Partner-specific adaptation in disfluency processing

Si On Yoon (sion-yoon@uiowa.edu)
Department of Communication Sciences and Disorders, University of Iowa
Iowa City, IA 52242 USA

Sarah Brown-Schmidt (sarahbrownschmidt@gmail.com)
Department of Psychology & Human Development, Vanderbilt University
Nashville, TN 37235 USA

Abstract

Disfluency leads listeners to expect an upcoming reference to unfamiliar objects. In two experiments, we examined if this expectation is adapted based on the way disfluency has been used in the discourse. Participants listened to instructions to look at an object on a screen containing familiar and novel images. We manipulated the co-occurrence of disfluency and reference to novel vs. familiar objects. In the predictive condition, disfluent expressions referred to novel objects, and fluent expressions referred to familiar objects. In the non-predictive condition, fluent and disfluent trials referred to either familiar or novel objects. Participants' gaze revealed that listeners more readily predicted familiar images for fluent trials and novel images for disfluent trials in the predictive condition than in the non-predictive condition. Listeners adapted their expectations about upcoming words based on recent experience with disfluency. Disfluency is not invariably processed, but is a cue adapted within the local context.

Keywords: Speech disfluency; Eye-tracking; Adaptation; Partner-specific processing

Introduction

Listeners are known to adapt to various aspects of linguistic input, such as speakers' speech sound, lexical choice or syntactic structures, based on the statistics of the recent linguistic experience (Bradlow & Bent, 2008; Fine et al., 2013; Trude & Brown-Schmidt, 2012; see also Harrington Stack, James, & Watson, 2018). For example, when listeners are exposed to the speech of two talkers, one with an unfamiliar regional dialect of American English and the other without the dialect, they process critical speech sounds differently based on the speaker's dialect (Trude & Brown-Schmidt, 2012). This adaptation to speech perception has been shown even in individuals with amnesia who have severe declarative memory impairment (Trude, Duff, & Brown-Schmidt, 2014). Adapting to different sources of variations is likely crucial to enhance efficiency in language processing which is faced with the challenge of rapidly interpreting speech despite substantial variability between and within talkers. While an extensive amount of work has examined adaptation to linguistic input, less explored is whether listeners adapt to paralinguistic properties of the input, such as a speaker's prosody or their use of disfluency. Here, we focus on listener's adaptation to a paralinguistic input: the way speakers produce disfluency.

Speakers are often disfluent in everyday conversation Brennan & Schober, 2001; Bortfeld, et al., 2001). Speakers often become hesitant or disfluent by producing a filler word (e.g., "Look at *thee… uh*…") or being silent between words. It is estimated that the rate of disfluencies in spontaneous speech is considerable, 6 words per every 100 words (Fox Tree, 1995). Numerous studies have shown that disfluency is often found when speakers encounter difficulty in planning utterances or retrieving lexical items (Clark & Fox Tree, 2002; Clark & Wasow, 1998; Ferreira, 1991; Fraundorf & Watson, 2013; Jaeger, 2010; Smith & Clark, 1993). Although disfluency (e.g., "um" or "uh") are not typically considered lexical items (cf, Fox Tree, 2001), disfluencies signal information, and listeners actively process disfluency, making predictions about what comes next; Listeners expect disfluent descriptions to refer to discourse-new entities or entities that are hard to describe (Arnold, Hudson Kam, & Tanenhaus, 2007; Arnold, et al., 2004; Barr & Seyfeddinipur, 2010).

Previous evidence suggests that the link between disfluency and discourse novelty is not simply due to tracking of co-occurrence statistics, but also integrates speaker information and their perspective. For example, listeners interpret different speakers' disfluencies with respect to each speaker's knowledge state (Barr & Seyfeddinipur, 2010). They also attenuate the "disfluency=new reference" prediction when they believe they are listening to a person with anomia who has difficulty naming familiar objects and frequently becomes disfluent (Arnold et al., 2007). Listeners also suspend their "disfluency=new" expectation when a second, naïve listener joins an ongoing conversation, reflecting the tendency of speakers to become disfluent in such situations, even when referencing familiar objects (Yoon & Brown-Schmidt, 2014). Thus, listeners make situation-specific inferences and interpret disfluency accordingly (see also Heller, et al., 2014).

