Semantic Adaptation in Gradable Adjective Interpretation*

Ming Xiang[a], Alex Kramer[b], Christopher Kennedy[a]

[a] *Department of Linguistics, University of Chicago*

[b] *Department of Linguistics, University of Michigan*

Abstract

Previous studies on learning and adaptation have largely focused on speech perception and syntactic parsing, but much less is known about whether and how language users adjust their semantic representation after being exposed to other individuals' utterances. The current study focuses on the interpretation of gradable adjectives — expressions with highly context-dependent interpretations — and investigates how individuals adjust their thresholds of application after exposure to utterances of the same expressions by other language users. Three experiments provide novel evidence to support robust and rapid semantic adaptation for gradable adjectives, the effect of which is modulated by the types of utterances and the class of adjectives participants were exposed to, as well as the communicative goal of the linguistic task and the identity of the communicative partner. We propose a single unified probabilistic belief update analysis to account for all of the observations. Under our account, threshold adaptation naturally falls out as the result of a listener-speaker coordination process, which is guided by general principles of pragmatic reasoning. The current empirical findings and theoretical proposals also find parallels in the perceptual learning and speech adaptation domain, suggesting a domain general mechanism of learning and adaptation at multiple levels of linguistic representation.

Keywords: semantic adaptation, communicative coordination, gradable adjectives, threshold of application, Bayesian pragmatic reasoning, probabilistic belief update

*DRAFT June 8, 2020; please do not cite without permission.

Email addresses: mxiang@uchicago.edu (Ming Xiang), arkram@umich.edu (Alex Kramer), ck@uchicago.edu (Christopher Kennedy)

1. Introduction

Language communication requires successful mapping between form and meaning. Although there are systematic grammatical constraints that regulate how a linguistic signal is mapped to meaning, it is well-known that the form-meaning mapping is often highly context-dependent, and exhibits a substantial degree of speaker variability. For example, at the phonetic level, one person's /p/ sound may be acoustically indistinguishable from another person's /b/ sound. At the lexical level, what counts as tall for one speaker can vary in different contexts, and further more what is tall may vary for a single speaker in different contexts, and for different speakers in the same context. Observations of this sort suggest that in order to successfully communicate, a listener needs to not only know *what* a linguistic signal can potentially mean, but also *how* different speakers can use the same signal to mean different things. In other words, language users need to develop strategies to adapt to different ways of talking, and coordinate accordingly with different conversational partners.

Empirical studies for adaptation have been carried out for different levels of linguistic representations. A particularly fruitful area of active research is in speech perception. When mapping acoustic signals to phonetic and phonological representations, listeners need to deal with variabilities from situation to situation, and from talker to talker. There is a growing body of work showing that listeners can make quick and flexible perceptual adjustment for specific talkers and situations (Creel et al., 2008; Kraljic and Samuel, 2005; Pisoni and Levi, 2012). Listeners change their phonetic category classification as a consequence of perceptual learning after repeated exposure to a given acoustic stimulus (Norris et al., 2003; Samuel, 1986; Vroomen et al., 2007; Kleinschmidt and Jaeger, 2015). At a different representational level, syntactic adaptation, at least in the form of syntactic priming, is also well attested. Repeated exposure to a given syntactic structure triggers more subsequent production of the similar structure, as evidenced by both laboratory studies (Bock, 1986; Pickering and Branigan, 1998; Jaeger and Snider, 2013) and corpus data (Gries, 2005).

Compared to the research in speech and syntactic domains, however, adaptation at the lexical semantic level is much less understood. A few previous studies looked at lexical entrainment on object reference, showing that listeners keep good track of the specific ways other interlocutors refer to an object (Brennan and Clark, 1996; Metzing and Brennan, 2003). A recent study on quantifier interpretation by Yildirim et al. (2016) also showed that a listener would change her belief about whether a particular speaker would use *some* or *many* to describe a certain quantity of items after the listener was exposed to the speaker's utterances. The existing findings lend strong support to the view that listeners can learn and store talker-specific information, and that becomes part of the contextual representation. But there is little direct evidence that listeners also (temporarily or long term) change their own lexical semantic representations as a result of coordinating with their interlocutors. Furthermore, the underlying mechanism that mediates semantic adaptation also remains an open question.

Our focus in this paper is adaptation in the interpretation of gradable adjectives such as *tall*. As is well known, what it means to be tall can vary from context to context: a tall candle is shorter than a tall tree, and a particular gymnast may be judged tall in a context involving other gymnasts, but not tall in a context involving gymnasts and basketball players. The standard analysis of gradable adjectives in linguistic semantics treats them as denoting "threshold-dependent" properties, such that *tall*, for example, denotes the property of having a height that is at least as great as a threshold of height (θ), whose value may vary (see e.g. Lewis 1970; McConnell-Ginet 1973; Cresswell 1976; Klein 1980; von Stechow 1984; Barker 2002; Kennedy 2007; Lassiter and Goodman 2013; Qing and Franke 2014a and many others). Sometimes the value of θ is explicitly specified as part of semantic composition, e.g. by a measure phrase (as in *"this candle is six inches tall"*) or a comparative construction (*"this candle is taller than this wine bottle"*). But when a gradable adjective is used in its unmarked, "positive" form (as in *"this candle is tall"*), the value of θ is both implicit and uncertain, and must be inferred.[1] Whether a particular object counts as tall in a particular context

[1]Our use of terms like "degree" and "threshold" in characterizing gradable adjectives should not be taken as a commitment to specific assumptions about the lexical semantics of gradable adjectives, e.g. that such expressions crucially involve reference to particular kinds of abstract objects or mental representations, with associated metaphysical or cognitive commitments. Instead, we use this terminology as a means of characterizing in a general and hopefully intuitive way what any descriptively adequate semantics must

of utterance, then, depends not only on the object's actual height, but also on decisions about the value of θ, and successful communication with a gradable adjective like *tall* thus involves coordination between interlocutors both on the height of the object described and on the implicit threshold for *tall*. Our goal in this paper is to ask whether individuals' decisions about threshold values change over time through exposure to other individuals' use of gradable adjectives, i.e. whether we find evidence for threshold adaptation. And if we do find evidence for adaptation, we wish to know what what it looks like, and what factors are responsible for it.

A further theoretical goal of the current study is to establish a close parallel between adaptation behavior at the lexical semantic level and adaptation at the level of speech perception. In the empirical domain of gradable adjectives, as discussed above, an important part of the research question is how a language user decides where to draw, on a continuous scale of degrees, an implicit threshold such that objects ordered on the scale could be classified as either belonging to certain category (e.g. the category of tall candles) or not. Framed in this way, the question of interpreting gradable adjectives bears some resemblance to the question of phonetic categorization in speech perception, such as how a listener decides the boundary between a \p\and a \b\categories on a VOT continuum. As mentioned earlier, adaption at the speech level has been thoroughly studied. A finding that is particularly relevant for the current purpose is that listeners adjust their categorical perception boundaries after being exposed to an ambiguous acoustic stimulus that is labeled as belonging to a certain category or a prototypical stimulus from a certain category. But the direction of the adaptation effect under these two types of exposure is different (Norris et al., 2003; Samuel, 1986; Vroomen et al., 2007). In the current study, we will look at how a listener shifts her threshold for adjective interpretation after being exposed to ambiguous or prototypical stimuli. We will

---

be committed to: that gradable adjectives categorize objects in terms of where they rank along (possibly multidimensional) orderings such as height, weight, beauty, intelligence and so forth; that they support different categorizations in different contexts of use; and that these categorizations are sometimes made explicit by other linguistic expressions (such as measure phrases or comparatives) and are sometimes implicit. It is this last case that we are interested in here.

in particular adopt the exposure-testing paradigm used in Vroomen et al. (2007). As we will show below, this paradigm not only allows us to evaluate how people adjust their thresholds of adjectives after repeated exposure to another speaker, but it also provides us with an opportunity to assess the time course of the adaptation behavior, i.e. how quickly people adapt.

To preview, the basic adaptation behavior in adjective interpretations (Experiment 1) will be very similar to the speech perception findings in the literature. To account for the observed adaptation behavior in adjective interpretations, we propose a mechanism that assumes a listener who probabilistically updates her beliefs about adjective thresholds based on experience. The belief update account developed here for gradable adjectives, when construed more broadly, is very much in line with the Bayesian belief update account of speech perception proposed by Kleinschmidt and Jaeger (2015). After presenting the empirical evidence and theoretical account for the basic adaptation effect, we will further demonstrate in two additional experiments (Experiment 2 and 3) that there is a general pragmatic constraint modulating the adaptation of adjective interpretations: the effect size of adaptation is affected by whether the listener perceives the speaker as having a shared communicative goal.

The basic structure of the exposure-testing paradigm consists of three phases. First, in the pre-calibration phase, we collected participants' judgments about whether a gradable adjectives accurately characterize an object from a scale, i.e. judgments about whether statements like *"this candle is tall"* are true of different candles. Next, in the exposure phase, we exposed participants to other individuals' judgments about objects from the ambiguous or prototypical regions of the same scale. And finally in the post-calibration phase, we collected participant judgments a second time, to determine whether their truth judgments about identical objects — and therefore adjectival thresholds — changed after exposure, and so indicated threshold adaptation. In order to build as comprehensive an empirical picture as possible, we looked for adaptation effects in three gradable adjectives, each of which was taken from one of three semantic classes of gradable adjectives that are distinguished based on the kinds of thresholds they use. The first, *tall*, comes from the class of RELATIVE gradable adjectives, whose thresholds are highly variable and context-dependent, and typically result in meanings that characterize an object as having something like a "significant" or "above average" degree of the relevant property. This class includes most dimensional adjectives like *tall, heavy* and *big*, but also evaluative terms like *smart, lazy* and *beautiful*, normative predicates like *good* and *bad*, experiential predicates like *fun* and *tasty*, and many others. The other two adjectives, *bent* and *plain*, come from the class of ABSOLUTE gradable adjectives: adjectives with default MINIMUM THRESHOLDS and adjectives with default MAXIMUM THRESHOLDS. Minimum adjectives are exemplified by adjectives like *bent, striped* and *open*, which have uses that characterize objects as having a non-zero degree of the relevant property. For example, a nail can be considered as bent as long as it has a non-zero degree of bend. Maximum adjectives are exemplified by adjectives like *plain, straight* and *closed*, which have uses that characterize objects has having maximal degrees of the relevant property. For instance, a straight nail, strictly speaking, is a nail that is absolutely straight. [2] Absolute adjectives also allow for variation in thresholds, but the variation is much more limited than for relative adjectives (see e.g. Pinkal, 1995; Rotstein and Winter, 2004; Kennedy and McNally, 2005; Kennedy, 2007; Toledo and Sassoon, 2011; Lassiter and Goodman, 2013; Qing and Franke, 2014a; Burnett, 2016). Looking at these three classes together therefore gives us a broader empirical picture than looking at one type of adjective alone.

In the following sections, we will present three experimental studies which provide evidence that an individual's decisions about how to resolve semantic uncertainty in the meaning of a gradable adjective in the positive form — how to fix the value of the adjective's threshold of application — are influenced by the exposure to another individual's use of the same expression in a communicative exchange. We will see that the particular pattern of adaptation depends both on the degree to which the object described manifests the degree of the relevant gradable property, and on the prior threshold distribution for the predicate (as

[2] Relative adjectives, in contrast, cannot have either minimum or maximum threshold interpretations. A minimum threshold interpretation of *tall*, for example, would be a meaning equivalent to *have height*, while a maximum threshold interpretation would presuppose a unique maximum height, and would characterize an object as having that height. But both of these interpretations are non-sensical.

exemplified by the three classes of adjectives). We will, nevertheless, argue that a single, general belief-update mechanism, geared towards maximizing coordination on the degree to which an object possesses a gradable property in a communicative task, can derive all the observed patterns.

2. Experiment 1

2.1. Material

We created images that depict scales for the adjectives *tall*, *bent* and *plain*, each of which exemplifies one of the three threshold-based classifications discussed in Section 1. For each scale, we created images along a five-point continuum, with scale point one corresponding to the minimum degree of the relevant adjectival property, and scale point five corresponding to the maximum degree of the relevant adjectival property. For the *tall* scale, the height of a candle gradually increased in height from shortest at point one to tallest at point five; for the *bent* scale, a nail gradually increased in bend from straight at point one to significantly bent at point five; and for the *plain* scale, a pillow gradually increased from very spotted at point one to devoid of spots at point five. Figure 1 illustrates the three sets of scalar continua.

2.2. Procedure and participants

The experiment was conducted on Amazon Mechanical Turk. The web-based experimental procedure was implemented using codes adapted from the study in Kleinschmidt and Jaeger (2015)[3]. We recruited four different groups of participants from Mechanical Turk, with 30 participants in each group. For each adjective scale, every participant completed three phases in the following sequence: **pre-calibration phase**, **exposure/test phase**, and **post-calibration phase**. The four groups were distinguished by the exposure/test phrase they received, as will be explained below. For each group, the testing on the three adjective

[3]The source code was adapted from http://hlplab.wordpress.com/2013/09/22/phonetics-online/.

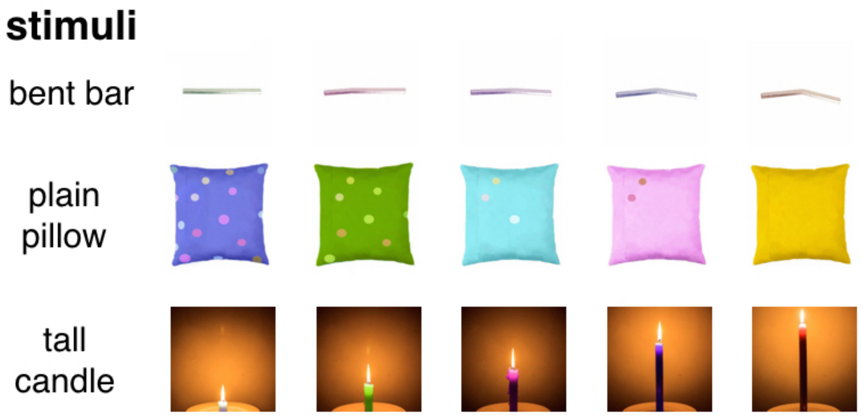

Figure 1: Five-point scalar continua.

scales was carried out in separate blocks. A participant finished all three phases for one adjective scale before moving on to the next one. The presentation order of the three adjective blocks was randomized for each participant. In addition to the three critical adjective blocks, we also included a fourth block that tested participants on a completely different scale: the quantitative scale of numerosity. These trials served as fillers, and we will not discuss them further. All groups of participants were tested with exactly the same procedure; the only difference between groups was the exposure block they saw in the second, exposure/testing phase of the experiment. Figure 2 provides a schematic overview of the study, which we now explain in more detail.

