Running head: PRAGMATIC INFERENCE MODULATION

This is a manuscript currently under review. Please don't cite this without the author's permission. Comments and feedback are greatly appreciated!

Information integration in online modulation of pragmatic inferences during language comprehension

Rachel A. Ryskin[1], Chigusa Kurumada[2], and Sarah Brown-Schmidt[3]

1. Dept. of Brain and Cognitive Sciences, Massachusetts Institute of Technology

2. Dept. of Brain and Cognitive Sciences, University of Rochester

3. Dept. of Psychology & Human Development, Vanderbilt University

Abstract

Upon hearing a scalar adjective in a definite referring expression such as "*the big*…," listeners typically interpret the speaker to be referring to an item in a contrast set denoted by the scalar adjective, such as a big glass in the context of a smaller glass. This inference guides online visual search for a referent, resulting in anticipatory eye-movements to the target object when a contrast item is present. Recent studies have suggested that these inferences are malleable, such that the likelihood of an inference reflects environmental statistics of felicitous language use. These results point to distributional learning about the contexts in which scalar adjectives occur (i.e., that scalar adjectives are typically used in contrastive settings) as a key mechanism underlying the inference process. In a series of eye-tracking experiments, we explore the nature of the evidence necessary for distributional learning in language comprehension, focusing on 1) modulation of pragmatic inferences based on bottom-up information and 2) the speaker-specificity of this process. Listeners were exposed to speakers who either used size adjectives felicitously (e.g., "the big dog" in the context of a big and a small dog) or infelicitously (e.g., "the big dog" in a context with one dog). We find that massive exposure to bottom-up evidence can trigger modulation of contrastive inferences but there was little evidence of speaker-specificity. These findings suggest that listeners may evaluate observed evidence against their prior assumptions about pragmatic language use.

In communication, referring expressions are shaped by properties of the intended referent as well as the context that the referent appears in. Definite referring expressions are thought to uniquely identify an intended referent within the relevant context, broadly construed (see Crain & Steedman, 1985; Roberts, 2003), and as a result, referential form reflects what else can be referred to in the same context (Olson, 1970; Osgood, 1971; Pechmann, 1989). In a context with a single dog, the bare noun phrase "the dog" would suffice to achieve reference. By contrast, in a situation with multiple dogs, the speaker would need to provide additional information, for example, through the use of a modified noun phrase, e.g., "*the fluffy dog"* (see Brown-Schmidt & Konopka, 2011; Davies & Katsos, 2013; Nadig & Sedivy, 2002; Ryskin, Benjamin, Tullis, & Brown-Schmidt, 2015).

Different classes of modifiers vary in their degree of contextual sensitivity. Experimental studies of referential form (Belke, 2006; Brown-Schmidt & Konopka, 2011; Brown-Schmidt & Tanenhaus, 2006; Nadig & Sedivy, 2002; Sedivy, 2005) show that scalar modifiers such as *tall, big,* and *skinny* are often prompted by the presence of a scalar contrast in the relevant context (e.g., a context with a big dog and a small dog), and are much less likely in contexts with only a single member of the class denoted by the noun (e.g., a context with a single dog). In addition, the meaning of a scalar adjective is shaped by the noun it describes, where "small" evokes a very different point on the scale when used to characterize a "small dog" compared to a "small building" (Kennedy & McNally, 2005). By contrast, color modifiers such as *green,* or *aqua* are often used attributively, and thus are much more common in situations where mentioning color is not necessary to uniquely identify the referent (see Donnellan, 1966 for a discussion of

the attributive use of definite reference). In addition, the meaning of color adjectives is thought to be less dependent on the context and the object it describes (see Sedivy, 2003 for discussion).

The fact that referential form in general is highly contextually dependent is thought to influence not only the way in which speakers design their referring expressions, but the moment-by-moment interpretation of language as well. Classic studies show that the presence of multiple candidate referents prompts listeners to interpret an ambiguous prepositional phrase such as "…*the frog on the napkin*…" as a modifier; in the absence of competing referents, noun phrase modification is not required and the prepositional phrase is more likely to be interpreted as an argument of the verb (Novick, Thompson-Schill, & Trueswell, 2008; Snedeker & Trueswell, 2003; Tanenhaus, et al., 1995).

The fact that prenominal scalar adjectives are highly context sensitive makes them a good candidate for investigating contextual influences in language comprehension. Along these lines, Sedivy, Tanenhaus, Chambers, and Carlson (1999) asked whether listeners use the presence of a scalar adjective in an unfolding noun phrase as a cue that the referent would be a member of a contrast set denoted by the adjective. Sedivy and colleagues evaluated this idea by examining the interpretation of scalar-modified noun phrases such as "*the tall glass*" in the following two contexts: The first context contained a pair of items matching the head noun that contrasted along the scalar dimension denoted by the adjective (e.g., a tall glass and a short glass), a competitor object that was consistent with the adjective but not in a contrast set (e.g., a large pitcher), and an unrelated item (e.g., a key). In the second context, the size-contrasting item (e.g. short

glass) was replaced with an unrelated item (e.g., a file folder). Sedivy, et al. found that when interpreting scalar-modified noun phrases, listeners looked at the intended referent (tall glass) more quickly when the size-contrasting object (short glass) was present in the display, compared to when it was replaced by an unrelated item (file folder). This finding demonstrates that interpretations of a prenominal scalar adjective is facilitated by the presence of a relevant scalar-contrast in the display.

While this finding suggests that scalar adjectives have predictive validity, it is not clear to what extent they require a contextually supported pragmatic inference about the speaker's referential intent. Grodner and Sedivy (2011) reasoned that if the contrast effect was contextually supported that it could be attenuated if the speaker deviated from the normal conversational usage of scalars, and therefore was unlikely to use scalar adjectives with the intent to highlight a contextual contrast. If, on the other hand, the scalar-elicited search for a contrast set is more automatic or directly tied to semantic properties of scalar adjectives, it should be impermeable to circumstantial factors. To tease apart these two accounts, Grodner and Sedivy followed up on Sedivy et al. (1999), with a task in which participants heard instructions such as "Pick up the tall glass", produced either by a reliable or an unreliable speaker. The reliable speaker used language in a conventional manner whereas the unreliable speaker demonstrated an idiosyncratic use of scalar adjectives and otherwise infelicitous use of the language in general (this variable was manipulated between subjects). In the unreliable speaker condition, participants were told that the instructions were recorded by a speaker with "an impairment that caused language and social problems," the speaker mislabeled objects and referred to inappropriate locations, and was consistently over-informative in the use

of size adjectives (e.g., "the tall glass" in a context with a single glass). In the reliable speaker condition, no information was provided about the speaker's mental status, the speaker made no labeling errors, and used scalar adjectives appropriately (e.g., "the tall glass" in a context with a tall and short glass). Analysis of eye-gaze as participants interpreted *"the tall glass…*" showed that participants in the reliable speaker condition made more fixations to the intended referent (e.g., tall cup) when a contrast item (e.g., short cup) was present, replicating Sedivy et al. (1999), while listeners in the unreliable speaker condition did not. This result suggests that participants can suspend their contrastive interpretation when the current speaker uses this form in an unconventional way, and along with other findings (Sedivy, 2003), supports the idea that the contrastive interpretation of scalar adjectives is a contextually-sensitive pragmatic effect.

Several questions, however, remain about the locus of this contextual sensitivity. One question is whether this reflects a speaker-specific effect, or a general adjustment to the unconventionality of the local context. A variety of findings in the language processing literature show that listeners adjust the way they use and understand language with specific partners (Brown-Schmidt, 2009; Creel, Aslin, & Tanenhaus, 2008; Graham, Sedivy, & Khu, 2014; Matthews, Lieven, & Tomasello, 2010; Metzing & Brennan, 2003; Yoon & Brown-Schmidt, 2014). However, given that participants in Grodner and Sedivy's (2011) experiment only interacted with a single partner leaves open the possibility that the observed adjustments were in reaction to the context more generally, and not to the specific partner (see Brown & Dell, 1987; Creel, 2014, for related discussion).

A second open question concerns the nature of the observed modulation. In Grodner and Sedivy's (2011) design, an explicit instructional manipulation was used to convey that the speaker's use of language may be idiosyncratic in unpredictable ways. This top-down manipulation may have supported learning about characteristics of the unconventional speaker, either by drawing attention to the disconnect between referential choices and context (e.g., Arnold, Hudson Kam, & Tanenhaus, 2007), or possibly by providing a scaffolding to support learning (e.g., Bransford & Johnson, 1972). Note that Grodner and Sedivy (2011) do mention a follow-up experiment that did not use an explicit instruction and reported a null effect. In the present research, we delve into this issue by manipulating the types and amount of the bottom-up input given to the listener. Listeners have been shown to track the relative frequencies of a speaker's linguistic choices in various domains (e.g., Creel et al., 2008; Creel, 2014; Creel & Tumlin, 2011; Fine et al., 2013; Fine & Jaeger, 2013; Wells, Christiensen, Race, Acheson, & MacDonald, 2009). Yet, little is known about whether listeners track pragmatic choices in the same way. If distributional learning about the contexts in which the speaker uses a scalar adjective supports pragmatic adjustments, then in principal this adjustment should be observable in the absence of a strong top-down cue that the speaker is infelicitous or unconventional.

In what follows we present four experiments that build upon the findings of Grodner and Sedivy (2011), addressing two major remaining unanswered questions, 1) whether the bottom-up input alone can modulate *online* language processing and 2) whether eye-movements can be modulated for two speakers simultaneously in one discourse. To anticipate our results, we find that massive exposure to bottom-up evidence

can trigger modulation of contrastive inferences but there was little evidence of speaker-specificity. We highlight implications of our results and discuss possible learning mechanisms that support online pragmatic inferences in linguistic communication.

Experiment 1a

The aim of Experiment 1a is to test whether listeners modulate their pragmatic inferences based on bottom-up exposure alone. We test this by examining eye fixations during the interpretation of scalar adjectives after exposure to felicitous or infelicitous speakers using scalar cues to contrast.[1]

Method

Participants. A total of 40 students from the University of Illinois at Urbana-Champaign were given partial course credit upon participation in the experiment. All participants were fluent speakers of American English with normal or corrected-to-normal vision and normal hearing.