Less clear is if the expectation of disfluency referring to novelty can be adapted based on how disfluency is used in the current discourse context. In a recent study, Bosker, et al., (2019) exposed listeners to disfluencies in a typical predictive context based on the lexical frequency of the noun (disfluent→low frequency; fluent→high), or the reverse. They found that when processing disfluency, participants were more likely to look at low frequency nouns in the typical condition; this tendency was attenuated when the contingencies were reversed and disfluency was predictive of reference to high frequency nouns. This work provides initial evidence for the idea that the way in which disfluencies signal

upcoming meaning can be flexibly adapted over brief time-scales.

In the present work, we first test the malleability of the processing of disfluency by manipulating the relationship between disfluency and reference to novel vs. familiar objects (Experiment 1). If disfluency is invariably processed, local context should not affect disfluency expectations. Alternatively, if disfluency is flexibly adapted, listeners' expectations will differ depending on the use of disfluency in the local context. In Experiment 2 we ask the novel question of whether adaptation to disfluency reflects simple adaptation to co-occurrence statistics in the local context, or instead, if this adaptation is a partner-specific process.

Experiment 1

Participants

Fifty-four undergraduates at Vanderbilt University participated in the experiment in return for either cash payment or partial course credit. Participants were all native speakers of North American English and normal or corrected-to-normal hearing and vision.

Materials and procedure

Participants performed a referential communication task (Krauss & Weinheimer, 1996), sitting at a table in front of a monitor. On each trial, they followed pre-recorded instructions to look at an object on a screen (e.g., "Look at the…"). The participants' eye movements were monitored during the task with an Eyelink 1000 (SR Research, Ontario, Canada) desktop-mounted eye-tracker. It sampled eye movements monocularly at 1,000 Hz.

In the task, four images were presented on the screen – two familiar images and two novel images (Figure 1). Each object type was shown in two colors. Images were adapted from previous studies (Arnold, Hudson Kam, & Tanenhaus, 2007; Brown-Schmidt, 2009a, Yoon & Brown-Schmidt, 2018).

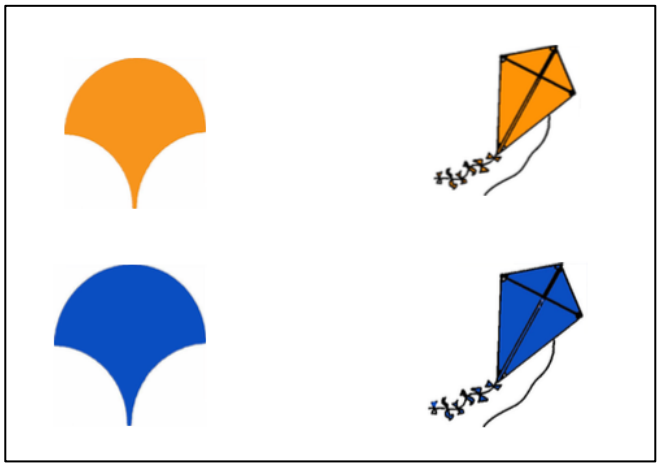

Figure 1. Example scene display.

The pre-recorded instructions were always the same format; the color of the target was described first, and then the label or the description of the target followed (e.g., "the orange kite"). Two factors were manipulated: Fluency (within-subjects) and predictability (between-subjects). The pre-recorded instructions were fluent ("Look at the orange kite.") for half of trials and disfluent ("Look at thee… uh… orange kite.") for the other half. The phrases "Look at the/thee…uh [color]" were cross-spliced, so that the auditory stimuli before the critical noun was the same in the predictive and non-predictive conditions. In fluent expressions, the average adjective onset (e.g., orange) was 623ms and the average noun onset (e.g., kite) was 1008m after onset of "Look". In disfluent expressions, the average adjective onset was 2340ms and average noun onset 2747ms after "Look".

In the predictive condition, disfluent expressions always referred to the novel image and fluent expressions to the familiar image. In the non-predictive condition, fluent and disfluent trials refer to either the novel or familiar image.

There were 96 critical trials that were repeated twice in the same disfluency condition (a total of 192 trials, no filler trials). Trials were randomly presented and the entire task took approximately an hour.