Each adjective block began with an initial **pre-calibration phase**, in which participants were presented with randomly selected images from each of the five adjectival scale positions, and were asked to make a yes-no binary judgment about whether the object in the image had the property named by the adjective. For example, a participant would be presented with one of the five images of a candle and asked *"Is this tall?"*, and would answer "yes" or 'no.' Given the semantics of gradable adjectives presented in Section 1, a "yes" response indicates that the value of the implicit threshold θ is lower than the degree to which the

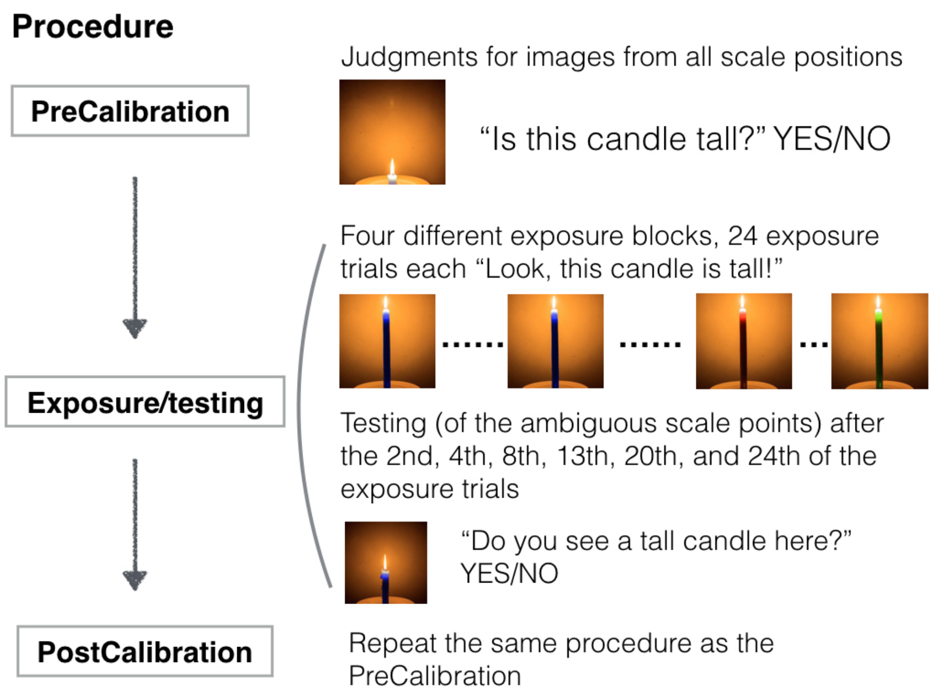

Figure 2: The experimental procedure with one adjective scale. All adjective scales have the same 3-phase procedure.

described object has the property in question, while a “no” response indicates that the value of θ is above this degree.

For each adjective scale, participants were presented with 30 trials total, with the image from scale position 3 repeated 10 times, the image from scale position 2 and 4 repeated 7 times each, and the image from scale position 1 and 5 repeated 3 times each. For each participant, a logistic regression was performed to determine which scale position was the most ambiguous for that participant; the image from this scale position was later used for this participant's exposure/test phase if they were in either the AmbiguousPositive or AmbiguousNegative exposure groups, as described below. For the *tall candle* scale, the ambiguous point falls on scale points 2, 3 or 4. For *plain pillow*, the ambiguous point was predominantly on scale point 4, and for *bent bar* predominantly on scale point 2.

After the pre-calibration phase, the participants moved to the **exposure/test phase**. In this phase, each participant was presented with an image of an object from one of the scale positions (either the ambiguous position on the scale or the scale end positions, see below), paired with an utterance by a female, native American English-speaking talker. In the first exposure trial, the talker began by uttering a sentence that established a conversational goal, e.g. *"I need a tall candle for a party,"* and then uttered a sentence that described the object in the image as either having or not having the relevant property: *"Look, this candle is tall!"* (positive polarity) or *"But this candle is not tall"* (negative polarity). Participants were then presented the same image 23 more times for a total of 24 trials, with slight variations of the crucial utterance for variability (*"Look, this candle is also tall!"*, *"Too bad, this candle is not tall."*, etc.). The color of the image was also manipulated to vary from trial to trial, but everything else remained the same throughout the 24-trial exposure sequence. Crucially, for each participant, the exposure image always came from the same scale position, and the utterance paired with the image always had the same polarity. The purpose of repeating the same exposure image/utterance pairs this many times is that it provided us with an opportunity to track the time course of adaptation.

At six different points of the exposure sequence — trial numbers 2, 4, 8, 13, 20, and 24 — we interrupted the participants with test trials. In the test trials, each participant was presented with the image from their unique ambiguous scale position, previously identified during the pre-calibration phase as described above, and the participant was asked to make a yes/no judgment as to whether the image satisfied the relevant adjective property. The participant was also asked to make the same judgment about two additional images: one from the scale point immediately above the ambiguous scale position, and one from the scale point immediately below it. After participants finished each test trial, the exposure sequence continued. To keep participants attention during the exposure sequence, in four different locations during the exposure sequence, a red or blue "+" symbol was displayed in the center of the screen for 500ms, and participants were asked what color they saw after the cross symbol disappeared from the screen. As mentioned earlier, there were four groups of participants. The overall procedure described above for the exposure/test phase was identical for all four participant groups. But two factors distinguished the four participant groups: 1) the scale position of image they were shown during exposure (AMBIGUOUS vs. PROTOTYPICAL) and 2) the polarity of the associated utterance (POSITIVE vs. NEGATIVE). Participants in the AMBIGUOUSPOSITIVE group were exposed to the image from their own most ambiguous scale point on the five-step scale continuum, and heard utterances in which the talker characterized the object in the image as having the property in question (*"This candle/nail/pillow is tall/bent/plain"*). Participants in the AmbiguousNegative group were also exposed to the image from their own most ambiguous scale point, but heard the talker describe the object as *not* having the property in question (*"This candle/nail/pillow is not tall/bent/plain"*). Participants in the PrototypicalPositive group were presented with images from scale position five (the highest scale position), and heard associated talker describe the object as having the property in question (*"This candle/nail/pillow is tall/bent/plain"*). And finally, participants in the PrototypicalNegative group were presented with images from scale point position one (the lowest scale position), and heard the associated talker describe the object as *not* having the property in question (*"This candle/nail/pillow is not tall/bent/plain"*). The Prototypical groups were so labeled because the talker's description of the images perfectly matched the participants' judgments on these positions during the pre-calibration phase. That is, an image from scale position five was always judged to be true for a given property, and an image from scale position one was always judged to be false for that adjectival property.

Finally, after the participants completed the exposure/test phase, they moved to the post-calibration phase, which was identical in all respects to the pre-calibration phase. For each participant, the experiment took about 30 minutes to complete.

2.3. Analysis and results

Our data analysis focused on two questions. First, we evaluated whether there are any changes in participants' judgments in the post-calibration phase compared to the pre-calibration phase, and how the changes (if any) were conditioned by the exposure trials. The post-calibration phase involved a task identical to the pre-calibration phase, thus any changes in participants' responses from pre- to post-calibration would indicate an adjustment of their threshold calculation for the adjective being tested, an adjustment triggered by exposure to the image/utterance pairs during the exposure phase. Second, we examined the time course over which the exposure-induced change developed. This was done by analyzing the 6 testing trials obtained at positions 2, 4, 8, 13, 20 and 24 of the exposure sequence. Since the results turned out to qualitatively similar for the three adjectives we investigated (with a couple interesting differences that we discuss below), we will first present the results for the trials involving the relative adjective *tall*, and then turn to the results for the maximum adjective *plain* and the minimum adjective *bent*.

2.3.1. The relative adjective tall

Effects of adaptation. The average number of "yes" responses from the pre-calibration and post-calibration phases, as well as the difference scores between them at each scale position, are presented in Figure 3. Our statistical analysis focused on two effects: the interaction between Exposure Condition (i.e. the four exposure groups) and Calibration Phase (i.e. pre vs. post-calibration); and the main effect of Calibration. For the interaction effect, we conducted a likelihood-ratio test between two mixed-effects logistic regression models: the first contained in its fixed effects the main effects for the two factors and also the interaction between the two, and the second is identical but without the interaction. Both models contained in their random effects the random intercepts for participants. In both models, the factors are sum-coded. Model comparison revealed a significant interaction ($p < .00001, df = 3$). The main effect of Calibration was also evaluated by a likelihood-ratio test between two mixed effects logistic models that only differed in whether the main effect of Calibration was included in their fixed effect structure. Following Levy (2018), when testing the main effect of Calibration in the presence of an interaction effect, we converted the factor Exposure Condition to a sum-coded numerical representation. The likelihood-ratio test showed a significant main effect for Calibration ($p < .05, df = 1$). For a summary of these results, see Table 1.

These results indicate a general effect of adaptation: participants' evaluation criteria for judging whether a given object is tall or not changed after they were exposed to the judgment of another talker. What is particularly interesting is that the direction of the change in participants' acceptance judgments is determined by the exposure type, as indicated by the significant interaction between Calibration and Exposure groups. As shown in Figure 3, af-

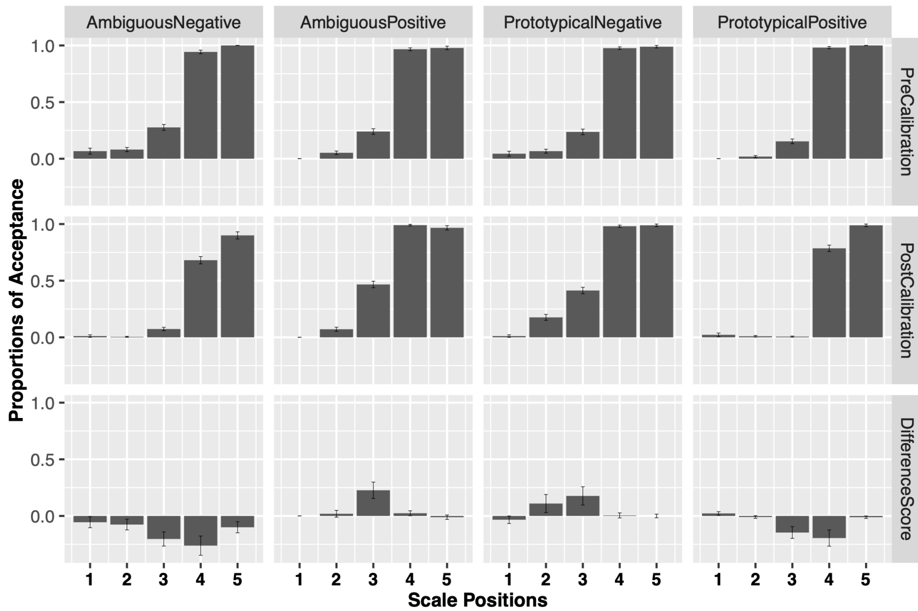

Figure 3: Comparison of pre- and post-calibration "yes" responses for relative adjective *tall*

ter the PrototypicalNegative and AmbiguousPositive exposure phases, participants provided more "yes" responses in the post-calibration phase compared to their pre-calibration phase, indicating a downward shift of of their threshold for *tall*. For the Prototypical-Positive and AmbiguousNegative exposures, participants provided more "no" responses in their post-calibration judgments, indicating their threshold for *tall* was shifted upward compared to the pre-calibration phase.

Time-course of adaptation. When assessing the development of the adaptation behavior over time, we analyzed the judgment data obtained during the exposure phase. As noted above, participants were tested after trials 2, 4, 8, 13, 20 and 24 of the testing/exposure phase. At these positions, participants made a Yes/No judgment on three testing trials: the image from the most ambiguous scale point, and the images immediately below and above that scale point. We chose to test these images because we took them to represent the *ambiguity region* for each participant's evaluation criteria, and we anticipated that they would be the most susceptible to the influence of the exposure trials.

Following the analysis method in Vroomen et al. (2007), we first compared data from the

AmbiguousPositive exposure group with the data from the AmbiguousNegative group. Recall that for these two groups of participants, during exposure, they were presented with the same ambiguous image, but the utterances they were exposed to had different polarity. At each of the six points, a difference score was calculated between the two exposure groups, collapsing over the acceptance judgments participants made over the multiple testing images at each point. Since the results from the post-calibration phase showed that after the AmbiguousPositive exposure, there was an overall increase in "yes" responses, in contrast to the overall decrease of "yes" responses after the AmbiguousNegative exposure, a significant positive difference between the two (i.e. subtracting responses after the Ambiguous Negative exposure from those after the Ambiguous Positive exposure, and comparing that difference with zero) would indicate a substantial adaptation effect. As shown in Figure 4, a significant effect already appeared at the earliest position we tested, the 2nd exposure trial.

We also compared data from responses after the PrototypicalPositive exposure and responses after the PrototypicalNegative exposure trials. For this comparison, the results from the post-calibration phase showed that after the PrototypicalPositive exposure, there was an overall decrease in the "yes" responses, in contrast to the overall increase of "yes" responses after the PrototypicalNegative exposure. Here a significant negative difference between the two (i.e. subtracting responses after the PrototypicalNegative exposure from those after the PrototypicalPositive exposure) indicates a significant adaptation effect has taken place. Figure 4 again shows that such an effect already appeared after the 2nd exposure trial.

