Speaker-specific adaptation to variable use of uncertainty expressions

Sebastian Schuster and **Judith Degen**
{sebschu, jdegen}@stanford.edu
Department of Linguistics, Stanford University
Stanford, CA 94305, USA

Abstract

Speakers exhibit variability in their choice between uncertainty expressions such as *might* and *probably*. Recent work has found that listeners cope with such variability by updating their expectations about how a specific speaker uses uncertainty expressions when interacting with a *single speaker*. However, it is still unclear to what extent listeners form speaker-specific expectations for *multiple speakers* and to what extent listeners are adapting to a situation independent of the speakers. Here, we take a first step towards answering these questions. In Experiment 1, listeners formed speaker-specific expectations after being exposed to two speakers whose use of uncertainty expressions differed. In Experiment 2, listeners who were exposed to two speakers with identical use of uncertainty expressions formed considerably stronger expectations than in Experiment 1. This suggests that listeners form both speaker-specific and situation-specific expectations. We discuss the implications of these results for theories of adaptation.

Keywords: psycholinguistics; semantics; pragmatics; adaptation; uncertainty expressions

Introduction

Speakers exhibit considerable production variability at all levels of linguistic representation (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Weiner & Labov, 1983; Finegan & Biber, 2001). This includes variation in lexical choice to describe a world state. For example, Yildirim, Degen, Tanenhaus, and Jaeger (2016) found that when asked to describe a scene with a candy bowl in which approximately half of the candies were green and half of the candies were blue, some participants judged “Some of the candies are green” to be the more appropriate utterance to describe the scene than “Many of the candies are green”, while others displayed the opposite pattern.

Schuster and Degen (2018) found that participants exhibit similar production variability when describing an event with an objective event probability of 60%: Some participants judged the event to be best described with a sentence containing the uncertainty expression *might* (“You might get a blue gumball”) whereas others judged a sentence with *probably* (“You’ll probably get a blue gumball”) more appropriate.

Such variability poses a challenge to a listener who aims to know what the world is like that the speaker is describing. When confronted with two speakers who use the same expression to convey different states of the world or who use different expressions to convey the same state of the world, listeners are doomed to draw the wrong inferences about the actual state of the world unless they track how individual speakers use language. Recent work suggests that listeners deal with this kind of variability by adapting to it (e.g., Norris, McQueen, & Cutler, 2003; Kraljic & Samuel, 2007; Bradlow & Bent, 2008; Kamide, 2012; Kleinschmidt & Jaeger, 2015; Fine & Jaeger, 2016; Roettger & Franke, submitted) and that in interaction, they learn how speakers choose among alternative utterances. In the domain of quantifiers, Yildirim et al. (2016) showed that listeners update their expectations about how a specific speaker uses the quantifiers *some* or *many* after being briefly exposed to a specific speaker. In line with their results, Schuster and Degen (2018) found that listeners update their expectations of how a specific speaker uses the uncertainty expressions *might* and *probably* to describe different event probabilities after a brief exposure phase. Participants who were exposed to a “*confident*” speaker, who used *probably* to describe the 60% probability event, expected the use of *probably* with a wider range of probabilities; participants who were exposed to a “*cautious*” speaker, who used *might* to describe the 60% probability event, expected the use of *might* with a wider range of probabilities.

The processes that lead listeners to update their expectations during semantic adaptation are poorly understood. In particular, it remains a largely open question to what extent listeners form speaker-specific expectations when interacting with multiple speakers. Some evidence for speaker-specific adaptation comes from the referring expressions literature. Metzing and Brennan (2003) found that participants exhibited a slowdown in resolving referring expressions when a confederate started referring to an object with a new expression after establishing a conceptual pact, but did not find such a slowdown when a new confederate was using a different referring expression than the original confederate.

Most closely related to our work, Yildirim et al. (2016) found that listeners form speaker-specific production expectations after being exposed to two speakers who used different quantifiers to describe a scene with a candy bowl in which half of the candies were green. While this suggests that listeners should also form speaker-specific expectations about the use of uncertainty expressions, there is evidence from other linguistic domains that speaker-specific adaptation is limited to specific items. For example, Kraljic and Samuel (2007) found that listeners adjust their phonemic representations for the fricatives /s/ and /sh/ to multiple speakers whereas listeners adjusted their phonetic representations for stop consonants such as /d/ and /t/ only to the most recent conversational partner. It could therefore be that speaker-specific adaptation in other linguistic domains is also limited to specific items and that listeners do not form speaker-specific expectations for the use of uncertainty expressions.

Further, Yildirim et al. (2016) observed that the adaptation effect was considerably smaller when they exposed participants to two speakers with opposing biases as compared to only exposing participants to one speaker and comparing the adaptation effect between groups. There seem to be two likely explanations for this observation. First, it could be that due to memory limitations, listeners were unable to keep track of the exact statistics of each speaker's utterances. Since everything about the context except the speaker identity stayed constant throughout the experiment, it could be that listeners had difficulty separating their experiences with the two speakers in memory (see Horton and Gerrig (2005) for a similar account of memory limitations affecting audience design). Second, it could be that listeners were tracking the statistics of the individual speakers as well as the overall statistics in the experimental situation and their post-exposure expectations were a combination of their speaker-specific expectations as well as their expectations about the situation.