Predictions

If the interpretation of disfluency as a signal to upcoming meaning flexibly adapts in response to signal→meaning contingencies in the local environment, the interpretation of disfluency should differ across conditions. Listeners should look at the target image more before they hear the critical noun for both fluent and disfluent trials in the predictive condition vs. in the non-predictive condition. In other words, upon hearing the color adjective, they will look at the color-matching novel image more than the familiar image in disfluent trials, and they will look at the color-matching familiar image more than the novel images in fluent trials.

Alternatively, if listeners do not adapt to the way how the speaker uses disfluency, their gaze would not differ between the two conditions for both fluent and disfluent trials.

Results

The latency between the onset of "Look" and the critical color adjectives was significantly longer for disfluent trials (2340ms) than fluent trials (623ms), making them difficult to directly compare. Thus, test trials were analyzed for fluent and disfluent trials separately, using mixed-effects models (see Yoon & Brown-Schmidt, 2014). When the maximal model did not converge, random slopes were removed from the model one at a time until convergence. The a-priori critical time window was from 200ms to 700ms after the onset of the color adjective, reflecting listeners' predictive processing prior to hearing the critical noun.

Fluent Trials We analyzed listeners' eye movements following the onset of the adjective (Figure 2). During the critical time window (200-700ms after the onset of the color adjective), listeners did not hear the critical noun (e.g., kite),

but had to make a prediction based on the color adjective (Either the familiar object or the novel object of the same color). We compared fluent trials referring to the familiar image in the predictive condition and in the non-predictive conditions.

Listeners' processing of fluent expressions was analyzed in a mixed-effects model that included predictability (predictive vs. non-predictive) as a fixed effect (Table 1). The dependent measure was the proportion of looks to the target image. The model revealed a significant main effect of predictability ($t=2.74$, $p<.05$). This finding suggests that listeners readily looked at the upcoming referent – the color-matching familiar image – more in the predictive condition than in the non-predictive condition.

Disfluent Trials As in the analysis of fluent trials, listeners' eye movements from the onset of the adjective was analyzed: from 200ms to 700ms after the adjective (Figure 2). Disfluent trials referring to the novel image in the predictive condition and the non-predictive condition were compared.

A mixed-effects model included predictability as a fixed effect (Table 2). The dependent measure was the proportion of looks to the target. The model revealed a significant effect of predictability ($t=2.05$, $p<.05$). Consistent with the finding in fluent trials, listeners predicted the color-matching novel image more as an upcoming referent following disfluency in the predictive condition than in the non-predictive condition.