To statistically evaluate the development of the adaptation effect over time, for each of the two comparisons represented in Figure 4 (i.e. one comparison between the responses after the two kinds of ambiguous exposures and another comparison between the responses after the two kinds of prototypical exposures), we conducted logistic mixed effects models to test for the main effect of exposure type and the interaction between the two. These two effects were important since our primary question was whether participants' judgments at the testing trials would differ under different exposure utterances and whether the influence of exposure

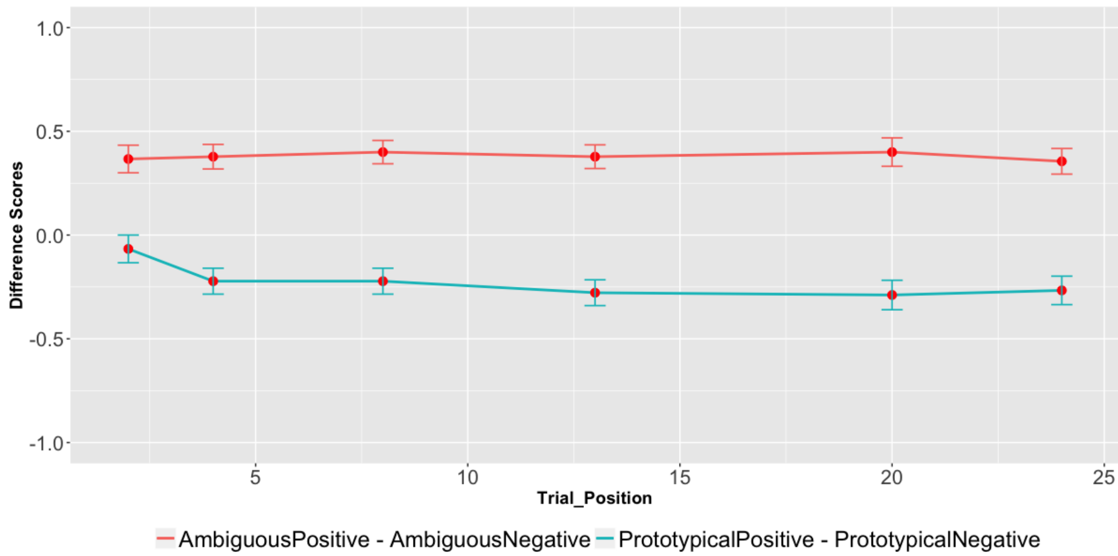

Figure 4: The time-course of adaptation for *tall*

utterance was modulated by the number of exposure trials participants heard. The effect of the location of the testing trial alone was not of theoretical interest to us, and therefore we do not report it below. For each comparison, the full model included the exposure type, location of the testing trial, and their interaction as fixed effects, and the maximum random effect structure that led to successful model convergence. The exposure type predictor was treatment coded, with the AmbiguousNegative and the PrototypicalNegative exposure coded as the reference baseline relative to the Positive exposure counterparts. The location of the testing trial was coded as a continuous numeric predictor.

For trials tested after the AmbiguousPositive and AmbiguousNegative exposures, we found a significant effect of Exposure type ($Est = 2.03 \pm 0.39, z = 5.15$), but there was no effect for the interaction between Exposure type and the location of the testing trial ($Est = 0.005 \pm 0.02, z = 0.24$). These results suggest that there was a robust adaptation effect after exposure, which appeared as early as the first location we tested (i.e. after the 2^{nd} exposure trial). The effect stayed stable throughout the exposure phase. In other words, more exposure trials did not change the size of the adaptation effect.

For trials tested after the PrototypicalPositive and PrototypicalNegative exposures, there was a trend towards an effect of Exposure type ($Est = -0.6 \pm 0.31, z = -1.9$), there was also an interaction between Exposure type and the location of the testing trial ($Est = -0.07 \pm 0.02, z = -2.8$). This suggests that the number of exposure trials had an effect on adaptation. A separate analysis on the the first and second testing location (i.e. testing carried out after exposure trials 2 and 4) found that there was no adaptation effect after trial 2 ($Est = -0.27 \pm 0.3, z = -0.89$), but there was an effect after trial 4 ($Est = -0.91 \pm 0.31, z = -2.96$). After removing data from the earliest testing locations (trial 2), for the rest of the five testing locations, there was also a robust main effect of exposure type ($Est = -1.35 \pm 0.5, z = -2.7$), but there was no longer an interaction between exposure type and testing locations ($Est = -0.02 \pm 0.02, z = -0.8$). These results suggest that although the adaptation effect did not appear at the earliest testing location after the prototypical exposure, it nonetheless appeared after 4 exposure trials, and stayed stable after that. More exposure trials did not change the size of the adaptation effect. For a summary of the time course effects, see Table 2.

2.3.2. The absolute adjectives bent *and* plain

In Figure 5 and Figure 6 we present the acceptance rating results from the pre-calibration and post-calibration phases for the two absolute adjective scales *bent* and *plain*. The differences between the two phases are also presented in these figures. The data analysis procedure for *bent* and *plain* was identical to *tall*. For *bent*, there is a significant main effect of Calibration Phase ($p < 0.0001, df = 1$); and there is also a significant interaction between Calibration Phase and Exposure type ($p < 0.0001, df = 3$). For *plain*, there is no main effect of Calibration Phase ($p > 0.4, df = 1$); but there is a significant interaction between Calibration Phase and Exposure type ($p < 0.0001, df = 3$).

The time course results for *bent* and *plain* are presented in Figure 7. The analyses procedures were identical as the procedure for *tall*.

For trials tested after the AmbiguousPositive and AmbiguousNegative exposures, we found a significant main effect of Exposure type for both *bent* and *plain* (for *bent*: $Est = 8.84 \pm 2.53, z = 3.5$; for *plain*: $Est = 6.87 \pm 1.55, z = 4.4$). There was no interaction between Exposure type and the location of the testing trial for both adjective scales (for

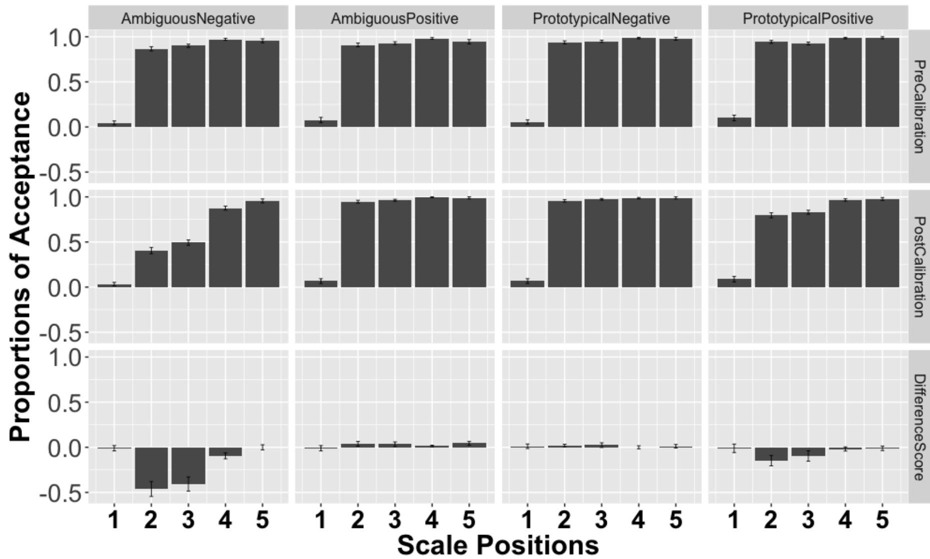

Figure 5: Comparison of pre- and post-calibration "yes" responses for minimum threshold absolute adjective *bent*

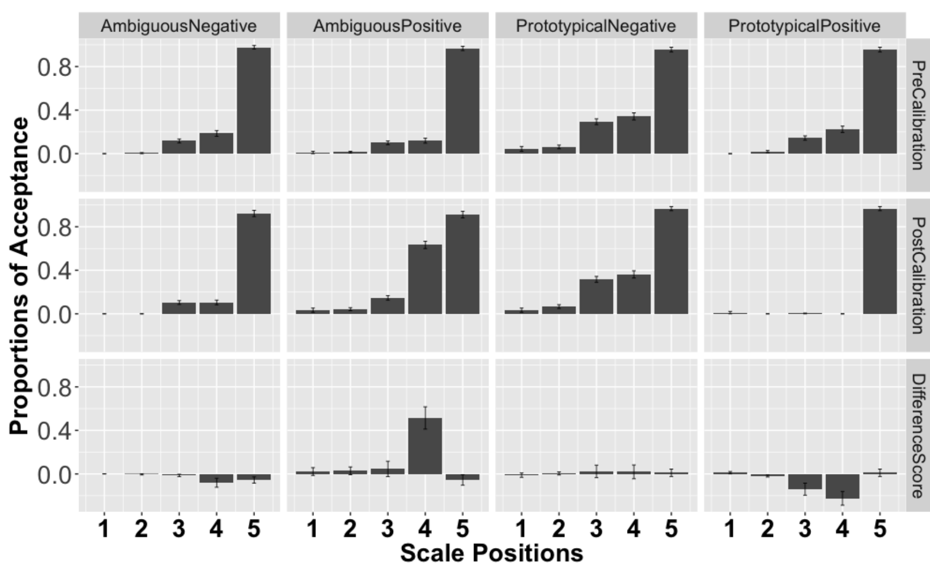

Figure 6: Comparison of pre- and post-calibration "yes" responses for maximum threshold absolute adjective *plain*

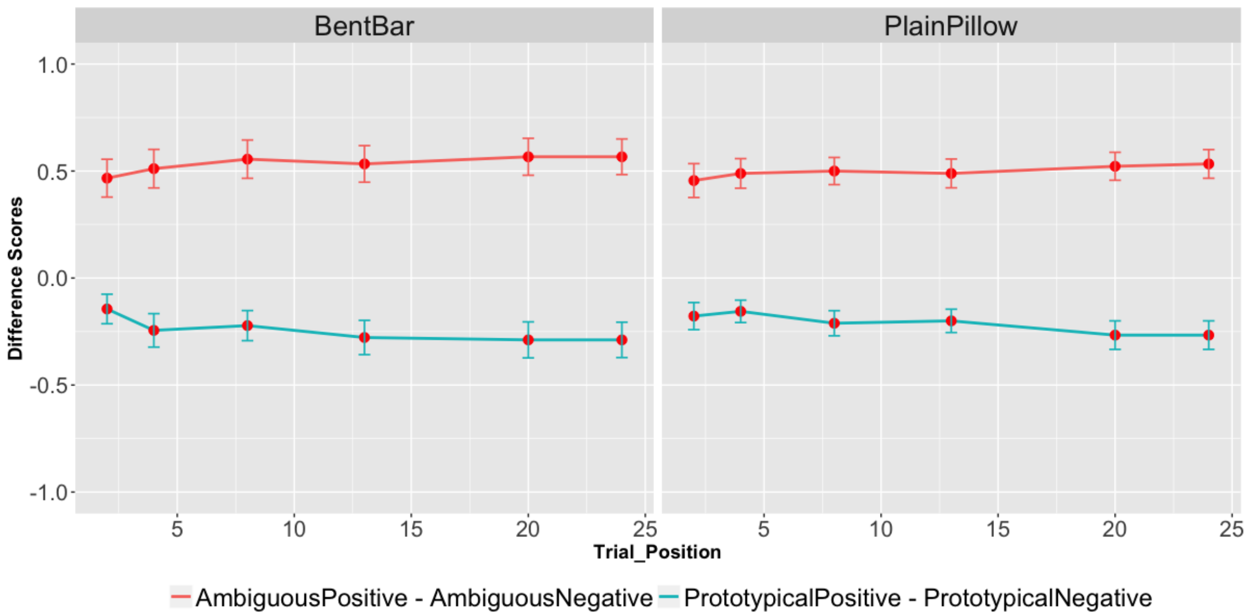

Figure 7: The time-course of adaptation for *bent* and *plain*

bent: $Est = 0.22 \pm 0.23, z = 0.96$; for *plain*: $Est = -0.02 \pm 0.07, z = -0.22$). Further tests showed that for both *bent* and *plain* the adaptation effect appeared at the earliest testing location (i.e. after the 2^{nd} exposure trial, for *bent*: $Est = 9.04 \pm 3.24, z = 2.8$; for *plain*: $Est = 4.12 \pm 0.97, z = 4.2$).

For trials tested after the PrototypicalPositive and PrototypicalNegative exposures, we found no significant effects for *bent* (Exposure type: $Est = -0.71 \pm 1.53, z = -0.5$; interaction between the two factors: $Est = -0.64 \pm 0.47, z = -1.3$). For *plain*, although there was an interaction effect (Exposure type: $Est = 0.27 \pm 2.14, z = 0.13$; interaction between the two factors: $Est = -0.44 \pm 0.19, z = -2.3$), further testing at each of the 6 testing locations showed very little support for robust adaptation, since 5 out of the 6 locations did not show an effect of Exposure type ($ps > .1$). By and large, although Figure 7 showed a trend of an adaptation effect after the prototypical exposure trials, the effect is not statistically robust.

All together, the maximum threshold adjective *plain* and the minimum threshold adjective *bent* showed an effect of adaptation, which appeared at the earliest testing position and did not change afterwards.