Procedure. Participants first completed a brief training task in which they saw four shapes on the computer screen, and heard an instruction to click on one of them. To provide a believable context, they were told to imagine that "little Joe" and his mom are playing a game on her computer and their job is to listen to her speech and click on the pictures that she is talking about. There were two instructions per set of four shapes (e.g., "Show me the red dotted triangle, Now show me the blue triangle."). Participants were told to execute the action by clicking on the shape. This was followed by the test phase

[1] A secondary goal of Experiment 1a was to test another question relating to the modulation of online interpretations of scalar adjectives based on prosodic cues. For the sake of clarity, and because those results were equivocal, we do not discuss those data in the main text of this paper but a summary can be found in the supplemental material available online.

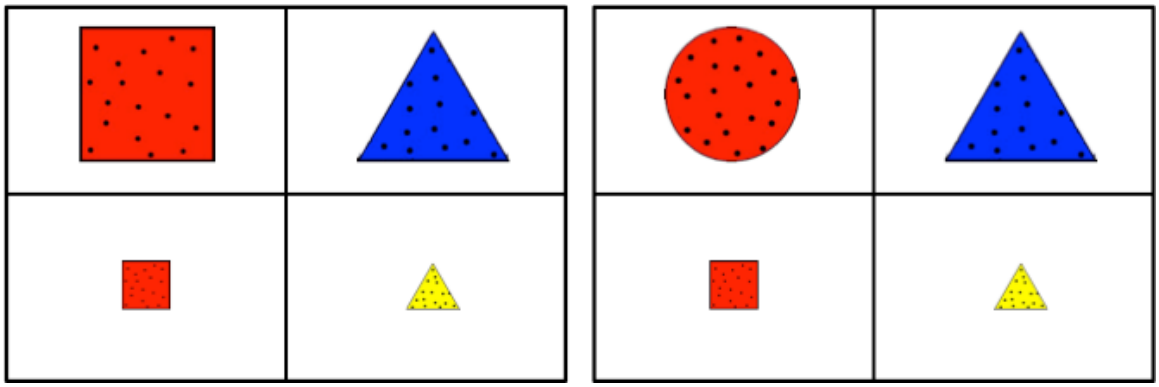
during which participants heard audio instructions such as, “Click on the small dog.”, and clicked on the corresponding object in a display with four pictures. Participants were informed that these sentences were recorded by the same speaker as the sentences in the first part. During the subsequent test phase, their eye movements were tracked using an EyeLink-1000 desktop mounted eye-tracker, using a sampling rate of 1000 Hz. Stimuli were presented on a 1920 by 1080 pixel display using the Psychtoolbox-3 extension (Brainard, 1997; Kleiner, Brainard, Pelli 2007; Pelli, 1997) for Matlab. The experimental session lasted about 20 minutes.

Materials.

Training phase. The training task consisted of 12 trials in which participants saw a 2 x 2 grid of shapes (squares, triangles, or circles) with different combinations of colors (red, blue, or yellow), sizes (big or small), and patterns (checkers, dots, or stripes). Each training trial consisted of a grid with four shapes and a set of two paired instructions such as “Find me the circle. Now, find me the large blue triangle.” There were four different carrier phrases: “Show me the X.”, “Point to the X.”, “Find me the X.”, and “Where’s the X?”. The order of the training trials was randomized for each participant separately within the Matlab program. Participants were randomly assigned to one of two (between-subjects) training conditions: Felicitous-Scalars, Infelicitous-Scalars.

During the training phase, the four objects varied in shape, color, and size, but not in pattern (e.g., all the objects on the screen were dotted but they were different in shape and color, see Figure 1). On every trial, two of the shapes were big and two were small. Participants heard pairs of instructions, such as “Show me the circle. Now, show me the large yellow square.” The second instruction always contained a scalar adjective. Prosody

was neutral across all the Scalar training sentences. Only one pair of instructions was recorded per display and Felicity was manipulated via the properties of the visual display (see Figure 1). In the Felicitous-Scalars condition, twelve trials contained felicitous use of scalar adjectives (e.g., "Show me the blue triangle. Now, show me the small square.") when a size contrast was present in the display (e.g., a small red square, a large red square, a large blue triangle, and a small yellow triangle). In the Infelicitous-Scalars training condition, twelve trials contained infelicitous scalar adjectives (e.g., "Show me the blue triangle. Now, show me the small square."), in which the scalar adjective was superfluous with respect to the goal of reference because the target item was uniquely identifiable without it. The target was either a singleton (e.g., there is only one square on the screen) or distinguished by color but not size (e.g., "Point to the triangle. Now, point to the small red circle." when there are two circles and both of them are small).

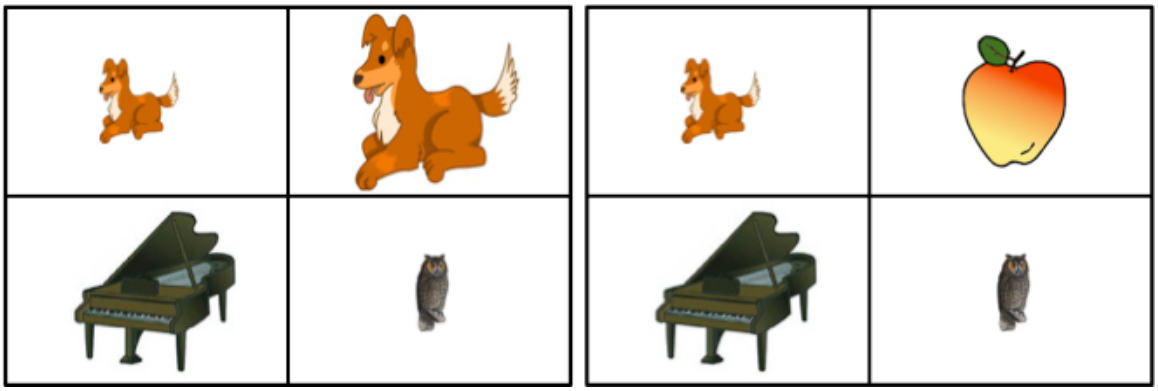


Felicitous training condition Infelicitous training condition

Figure 1. Experiment 1a: Example Scalar training conditions corresponding to the instruction, "Show me the blue triangle. Now, show me the small square."

Test Phase. The test phase consisted of 50 trials (20 critical and 30 filler trials) in which participants saw a 2 x 2 grid of pictures of animals and common objects in conjunction with an auditory instruction about which item to click on (Figure 2). The test

trials were identical across the Felicitous-Scalars and the Infelicitous-Scalars conditions. All participants saw ten critical trials in the Contrast condition in which the instruction contained a scalar adjective (e.g., “Click on the small dog.”) and the display contained a pair of size-contrasted items, one of which was the target item (e.g., a small dog and a large dog), and two distractors (e.g., a large piano and a small owl). Another ten critical trials comprised the No Contrast condition in which the auditory stimulus contained a scalar adjective (e.g., “Click on the small dog.”) and the display contained a large or small target object, a different object contrasting in size with the target, a large distractor, and a small distractor (e.g., a small dog, a large apple, a large piano, and a small owl).



Contrast No Contrast

Figure 2. Experiment 1a: Example trials from test phase (Contrast vs. No Contrast condition) accompanied by the instruction, “Click on the small dog.”

Ten of the filler trials belonged to the Contrast control condition in which the instructions did not contain a scalar adjective (e.g., “Point to the dog.”). The display contained a large or small target object, a distractor object contrasting in size with the target, and a pair of size-contrasted distractor items (e.g., a small dog, a large flag, large scissors, small scissors). Ten of the filler trials were in the Other Contrast condition in

which the instructions contain a non-scalar contrast (e.g., "Point to the leather jacket."). The display contained a target object, a contrasting object, a large distractor, and a small distractor (e.g., a leather jacket, a rain jacket, a large hydrant, and a small dollar). The last ten filler trials were the Other Contrast Control condition in which the instructions did not contain a contrast (e.g., "Point to the bike"). The display contained a target item, an item contrasting in size with the target, and a pair of contrasting items (e.g., a small bike, a large glass, a lead pencil, and a coloring pencil). See Appendix A for a summary of trials in all experiments.

Two counterbalancing lists were created to allow target items (e.g., small dog) to appear both in the Contrast and No Contrast conditions across subjects. Target items were never repeated for a given subject, but they could reappear as distractor items. Participants were randomly assigned to a counterbalancing list.

Results

Interpretation of the scalar adjective was indexed by the proportion of eye-movements that participants made to the target item as they interpreted the critical instructions, which consisted of a scalar adjective and a noun (e.g., *Click on the small dog*). A fixation was coded as a target-fixation if the x,y fixation-coordinates landed on the target object (e.g., *the small dog*), or on the white space in the quadrant of the screen surrounding it (this buffer space did not overlap with any other object). A plot of the full time-course of target fixations by conditions can be seen in Appendix B. Target fixations were measured in a time window (average duration 726 milliseconds) that began 200 milliseconds after the onset of the adjective (e.g., *small*) and ended 200 milliseconds after the average offset of the noun (e.g., *dog*; the average duration of the noun was 475

milliseconds). The 200 millisecond delay was necessary due to the time needed to program and launch an eye movement (Hallett, 1986). The average proportions of target fixations within this time-window across Felicity and Contrast conditions are shown in Figure 3.