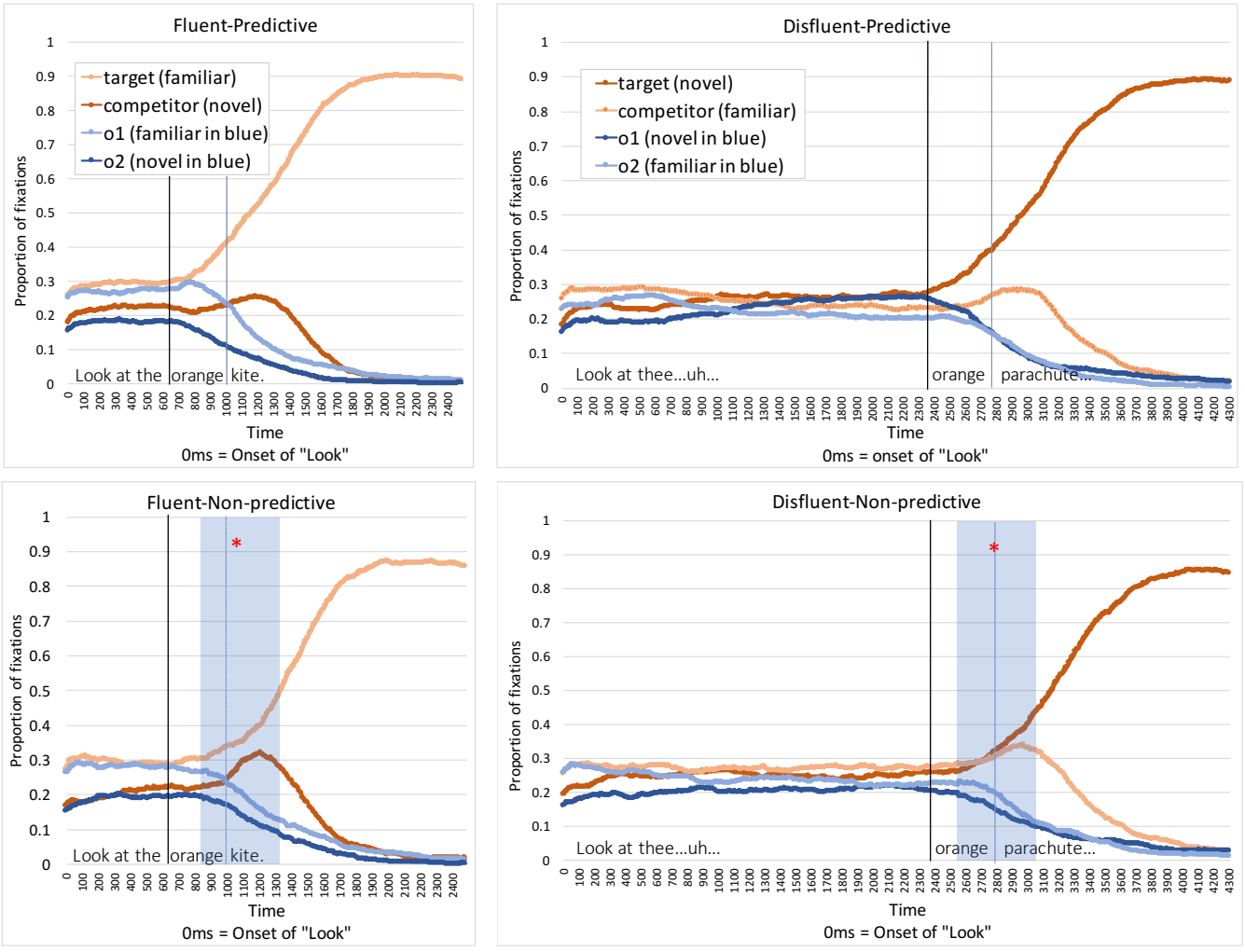

Figure 2. The proportion of fixations for fluent expressions (left) and disfluent expressions (right) in Experiment 1. In fluent expressions, the average onset of the critical adjective (e.g., orange) is 623ms and the average onset of the critical noun (e.g., kite) is 1008ms. In disfluent expressions, the average onset of the critical adjective is 2340ms and the average onset of the critical noun is 2747ms. The solid vertical line shows the average onset of the adjective and the dotted line shows the average onset of the noun. The blue shading indicates the critical analysis window.