2.3.3. Summary of results

We present a summary of the major effects for all three adjective scales in Table 1 and Table 2.

	tall	*bent*	*plain*
(Pre- vs. Post-) Calibration	<.05	<.0001	-
Calibration x Exposure Group	<.0001	<.0001	<.0001

Table 1: Experiment 1: Summary of the effects of adaptation

	tall		*bent*		*plain*	
	Ambig	Proto	Ambig	Proto	Ambig	Proto
Exposure Group	<.0001	<.06	<.001	-	<.0001	-
Exposure Group x Test Location	-	<.01	-	-	-	<.05

Table 2: Experiment 1: Summary of the time-course of adaptation

Qualitatively speaking, the four exposure groups had a similar effect on all adjective scales. The common pattern emerged from Figure 3, 5, and 6 is that for all of them, the AmbiguousNegative and the PrototypicalPositive exposure trials lead to more "no" responses in the post-calibration phase, and the AmbiguousPositive and the PrototypicalNegative exposure trials lead to more "yes" responses in the post-calibration phase. Quantitatively speaking, however, it is also true that the effect of each exposure group is not identical for the three types of adjectives. In particular, for *tall*, all exposure groups showed substantial effects, but for *bent* and *plain*, the effects are more dependent on which exposure trials were presented. We performed additional logistic regression models on data from all the adjective scales together. The full model included the 4 exposure groups, 2 calibration phases, and 3 adjective scales and all the interaction terms as predictors. Another model removed the three-way interaction between the three factors. Model comparison between these two models showed a significant interaction between the three factors ($p < .00001$). For the time course analyses, recall again that the testing trials at different temporal point during the exposure sequence targeted the ambiguous region on each adjective scale. For the relative adjective scale (i.e. *tall*), participants' judgments about these ambiguous objects changed quickly both after the Ambiguous and after the Prototypical exposure trials. For the absolute adjective scales (i.e. *bent* and *plain*), however, changes were only observed after the Ambiguous but not the Prototypical exposure trials. In the discussion section, we propose an explanation for this difference in adaptation effects based on exposure condition.

3. Accounting for adaptation through probabilistic belief update

As discussed in section 1, we assume a standard semantics for gradable adjectives such that an utterance like *"that candle is tall"* is true just in case the height of the candle exceeds a contextual threshold of height. More generally, an utterance of the form *"x is A"* is true in a context c just in case $\theta \leq \delta_A(x)$, where $\delta_A(x)$ is the degree to which x is A and θ is the threshold for A in c. Based on this, when a participant starts giving more "yes" responses to the question *"Is x tall/bent/plain?"* in the post-calibration phase relative to the pre-calibration phase, it means they have chosen a less stringent θ, widening the application condition of the adjectival predicate. Analogously, when a participant starts giving more "no" responses to the question *"Is x tall/bent/plain?"* in the post-calibration phase, it means they have chosen a more stringent θ, narrowing the application condition of the adjectival predicate. In our experiments, the former situation occurred after participants were exposed to the AmbiguousPositive and the PrototypicalNegative exposure trials, and the latter occurred as a result of the AmbiguousNegative and the PrototypicalPositive exposure trials (see Figure 3, 5, and 6). This shift in thresholds, under the influence of specific exposure trials, is the core of the adjective adaptation effect we aim to explain.

In a nutshell, we propose that threshold adaptation is the by-product of the process of establishing listener-talker coordination on facts about objects in communicative exchanges involving semantic uncertainty. In our study, this involved coordination on the degree to which an object under discussion manifests the gradable property named by a gradable adjective, e.g. coordination on the height of a particular candle, the height of which best exemplifies what means for the predicate *tall* to be true. The proposal we outline below is couched upon the recent development in the Rational Speech Act framework and other related models for modeling pragmatic reasoning in a communicative context (Frank and Goodman, 2012; Goodman and Frank, 2016; Lassiter and Goodman, 2013, 2015, 2017; Qing and Franke, 2014a,b). As in all communicative exchanges, the coordination between interlocutors takes place in the context of a particular conversational goal, which was made explicit in our task by the talker's first utterance: *"I need a tall candle for a party,"* etc. Given an assumption of the talker's sincere participation in the conversational interaction, the listener may then take her subsequent utterances about particular objects to be geared towards coordinating on facts about these objects that are relevant to achieving the conversational goal, e.g. what the candle's height is and whether that height is a good exemplar as a tall candle in the context of a party.

Upon hearing the talker's utterance *"This candle is tall"*, the listener knows something about the height of the candle, since the image of the candle is presented to the participants, and he also knows that this candle is considered tall by the talker. What was not explicitly given to the listener is the talker-specific threshold for *tall*. But the listener already has sufficient information to reason about the threshold in the talker's mind. This is so because the listener can make the pragmatic inference that the talker, being a cooperative communicative partner, engages in making her utterances maximally informative (Grice, 1975). The talker's production of an utterance using a semantically uncertain gradable adjective, such as *"that candle is tall"*, is conditioned on the height h of the candle she is referring to, and her decision about the threshold θ, which represents the cut-off between heights that count as tall and those that do not. Following the recent development of probabilistic pragmatic inferences (Frank and Goodman, 2012; Goodman and Frank, 2016; Lassiter and Goodman, 2013, 2015, 2017), the probability of a speaker making an utterance (e.g. her choosing to say *"x is tall"*) is determined by the utility of the utterance, which in simple scenarios is largely determined by the informativity of the utterance to the listener.[4]. We define this in equation (1), with h representing height, and θ representing the threshold.

[4]Strictly speaking, the utility of an utterance reflects trade-off between informativity of the utterance and the cost of producing it. We will focus on informativity and not cost in this paper.

$$P_{talker}(\text{“}x\ is\ tall\text{”} \mid h, \theta) \propto exp(\lambda(informativity(\text{“}x\ is\ tall\text{”}, h, \theta))) \quad (1)$$

The informativity of an utterance can be further defined as the negative surprisal of a *listener*'s updated probabilistic belief about h given the utterance and the threshold (equation 2). In particular, the listener's belief update is conditioned by the truth conditions of the given utterance (equation 3, where $[\![\phi]\!]^{\theta}$ is the semantic value of ϕ relative to θ, in this case a truth value).[5] In other words, before hearing any utterance, a listener may initially have a probability distribution over different heights as to how likely each height is the one that the speaker may refer to. But upon hearing the utterance, the listener will update his belief such that those heights that will make the utterance false are now removed from consideration, and the probability mass is redistributed among those heights that make the utterance true. Combining equations (1) to (3), the utterance that a speaker chooses to use is the one she believes will maximize the probability of the target degree of height — the height of the referent in *her* mind — in the *listener's* mind.

$$informativity = -log(\frac{1}{P_{Listener}(h \mid \text{“}x\ is\ tall\text{”}, \theta)}) = log(P_{Listener}(h \mid \text{“}x\ is\ tall\text{”}, \theta)) \quad (2)$$

$$P_{Listener}(h \mid \text{“}x\ is\ tall\text{”}, \theta) = P_{Listener}(h \mid [\![x\ is\ tall]\!]^{\theta} = TRUE) \quad (3)$$

Given the discussion above, when a speaker says *“that candle is tall”*, she intends to communicate a particular degree of height h in her mind as the candle she needs. If the communication is successful, that particular height should have the highest probability in the listener's updated posterior probability distribution over all possible heights for candles. What is at stake here, however, is that this communication is mediated by the truth condition of the

[5] This is essentially the *literal listener* in the Rational Speech Acts model (Frank and Goodman, 2012; Goodman and Frank, 2016). In the original RSA model, there are multiple levels of recursive reasoning when computing the posterior belief of a listener. For the current paper, however, the simple literal listener level suffices to demonstrate how the basic adaptation process operates.

utterance. But the truth conditions of an utterance with gradable adjectives are dependent on the particular value of the threshold θ. If a speaker and a listener do not share similar assumptions about what the likely θ values are, there will be misalignment between the information the speaker wants to deliver and the information the listener receives.

In the exposure phase of the current study, participants were presented with the image of the candle that the other talker was referring to. The listener, under the belief that the speaker is cooperative and intends to be informative, could make the inference that the speaker believes the height of the candle presented to the listener should have a relatively high probability per the equation in (3). This is an important cue for the listener. A cooperative listener, in an effort to be more aligned with the talker, should decide to boost the probability of the target h, since that would be the safest strategy to ensure speaker-listener coordination. Given equation (3), there are at least two different ways to boost the probability of the target h. First, in situations in which there is uncertainty as to whether the target h makes the utterance true, as in our Ambiguous trials, the listener can adjust their thresholds such that it becomes more likely that the utterance is true when applied to target height h. Second, if there is no uncertainty as to whether the utterance is true for the target h, as in our Prototypical trials, the only way to increase the probability of h is to ensure that there are fewer alternative heights such that the utterance is true of them. In other words, listeners should adjust their thresholds to decrease the probability that the utterance is true for objects with heights other than h. We will show below that these two strategies account for the adaptation behavior for the Ambiguous and Prototypical exposure groups respectively.

To recap, we propose a single unified mechanism to account for the adaptation behavior in gradable adjective interpretation. For the rest of this section, we apply the general proposal outlined above to each experimental condition. The key mechanism accounting for all behavioral patterns is that a listener, after the exposure to another individual use of a gradable adjective, aims to increase the probability of the target object as the one that best manifests the speaker's intention. Since this belief update is conditioned by the semantic truth condition of the utterance, which are dependent on the latent threshold variable, the listener must ultimately adjust their beliefs about the threshold. In our discussion below, we focus on explaining how the thresholds change based on the exposure. We will first explain the adaptation for relative adjectives, and then extend the discussion to absolute minimum and maximum adjectives.

3.1. Threshold adaptation in relative adjectives

We first make the assumption that listeners (and speakers) hold probabilistic beliefs about a threshold distribution, instead of making fixed point estimates (Lassiter and Goodman, 2013; Qing and Franke, 2014a). Before a listener is exposed to other individual's utterances, she has a prior probability distribution over possible threshold values. When adaptation occurs, she will update this prior distribution based on the exposure. For relative adjectives like *tall*, we make a simple assumption that a listener's prior beliefs about the threshold follows a normal distribution over heights (Qing and Franke, 2014a). Given this distribution and the basic semantics of the gradable adjectives, for any object x with a particular height ($height(x)$), the probability that *"x is tall"* is (judged to be) true in a particular context of utterance is just the probability of θ falling below $height(x)$. This amounts to the cumulative probability under the threshold distribution up to $height(x)$, as shown in (4).

$$P(\theta \leq height(x)) = \int_{0}^{height(x)} P(\theta)\, d\theta \tag{4}$$

The leftmost plot in Figure 8 represents an hypothetical listener's prior beliefs about threshold distribution for a range of heights such as those of the candles in Experiment 1. Given these priors, the probability that an object x with height h_3 counts as tall — i.e., the probability that $\theta \leq height(x)$, with $height(x) = h_3$ — is around 0.5, as shown by the probability mass in the shaded area under the black curve of the θ-distribution. h_3 thus corresponds to the height of an ambiguous token for this individual.[6] h_1, on the other hand, corresponds to

[6]For the purpose of illustration, we will use scale point three here to represent the ambiguous token, but recall that the actual ambiguous token was determined individually for each participant based on their responses in the pre-calibration phase. The prototypical negative and prototypical positive scale tokens, in contrast, always corresponded to scale points one and five, respectively.

the height of a prototypical negative token for this individual, since the probability mass to the left of h_1 under the black curve is close to zero and there is therefore an exceedingly low probability of θ falling below h_1. And finally, h_5 corresponds to the height of a prototypical positive token, since the probability of θ falling below h_5 is extremely high.

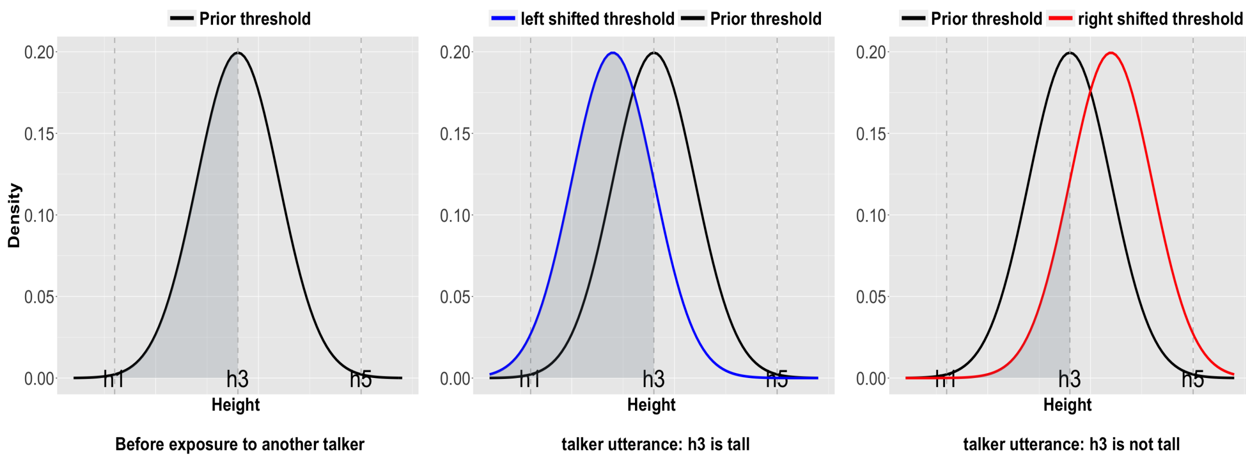

Figure 8: Changes of the listener's threshold distribution for **tall**. The three distributions have different means but the same sd. **Left**: θ distribution before exposure trials. **Middle** and **Right**: θ distribution after the AmbiguousPositive and AmbiguousNegative exposure trials

Taking the leftmost plot in Figure 8 as the listener's prior, we now ask what happens when he is exposed to a talker who asserts of an object x with height h_3 that *"x is tall,"* as was the case for the participants in the AmbiguousPositive exposure group in Experiment 1. Given the simple model of communication with gradable adjectives outlined above, the goal of the communicative exchange is to achieve listener-talker coordination on the height of the object described, which we cash out as increasing the listener's posterior probability for h_3. This value is conditioned on the truth of the utterance, but for an ambiguous token the probability that the utterance is true given the listener's prior for θ is only .5. Given the assumption of speaker sincerity, the listener should boost this value upon hearing the utterance, and can do so by adopting a new distribution for θ which is shifted towards the lower end of the scale, as in the middle plot in Figure 8. The blue curve represents the downward-shifted θ distribution, under which there is a now larger probability mass to the left of h_3 than there was under the black curve that represents the baseline. There is a corresponding increase in the probability that h_3 counts as tall, and so an increase in the listener's posterior belief that h_3 is the height of the candle referred to.