In this work, we build on the recent work by Schuster and Degen (2018) on adaptation to variable use of uncertainty expressions and take a first step towards investigating the nature of semantic adaptation in response to multiple speakers. In particular, we aim to answer the following two questions:

1. Do listeners form speaker-specific production expectations when they are exposed to speakers whose use of uncertainty expressions differ?
2. Do listeners form situation-specific production expectations independent of speaker identity?

In Experiment 1, we address question 1 by exposing listeners to two speakers whose use of uncertainty expressions differs. In Experiment 2, we expose listeners to two speakers whose use of uncertainty expressions is the same. We compare adaptation effect sizes across experiments to address question 2.

Experimental paradigm

In both of our experiments, we build upon the semantic adaptation paradigm used in Schuster and Degen (2018), which we briefly review here. This paradigm is a classic exposure-and-test paradigm which has been used to study adaptation across several linguistic domains (e.g., Norris et al., 2003; Kleinschmidt & Jaeger, 2015; Yildirim et al., 2016). As shown in Figure 1, each trial shows an adult sitting behind a table with a gumball machine on it. The gumball machine is filled with orange and blue gumballs. Next to the table, there is a child who is requesting a blue or an orange gumball with the utterance "I want a blue/an orange one". Participants are told that the gumball machine is too high up for the child to

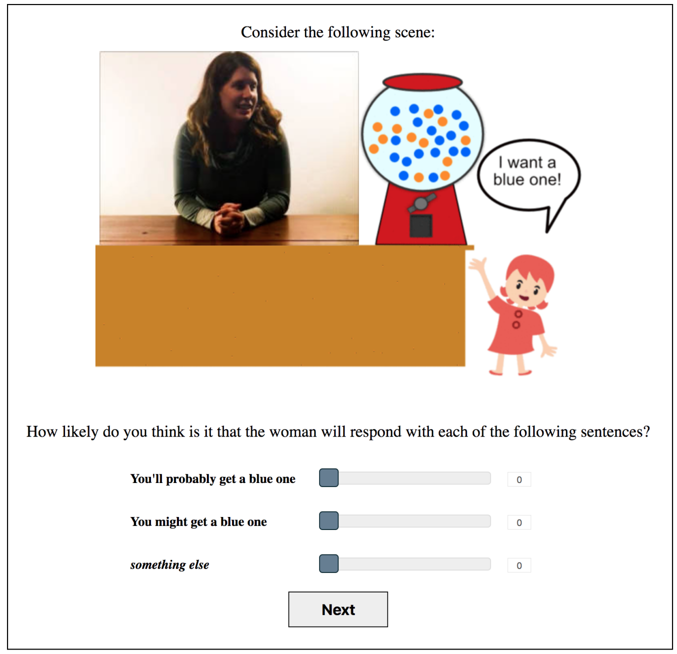

Figure 1: Example post-exposure test trial. On exposure trials the rating scales were absent, and the image of a speaker was replaced by a video of a speaker producing an utterance.

see and that only the adult can see the contents of the gumball machine.

On each exposure trial, participants watch a short video clip in which the adult responds to the child with an utterance like "You might get a blue one". Across trials, the proportion of gumballs as well as the response by the adult vary.

On each test trial (Fig. 1), participants are shown a static scene in which they only see a picture of the speaker from the exposure trials. On these trials, participants are asked to provide ratings of how likely they think it is that the speaker would use the two provided utterances or some other utterance. Across trials, the proportion of blue and orange gumballs as well as the color of the gumball that the child is requesting (the target color) varies.

Experiment 1: Different speaker types

In Experiment 1, we exposed participants to two different speakers who use the uncertainty expressions *might* and *probably* differently. The primary purpose of this experiment was to test whether listeners form speaker-specific utterance choice expectations. Procedure, materials, analyses and exclusions were pre-registered on OSF (https://osf.io/qnspg).

Methods

Participants We recruited 104 participants on Amazon Mechanical Turk. Participants had to have a US-based IP address and a minimal approval rating of 95%, and they were paid $4.75 (approximately $12–$15/hr).

	MIGHT		PROBABLY		BARE	
	n	*p*	*n*	*p*	*n*	*p*
cautious	**10**	**60%**	5	90%	5	100%
confident	5	25%	**10**	**60%**	5	100%

Table 1: Number of exposure trials (*n*) per utterance (MIGHT, PROBABLY, BARE) and associated proportion of target gumballs (*p*) in the cautious vs. confident speaker block. Critical trials bolded.

Materials and procedure In the first part of the experiment, participants saw 40 exposure trials in two blocks. As mentioned above, each trial showed a child requesting a blue or orange gumball, a gumball machine with blue and orange gumballs, and a video of an adult male or female speaker. The speaker always produced one of the following six utterances:

- You'll get a blue/orange one (BARE)
- You might get a blue/orange one (MIGHT)
- You'll probably get a blue/orange one (PROBABLY)

The number of trials with each of these utterances as well as the gumball proportions varied across the two blocks (see Table 1 for an overview). Filler trials with the bare form were included to provide evidence that the speaker is generally cooperative. One of the blocks always showed a female speaker and the other block always showed a male speaker. Both speakers were from the East Coast and native speakers of American English. The order of blocks and the speaker assignment to blocks was counterbalanced across participants.