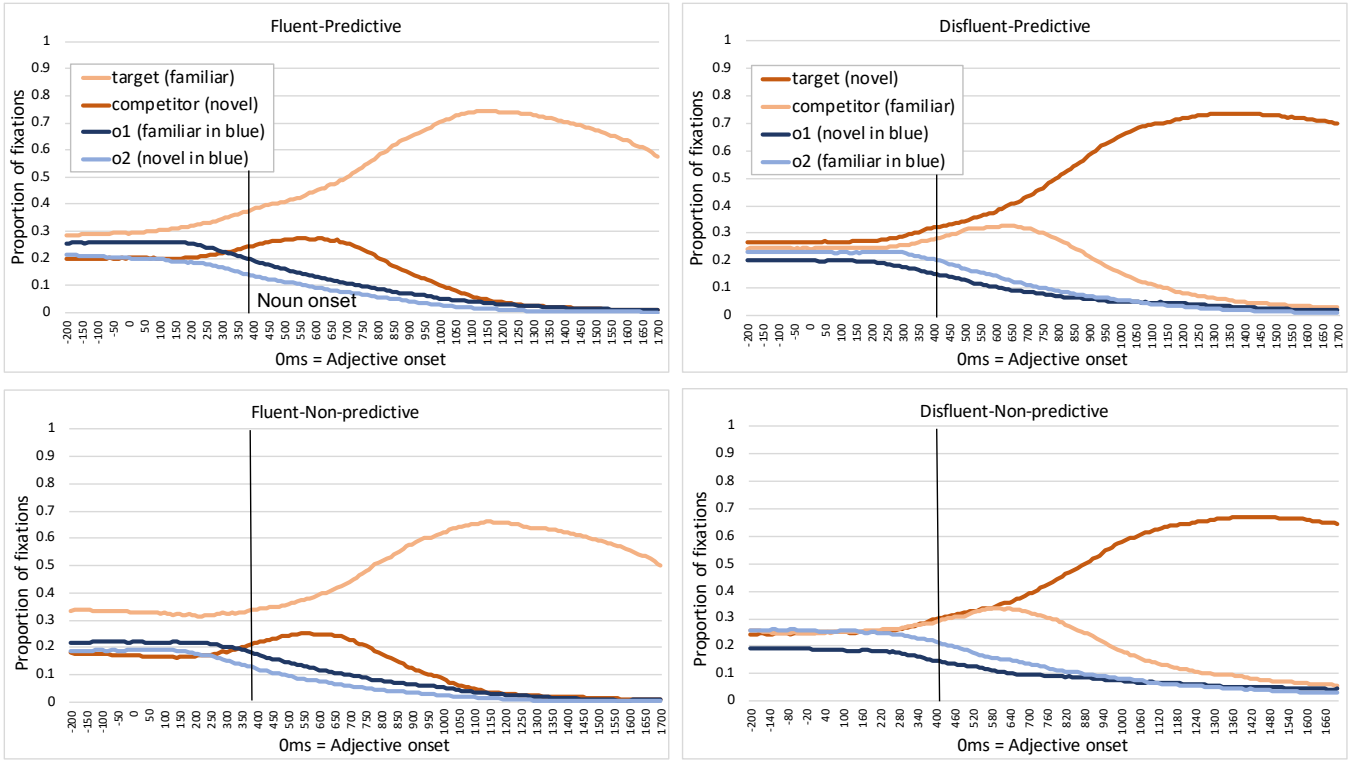

Figure 3. The proportion of fixations for fluent expressions (left) and disfluent expressions (right) in Experiment 2.

Table 1: Fluent trials: mixed effect model with predictability (non-predictive vs. predictive) as a fixed effect. The dependent measure was the proportion of looks to the target in fluent trials in Experiment 1.

	Estimate	SE	t-value	p-value
(intercept)	0.42	0.02	25.24	<.001
Predictability	0.08	0.03	**2.74**	**0.008**
Random effects	Variance	SD		
Subject (intercept)	0.01	0.10		
Item (intercept)	0.003	0.06		
Residual	0.15	0.39		

Table 2: Disfluent trials: Mixed effect model with predictability (non-predictive vs. predictive) as a fixed effect. The dependent measure was the proportion of looks to the target object in disfluent trials in Experiment 1. Values in bold indicate significant results.

	Estimate	SE	t-value	p-value
(intercept)	0.38	0.02	18.47	<.001
Predictability	0.08	0.04	**2.05**	**0.045**
Random effects	Variance	SD		
Subject (intercept)	0.02	0.13		
Item (intercept)	0.004	0.06		
Residual	0.15	0.38		

Experiment 2

In Experiment 1, the analysis of gaze revealed that the tendency to interpret "disfluency=new" was attenuated in the non-predictive condition for disfluent trials compared to the tendency in the predictive condition. Listeners also readily predicted upcoming referents – familiar images – in the predictive condition than in the non-predictive condition for fluent trials. These findings suggest that listeners flexibly adapt to how a speaker uses disfluency based on recent experience with the speaker. However, in the natural world talkers naturally vary in their use of disfluency, indicating that the signal strength of disfluency → something hard or new likely varies in a talker-specific manner. It is well established that listeners quickly adapt to specific talkers' acoustic or linguistic features (e.g., a speaker's accent or syntactic structure). By contrast, if and how listeners adapt to paralinguistic cues of a specific talker have been less well explored. Specifically, whether and how listeners adapt to partner-specific non-linguistic cues, such as disfluency, before hearing critical linguistic information is less understood, although adaptation of both linguistic and non-linguistic cues facilitate listeners' comprehension processing. Thus, in Experiment 2, we examined if listeners' adaptation to disfluency is partner-specific when they interact with two talkers simultaneously; one speaker producing disfluency in a predictive way and the other speaker producing disfluency

in a non-predictive way. To test this question, predictability was manipulated *within*-subjects in Experiment 2.