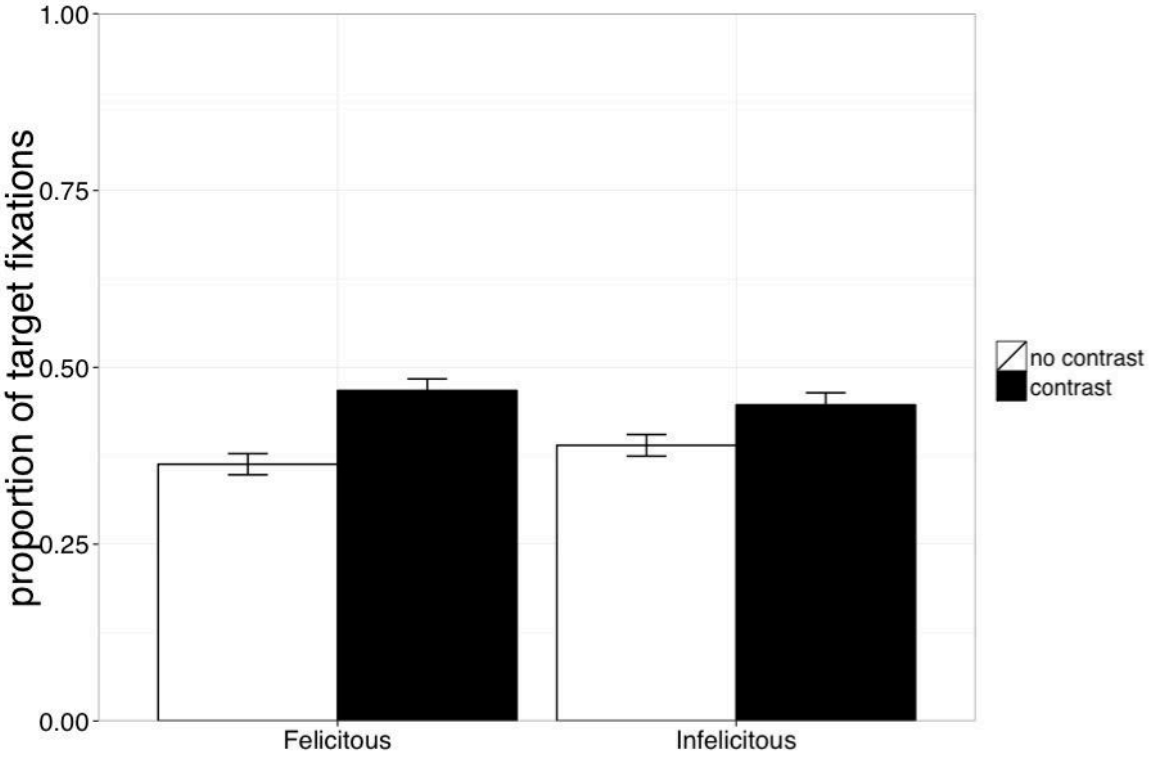


Figure 3. Experiment 1a: Average proportions of target fixations during interpretation of the scalar adjective and noun (e.g., *Click on the small dog*) by Felicity and Contrast conditions.

The trial-level proportions of target fixations were first transformed using the empirical logit transformation and then analyzed in a multilevel linear regression, using the *lme4* software package in R (Bates, Maechler, Bolker, & Walker, 2015). Felicity and Contrast along with their interaction were entered as fixed effects (Table 1) with participants and items as random effects. All fixed effects were coded with mean-centered contrast codes. When the maximal random effects structure justified by the

design did not converge or the model was overfit, random slopes with the least variance were removed until convergence was achieved. The final model included random intercepts for participants and items. Model comparison was used to assess the significance of effects.

There was a main effect of Contrast, such that participants made more target fixations in the Contrast condition than the No Contrast condition (t= -3.4). There was no significant interaction of Contrast and Felicity (t=-0.25). The main effect of Felicity was also not significant (t=-0.25).

Fixed Effects				
	β	Std. Error	t-value	p-value
(Intercept)	0.00	0.07	0.02	
Contrast condition	-0.17	0.05	-3.41	**$1.10e^{-3}$**
Felicity condition	-0.02	0.06	-0.25	0.80
Contrast x Felicity	-0.01	0.04	-0.25	0.82
Random Effects				
Groups		Variance	Std. Dev.	
Participants	(Intercept)	0.07	0.27	
Items	(Intercept)	0.06	0.24	
Residual		0.85	0.92	
Observations: 800; Items: 40; Participants: 40				

Table 1. Experiment 1a: Results of the linear mixed-effects model of target fixation proportions

Discussion

During the interpretation of scalar adjectives, participants made more fixations to targets that were in a contrast set, consistent with results from Sedivy et al., (1999). The felicity manipulation, on the other hand, yielded no significant difference between the two conditions. One possible source of this null effect may be the fact that the Infelicitous training trials consisted only of over-informative instructions. It has been reported that naturalistic adjective use contains a large amount of instances in which an adjective is not

strictly necessary with respect to the goal of unique reference (Brown-Schmidt & Konopka, 2011). Over-informative adjectives may not impair on-line language processing (Arts, Maes, Noordman, & Jansen, 2011; Davies & Katsos, 2009; Levelt, 1989; cf., Engelhardt, Demiral, & Ferreira, 2011), and may reflect natural properties of utterance formulation (Belke, 2006; Pechmann, 1989). Furthermore, because the three shapes (circles, square, triangle) were repeated across training trials, the use of scalars that were over-informative for a given visual display may have been attributed to a tendency to lexically differentiate the currently observed shapes from those seen on previous trials (Van der Wege, 2009; Yoon & Brown-Schmidt, 2014) or to a looser definition of the comparison class—for example, the comparison class for a triangle may be other shapes in general (e.g., "*small* square" to contrast with a larger circle). Thus, due to speakers' tendency to over-use adjectives in general, and the potential attribution of over-modification to other communicative goals, participants may not have perceived our over-modified instructions as truly infelicitous (Engelhardt et al., 2006; Pogue et al. 2016). Indeed, post-hoc norming (see Appendix C for details) of the naturalness of instructions paired with their corresponding displays (on a scale of 1 to 5) revealed that participants rated Infelicitous instructions (M= 4.83) as equally natural compared to Felicitous instructions (M= 4.88).

Experiment 1b

The aim of Experiment 1b is to test whether the modulation of pragmatic inference is dependent on the type of infelicitous cue used to convey a speaker's unreliability. In Experiment 1a, the over-informative scalars may not have been perceived

as infelicitous because other size contrasts existed in the displays and speakers' general proclivity to over-use adjectives. In Experiment 1b, infelicity is made more salient by making all the shapes in the training displays the same size, invalidating the uses of size adjectives.

Methods

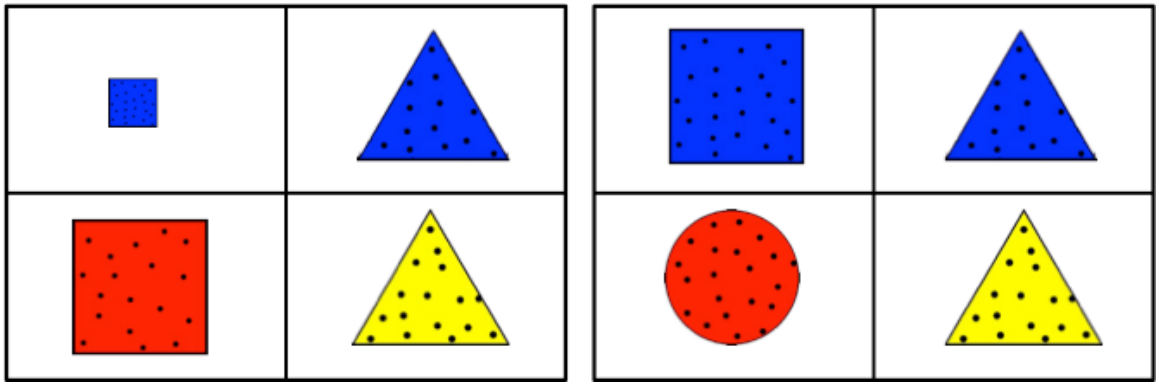
Participants. A total of 68 students from the University of Illinois at Urbana-Champaign were given partial course credit upon participation in the experiment. All participants were fluent speakers of American English with normal or corrected-to-normal vision and normal hearing.

Procedure & Materials. The procedure and materials were identical[2] to Experiment 1a except that the training trials in Experiment 1b were designed such that the Infelicitous training trials were more strongly infelicitous than in Experiment 1a. The scalar adjectives were over-informative relative to the entire set of items in the display because all four items were the same size (see Figure 4). By contrast, in Experiment 1a, the Infelicitous training trials were only over-informative relative to one pair of items—there were always both big and small shapes in the display.

The auditory stimuli were identical across Felicitous and Infelicitous conditions and Felicity was manipulated through the visual display, by switching one (or two) of the items. The Felicitous training condition was identical to Experiment 1a. In the Infelicitous training condition, six of the trials included a second instruction with a pre-nominal scalar adjective (e.g., "Point to the small circle.") paired with a display that made

[2] Due to experimenter error, participants in Experiment 1b were eye-tracked with a different sampling rate, 250 Hz, except for 2 participants who were tracked at 1000 Hz. Due to the comparatively large time windows of analysis, there is no reason to believe that the lower sampling rate affected the results. Note that Sedivy, et al. (1999) recorded data at a much lower sampling rate – 60hz.

the scalar redundant (e.g., a small red circle, a small blue triangle, a small yellow triangle, a small blue square). Another three trials contained a post-nominal scalar adjective (e.g., "Show me the circle that's large") paired with a display in which the scalar adjective is redundant (e.g., large red circle, large yellow square, large blue triangle, large red square). The last three training trials contained a scalar adjective (e.g., "Where's the small square?") and the display contained four shapes the size of which was inconsistent with the scalar adjective used in the audio stimulus (e.g., large blue square, large blue triangle, large red circle, and large yellow triangle). See Appendix A for a summary of trials in all experiments.



Felicitous Infelicitous

Figure 4. Experiment 1b: Example training trials across conditions, accompanied by the instruction, "Where's the yellow triangle? Now, where's the small square?"

Results

As in Experiment 1a, interpretation of the scalar adjective was indexed by the proportion of eye-movements that participants made to the target item as they interpreted the critical instructions, which consisted of a scalar adjective and a noun (e.g., *Click on the small dog*). Fixations were coded in the same way as Experiment 1a. A plot of the full

time-course of target fixations by conditions can be seen in Appendix B. Target fixations were measured in the same time window as Experiment 1a (average duration 726 milliseconds). The average proportions of target fixations in this time window, across the Felicity and Contrast conditions are shown in Figure 5.