In a completely analogous way, a listener exposed to an individual who asserts that *"x is not tall"* when x has the height h, as was the case for the AmbiguousNegative exposure group, should shift his threshold distribution to the right, increasing $P(\theta > height(x))$, the probability that the negated utterance is true.[7] This update is illustrated by the red curve in the rightmost plot in Figure 8, in which the probability mass to the left of h_3 is smaller under the red curve than under the prior baseline black curve, decreasing the probability that $\theta \leq height(x)$ (and hence increasing the probability $P(\theta > height(x))$). The end result is that, upon hearing an utterance about an ambiguous token on a scale, a listener updates her threshold distribution in the direction that increases the probability that the utterance is true; which direction this is depends on whether utterance is an affirmation or denial that the object has the relevant property: the former results in threshold lowering, the latter in threshold raising.

The schematic demonstration in Figure 8 shifts the threshold distribution only by moving the means, but the variance stays the same. In principle, simultaneously shifting the mean and reducing the variance would achieve the same goal, as shown in Figure 9. It is possible that there are different strategies for changing the threshold distribution. Our proposal only predicts the qualitative trend of the data; we do not have precise quantitative estimates of the parameter changes. However, we note that shrinking the variance appears to be more in line with intuition. As shown in Figure 8, for the AmbiguousPositive exposure group (middle plot), the posterior blue threshold distribution derived by only shifting the means predicts that in the post-calibration phase, the listener should increase the acceptance rate for *every* scale point. That is to say, she should give more *"Yes, it is tall"* responses not only for the ambiguous scale points in the middle range of the scale, but also for the extreme ends of the scale. This could lead to incorrect predictions about extreme values, however. Intuitively, if a listener is certain that a prototypically short candle should not be called *tall*

[7] If *x is tall* is true just in case $\theta \leq height(x)$, then *x is not tall* is true just in case $\theta > height(x)$.

— for example a candle with height $h1$ on the scale — her judgment about this item should stay relatively stable from the pre-calibration to the post-calibration phase (see Figure 3). But under the blue curve in the middle plot in Figure 8, even $h1$ has a non-trivial probability to be considered *tall*. Similarly, for the AMBIGUOUSNEGATIVE exposure group (right-most plot), the posterior red threshold distribution derived from only shifting the means predicts that in the post-calibration phase listeners would reduce the acceptance rate for *every* scale point, include the tallest candle (e.g. $h2$ on the scale). This, again, is not intuitively appealing. And in fact, the empirical data in Figure 3 shows that in the two AMBIGUOUS exposure groups, listeners were most likely to change their judgments about the scale points in the middle range of the scale from the pre-calibration to the post-calibration phase, but not as much for the extreme ends of the scale. It appears that when a talker makes an assertion about an ambiguous scale point, listeners are willing to take that as evidence to update their beliefs about the ambiguous region of the scale, but they are more conservative when it comes to changing beliefs about the region for which they already had a relatively high degree of certainty. Shrinking the variance of their updated threshold distribution, as shown in Figure 9, more adequately models this aspect of the empirical data.

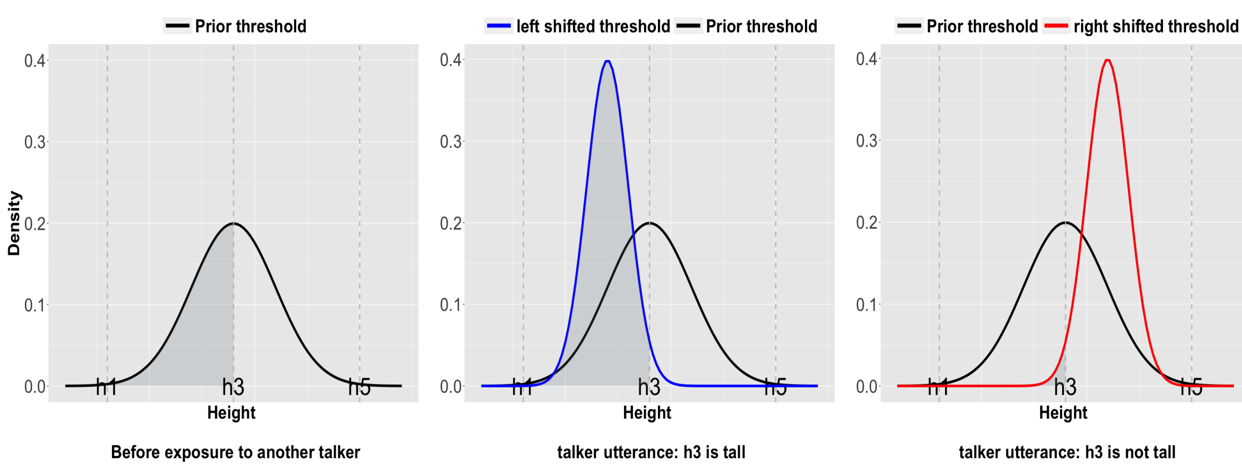

Figure 9: Same as Figure 8. But the shifted θ distributions have both different means and sds from the original distribution.

The preceding discussion showed how listener-talker coordination on the height of an object from the ambiguous region of a scale triggers shifting of the threshold, on the listener's part, in a way that boosts the probability that the utterance is true of the ambiguous object, resulting in the qualitative pattern of threshold adaptation that we observed for the AmbiguousPositive and AmbiguousNegative exposure groups. Now we show that the same pragmatic model can also derive the qualitatively distinct pattern of adaptation that we observed for subjects in the PrototypicalPositive and the PrototypicalNegative exposure groups. Recall that participants in the former group were exposed to positive utterances (*"this candle is tall"*) about the object at scale point five (the tallest candle), and participants in the latter group were exposed to negative utterances (*"this candle is not tall"*) about the object at scale point one (the shortest candle). For these participants, unlike those in the Ambiguous exposure groups, there was no uncertainty about the truth of the talker's utterance: as shown by data in the pre-calibration phase, their acceptance rate of *"this candle is tall"* at scale positions five and one were at ceiling and floor, respectively. Coordination on the height of a prototypical object cannot, therefore, be achieved by modifying the distribution of θ to increase the probability that the utterance is true; instead, coordination is achieved by modifying the distribution of θ in such a way as to reduce the probability that the utterance is true of objects with other heights.

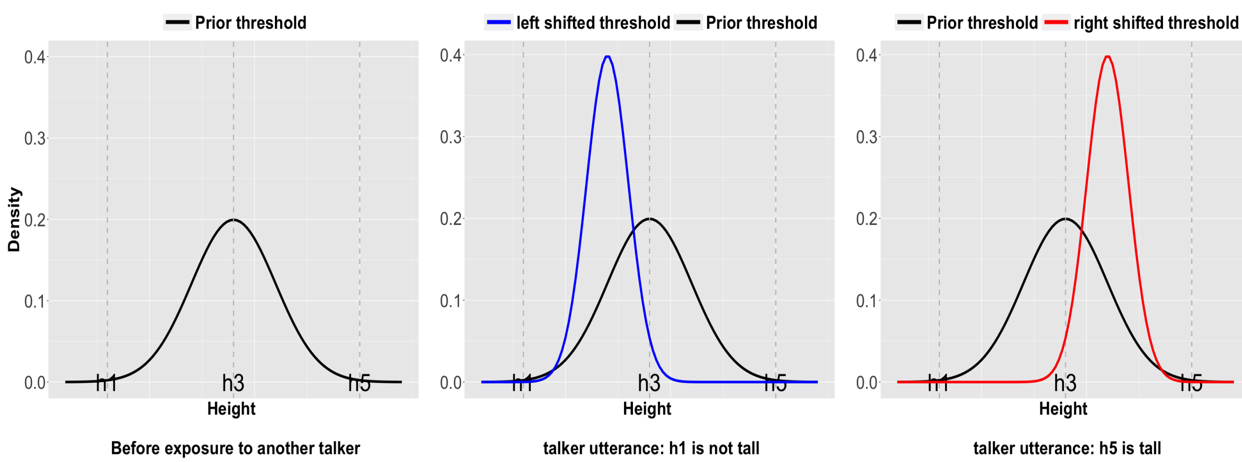

Figure 10: Threshold shift after the prototypical stimulus for **tall**. Both the mean and the variance change. **Left**: θ distribution before exposure trials. **Middle** and **Right**: θ distribution after the PrototypicalNegative and PrototypicalPositive exposure trials

In the PrototypicalPositive scenario, this means shifting the prior threshold distribution towards the upper end of the scale, as shown in the righthand plot of Figure 10. Such a shift does not increase the probability that the utterance is true of the prototypical positive object with height h_5, but it does reduce the probability that the utterance is true of objects with lower heights, in effect narrowing the application conditions of the predicate, and in particular ruling out objects in other region of the scale that could have been previously considered true for the utterance *"this candle is tall"*. For instance, the probability that height h_3 is true is much smaller now under the red θ distribution than under the prior baseline black θ distribution. After removing more candidates from consideration, the probability of h_5 being the target height intended by the speaker will in effect be boosted. And in the PrototypicalNegative scenario, the prior threshold distribution is shifted towards the lower end of the scale, as in the middle plot in Figure 10. This does not change the probability that the negative utterance *"this candle is not tall"* is true of the prototypical negative object with height h_1, but it does reduce the probability that it is true of objects higher on the scale. For instance, under the blue θ distribution, the utterance *"x is not tall"* is much less likely to be true for the height h_3 than under the prior baseline black θ distribution. This has the effect of widening the application conditions of the non-negated form of the predicate.

To summarize, after the AmbiguousPositive and the PrototypicalNegative exposure trials, participants shifted their threshold distribution downwards on the scale, widening the application condition of the adjective predicate. As a result, in the post-calibration phase participants gave more "yes" responses to the question "Is x ADJ". Meanwhile, after the AmbiguousNegative and the PrototypicalPositive exposure trials, participants shifted their threshold distribution upwards on the scale, narrowing the application condition of the adjective predicate, leading to more "no" responses in the post-calibration phase.

3.2. Threshold adaptation in absolute adjectives

The exact same pragmatic mechanism that we used to account for adaptation in relative adjectives can explain the adaptation behavior for absolute adjectives, i.e. minimum adjectives such as *bent* and maximum adjectives such as *plain.* In brief, upon hearing an utterance "x is bent/plain", a listener will make the inference that, per equation 3 in section 3, the utterance is meant to communicate a high probability of the bentness or plainness degree manifested by the target object. To align with the speaker's intent, the listener will shift their prior threshold distribution downwards on the scale after the AmbiguousPositive and the PrototypicalNegative exposure trials, and upwards after the Ambiguous-Negative and the PrototypicalPositive exposure trials. So far, the principles that govern the threshold shift is completely identical as relative adjectives. There is a key difference, however, between absolute and relative adjectives concerning the prior threshold distribution: for relative adjectives, the distribution is most plausibly treated as normal; for absolute adjectives, it is skewed towards one end of the scale or the other. For a minimum threshold absolute adjective like *bent*, the distribution peaks at the lower end of the scale; for a maximum threshold adjective like *plain*, the threshold peaks at the maximum end of the scale. Here we follow Qing and Franke (2014a) and use a beta(a,b) distribution to represent the θ prior for absolute adjectives. The properties of the prior have interesting consequences for the threshold adaptation behavior.

Let us begin by examining the adaptation profile of the minimum threshold adjective *bent*. The predictions of our pragmatic model for the Ambiguous and Prototypical exposure groups are schematically represented in Figures 11 and 12, respectively. The leftmost plot in each figure represents the prior θ distribution, in which there is a peak at the lower end of the scale. This threshold distribution sufficiently explains the pre-calibration judgments, in which we saw that subjects systematically answered "no" to the question *"Is this bent?"* for the object at scale point one, and then immediately started to answer "yes," from scale point two. For the purpose of demonstration, we take b_2 as the degree of bend for the ambiguous token; b_1 and b_5 in turn correspond to the degrees of prototypical negative and prototypical positive tokens, as before.

With the Ambiguous exposure trials (Figure 11), the θ distribution would shift downwards with a reduced variance after the AmbiguousPositive exposure trials. But since the majority of the probability mass is already highly concentrated at the lower end of the scale in the prior θ distribution, the new distribution (blue) is almost identical as the prior distribution (black). But when the θ distribution is shifting upwards, as for the AmbiguousNegative trials, the new distribution (red) could potentially be very different from the prior distri-

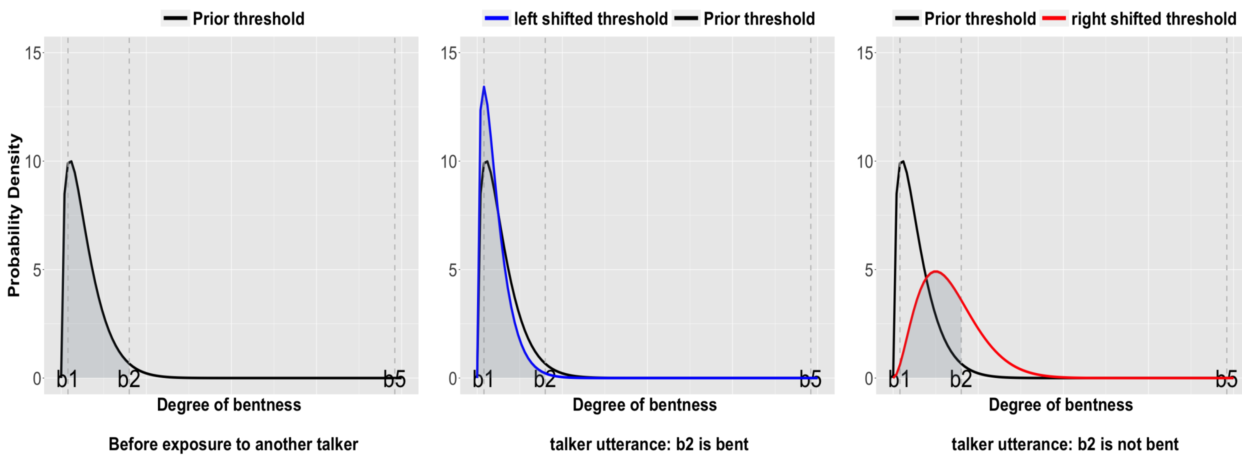

Figure 11: Threshold shift for **bent**, after the **Ambiguous** exposure trials. **Left**: θ distribution before exposure trials. **Middle** and **Right**: θ distribution after the AmbiguousPositive and AmbiguousNegative exposure trials