Participants

Fifty-four undergraduates at Vanderbilt University participated in the experiment in return for either cash payment or partial course credit. Participants were all native speakers of North American English and normal or corrected-to-normal hearing and vision.

Materials and procedure

The material and the procedure of Experiment 2 were identical to Experiment 1, with the following exception: the manipulation of predictability was within-subjects. Two pre-recorded voices were introduced: one female and one male voice. One voice produced disfluency in a non-predictive way and the other voice produced disfluency in a predictive way (which voice was predictive was counterbalanced across persons). Two voices were randomly presented. There were 96 critical targets that were repeated four times in the same disfluency condition (a total of 384 trials, no filler trials). The color of competitors changed across four repetitions to avoid a learning effect.

Predictions

We predict that if listeners are able to adapt to disfluency in a talker-specific pattern, listeners would look at the color-matching upcoming referent more upon hearing the color adjective in the predictive condition vs. in the non-predictive condition, consistent with the results of Experiment 1; They should look at the novel image more for disfluent trials and look at the familiar image more for fluent trials in the predictive condition than in the non-predictive condition.

Alternatively, if listeners do not adapt their expectations about upcoming words according to a recent experience with each talker, their gaze would not be different regardless of the identity of the talkers for both fluent and disfluent trials.

Results

Consistent with the analyses in Experiment 1, test trials were analyzed for fluent and disfluent trials separately.

Fluent Trials Listeners' eye movements following the onset of the adjective were analyzed. The critical time window was from 200ms to 700ms after the onset of the color adjective (Figure 3). Fluent trials referring to the familiar image in the predictive and non-predictive conditions were compared.

A mixed-effect model included predictability (predictive vs. non-predictive) as a fixed effect (Table 3). The dependent measure was the proportion of looks to the target image. The model revealed a significant main effect of predictability ($t=2.61$, $p<.05$). Before hearing the critical noun, listeners looked at the familiar image more in the predictive condition than in the non-predictive condition.

Disfluent Trials Listeners' eye movements following the onset of the adjective were analyzed: from 200ms to 700ms after the onset of the color adjective (Figure 3). Disfluent trials referring to the novel image in the predictive and non-predictive conditions were compared.

A mixed-effect model included predictability (predictive vs. non-predictive) as a fixed effect was used to examine listeners' processing of disfluent expressions (Table 4). The dependent measure was the proportion of looks to the target image. A significant main effect of predictability ($t=2.607$, $p<.05$) showed that listeners' adaptation was tailored to a specific partner; they predicted the novel referent more in the predictive condition than in the non-predictive condition.

Table 3: Fluent trials: Mixed effect model with predictability (non-predictive vs. predictive) as a fixed effect. The proportion of looks to the target in fluent trials in Experiment 2. Values in bold indicate significant results.

	Estimate	SE	t-value	p-value
(intercept)	0.38	0.02	25.32	<.001
Predictability	0.04	0.02	**2.61**	**0.01**
Random effects	Variance	SD		
Subject (intercept)	0.01	0.10		
Predictability	0.006	0.08		
Item (intercept)	0.006	0.08		
Predictability	0.004	0.07		
Residual	0.16	0.39		

Table 4: Disfluent trials: Mixed effect model with predictability (non-predictive vs. predictive) as a fixed effect. The proportion of looks to the target object in disfluent trials in Experiment 2. Values in bold indicate significant results.

	Estimate	SE	t-value	p-value
(intercept)	0.33	0.02	17.97	<.001
Predictability	0.03	0.01	**2.607**	**0.01**
Random effects	Variance	SD		
Subject (intercept)	0.02	0.13		
Predictability	0.001	0.03		
Item (intercept)	0.003	0.06		
Residual	0.15	0.39		

General Discussion

The results show that listeners are able to learn and use their knowledge about individual speakers' use of disfluency based on recent experience with them. Even before hearing the critical noun, listeners predicted the upcoming referent, looking at the familiar image more for fluent trials and the novel image more for disfluent trials when the speaker produced disfluency in a predictive way compared to when the speaker produced disfluency in a non-predictive way. This adaptation was partner-specific in that listeners interpreted speakers' disfluency differently based on the way disfluency had been used by a specific speaker.