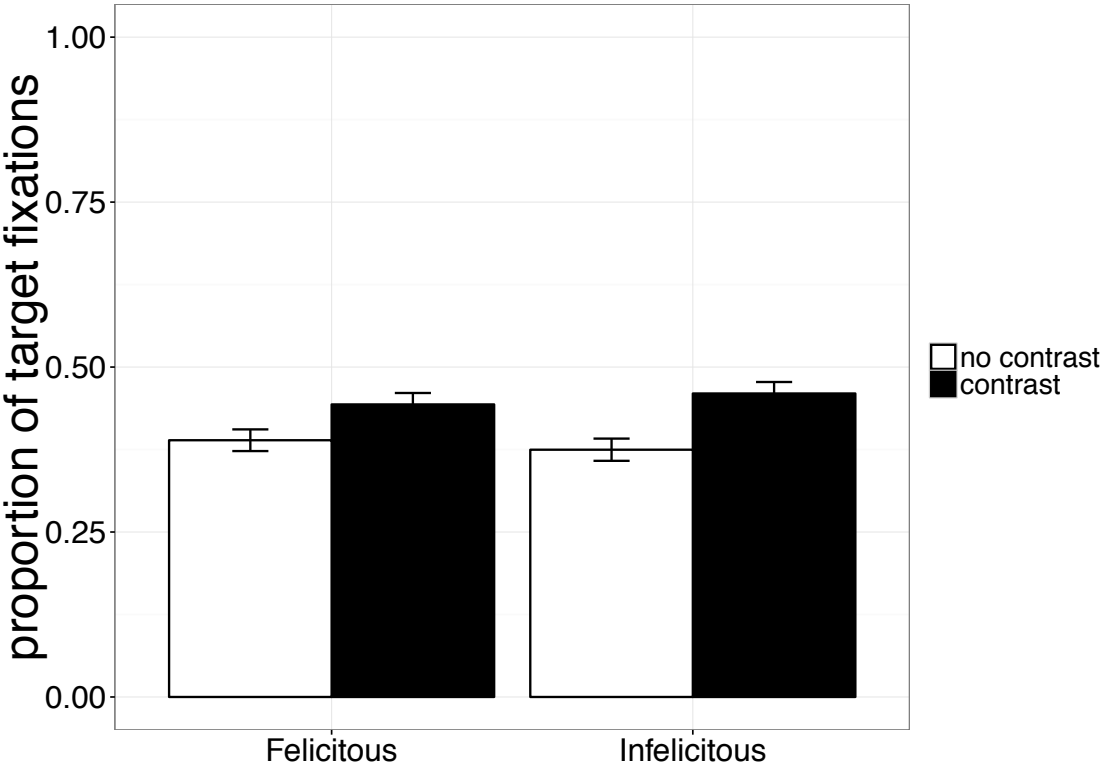


Figure 5. Experiment 1b: Average proportions of target fixations during interpretation of the scalar adjective and noun (e.g., *Click on the small dog*) by Felicity and Contrast conditions.

As in Experiment 1a, the trial-level proportions of target fixations were first transformed using the empirical logit transformation and then analyzed in a multilevel linear regression. Felicity and Contrast along with their interaction were entered as fixed effects (Table 2) with subjects and trials as random effects. All fixed effects were coded with mean-centered contrast codes. The same procedure from Experiment 1a was used to determine the random effect structure and to compare models. The final model included

random intercepts for participants and items. There was a main effect of Contrast, such that participants made more target fixations in the Contrast condition than the No Contrast condition (t= -2.84). There was no significant interaction of Contrast and Felicity (t=0.04), nor a main effect of Felicity (t=-0.84).

Fixed Effects				
	β	Std. Error	t-value	p-value
(Intercept)	0.00	0.06	0.00	
Contrast condition	**-0.11**	**0.04**	**-2.84**	**$5.48e^{-3}$**
Felicity condition	0.00	0.06	0.04	0.97
Contrast x Felicity	-0.03	0.03	-0.84	0.40
Random Effects				
Groups		Variance	Std. Dev.	
Participants	(Intercept)	0.17	0.41	
Items	(Intercept)	0.04	0.19	
Residual		0.79	0.89	
Observations: 1360; Items: 40; Participants: 68				

Table 2. Experiment 1b: Results of the linear mixed-effects model of target fixation proportions

Discussion

As in Experiment 1a, during the interpretation of scalar adjectives participants made more fixations to targets that were in a contrast set, replicating Sedivy et al., (1999). This contrast-contingent target preference was not diminished when participants were exposed to a speaker who used scalar adjectives in a contextually infelicitous manner. These results may suggest that listeners do not discount scalar adjectives as a cue to a contextual contrast solely based only on the exposure to infelicitous uses of these adjectives.

Alternatively, it is possible that the experimental design used in 1a and 1b may not have provided suitable circumstances to observe speaker-specific modulation of online interpretation of scalar adjectives. It may be that participants altered their

pragmatic inferences during the training phase but it did not transfer to the test phase, because the context change was too abrupt. Even though we tried to ensure the continuity of the two phases by instructing the participant that the speaker remained the same across the training and test phases, these phases differed in two important ways: (i) the training phase consisted of trials with colored geometric shapes while the test phase consisted of trials with more complex images, (ii) participants' eye-movements were monitored during the test phase but not the training phase (and as a result, a calibration occurred in between the two phases). Memory retrieval is contextually sensitive, and changes in context from learning to test can impair retrieval of previously-learned information (Godden & Baddeley, 1975; Smith & Vela, 2001; cf. Eich, 1985, Mulligan, 2011). As a result, the contextual changes between training and test might have limited transfer of learning, mitigating whatever effect of training there was in the first place.

Furthermore, the amount of evidence that the speaker was infelicitous may have been insufficient to elicit the predicted speaker-specific changes of interpretations. Across the entire experimental session, the proportion of infelicitous uses of a pragmatic cue was only 36% in the Infelicitous conditions[3]. Perhaps the large number of felicitous trials (including 30 felicitous fillers) at test counteracted any adaptation that resulted from the training phase. Experiment 2 was conducted to address these concerns.

Experiment 2

The goal of Experiment 2 is to provide a more suitable environment for observing changes in the interpretation of scalar adjectives. To achieve this, the contexts during

[3] The training phase consisted of 12 infelicitous trials and the test phase contained 10 infelicitous trials (No Contrast trials), for a total of 22 infelicitous trials out of a grand total of 62 trials across training and test (22/62= 0.355).

training and test were made more similar in multiple ways: (i) training, test, and filler trials were randomly intermixed, (ii) eye-tracking occurred during all trials and the calibration was done at the very beginning of the experimental session, (iii) the types of images used in training and test were the same. Additionally, in the Infelicitous condition in Experiment 2, infelicitous sentences constitute the vast majority (93%) of what the participant is exposed to. Finally, the longer testing period in this experiment, where order of trial and item are not confounded due to randomization of the trial order, will allow us to examine the potential emergence of the modulation effect as the evidence of the speaker's infelicitous scalar adjective use accumulates. In other domains, where evidence of learning new probabilities for a speaker's linguistic choices has been obtained (e.g., syntactic alternatives in Fine et al., 2013), the gradual change in processing time in response to evidence available in the environment provides compelling support for an error-based, implicit learning account of these changes (Chang, Dell, & Bock, 2006).

Method

Participants. Sixty-four undergraduate students at the University of Illinois at Urbana-Champaign participated in this experiment in exchange for partial course credit. Participants had normal or corrected-to-normal vision and spoke English fluently.

Procedure. Participants listened to instructions such as "Click on the big dog" and viewed 4-picture displays on a desktop computer while their eye-movements were monitored using an Eyelink-1000 desktop-mounted eye-tracker at 1000hz. They were instructed to click on the item that best matched what the speaker said and, if they weren't sure, to just use their best guess. This instruction was included because some of

the instructions were globally ambiguous as a result of our felicity manipulations. As opposed to Experiment 1, there was no cover story about where the sentences came from. This was because the background story might bias listeners to invoke unique assumptions about the properties of child-directed speech (e.g., frequent non information seeking questions with pedagogical intentions). We chose to keep the focus here on examining whether listeners learn about the pragmatic competence of a speaker in a situation where the default assumption is that the speaker will be engaging in cooperative interactions. The entire experiment lasted about 30 minutes.

Materials. Each participant saw 80 training trials (Figure 6), 40 test trials (Figure 7), and 180 fillers, for a total of 300 trials. The order of trials was randomized for each subject with the constraint that the first three trials were filler trials. Participants were randomly assigned to the Felicitous or Infelicitous condition. Training trials differed by Felicity condition. Forty of the training trials were Contrast trials. The instruction about which object to select always contained a scalar adjective (e.g., "Click on the big briefcase"). In the Felicitous condition, the adjective "big" was justified: the Contrast displays (Figure 6a) contained a target item (e.g., a big briefcase), an item that was in a size contrast pair with the target (e.g., a small briefcase), a large distractor item (e.g., a big lizard) and a small distractor item (e.g., a small car seat). In the Infelicitous condition, the adjective "big" was not justified: the Contrast displays (Figure 6c) contained a target item (e.g., a big briefcase), a distractor item that was NOT in a size contrast with the target (e.g., a small calculator), a large distractor item (e.g., a big lizard) and a small distractor item (e.g., a small car seat).

The other 40 training trials were No Contrast trials. On these no-contrast trials, the instruction about which object to select never contained a scalar adjective (e.g., "Click on the briefcase"). In the Felicitous condition, the lack of an adjective was justified: the No Contrast displays (Figure 6b) contained a target item (e.g., a big briefcase), a distractor item that was *not* in a size contrast with the target (e.g., a small calculator), a large distractor item (e.g., a big lizard) and a small distractor item (e.g., a small car seat). In the Infelicitous condition, the lack of an adjective was not justified: the No Contrast displays (Figure 6d) contained a target item (e.g., a big briefcase), an item that was in a size contrast pair with the target (e.g., a small briefcase), a large distractor item (e.g., a big lizard) and a small distractor item (e.g., a small car seat). Thus, for participants in the Infelicitous condition, Contrast training trials were infelicitous because they contained a contrastive adjective in the absence of a size contrasting pair of items and No Contrast training trials were infelicitous because no size adjective was used when it was necessary in order to disambiguate between two differently-sized but otherwise identical items.

Felicitous Training condition

"Click on the big briefcase."
a. Contrast condition

"Click on the briefcase."
b. No Contrast condition

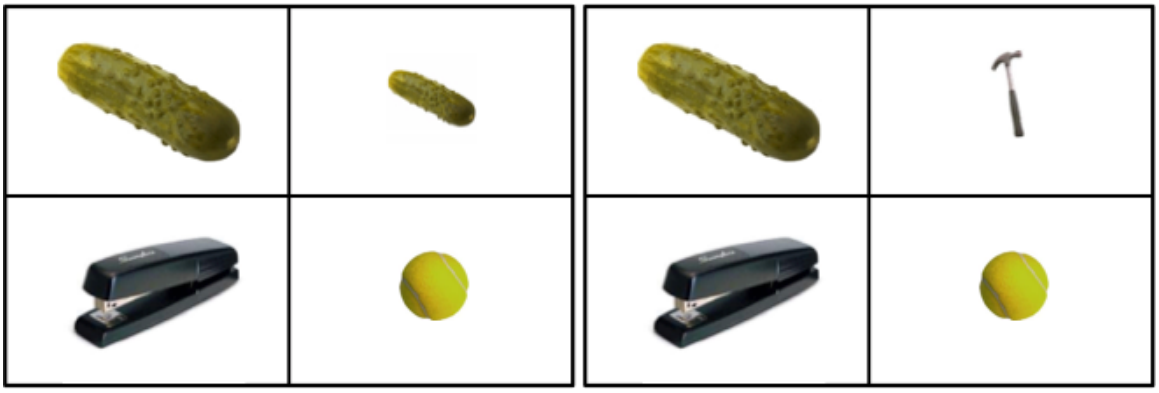
Infelicitous Training condition

"Click on the big briefcase." c. Contrast condition

"Click on the briefcase." d. No Contrast condition

Figure 6. Experiment 2: Example training trials across Felicity and Contrast v. No Contrast conditions.