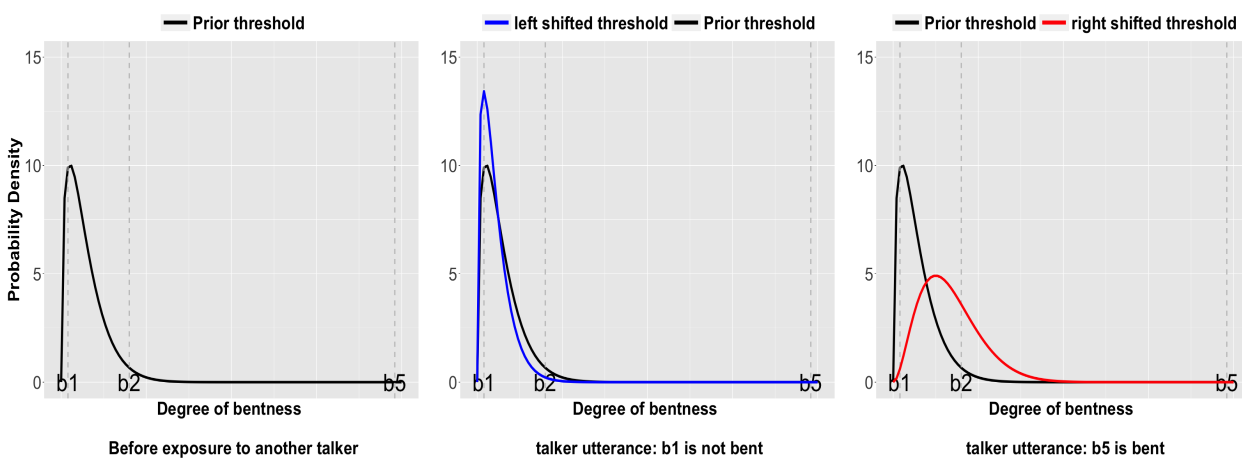

Figure 12: Threshold shift for **bent**, after the **Prototypical** exposure trials. **Left**: θ distribution before exposure trials. **Middle** and **Right**: θ distribution after the PrototypicalNegative and PrototypicalPositive exposure trials

bution (black). Analogously, with the Prototypical exposure trials, the downward shift of the θ distribution after the PrototypicalNegative trials leads to little change in the threshold, but the upward shift of θ after the PrototypicalPositive trials results in substantial changes. What this means is that there should only be minimal effect of adaptation for participants in the AmbiguousPositive and the PrototypicalNegative exposure groups, but there should be a sizable adaptation effect for the other two exposure groups. This is exactly what the experimental results showed in Figure 5. This result reflects the intuition that there is a relatively small degree of uncertainty about the application condition of *bent* to start with. A minimal degree of bend is sufficient for something to be considered as bent. It is therefore not easy to further widen the application condition. But it is possible to narrow the application condition, i.e. shifting the threshold upwards and making the application criterion more stringent.

For the maximum threshold adjective *plain*, the predicted adaptation behavior is schematically shown in Figure 13 and Figure 14 for the Ambiguous and Prototypical exposure groups, respectively. Prior to exposure, the threshold distribution is heavily concentrated on the upper end of the scale (the left-most plot in Figure 13 and 14). For a pillow to be considered plain, it demands a very high degree of plainness, as confirmed by the judgment results in the pre-calibration phase in Figure 6. In the schematic demonstrations in Figure 13 and 14, p_1 represents the plainness of a prototypically negative token (in our study, a pillow with lots of spots); p_5 represents the prototypically token (a pillow with no spots), and here we use p_4 to exemplify the (slightly) ambiguous token.

For the AmbiguousPositive and the PrototypicalNegative exposure groups, we expect lowering of the prior threshold distribution (i.e. the blue distributions in the two figures). This should result in more acceptance of tokens on the ambiguous region of the scale to be considered plain. This was confirmed by the increased number of "yes" responses in the post-calibration phase, compared to the pre-calibration phase, in Figure 6, but only for the AmbiguousPositive group. There was virtually no change in the responses in the PrototypicalNegative exposure group, perhaps because the initial threshold distribution was already skewed too far towards the maximum to be influenced by a negative statement about

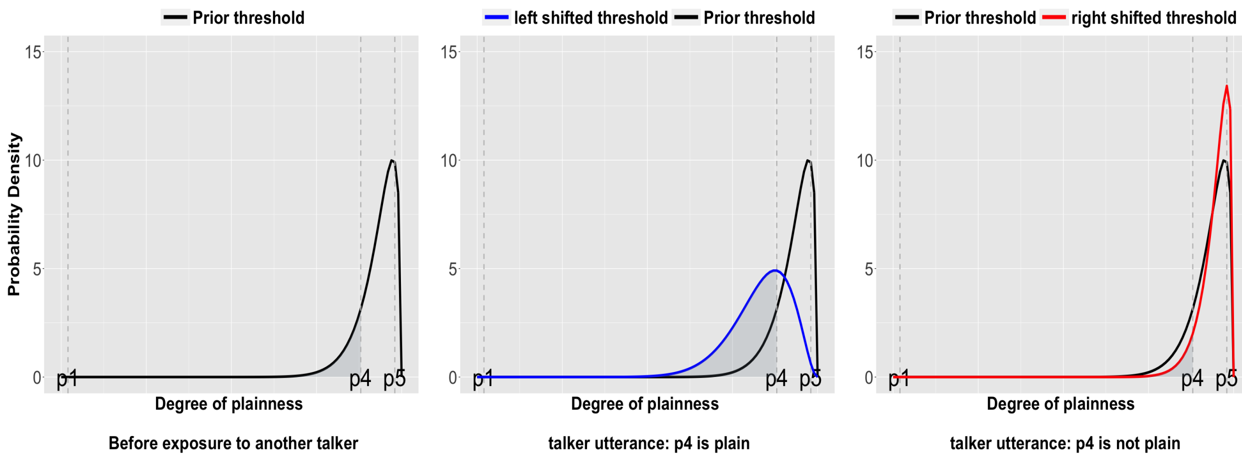

Figure 13: Threshold shift for **plain**, after the **Ambiguous** exposure trials. **Left**: θ distribution before exposure trials. **Middle** and **Right**: θ distribution after the AmbiguousPositive and AmbiguousNegative exposure trials

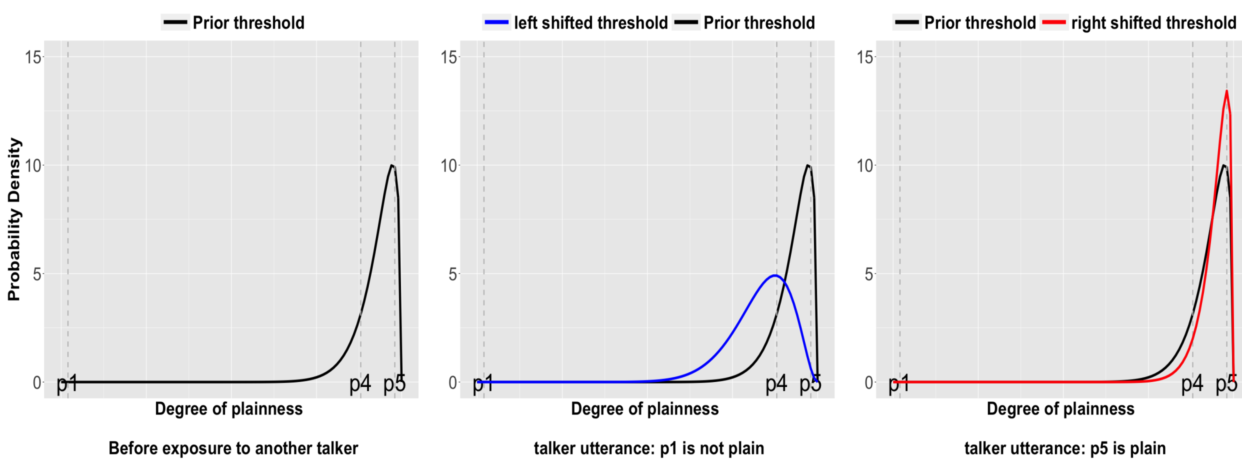

Figure 14: Threshold shift for **plain**, after the **Prototypical** exposure trials. **Left**: θ distribution before exposure trials. **Middle** and **Right**: θ distribution after the PrototypicalNegative and PrototypicalPositive exposure trials

an object at the lower end of the scale. For the AmbiguousNegative and the PrototypicalPositive exposure groups, we expect a raising of the threshold and some reduction in the variance. There is not a lot of room for change since the prior distribution is already skewed towards the upper maximum, but the empirical judgments in the pre-calibration phrase in Figure 6 showed a small degree of uncertainty for scale positions just below the maximum scale point. Further raising the threshold (i.e. the red distribution) resulted in a decrease in acceptance of ambiguous tokens — essentially an effect of "precisification" of the maximum threshold predicate (Pinkal, 1995; Lasersohn, 1999; Kennedy, 2007; Syrett et al., 2010).

3.3. Summary

Summarizing so far, we have proposed that threshold adaptation in gradable adjectives is the by-product of a pragmatic process of establishing listener-talker coordination on the degree to which particular objects possess the property named by the adjective used, in the service of moving towards a particular communicative goal. We have shown that the qualitatively distinct patterns of adaptation that we observed for different exposure groups can be accounted for through evaluating the truth conditions of the talker's utterance and the corresponding threshold distributions. The difference between three classes of adjectives can be explained by the differences in their prior threshold distributions.

A crucial feature of the model is the assumption that the interaction is communicative: our model assumes that the the listener's adjusts her threshold distribution based on her assumption that the talker's utterance is sincere, and geared towards the communicative goal of identifying an object that meets her needs, in virtue of the degree to which it manifests some property. In two follow-up experiments, we introduced two potential disruptions in the listener's assumption that the talker's utterances were sincere and goal-oriented in this way. In Experiment 2, we replaced the human talker with synthesized speech, but retained the talker's explicit statement of conversational goals, and in Experiment 3, we replaced the human talker with synthesized speech and also eliminated any explicit mention of conversational goals.

4. Experiment 2

4.1. Materials, procedure and participants

The stimuli and procedure for Experiment 2 were identical to Experiment 1. The only change is that we used a computer synthesized voice as the talker instead of a native English speaker, to test whether a listener would be less likely to adapt to an atypical talker — in this case a non-human — about whom they might not be willing to grant the same degree of communicative agency as a typical talker. Otherwise, the format was the same; in particular, the synthesized voice began the exposure phase by stating an explicit conversational goal: *"I need a tall candle for a party."*, etc. A total of 120 native English speakers participated in the experiment on Amazon Mechanical Turk, with 30 participants in each exposure group.

4.2. Analysis and results

The averaged results from the pre-calibration, post-calibration and the testing trials during the exposure sequence are plotted in Figure 15. The general patterns look similar to Experiment 1. The statistical analysis, which was also identical to Experiment 1, provided clear evidence that participants' judgments did in fact change after different exposure conditions, as shown in Table 3.

	tall	*bent*	*plain*
Calibration (Pre vs. Post)	$<.01$	$<.0001$	$<.01$
Calibration x Exposure Group	$<.0001$	$<.0001$	$<.0001$

Table 3: Experiment 2: summary of the effects of adaptation

Figure 16 plots the incremental time-course of the adaptation effect, and the results of the statistical analyses are presented in Table 4. We did not find an adaptation effect in the Prototypical exposure cases for *bent* and *plain*, but there was a robust adaptation effect for the rest of the cases. For the time-course results, we also did further testing when there appeared to be an interaction between exposure group and the location of the testing trial. For *tall* in the Prototypical exposure, the interaction was mainly driven by the first

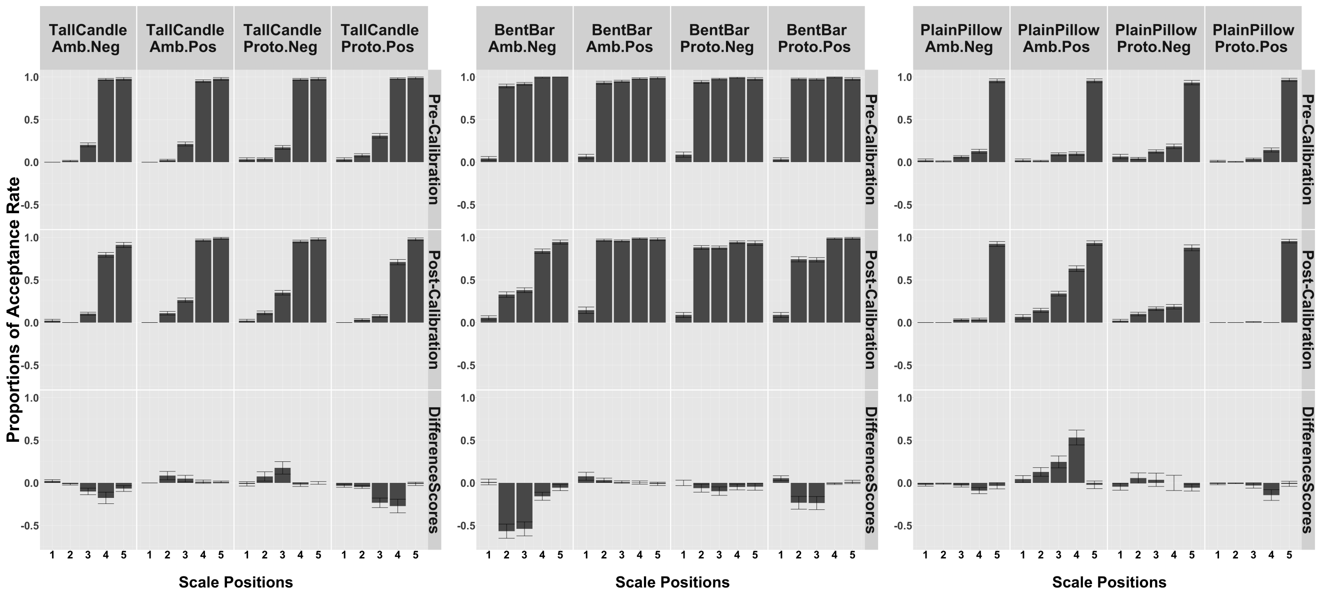

Figure 15: Experiment2: Acceptability rating from the pre-calibration and the post-calibration phases, and their difference scores. **Left**: *tall*; **Middle**: *bent*; **Right**: *plain*.

