significant change during childhood (e.g., Liu et al., 2009). From this perspective, adaptive normalization during language comprehension likely improves during early development. A deeper understanding of how adaptive normalization during language comprehension takes place will require a research approach that brings together cognitive, social, and developmental perspectives, among others.

This work thus adds to the growing evidence that successful language processing requires interlocutors to model each other's communicative intentions and not merely to derive an accurate representation of the input. Prior work has already established that filled pauses help listeners better understand the speaker's message (Arnold et al., 2004, 2007) and may also help listeners detect a speaker's attempt at deception (Loy et al., 2017). The current results build on

this knowledge by showing that listeners do not always take even the lexical input of the speaker at face value, but instead take advantage of cues that allow them to assess the speaker's intention. Our results highlight the independent contribution of two of these cues, such that when a speaker signals uncertainty and then says something implausible, the listener has sufficient evidence justifying the assumption that the speaker did not say what she meant. In scenarios such as this, the listener will derive a model of what the speaker intended to say and will mentally repair the utterance, even before the repair is initiated by the speaker.

Context of the Research

Recent work in the areas of sentence processing in particular and cognitive science more generally has focused heavily on the extent to which humans anticipate and actively predict upcoming information. The current line of work was inspired by our hypothesis that listeners might rely on relevant cues from the speaker and use these cues to generate top-down predictions about what the speaker intended to say that can override an error in the bottom-up speech input. The current results offer robust support in favor of this hypothesis and suggest that the approach used here might be especially useful in future studies examining predictive processing. These fields have also recently started to emphasize the need to understand cognitive processes more naturalistically, taking into consideration the normal contexts in which language comprehension takes place and the normal variation in the form and quality of the linguistic signal (for recent discussion, see Hasson et al., 2018). These results help us understand why listeners don't simply give up when they encounter "noisy" utterances that deviate significantly from the idealized stimuli typically presented in psycholinguistic experiments: Listeners comprehend such utterances because the language system includes mechanisms for identifying words that seem

inconsistent with the speaker's communicative intention and replacing them with words that better reflect what the speaker likely meant to say.

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*, 419-439.

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

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

Arnold, J. E., Tanenhaus, M. K., Altmann, R., & Fagnano, M. (2004). The old and thee, uh, new. *Psychological Science, 15*, 578–581.

Barton, S. B., & Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & Cognition, 21*, 477-487.

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

Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language, 34*, 383-398.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology, 42*, 368-407.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1-39.

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*, 84-107.

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition, 105*, 658-668.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology, 47*, 164-203.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*, 11-15.

Ferreira, F., & Lowder, M. W. (2016). Prediction, information structure, and good enough language processing. *Psychology of Learning and Motivation, 65*, 217-247.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language, 34*, 709-738.

Fraundorf, S. H., & Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language, 65*, 161-175.

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, 110, 8051-8056.*

Ginzburg, J., Fernández, R., & Schlangen, D. (2014). Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics, 7,* 1-64.

Grodner, D. J., & Sedivy, J. C. (2007). The effect of speaker-specific information on pragmatic inferences. In E. Gibson & N. Perlmutter (Eds.), *The processing and acquisition of reference.* Cambridge, MA: MIT Press.

Hasson, U., Egidi, G., Marelli, M., & Willems, R. M. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition, 180,* 135-157.

Honal, M., & Schultz, T. (2003). Correction of disfluencies in spontaneous speech using a noisy-channel approach. In *Proceedings of the 8th Conference on Speech Communication and Technology* (pp. 2781-2784).

Johnson, M., & Charniak, E. (2004). A TAG-based noisy-channel model of speech repairs. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (pp. 33-39).

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31,* 32-59.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2016). lmerTest: Tests in linear mixed effects models. R package version 2.0-33.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211-240.

Lau, E. F., & Ferreira, F. (2005). Lingering effects of disfluent material on comprehension of garden path sentences. *Language and Cognitive Processes, 20,* 633-666.

Liu, D., Sabbagh, M. A., Gehring, W. J., & Wellman, H. M. (2009). Neural correlates of children's theory of mind development. *Child Development, 80,* 318-326.

Lowder, M. W., & Ferreira, F. (2016a). Prediction in the processing of repair disfluencies. *Language, Cognition and Neuroscience, 31,* 19-31.

Lowder, M. W., & Ferreira, F. (2016b). Prediction in the processing of repair disfluencies: Evidence from the visual-world paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42,* 1400-1416.

Loy, J. E., Rohde, H., & Corley, M. (2017). Effects of disfluency in online interpretation of deception. *Cognitive Science, 41,* 1434-1456.

Pickering, M. J., & Garrod, S. (2014). Self-, other-, and joint monitoring using forward models. *Frontiers in Human Neuroscience, 8,* 132.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21,* 211-232.

Shriberg, E. E. (1996). Disfluencies in SWITCHBOARD. *Proceedings of International Conference on Spoken Language Processing*, Addendum, pp. 11-14, Philadelphia, PA.

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language, 32,* 25-38.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research, 29*, 557-580.

Tanenhaus, M. K., Spivey, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632-1634.

Zwarts, S., Johnson, M., & Dale, R. (2010). Detecting speech repairs incrementally using a noisy channel approach. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 1371-1378.

Author Note

This work was supported in part by grants from NICHD (F32 HD084100) awarded to M.W.L and NIDCD (R56 DC013545) awarded to F.F. The findings reported in this article were presented at the 29th annual CUNY Conference on Human Sentence Processing (2016), as well as the 57th annual meeting of the Psychonomic Society (2016).

We thank Susan Ahmed, Matthew Chan, Rajneet Gurai, Nimitha Kommoju, Kellie McDonald, and Alba Peris-Yague for assistance in conducting the experiment.