Test trials were identical across Felicitous and Infelicitous conditions. Twenty of the test trials were Contrast trials (Figure 7a.). The instruction about which object to select always contained a scalar adjective (e.g., "Click on the big pickle"). The display contained a target item (e.g., a big pickle), an item that was in a size contrast pair with the target (e.g., a small pickle), a large distractor item (e.g., a big stapler) and a small distractor item (e.g., a small tennis ball). The other 20 test trials were No Contrast trials (Figure 7b.). The instruction also always contained a scalar adjective (e.g., "Click on the big pickle"), however in the No Contrast trials the adjective was infelicitous. The display contained a target item (e.g., a big pickle), a distractor item that was in a size contrast with the target (e.g., a small hammer), a large distractor item (e.g., a big stapler) and a small distractor item (e.g., a small tennis ball).



a. Contrast b. No Contrast

Figure 7. Experiment 2: Example test trials in the Contrast and No Contrast conditions. Corresponding audio instruction for both conditions: “Click on the big pickle.”

Filler trials also differed by Felicity condition. Sixty filler trials were Contrast Control trials in which the instructions did not contain a scalar adjective (e.g., “Click on the cake.”) and two of the distractors were in a size contrast pair. In the Felicitous condition, the displays contained a target item (e.g., a big cake), a distractor item (e.g., a small toothbrush), and a pair of size-contrasted distractor items (e.g., a big burger and a small burger). In the Infelicitous condition, the single distractor item was replaced by a contrast for the target (e.g., a small cake), rendering the instruction globally ambiguous (e.g., “Click on the cake” when there are two cakes in the scene). A pair of size-contrasted distractor items (e.g., a big burger and a small burger) was also present in the display. Another sixty filler trials were Other Contrast trials. The audio instructions contained a non-scalar contrastive adjective, such as, “Click on the glazed doughnut.” In the Felicitous condition, the display contained a target item (e.g., a glazed doughnut), a contrasting item of the same category as the target (e.g., a powdered doughnut), and two distractors (e.g., a big muffin pan and a small saddle). In the Infelicitous condition, the display contained a target item (e.g., a glazed doughnut), but no contrasting item. There

were three distractors (e.g., a medium rug, a big muffin pan, and a small saddle). The last sixty filler trials were Other Contrast Control trials. The audio instructions did not contain an adjective (e.g., "Click on the couch"). In the Felicitous condition, the display contained a target item (e.g., a small couch), a distractor item (e.g., a big cheese plate), and a pair of items from the same category (e.g., a glazed doughnut and a powdered doughnut). In the Infelicitous condition, the display contained a target item (e.g., a small couch), a size-contrasted item (e.g., a big couch), and a pair of items from the same category (e.g., a glazed doughnut and a powdered doughnut). See Appendix A for a summary of trials in all experiments.

Two counterbalancing lists were created to allow target items (e.g., big pickle) to appear both in the Contrast and No Contrast conditions across subjects. Target items were never repeated for a given subject, but they could reappear as distractor items. Participants were randomly assigned to a single counterbalancing list.

Results

Interpretation of the scalar adjective was indexed by the proportion of eye-movements that participants made to the target item as they interpreted the critical instructions, which consisted of a scalar adjective and a noun (e.g., *Click on the big pickle*). Fixations were coded in the same way as Experiment 1. A plot of the full time-course of target fixations by conditions can be seen in Appendix B. Target fixations were measured in a time window (average duration 1197 milliseconds[4]) that began 200 milliseconds after the onset of the adjective (e.g., *big*) and ended 200 milliseconds after

[4] This duration differs from the one in Experiment 1 because the auditory stimuli for this experiment were expanded and recorded by a different speaker (the first author).

the average offset of the noun (e.g., *pickle*; average duration of the noun 564 milliseconds). Average proportions of target fixations across Felicity (Felicitous vs. Infelicitous) and Contrast (Contrast vs. No Contrast) conditions are shown in Figure 8.

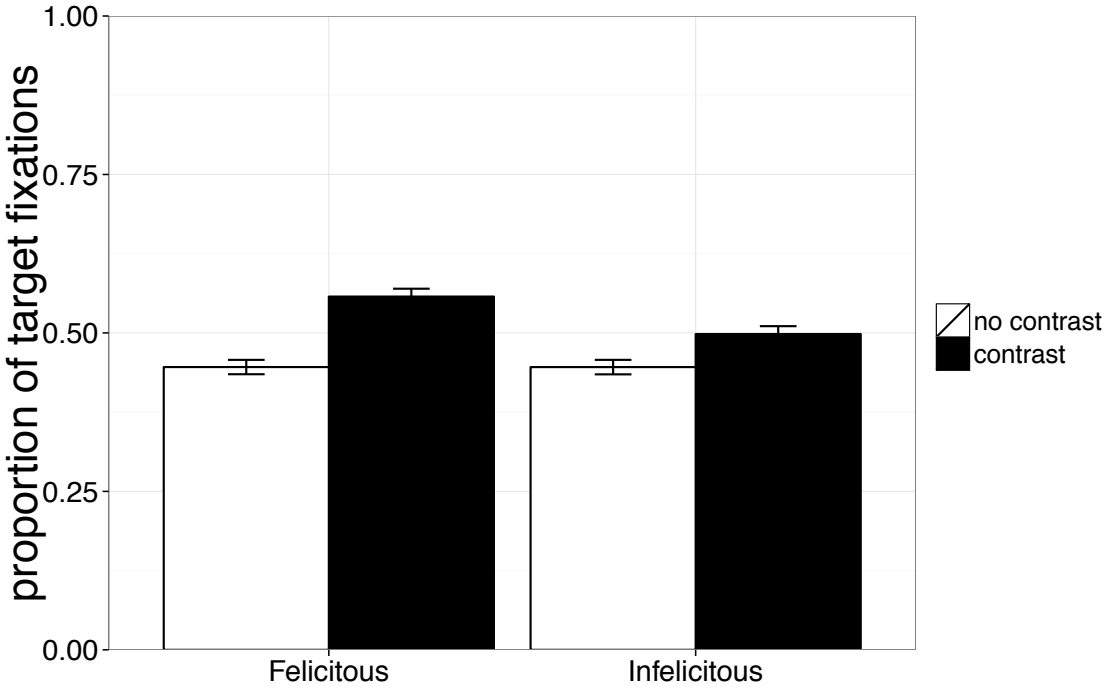


Figure 8. Average proportions of target fixations during interpretation of the scalar adjective and noun (e.g., *Click on the big pickle*) by Felicity and Contrast conditions.

The same procedure from Experiment 1 was used to determine a random effect structure and to compare models (Table 3). The final model included fixed effects for Felicity, Contrast, and their interaction, as well as random intercepts for participants and items. There was a main effect of Contrast, such that participants made more target fixations in the Contrast condition than the No Contrast condition (t= 6.04). Importantly, a significant interaction of Contrast and Felicity was observed (t= -2.2). Mixed effects regression analyses on subsets of the data, split by Felicity condition, revealed a large Contrast effect in the Felicitous Speaker condition (β= -0.20 , SE= 0.03, t= -6.62, $p< 5e^{-}$

^{08}), and a reduced effect of Contrast in the Infelicitous Speaker condition (β= -0.09 , SE= 0.04, t= -2.49, p< 0.05).

Fixed Effects				
	β	Std. Error	t-value	p-value
(Intercept)	$8.04e^{-16}$	0.04	0	
Contrast condition	**0.15**	**0.02**	**6.04**	**$2.66e^{-8}$**
Felicity condition	-0.05	0.04	-1.51	0.13
Contrast x Felicity	**-0.05**	**0.02**	**-2.20**	**0.03**
Random Effects				
Groups		Variance	Std. Dev.	
Items	(Intercept)	0.02	0.14	
Participants	(Intercept)	0.05	0.21	
Residual		0.91	0.95	
Observations: 2520; Items: 80; Participants: 63				

Table 3. Experiment 2: Results of the linear mixed-effects model of target fixation proportions

In order to understand whether this modulation emerges over time, we examine the data patterns across the two halves of the experiment (Table 4). In the first half of the experiment, participants encountered approximately 40 training trials, 20 test trials, and 90 fillers. As in the main analysis, there was a main effect of Contrast and a significant interaction of Contrast and Felicity. In the second half of the experiment, there was a main effect of Contrast and a main effect of Felicity, such that participants made more fixations to the target in the Felicitous condition than the Infelicitous condition. There was no significant interaction between Felicity and Contrast.

First Half				
Fixed Effects				
	β	Std. Error	t-value	p-value
(Intercept)	0.05	0.04	1.31	
Contrast condition	**0.18**	**0.03**	**6.74**	**$1.2e^{-9}$**
Felicity condition	-0.02	0.04	-0.47	0.70
Contrast x Felicity	**-0.09**	**0.03**	**-3.55**	**$5.12e^{-4}$**

Random Effects				
Groups		Variance	Std. Dev.	
Items	(Intercept)	$1.24e^{-14}$	$1.11e^{-7}$	
Participants	(Intercept)	0.04	0.19	
Residual		0.82	0.91	
Observations: 1210; Items: 80; Participants: 63				
Second Half				
Fixed Effects				
	β	Std. Error	t-value	p-value
(Intercept)	-0.04	0.04	-0.97	
Contrast condition	0.11	0.03	3.44	**$7.06e^{-4}$**
Felicity condition	**-0.09**	**0.04**	**-2.00**	**0.042**
Contrast x Felicity	-0.02	0.03	-0.46	0.638
Random Effects				
Groups		Variance	Std. Dev.	
Items	(Intercept)	0.02	0.15	
Participants	(Intercept)	0.05	0.22	
Residual		1.01	1.00	
Observations: 1310; Items: 80; Participants: 63				

Table 4. Experiment 2: Results of the linear mixed-effects model of target fixation proportions for each half of the experiment

Discussion

The contrast-contingent target preference was diminished when participants were exposed to a speaker who used adjectives infelicitously. These results suggest that listeners can discount scalar adjectives as a cue to a contextual contrast based on the bottom-up information alone. Experiment 1 and 2 together suggest that such modulation of online interpretation of scalar adjectives requires a consistent context as well as a great deal of evidence that the speaker is unlikely to use an adjective in an informative manner.