testing location ((after the 2nd exposure trial). There is a significant exposure effect at this testing location ($Est = -0.8 \pm 0.3, z = -2.8$), though the effect is smaller than the rest of the testing locations. After removing the first testing location, the main effect of exposure remained ($p < .001$), but there was no interaction between exposure and the testing location ($p > .1$), suggesting that there was no increase in adaptation effect after the fourth exposure. Similarly, for *bent* and *plain* in the AMBIGUOUS exposure groups, the interaction between exposure and testing location was driven by the significant but relatively smaller effect at the first testing location. In both cases, after the fourth exposure trial, the exposure effect stayed stable and more exposure did not enhance the adaptation effect.

	tall		*bent*		*plain*	
	AMB	PROTO	AMB	PROTO	AMB	PROTO
Exposure Group	<.0001	<.001	<.001	-	<.05	-
Exposure Group x Test Location	-	<0.01	<.05	-	<.05	-

Table 4: Experiment 2: summary results for the time-course of adaptation

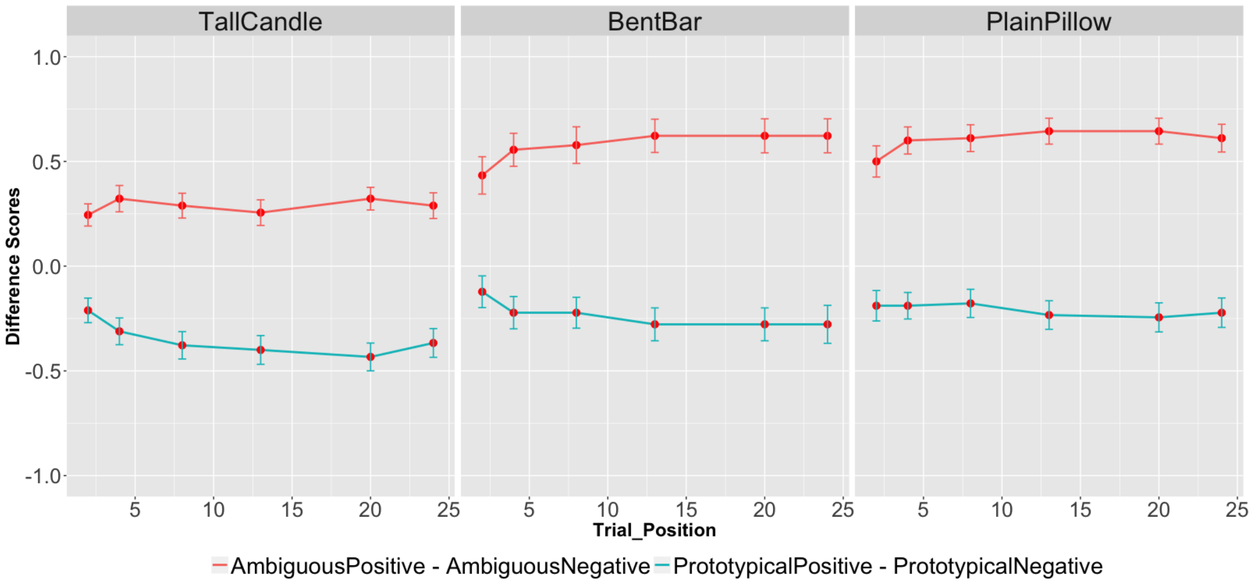

Figure 16: Experiment 2: incremental time course for the adaptation effect

4.3. Summary of Experiment 2

Overall, Experiment 2 resulted the same basic pattern of adaptation behavior as Experiment 1, suggesting that merely making the talker atypical does not change the underlying reasoning driving the adaptation effect. In Experiment 3, we made a further experimental manipulation to remove the communicative goal.

5. Experiment 3

5.1. Materials, procedure and participants

The experimental procedure was largely identical to Experiment 2, but with one additional modification. The speaker's voice was still generated by a speech synthesizer, but the explicit statement of a conversational goal was eliminated. Instead, after the pre-calibration phase, participants hear the following statement: *"We are testing a speech synthesizer that can imitate human voice. In this section you will hear some verbal statements made by this synthesizer."* The exposure phase then began directly with the synthesized voice's statements about particular objects (*"This candle is tall"*, etc.), without an initial specification of a conversational goal.

5.2. Analysis and results

The averaged results are plotted in Figure 17 and 18. The statistical analysis procedure is identical to the previous experiments. There is again clear evidence for an effect of exposure (see Table 5). For the incremental development of the exposure effect (Table 6), there was no effect for the prototypical exposure groups. The two interactions reported in Table 6 were again largely driven by the first testing location; and once the first testing location is removed, the number of exposure trials did not make a difference.

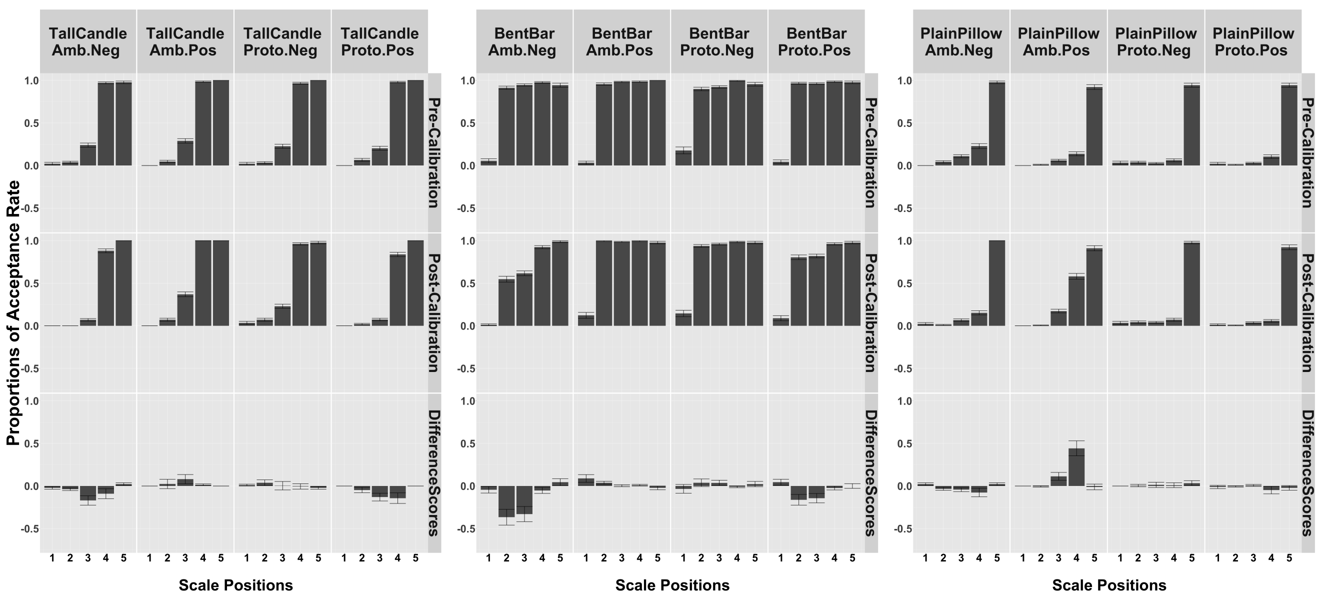

Figure 17: Experiment3: Acceptability rating from the pre-calibration and the post-calibration phases, and their difference scores. **Left**: TallCandle; **Middle**: BentBar; **Right**: PlainPillow.

	tall	*bent*	*plain*
Calibration (Pre vs. Post)	<.01	<.0001	<.05
Calibration x Exposure Group	<.001	<.0001	<.001

Table 5: Experiment 3: summary of the effects of adaptation

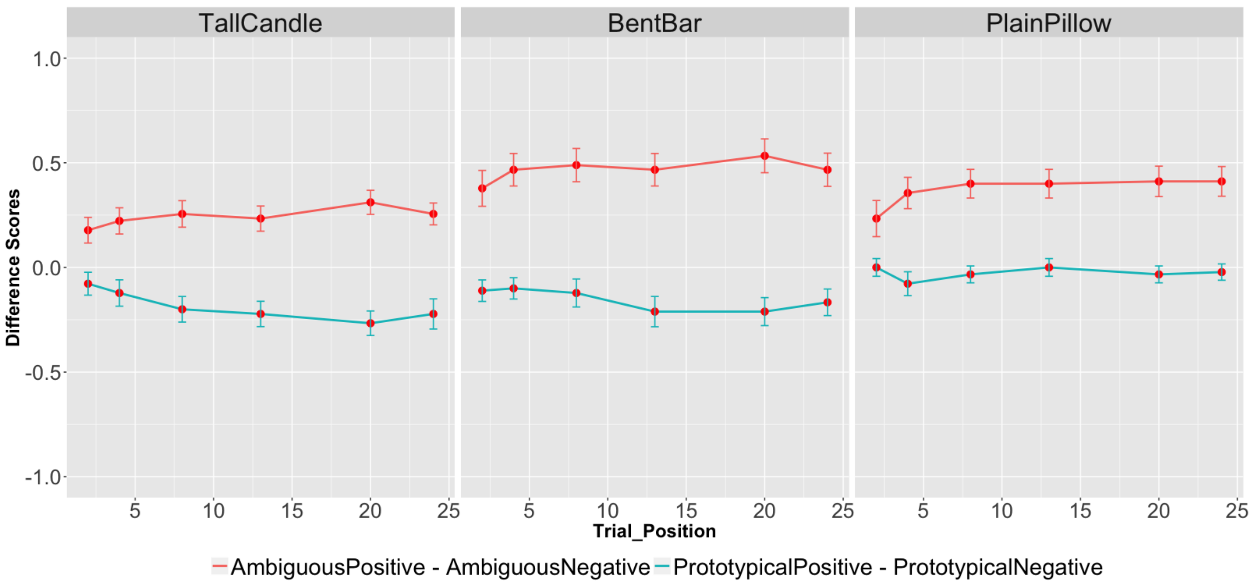

Figure 18: Experiment3: incremental time course for the adaptation effect

	tall		*bent*		*plain*	
	AMBIG	PROTO	AMBIG	PROTO	AMBIG	PROTO
Exposure Group	$<.01$	-	$<.01$	-	$<.001$	-
Exposure Group x Test Location	-	$<.05$	$<.05$	-	-	-

Table 6: Experiment 3: summary results for the time-course of adaptation

5.3. Summary of Experiment 3

Experiment 3 demonstrated that even synthesized speech statements about objects, uttered in the explicit absence of a conversational goal, can trigger threshold adaptation, suggesting 770 that the pragmatic mechanism underlying this effect is quite automatic. However, before drawing more fine-grained conclusions, we will need to compare the effect sizes of the three experiments, which we turn to below.

6. Comparing the effect size of the adaptation effect across three experiments

As we indicated in the results sections for Experiments 1-3, participants' judgments in the 775 post-calibration session were influenced by the exposure trials in all three experiments. Figure 19 presents the overall change from the pre-calibration to the post-calibration phase in listener judgments for the three experiments together.

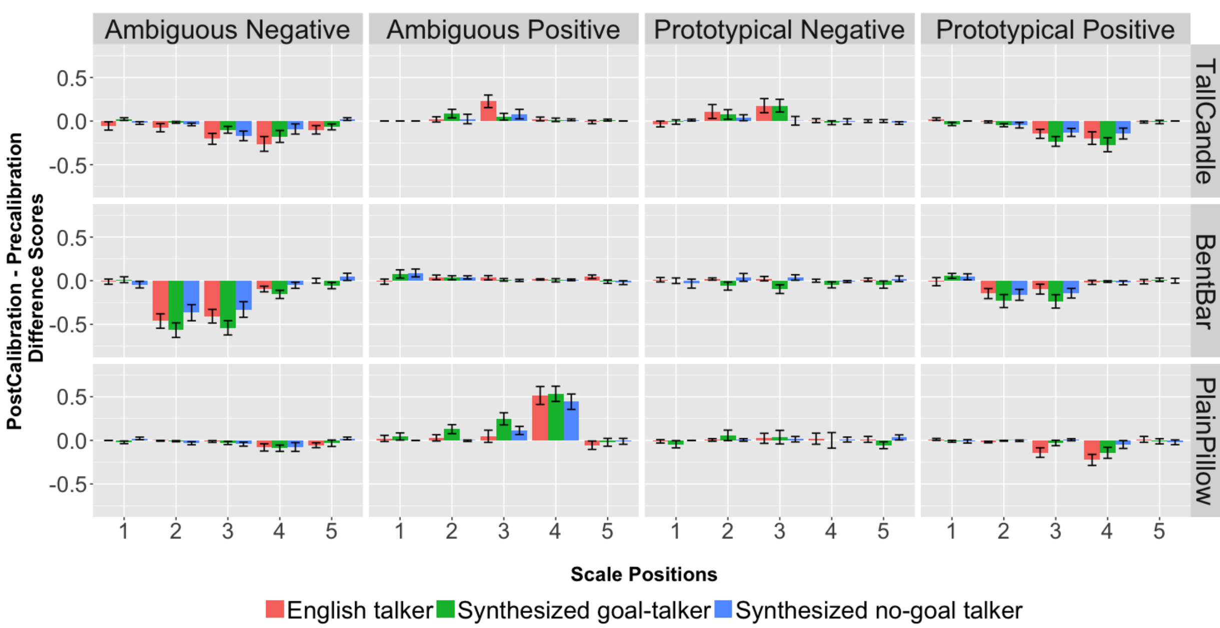

Figure 19: Difference scores, postcalibration-precalibration, for all three experiments