Even though the results in both Experiments showed that listeners learn speakers' disfluency and apply their knowledge in online language processing, their performance to predict an upcoming referent was notably better when interacting with one speaker (Experiment 1) compared to when interacting with two speakers simultaneously (Experiment 2). The difference in target fixations during the critical time window between the predictive and non-predictive conditions was ~ 15% in Experiment 1 and ~3% in Experiment 2 (Figure 2 and 3). Although adaptation to a paralinguistic input such as disfluency is a robust phenomenon, the smaller effect when interacting with two partners vs. one partner may be caused by higher cognitive costs required when simultaneously tracking two perspectives than one perspective (see also Ryskin et al., 2015). Further, participants in our experiments had to build up the representations of the speakers as the task unfolded, because they were not given any information about them. This bottom-up process of establishing representations of the speakers based on local information might require more cognitive resources compared to the top-down process of using representations of speakers (e.g., given information about the speaker – anomia; Arnold , et al., 2007). Another consideration is that in the non-predictive condition, it might be difficult for listeners to overcome their ordinary prediction of disfluency that they use in everyday life. Without any given explanation about speakers' characteristics (e.g., anomia), listeners may naturally attribute the speaker's disfluency to planning difficulty. When it is violated without any specific justification, overcoming this violated prediction could be difficult for listeners especially when only one of the speakers violated the prediction.

An open question is how listeners learn variations in paralinguistic cues and use this information in online language processing. Trude et al. (2014) showed that individuals with amnesia who have declarative memory impairment were able to learn speakers' accents, suggesting that speech adaptation does not require intact declarative memory. Further research is warranted to determine whether declarative memory is required in adaption to paralinguistic cues (e.g., disfluency, prosody, etc) as well as linguistic cues (e.g., speech sounds, syntactic structure).

In conclusion, we have shown that listeners are sensitive to the way speakers use disfluency in the local context and flexibly adjust how they process disfluency accordingly. Rather than processing disfluency based on simple associations between disfluency and novel referents, our findings point to flexible adaptations listeners make during the online processing of disfluency. Disfluency is not invariably processed, but instead a cue that is flexibly adapted within the local context. This study expands previous findings and shows that these paralinguistic adaptation effects can be talker-specific.

References

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,* 578–582.

Barr, D., & Seyfeddinipur, M. (2010). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes, 25,* 441–455.

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

Bostker, H. R., & van Os, M., Does, R., van Bergen, G. (2019). Counting 'uhm's: How tracking the distribution of native and non-native disfluencies influences online language comprehension. *Journal of Memory and Language, 106,* 189-202.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106,* 707-729.

Brennan, S. E., & Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language, 44,* 274-296.

Brown-Schmidt, S. (2009a). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language, 61,* 171-190.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition, 84,* 73–111.

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology, 37,* 201–242.

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language, 30,* 210–233.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS One, 8*, e77661.

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.

Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory & Cognition, 29,* 320-326.

Fraundorf, S. H., & Watson, D. G. (2013). Alice's adventures in um-derland: Psycholinguistic sources of variation in disfluency production. *Language and Cognitive Processes. 29,* 1083-1096.

Harrington Stack, C. M., & James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition, 46,* 864-877.

Heller, D., Arnold, J. E., Klein, N., & Tanenhaus, M. K. (2014). Inferring difficulty: Flexibility in the real-time processing of disfluency. *Language and Speech, 58,* 190-203.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage information density. *Cognitive Psychology, 61,* 23– 62.

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology, 4,* 343–346.

Ryskin, R. A., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General, 144,* 898–915.

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

Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during on-line speech perception. *Language and Cognitive Processes, 27,* 979-1001.

Trude, A. M., Duff, M. C., & Brown-Schmidt, S. (2014). Talker-specific learning in amnesia: Insight into mechanisms of adaptive speech perception. *Cortex, 54,* 117-123.

Yoon, S. O., & Brown-Schmidt, S. (2014). Adjusting conceptual pacts in three-party conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 919-937.

Yoon, S. & Brown-Schmidt, S. (2018). Aim Low: Mechanisms of audience design in multiparty conversation. *Discourse Processes, 55,* 566-592.