Of interest is that the Contrast effect was persistent, and was observed even in the Infelicitous condition in Experiment 2 despite overwhelming evidence that pragmatic cues were being used infelicitously (93% infelicitous sentences). Put another way, when

the speaker used a size adjective, 75% of the time she was referring to an item not in a contrast set (see Appendix A). Yet, listeners continued to anticipate a referent in a contrast set.

The analysis across halves of the experiment indicated that only half of the exposure that we provided was necessary for participants to begin modulating their pragmatic inferences in the Infelicitous condition. Furthermore, as the exposure to the Infelicitous speaker continued, participants decreased fixations to the target object across Contrast conditions, suggesting that the overwhelmingly infelicitous environment may have lead listeners to become generally less attentive to the task.

Experiment 3

The aim of Experiment 3 is to test whether the change in interpretation of contrastive adjectives observed in Experiment 2 can in fact be attributed to speakers, rather than to the experimental context more generally. Thus far experiments in this domain have used between-subject manipulations (Grodner & Sedivy, 2011), which makes it difficult to determine if comprehenders are changing their online processing with respect to speaker identities or to the task environment. To this end, we use a within-subject manipulation, in which we expose participants to two speakers, one who uses adjectives felicitously and the other who does not. If the reduction of anticipatory eye movements observed in Experiment 2 is speaker-specific, participants will suspend their contrast-contingent anticipatory fixations only when the infelicitous speaker is heard. A large body of literature suggests that language users track speaker-specific information and apply that knowledge in the service of comprehension efficiency (see Creel &

Bregman, 2011 for review). In domains related to referential resolution, these speaker-specific representations consist of various types of information such as what other individuals do and don't know (Brennan & Clark, 1996; Metzing & Brennan, 2003; Yoon & Brown-Schmidt, 2014), verb attachment preferences (Kamide, 2012), spatial viewpoint (Ryskin, Wang, & Brown-Schmidt, 2016), and how speakers tend to use specific types of modifiers such as prenominal adjectives (Pogue, Kurumada, Tanenhaus, 2016) and quantifiers (Yildirim, Degen, Tanenhaus, Jaeger, 2016). Thus, in the current task, it seems likely that listeners would track each speaker's adjective use and its felicity, and modulate their eye-movements in a speaker-specific manner. However, it is also possible that listeners do not keep track of the adjective use distribution in a strictly speaker-specific manner. Rather, listeners may average the recent input across speakers. In fact, one might argue that tracking the adjective use distribution of every speaker may not be an efficient strategy due to attentional and memory constraints (see Brown-Schmidt, Yoon, & Ryskin, 2015 for discussion). If this is the case, in the current experiment, we expect to replicate Sedivy, et al.'s (1999) Contrast effect with no difference between the two speakers. Recall that a rate of 36% infelicitous scalars was not sufficient to modulate this contrast effect (Experiment 1) – it was only with 93% infelicitous scalars did we observe a reduction in the Contrast effect (Experiment 2). Thus, if speaker-specific felicity is not tracked, a rate of 50% infelicitous scalars (averaged across the two speakers) would not likely be enough to overcome listeners' strong bias to expect cooperativity from the speakers, and we should observe a Contrast effect.

Method

Participants. Sixty-five undergraduate students at the University of Illinois at Urbana-Champaign participated in this experiment in exchange for partial course credit. Participants had normal or corrected-to-normal vision and spoke English fluently.

Procedure. The procedure was identical to that of Experiment 2, except that participants were informed that they would be hearing two speakers. Participants listened to instructions such as "Click on the big dog" and viewed 4-picture displays on a desktop computer while their eye-movements were monitored using an Eyelink-1000 desktop-mounted eye-tracker at 1000hz. They were instructed to click on the item that best matched what the speaker said and, if they weren't sure, to just use their best guess. The entire experiment lasted about 30 minutes.

Materials. Each instruction was recorded both by a male speaker and a female speaker. Each participant saw 80 training trials (40 produced by the male speaker and 40 by the female speaker), 80 test trials (40 produced by the male speaker and 40 by the female speaker), and 180 fillers (90 produced by the male speaker and 90 by the female speaker), for a total of 340 trials. The order of trials was randomized for each subject with the constraint that the first three trials were filler trials. Each participant received both Felicitous and Infelicitous training; which of the two voices was associated with the Felicitous trials and which was associated with Infelicitous trials was randomly determined for each participant. Felicity and Contrast in the training, filler, and test trials were manipulated in the same way as in Experiment 2 (Figures 6 and 7). The Infelicitous Speaker produced 40 Infelicitous training trials (Contrast and no Contrast) and 90 Infelicitous filler trials (Contrast control, Other Contrast, and Other Contrast Control). The Felicitous Speaker produced 40 Felicitous training trials (Contrast and no Contrast)

and 90 Felicitous filler trials (Contrast control, Other Contrast, and Other Contrast Control). Both speakers produced 40 unique test trials each (20 Contrast and 20 no Contrast), such that the power to detect the pragmatic modulation is the same for each speaker as it was in Experiment 2. See Appendix A for a summary of trials in all experiments. Two counterbalancing lists were created so that the two voices were paired with different items across the lists.

Results

Interpretation of the scalar adjective was indexed by the proportion of eye-movements that participants made to the target item as they interpreted the critical instructions, which consisted of a scalar adjective and a noun (e.g., *Click on the big pickle*). A fixation was coded as a target fixation in the same way as in Experiments 1 and 2. A plot of the full time-course of target fixations by conditions can be seen in Appendix B. Target fixations were measured in a time window (average duration 1135 milliseconds[5]) that began at the onset of the adjective (e.g., *big*) and ended 200 milliseconds after the average offset of the noun (e.g., *pickle* ; average duration of the noun 542 milliseconds). The time window was offset by 200 milliseconds due to the time needed to program and launch an eye movement (Hallett, 1986). Average proportions of target fixations across Felicity (Felicitous vs. Infelicitous) and Contrast (Contrast vs. No Contrast) conditions are shown in Figure 9.

[5] This duration differs from the one in Experiment 2 because it includes recordings from an additional (male) speaker.

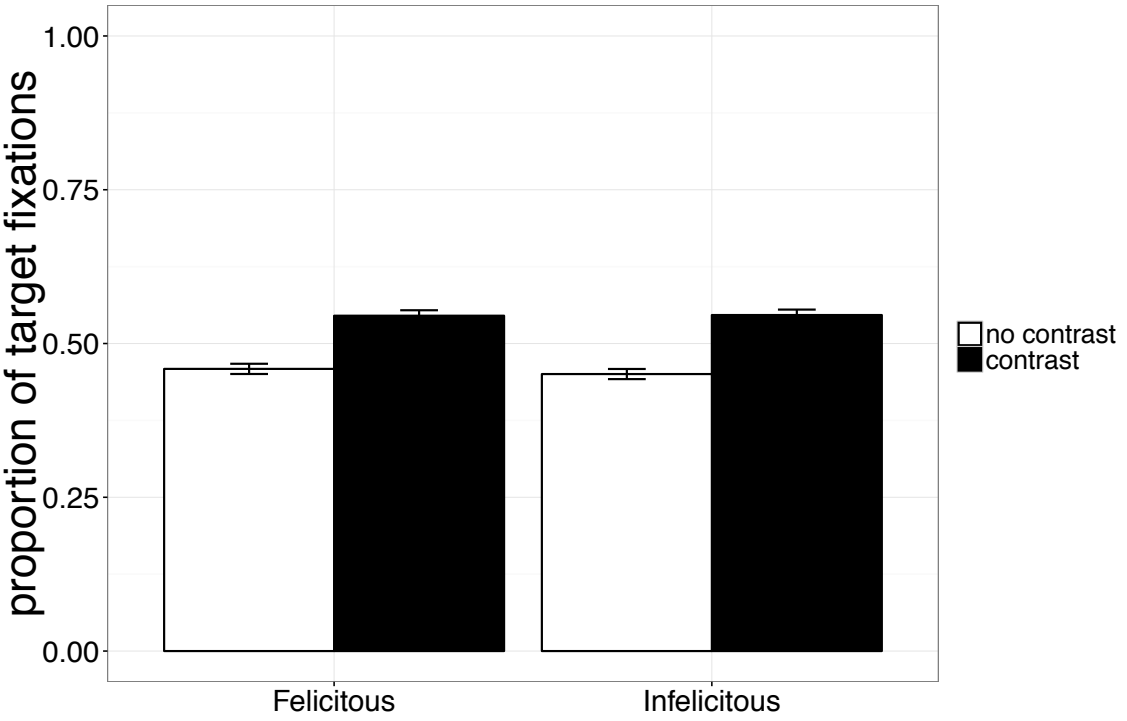


Figure 9. Average proportions of target fixations during interpretation of the scalar adjective and noun (e.g., *Click on the big pickle*) by Felicity and Contrast conditions.

The same procedure from Experiments 1 and 2 was used to determine the random effect structure and compare models (Table 5). The final model included random intercepts and random slopes for Felicity by items and random intercepts and random slopes by Felicity and Contrast by participants. There was a main effect of Contrast, such that participants made more target fixations in the Contrast condition than the No Contrast condition (t= 8.08). However, there was no significant main effect of Felicity (t=-0.08) and no significant interaction between Contrast and Felicity (t=-0.57).

Fixed Effects				
	β	Std. Error	t-value	p-value
(Intercept)	$-1.1e^{-3}$	0.02	-0.06	
Contrast condition	**0.19**	**0.02**	**8.08**	**$7.14e^{-15}$**
Felicity condition	$-1.71e^{-03}$	0.02	-0.08	0.97
Contrast x Felicity	-0.02	0.04	-0.57	0.51
Random Effects				

Groups		Variance	Std. Dev.	
Items	(Intercept)	0.01	0.10	
	Felicity	0.01	0.08	
Participants	(Intercept)	0.01	0.12	
	Contrast	4.2e-3	0.07	
	Felicity	4.7e-3	0.07	
Residual		0.32	0.56	
Observations: 5199; Items: 160; Participants: 65				

Table 5. Experiment 3: Results of the linear mixed-effects model of target fixation proportions

In order to examine the effects over time, we examine the data patterns across the two halves of the experiment (Table 6). In both halves of the experiment, as in the main analysis, there was a main effect of Contrast, no main effect of Felicity and no significant interaction of Contrast and Felicity.