Visual inspection suggests that although qualitatively similar changes appeared in all three experiments, the amplitude of the changes are not identical. To quantify the differences in 780 effect sizes between experiments, we calculated the effect sizes for each experiment. We used odds ratio as the metric for effect size. As an example, consider the judgment results in the post-calibration session for *tall*, after the AmbiguousNegative and AmbiguousPositive exposure trials. We already know these two ambiguous exposure groups produced the opposite effects in the post-calibration phase: participants produced more "no" responses after the AmbiguousNegative exposure trials, and more "yes" responses after the AmbiguousPositive exposure. Additionally, we also observe that, for each adjective scale, participants from the four exposure groups gave very similar judgments in the pre-calibration phase. We therefore focused our analysis on the judgments in the post-calibration phase. We set up a logistic regression to predict the "yes" responses in the post-calibration phases, with the exposure type as the predictor, and the model also included a random intercept by participants. The variable exposure type was treatment coded, with the AmbiguousNegative exposure as the baseline referent level. Taking the exponential function of the coefficient obtained for the AmbiguousPositive level, we derived the odds ratio between the two exposure groups as 3.1. This means that the odds of participants judging e.g. the utterance *"that candle is tall"* to be true (in the post-calibration phase) in the AmbiguousPositive exposure group is 3.1 times of the odds in the AmbiguousNegative group. We therefore obtained a measure that indexes how different participants' judgments were after exposure to two different utterances. A larger difference (as indexes by a larger odds ratio) indicates a larger impact from the exposure utterances, whereas an odds ratio of 1 means that different exposure utterances have no effect in modulating participants' judgments. [8]. Following the same procedure, for the two Prototypical exposure groups with *tall*, we obtain an odds ratio of 0.4. The odds ratio is less than 1 here because the logistic regression model is predicting the "yes" responses, and there was a decrease of "yes" responses from the baseline referent group (the PrototypicalNegative group) to the other group (the PrototypicalPositive group) in the model. For easier interpretability of odds ratios less than 1, we can look at the inverse: 1/0.4=2.5. This means that the odds of participants' judging *"that*

[8] An alternative is to conduct a logistic regression, for each adjective scale under each exposure condition, that predicts the change from the pre-calibration to the post-calibration phase. But that would require a calculation of 3 adjective x 4 exposure x 3 experiments =36 odds ratios, which makes the comparison between experiments more unwieldy.

candle is tall" to be true in the PrototypicalNegative exposure group is 2.5 times of the odds in the PrototypicalPositive exposure group. In Table 7 we present all the effect sizes derived in this way, with inverses in parentheses. The discussion below is based on odds ratios larger than 1 (i.e., for the prototypical exposure groups, we used the inverse odds ratio).

	tall		*bent*		*plain*	
	Ambig	Proto	Ambig	Proto	Ambig	Proto
Experiment 1	3.1	0.4 (2.5)	6.2	0.6 (1.7)	2.8	0.3 (3.3)
Experiment 2	1.8	0.4 (2.5)	9.2	0.7 (1.4)	5.6	0.5 (2)
Experiment 3	1.9	0.6 (1.7)	4.7	0.6 (1.7)	2.1	0.9 (1.1)

Table 7: Effect sizes for ratings in the post-calibration session, measured in odds ratio.

Some observations about the effect sizes are worth noting. Most significantly, Experiment 3 by and large produced the smallest adaptation effect. This suggests that removing the communicative goal from the experiment reduced the listener's adaptation to the speaker's threshold values. Between the other two experiments, the speaker with a synthesized voice (Experiment 2) generally has a slightly smaller effect than the speaker with a human voice (Experiment 1). This is likely due to the fact that a human listener may not consider a synthesized voice as a total communicative agent, and therefore adapts to a smaller degree. But there is an interesting exception to this overall trend: for the two absolute adjectives *bent* and *plain*, but not the relative adjective *tall*, the ambiguous exposure groups produced a larger adaptation effect in Experiment 2 than in Experiment 1. In fact, even in Experiment 3, in which the adaptation effect is generally weaker, there is a stronger effect for the absolute adjectives under the ambiguous exposure groups, but not for the relative adjective. While we do not have an account of this pattern at this time, we note that it may bear on hypotheses about why absolute adjectives show a much smaller degree of threshold variation and uncertainty than relative adjectives. There is an ongoing debate in the literature about whether this merely reflects the fact that the objects in their domain tend to have skewed (rather than normal) distributions, and that otherwise reasoning about thresholds is the same for both absolute and relative adjectives (Lassiter and Goodman, 2013), or whether this reflects the fact that absolute adjective thresholds are lexically determined, and variation emerges from a qualitatively different mechanism for "loose talk" — pragmatic weakening of maximum thresholds and strengthening of minimum thresholds (Kennedy, 2007; Qing and Franke, 2014a; Leffel et al., 2016; Burnett, 2016). One potential explanation of the pattern we observed for absolute adjectives in Experiments 2 and 3, which is more in line with this second approach to threshold variation than the first one, is that listener may think that a *machine* is less likely than a human to engage in loose talk.

7. General discussion

Taken together, our three experiments provide evidence that an individual's decisions about how to resolve semantic uncertainty in the meaning of a gradable predicate in the positive form — specifically, how to fix the value of the adjective's threshold of application, which determines the cutoff point that separates the objects it is true of from the objects it is false of — are influenced by the exposure to another individual's use of the same expressions in a communicative exchange. The adaptation effect happens very quickly, within as few as two exposure trials, and additional exposure trials do not substantially enhance it.

To account for the observed effects, we propose that a listener adjusts her threshold of application for a gradable adjective in a way that maximizes listener-talker coordination on the degree to which the object described by the adjective possesses the gradable property that the predicate encodes in a conversational interaction. We demonstrated that the distinct adaptation effects we observed for statements about ambiguous vs. prototypical objects, and using adjectives with different prior threshold distributions, can be derived from a probabilistic pragmatic model of communication with gradable adjectives that is based around the idea of coordination on degrees in communication. The different effect sizes in the three experiments support the conclusion that this effect is most robust when coordination is in the service of reaching an explicit conversational goal.

At the center of our proposal is the notion of informativity (see section 3). A speaker chooses to make the most informative utterance that can maximize the probability of the listener of receiving the intended message, and a listener interprets the utterance with the understanding that the speaker is being maximally informative. Conditioned by this pragmatic principle, linguistic exchanges not only provide information about the state of affairs in the world, they also provide information about the "state of the language" that conversational participants are using to communicate about the world (Lewis, 1979). In the case of gradable adjectives in particular, linguistic exchanges provide information both about the degree to which an object manifests some gradable property, *and* about the threshold of application for the predicate — what Barker (2002) calls *descriptive* and *metalinguistic* updates, respectively. Typical conversational interactions involve both descriptive and metalinguistic updates, but the observations about semantic adaptation in this study demonstrate language users' sensitivity to the latter in particular. During the exposure phase, when the talker said e.g. *"this candle is tall,"* listeners gained no new information about the height of the candle: that was already clear from the image on the screen. But the talker's utterance did provide crucial information about their use of the word *tall*, which listeners could then use to adjust their posterior beliefs about the threshold distribution.

The fact that people do not just interpret the content of an utterance but also use the utterance to make inferences about how their conversational partners use language is not a unique feature to semantic and pragmatic processing. As we mentioned in the introduction, there are close parallels at the level of speech perception. The current experimental paradigm was adapted and modified from Vroomen et al. (2007) and Kleinschmidt and Jaeger (2015). In both studies, it was found that when an ambiguous auditory token was embedded in an disambiguating context, for instance, when an acoustically ambiguous /b/ or /d/ segment was dubbed onto a face that articulated a /b/ or /d/ target and therefore made the disambiguating articulator information visually available, participants would recalibrate their perception of the auditory targets towards the direction of the visual information. However, if the auditory tokens themselves were non-ambiguous, the pairing of the auditory-visual information would shift participants' perceived speech category to the opposite direction. These findings are parallel to the current adaptation effects triggered by the ambiguous and prototypical exposure utterances respectively. The convergence between the speaker and listener in these different domains signals a more general mechanism for adaptation at different levels of linguistic processing. The probabilistic belief update model we adopted here for pragmatic reasoning is in line with a class of Bayesian models proposed for speech adaptation and perceptual learning in general (Kleinschmidt and Jaeger, 2015; Feldman et al., 2009). But it is worth noting that the effects we observed in the current study are modulated by shared communicative goals between interlocutors, at least to some degree. It is an open question whether the adaptation mechanism in speech perception is a more automatic process than the one guiding adaptation at higher linguistic levels.

8. Conclusion

This study provides novel experimental evidence showing rapid and robust adaptation effect in adjective interpretation. The direction and the size of the adaptation effect vary depending on the polarity of the utterances, the property of the object under discussion, the specific gradable predicate involved in the utterance, as well as the shared communicative goal (or the lack of it) between the interlocutors. But we argue that a single unified mechanism, based on a probabilistic belief update model for pragmatic reasoning, can successfully account for the seemingly diverse set of findings. Both the empirical findings and the theoretical proposal in the current study also find parallels at other levels of linguistic representation, in particular speech perception, opening up more future research directions for the general cognitive architecture of learning and adaptation.

References

Barker, C., 2002. The dynamics of vagueness. Linguistics and Philosophy 25 (1), 1–36.

Bock, K., 1986. Syntactic persistence in language production. Cognitive Psychology 31, 163–186.

Brennan, S. E., Clark, H. H., 1996. Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition 22 (6), 1482.

Burnett, H., 2016. Gradability in Natural Language: Logical and Grammatical Foundations. Oxford Studies in Semantics and Pragmatics. Oxford University Press, Oxford, UK.

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 (2), 633–664.

Cresswell, M. J., 1976. The semantics of degree. In: Partee, B. (Ed.), Montague Grammar. Academic Press, New York, pp. 261–292.

Feldman, N. H., Griffiths, T. L., Morgan, J. L., 2009. The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. Psychological review 116 (4), 752.

Frank, M., Goodman, N., 2012. Predicting pragmatic reasoning in language games. Science 336, 998.

Goodman, N. D., Frank, M. C., 2016. Pragmatic language interpretation as probabilistic inference. Trends in cognitive sciences 20 (11), 818–829.

Grice, H. P., 1975. Logic and conversation. In: Cole, P., Morgan, J. L. (Eds.), Syntax and Semantics, Vol. 3, Speech Acts. Academic Press, New York, pp. 41–58.

Gries, S. T., 2005. Syntactic priming: A corpus-based approach. Journal of psycholinguistic research 34 (4), 365–399.

Jaeger, T. F., Snider, N. E., 2013. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the primeâĂŹs prediction error given both prior and recent experience. Cognition 127 (1), 57–83.

Kennedy, C., 2007. Vagueness and grammar: The semantics of relative and absolute gradable predicates. Linguistics and Philosophy 30 (1), 1–45.

Kennedy, C., McNally, L., 2005. Scale structure and the semantic typology of gradable predicates. Language 81 (2), 345–381.

Klein, E., 1980. A semantics for positive and comparative adjectives. Linguistics and Philosophy 4, 1–45.

Kleinschmidt, D. F., Jaeger, T. F., 2015. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. Psychological review 122 (2), 148.

Kraljic, T., Samuel, A. G., 2005. Perceptual learning for speech: Is there a return to normal? Cognitive psychology 51 (2), 141–178.

Lasersohn, P., 1999. Pragmatic halos. Language 75 (3), 522–551.

Lassiter, D., Goodman, N. D., 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In: Semantics and linguistic theory. Vol. 23. pp. 587–610.

Lassiter, D., Goodman, N. D., 2015. How many kinds of reasoning? inference, probability, and natural language semantics. Cognition 136, 123–134.

Lassiter, D., Goodman, N. D., 2017. Adjectival vagueness in a bayesian model of interpretation. Synthese 194 (10), 3801–3836.

Leffel, T., Xiang, M., Kennedy, C., 2016. Imprecision is pragmatic: Evidence from referential processing. In: Semantics and linguistic theory. Vol. 26. pp. 836–854.

Levy, R., 2018. Using r formulae to test for main effects in the presence of higher-order interactions. arXiv preprint arXiv:1405.2094.

Lewis, D. K., 1970. General semantics. Synthese 22 (1-2), 18–67.

Lewis, D. K., 1979. Scorekeeping in a language game. Journal of Philosophical Logic 8 (1), 339–359.

McConnell-Ginet, S., 1973. Comparative constructions in English: A syntactic and semantic analysis. Ph.D. thesis, University of Rochester.

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 (2), 201–213.

Norris, D., McQueen, J. M., Cutler, A., 2003. Perceptual learning in speech. Cognitive psychology 47 (2), 204–238.

Pickering, M. J., Branigan, H. P., 1998. The representation of verbs: Evidence from syntactic priming in language production. Journal of Memory and language 39 (4), 633–651.

Pinkal, M., 1995. Logic and Lexicon. Kluwer, Dordrecht.

Pisoni, D. B., Levi, S. V., sep 2012. Representations and representational specificity in speech perception and spoken word recognition. Oxford University Press.

Qing, C., Franke, M., 2014a. Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In: Champollion, L., Szabolcsi, A. (Eds.), Proceedings of Semantics and Linguistic Theory 24. pp. 23–41.

Qing, C., Franke, M., 2014b. Meaning and use of gradable adjectives: Formal modeling meets empirical data. In: Proceedings of the Annual Meeting of the Cognitive Science Society. Vol. 36.

Rotstein, C., Winter, Y., 2004. Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers. Natural Language Semantics 12, 259–288.

Samuel, A. G., 1986. Red herring detectors and speech perception: In defense of selective adaptation. Cognitive psychology 18 (4), 452–499.

Syrett, K., Kennedy, C., Lidz, J., 2010. Meaning and context in children's understanding of gradable adjectives. Journal of Semantics 27 (1), 1–35.

Toledo, A., Sassoon, G. W., 2011. Absolute vs. relative adjectives - variance within vs. between individuals. In: Ashton, N., Chereches, A., Lutz, D. (Eds.), Proceedings of Semantics and Linguistic Theory 21. pp. 135–154.

von Stechow, A., 1984. Comparing semantic theories of comparison. Journal of Semantics 3, 1–77.

Vroomen, J., van Linden, S., De Gelder, B., Bertelson, P., 2007. Visual recalibration and selective adaptation in auditory–visual speech perception: Contrasting build-up courses. Neuropsychologia 45 (3), 572–577.

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.