First Half				
Fixed Effects				
	β	Std. Error	t-value	p-value
(Intercept)	0.03	0.03	0.83	
Contrast condition	-0.19	0.02	-7.94	**$1.97e^{-13}$**
Felicity condition	0.01	0.02	0.64	0.53
Contrast x Felicity	-0.01	0.02	-0.37	0.72
Random Effects				
Groups		Variance	Std. Dev.	
Items	(Intercept)	0.04	0.19	
Participants	(Intercept)	0.03	0.18	
Residual		0.86	0.93	
Observations: 1210; Items: 80; Participants: 63				
Second Half				
Fixed Effects				
	β	Std. Error	t-value	p-value
(Intercept)	-0.03	0.04	-0.80	
Contrast condition	-0.13	0.02	-5.75	**$3.01e^{-8}$**
Felicity condition	-0.02	0.02	-0.71	0.47
Contrast x Felicity	-0.01	0.02	-0.67	0.50
Random Effects				

Groups		Variance	Std. Dev.	
Items	(Intercept)	0.03	0.17	
Participants	(Intercept)	0.06	0.24	
Residual		0.93	0.97	
Observations: 1310; Items: 80; Participants: 63				

Table 6. Experiment 3: Results of the linear mixed-effects model of target fixation proportions for each half of the experiment

Discussion

While listening to two speakers, the contrast-contingent target preference was not diminished when participants were listening to a speaker who used adjectives infelicitously. This pattern held throughout the duration of the experimental session. This result suggests that listeners do not modulate pragmatic inferences speaker-specifically. Instead, they modulate their pragmatic references relative to the local environment, across speakers. Within the Infelicitous Speaker trials, 88% of trials were infelicitous, but across the experimental environment, only 50% of trials were infelicitous. In contrast, in Experiment 2, where modulation of pragmatic inferences was observed, 93% of trials were infelicitous in the Infelicitous condition. Alternatively, the presence of a Felicity by Contrast interaction in Experiment 2 and absence thereof in Experiment 3 may simply reflect a failure to replicate a rather small interaction effect (E2: $\beta_{\text{Felicity x Contrast}}$=-0.05) in the subsequent experiment (E3: $\beta_{\text{Felicity x Contrast}}$=-0.02).

General Discussion

In a series of eye-tracking experiments, we explored the nature of evidence necessary for listeners to modulate contrastive scalar inferences, which are indexed by rapid, anticipatory looks to sets of size-contrasted items in the presence of a size adjective

(Sedivy et al., 1999). In previous work (Grodner & Sedivy, 2011), listeners suppressed these inferences when faced with a speaker who was introduced to be pragmatically incompetent and prone to linguistic errors. This finding points to the intriguing possibility that one mechanism by which listeners navigate the variability and ambiguity inherent in much of reference resolution is by learning about the likelihood with which the speaker formulates linguistic expressions optimally given the context. However, the generalizability of these findings to everyday language use was tempered by the fact that both top-down and bottom-up information were available to these listeners—explicit characterization of a speaker's linguistic idiosyncrasies being rarely available in real-life settings.

In the present work, we examined the comprehender's ability to modulate pragmatic inferences based exclusively on bottom-up information, as well as the speaker-specificity of this process. Listeners were exposed to speakers who either used size adjectives felicitously (e.g., "the big dog" when a small dog and a big dog were present) or infelicitously (e.g., "the big dog" when only one dog was present). We found that bottom-up evidence was sufficient to trigger modulation of contrastive inferences but only with tremendous evidence and when the learning and testing environment were not differentiated. Furthermore, we did not find clear evidence to support the hypothesis that listeners model adjective use in a speaker-specific manner; instead, listeners seemed to take an average across speakers.

The role of prior assumptions

A noteworthy result from this work was the surprising persistence of the contrast-contingent anticipatory fixations. Even when exposed to an overwhelming 93% infelicitous use of scalar adjectives, participants fixated the target significantly more when it was part of a contrast set (though this effect was reduced relative to the case when instructions were mostly felicitous). The contrastive inference derived from size adjectives appears to be very resilient to new evidence that it is no longer valid in the present context. This persistence of the inference may be due to the expectations that participants bring to the experiment. In everyday language use, listeners experience mostly felicitous adjective use, as well as occasional instances of over-informativeness. Scalars in particular are unlikely to be used infelicitously (Brown-Schmidt & Konopka, 2011; Nadig & Sedivy, 2002; Ryskin et al., 2015; Tarenskeen, Broersma, & Geurts, 2015).[6] Moreover, in the current experiment listeners were given no top-down or explicit explanation as to *why* the speaker's adjective use may deviate from what is expected, which might have encouraged the perseverance of the otherwise efficient way of processing prenominal scalar adjectives. Overall, the instances of infelicitous adjective use experienced in the experimental session did not seem sufficient to counter a lifetime of experience with felicitous scalar adjectives.

This way of reasoning generates a number of testable hypotheses as to what determines the rate and degree of pragmatic modulations. For example, one could argue that inferences linked to more variably-used adjectives may be more susceptible to modulation. In particular, color adjectives are often used even when not strictly necessary

[6] In real-world settings, outside of the laboratory, scalars can be used when no contrast set is visually co-present (e.g., "Hand me that big coffee mug."); this may not constitute an infelicitous use of the modifier because the contrast might be derived from prior experiences with items similar to the referent (e.g., this particular coffee mug is over-sized relative to all previously seen mugs).

for disambiguation (Brown-Schmidt & Konopka, 2011; Sedivy 2003). Exposure to consistently infelicitous color adjectives may lead to complete suppression of a contrastive inference in the presence of a color adjective, because that inference is not as robust to begin with. Similarly, prosodic cues to contrast (e.g., L+H*; Ito & Speer, 2008; Watson et al., 2008) may be more variable by virtue of reflecting each speaker's realization of the intonational contour. Indeed, Kurumada, Brown, and Tanenhaus (under review) find that listeners suppress their contrastive interpretations of the L+H* accent after being exposed to a speaker who uses contrastive pitch accenting inappropriately.

On the other hand, one could also argue that listeners' familiarity with more variable uses of color adjectives may slow down the process of pragmatic modulation. In other words, listeners may not learn easily that a given environment is more likely to contain infelicitous adjective use because some amount of infelicity is consistently present in everyday language use. On this account, the exposure to an infelicitous use of a color adjective (e.g., *Point to the orange car* when there is only one car in sight) is unlikely to trigger a large error signal and therefore leads to only moderate learning (e.g., Chang et al., 2006). This is in line with the previous observation (Pogue et al. 2016) that over-informative utterances, which are more prevalent in the input and hence less surprising, are less likely to trigger speaker-specific modulation of pragmatic assumptions compared to under-informative utterances. Teasing these two, non-mutually exclusive, views apart will require a systematic investigation of the prior assumptions that listeners have about different types of contrastive cues and what role these play in the modulation process.

Semantic tuning and pragmatic generalization

Another question that arises from the current results concerns the extent to which they reflect a truly *pragmatic* process. The results are compatible with an account in which listeners *tune up* their semantic notions of the scalars, "big" and "small". In response to an overwhelming amount of evidence that these words do not highlight contextual contrast, listeners may begin to assume that the semantic meaning, not the usage, is altered. To argue that the observed changes in eye-movements are pragmatic in nature, one needs to show that the exposure to infelicitous scalars generalize not only to the same lexical items but to other pragmatic uses of language by the same speaker.

Recent studies have taken a step toward addressing these questions (Bott & Chemla, 2016; Pogue, Kurumada, & Tanenhaus, 2016). Pogue et al. (2016) first exposed listeners to two speakers who gave instructions in displays where scalar adjectives were necessary (e.g., a display with a large and a small chair, and two unrelated items). One speaker used scalar adjectives informatively (e.g., "Click on the big chair"), while the other consistently failed to use scalars even when needed (e.g., "Click on the chair"). No explicit commentary about the pragmatic capabilities of the speakers was provided. Participants were subsequently asked to make a judgment about which speaker was likely to have uttered a given sentence in a particular display. Pogue, et al. found that participants were more likely to attribute under-informative color-modified expressions (e.g., "Click on the red car" in a display with a large and small red car) to the previously-under-informative speaker. This finding lends support to the idea that listeners are able to track how different speakers use, or do not use, adjectives informatively and generalize this learning beyond directly experienced items. This generalization is not predicted if the

speaker is modifying their semantic expectations for given lexical items directly observed in the input.

What are we learning? Inferences about infelicity and unreliability of speakers

Consideration of the *generalization* of pragmatic modulation opens up a number of new research directions with respect to how listeners may accommodate variability across speakers in their pragmatic language use. One of the core questions concerns what is being learned in the face of unexpected, seemingly uncooperative, use of linguistic elements, such as prenominal adjectives. We can sketch out at least three different classes of inferences that listeners might make. One possibility is that the listener learns that the speaker is indifferent to the Gricean Cooperative Principle. If so, this predicts that the observed learning should extend to other domains of pragmatic language use, such as relevance of the speaker's comments, or the quantity of information that the speaker provides in other domains. A second possibility is that the learning is specifically about, say, prenominal modifiers and their mapping to context per se, such that what is learned is distributional information about how this speaker uses this adjective type in context. As a variant of this view, a learner might assume that the speaker's idiosyncrasy is restricted to adjective use and does not extend to other types of modifiers such as quantifiers.

A third class of possibilities is that the listener preserves the assumptions that the speaker is Gricean, but infers that their perspective or assumed common ground is distinct. For example, the listener may assume that the speaker is seeing something different or having a different experience of the world than the addressee. On this third possibility then, the listener's task is to uncover what this alternative perspective might

constitute. Grodner and Sedivy (2011) concluded that listeners suspended contrastive inferences in the face of an unreliable speaker. Yet, it is possible that contrastive inferences were not suspended, but instead that the listener was entertaining the possibility that the speaker had a different perspective. Take the following scenario. If a coffee shop customer says "I'll have the big muffin", in reference to a muffin in a display case with just a single muffin, the use of the modifier in this context need not prompt the inference that the customer is non-cooperative or non-rational. Instead it may lead the coffee shop owner to think that the muffin is larger than other muffins that the customer normally encounters. Thus, the granularity and bounds of pragmatic generalization reveal listeners' underlying beliefs about how relevant pragmatic uses of language may predict each other as well as how they may vary within and across speakers and contexts.

Talker (non)-specificity

The adaptation to the infelicitous speaker that was observed in Experiment 2 and in Grodner and Sedivy (2011) could have come about through a speaker-specific process – that is the learning about the infelicitous use of scalar adjectives was tied to that particular speaker. Another possibility is that the learning was instead tied to the situation more generally – a type of context-specific, rather than speaker-specific learning. The lack of a significant interaction with talker in Experiment 3 does not provide strong support for the idea that this learning was, in fact, speaker specific. Yet, given that this was a null effect, it does not rule out the possibility either. Perhaps, with a pair of speakers who were more distinctive (e.g., Horton & Gerrig, 2005), or with even more training exposure, speaker-specificity would have been revealed. If learning about the

pragmatic tendencies of speakers is indeed anchored to individuals, it would be consistent with a variety of other findings that listeners process speech in a speaker specific manner (e.g., Brown-Schmidt, 2009; Creel, Aslin, & Tanenhaus, 2008; Graham, Sedivy, & Khu, 2014; Matthews, Lieven, & Tomasello, 2010; Metzing & Brennan, 2003; Yoon & Brown-Schmidt, 2014).

On the other hand, contexts of language use potentially contain an infinite number of contextualized variables that could be linked in memory to the way language is used in that setting. These variables could include a variety of speaker-related variables (e.g., name, gender, accent, speech register), as well as environmental variables (e.g., location, time, room, smells). If adaptive processes in language comprehension are tailored to the local context, then a critical question is which of these variables become associated with the language behavior in question (in the present case, the infelicitous use of scalars). An alternative possibility, then, is that the adjustments observed here and in Grodner and Sedivy (2011) are not linked to the speaker or to any other contextual element in particular. This view, which is consistent with the results of Experiment 3, would suggest that the observed adaptations reflected adjustments to the conversational situation more generally (see Brown & Dell, 1987 for a related view in the case of language production). Teasing apart these possibilities – whether the observed learning is truly speaker specific (and we simply did not observe it), or generic, would likely benefit from further inquiries that examine the circumstances under which this learning does or does not generalize to new speakers and settings.

Conclusion

Grice's Cooperative Principle established the idea that human listeners are unlikely to assume that the speaker is being unreliable or infelicitous. Exploiting this feature provides the speaker with a variety of powerful means to convey their intentions and social meanings (e.g., jokes, lies, sarcasm). The persistence of contrastive inferences observed in all the experiments reported here is in line with this basic assumption that speakers rarely misuse scalar modifiers. Nevertheless, when given an overwhelming amount of evidence (Experiment 2) and/or an explicit instruction to counteract the usual expectations (Grodner & Sedivy, 2011), the listener can flexibly modify their online language comprehension so as not to be led astray. What emerges from these observations is a particularly intelligent, socially adaptive mechanism of language comprehension that can integrate multiple sources of information as well as distributional statistics from past experiences. An important goal for future research will be to investigate how these different types of information are evaluated and combined to support efficient and effective pragmatic communication.

References

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

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*, 361–374. doi:10.1016/j.pragma.2010.07.013

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-Effects models using {lme4}. *Journal Of Statistical Software*, *67*, 1–48. doi:10.18637/jss.v067.i01

Belke, E. (2006). *Visual determinants of preferred adjective order*. *Visual Cognition*, *14*, 261-294. doi:10.1080/13506280500260484

Bott, L. & Chemla, E. (2016). Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language, 91*, 117-140. doi:10.1017/CBO9781107415324.004

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision*, *10*, 433-436.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, *11*, 717–726. doi:10.1016/S0022-5371(72)80006-9

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *22*, 1482–1493. doi:10.1037/0278-7393.22.6.1482

Brown, P. M., & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, *19*, 441–472. doi:10.1016/0010-0285(87)90015-6

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, *61*, 171–190. doi:10.1016/j.jml.2009.04.003

Brown-Schmidt, S., & Konopka, A. E. (2011). Experimental approaches to referential domains and the on-line processing of referring expressions in unscripted conversation. *Information*, *2*, 302-326.

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*, 592–609. doi:10.1016/j.jml.2005.12.008

Brown-Schmidt, S., Yoon, S. O., & Ryskin, R. A. (2015). People as contexts in conversation. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 62, pp. 59–99). Academic Press. doi:10.1016/bs.plm.2014.09.003

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*, 234–272. doi:10.1037/0033-295X.113.2.234

Crain, S., & Steedman, M. (1985). On not being led up the garden path: the use of context by the psychological parser. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural Language Parsing: psychological, computational, and theoretical perspectives* (pp. 320–358). Cambridge University Press. doi:10.1080/10643389.2012.728825

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: the role of talker variation in lexical access. *Cognition, 106*, 633–64. doi:10.1016/j.cognition.2007.03.013

Creel, S. C. (2014). Preschoolers' flexible use of talker information during word learning. *Journal of Memory and Language*, *73*, 81–98. doi:10.1016/j.jml.2014.03.001

Creel, S. C., & Bregman, M. R. (2011). How talker identity relates to language processing. *Language and Linguistics Compass*, *5*, 190–204. doi:10.1111/j.1749-818X.2011.00276.x

Creel, S. C., & Tumlin, M. A. (2011). On-line acoustic and semantic interpretation of talker information. *Journal of Memory and Language*, *65*, 264–285. doi:10.1016/j.jml.2011.06.005

Davies, C., & Katsos, N. (2013). Are speakers and listeners "only moderately Gricean"? An empirical response to Engelhardt et al. (2006). *Journal of Pragmatics*, *49*, 78–106. doi:10.1016/j.pragma.2013.01.004

Donnellan, K. S. (1966). Reference and definite descriptions. *The Philosophical Review*, *75*, 281–304.

Eich, E. (1985). Context, memory, and integrated item/context imagery. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 764–770.

Engelhardt, P. E., Bariş Demiral, Ş., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, *77*, 304–314. doi:10.1016/j.bandc.2011.07.004

Fine, A. B., & Florian Jaeger, T. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science, 37*, 578–91. doi:10.1111/cogs.12022

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid Expectation Adaptation during Syntactic Comprehension. *PloS One, 8*, e77661. doi:10.1371/journal.pone.0077661

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: on land and underwater. *British Journal of Psychology, 66*, 325–331.

Graham, S. A, Sedivy, J., & Khu, M. (2014). That's not what you said earlier: Preschoolers expect partners to be referentially consistent. *Journal of Child Language*, *41*, 34–50. doi:10.1017/S0305000912000530

Grodner, D. J., & Sedivy, J. C. (2011). The effect of speaker-specific information on pragmatic inferences. (E. A. Gibson & N. J. Pearlmutter, Eds.) *The Processing and*

Acquisition of Reference. MIT Press. doi:10.7551/mitpress/9780262015127.001.0001

Hallett, P.E. (1986). Eye movements. In *Handbook of Perception and Human Performance*. Boff, K.R., Kaufman, L.; Thomas, J.P., Eds.; Wiley: New York, NY, USA, pp. 10.1–10.112.

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition, 96*, 127–42. doi:10.1016/j.cognition.2004.07.001

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, *58*, 541–573. doi:10.1016/j.jml.2007.06.013

Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition*, *124*, 66–71. doi:10.1016/j.cognition.2012.03.001

Kennedy C., & McNally L. (2005). Scale structure and the semantic typology of gradable predicates. *Language, 81,* 345-381.

Kleiner M., Brainard D., Pelli D., (2007) "What's new in Psychtoolbox-3?" *Perception 36*, ECVP Abstract Supplement.

Kurumada, C., Brown, M., & Tanenhaus, M. K. (under review). Adaptation and inferences in pragmatic interpretation of English contrastive prosody.

Levelt, W. M. (1989). *Speaking : from intention to articulation*. Cambridge, Mass.: MIT Press, c1989.

Matthews, D., Lieven, E., & Tomasello, M. (2010). What's in a manner of speaking? Children's sensitivity to partner-specific referential precedents. *Developmental Psychology*, *46*, 749–60. doi:10.1037/a0019657

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*, 201–213. doi:10.1016/S0749-596X(03)00028-7

Mulligan, N. W. (2011). Conceptual implicit memory and environmental context. *Consciousness and Cognition, 20*, 737–744.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, *13*, 329–336. doi:10.1111/1467-9280.00460

Novick, J. M., Thompson-Schill, S. L., & Trueswell, J. C. (2008). Putting lexical constraints in context into the visual-world paradigm. *Cognition*, *107*, 850–903. doi:10.1016/j.cognition.2007.12.011

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, *77*, 257–273.

Osgood, C. E. (1971). Where do sentences come from? In D. D. Steinberg & L. A. Jakobovits (Eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology*. Cambridge, MA: Cambridge University Press.

Pechmann, T. (1989) Incremental speech production and referential overspecification. *Linguistics, 27*, 89–110.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*. doi:10.1163/156856897X00366

Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under- and over-informative prenominal adjective use. *Frontiers in Psychology*, *6*, 1–18. doi:10.3389/fpsyg.2015.02035

Roberts, C. (2003). Uniqueness in definite noun phrases. *Linguistics and Philosophy*, *26*, 287–350.

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. doi:10.1037/xge0000093

Ryskin, R. A., Wang, R. F., & Brown-Schmidt, S. (2016). Listeners use speaker identity to access representations of spatial perspective during online language comprehension. *Cognition*, *147*, 75–84. doi:10.1016/j.cognition.2015.11.011

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*, 3–23.

Sedivy, J. C., K. Tanenhaus, M., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*, 109–147. doi:10.1016/S0010-0277(99)00025-6

Snedeker, J., & Trueswell, J. C. (2003). Using prosody to avoid ambiguity : Effects of speaker awareness and referential context. *Journal of Memory and Language*, *48*, 103–130.

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: a review and meta-analysis. *Psychonomic Bulletin & Review, 8*, 203–220.

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*, 1632–1634.

Tarenskeen, S., Broersma, M., & Geurts, B. (2015). Overspecification of color, pattern, and size: Salience, absoluteness, and consistency. *Frontiers in Psychology*, *6*:1703. doi:10.3389/fpsyg.2015.01703

Van Der Wege, M. M. (2009). Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, *60*(4), 448–463. doi:10.1016/j.jml.2008.12.003

Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. A. (2008). Interpreting pitch accents in online comprehension: H* vs. L+H*. *Cognitive Science*, *32*, 1232–44. doi:10.1080/03640210802138755

Wells, J. B., Christiansen, M. H., Race, D. S., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*, 250–271. doi:10.1016/j.cogpsych.2008.08.002

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, *87*, 128–143. doi:10.1016/j.jml.2015.08.003

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. doi:10.1037/a0036161

Acknowledgement

This material is based on work supported by National Science Foundation Grants NSF 12-57029 and NSF 15-56700 to Sarah Brown-Schmidt. We would also like to thank Sarah Bibyk and Geoffrey McKinley for help with stimulus recording.