Participants were instructed to watch what the speaker had to say to the child. The video started automatically after a 400ms delay and participants had the option to replay the video as often as they wanted. To advance, participants had to press a button which was disabled until the video had ended.

After the two exposure blocks, participants went through two test blocks. In each of the blocks they saw a picture of one of the two speakers with a gumball machine next to it, and again, a child requesting a blue or an orange gumball. On each trial, participants were asked how likely they thought it was that the adult would respond with MIGHT, PROBABLY or a blanket *something else* option. Participants indicated their expectations by distributing 100 points across these three options using sliders. In each block, participants provided ratings for scenes with 9 different gumball machines ranging from 0% to 100% blue gumballs. For each machine, participants provided four ratings in total, resulting in 36 trials per block. The order of blocks was counterbalanced such that half of the participants saw them in the same order as the exposure blocks whereas the other half saw them in opposite order.

Attention checks To verify that participants were paying attention to the video and the scenes, we included 14 attention checks: after 14 of the exposure trials, participants were shown two different gumball machines and were asked to choose the one they saw on the previous trial.

Exclusions We excluded participants who provided correct responses to fewer than 11 attention checks. Based on this criterion, we excluded 31 participants. We further excluded participants whose utterance ratings for the different event probabilities strongly correlated ($R^2 > 0.75$) with their mean utterance ratings across all event probabilities. This suggests that they provided approximately the same ratings independent of the observed scenes and indicates that they did not pay attention. This led to one additional exclusion. None of the results discussed below depend on these exclusions.

Analysis and predictions Intuitively, a more confident speaker uses PROBABLY for a larger and MIGHT for a smaller range of gumball proportions than a more cautious speaker. Therefore, if participants track these different uses, we expect their ratings of what they think a specific speaker is likely to say to depend on how that speaker used uncertainty expressions during the exposure phase. Following Yildirim et al. (2016) and Schuster and Degen (2018), we quantify this prediction by fitting a spline with four knots for each expression and each participant and computing the area under the curve (AUC) for the splines corresponding to each expression, block and participant. The area under the curve is proportional to how highly and for how large of a range of gumball proportions participants rate an utterance, so if an utterance is rated highly for a larger range of gumball proportions, the AUC will also be larger. We therefore test whether listeners update their expectations by computing the difference between the AUC of the spline for MIGHT and of the spline for PROBABLY for each test block for each participant.

Based on the results of the adaptation experiment with multiple speakers by Yildirim et al. (2016), we expect speaker-specific adaptation effects. We thus predict that the mean AUC difference will be bigger for the *cautious* speaker test blocks than for the *confident* speaker test blocks.

As a secondary analysis, we also investigate whether the order of exposure blocks (*confident* or *cautious* first), the assignment of speaker to speaker type (whether the male speaker was the *cautious* speaker or vice versa), or the order of the test blocks (same as exposure or reverse) has an effect on adaptation. We do not expect any of these factors to have an effect on adaptation.

Results and discussion

Figure 2 shows the mean utterance ratings of participants grouped by the two post-exposure test blocks. As this plot shows, participants expected the *confident* speaker to be more likely to use *probably* for lower event probabilities than the *cautious* speaker. This is also reflected in the AUC differences between the splines for MIGHT and of the splines for PROBABLY: As predicted, this difference was greater for the *cautious* speaker ratings than for the *confident* speaker ratings ($t(142) = 2.92, p < 0.01$).

For our secondary analysis, we fit a linear regression model to predict the AUC difference with speaker type, exposure

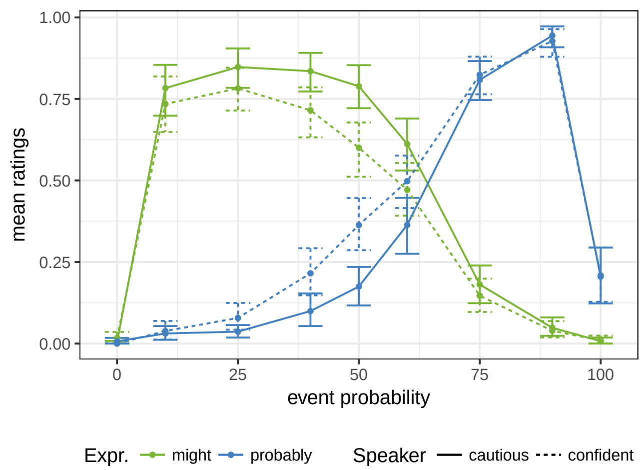

Figure 2: Mean utterance ratings for scenes with different event probabilities in Experiment 1. Error bars indicate bootstrapped 95% confidence intervals.

block order, speaker assignment, and test block order as predictors. Only speaker type is a significant predictor in this model (exposure block order: $\beta = 5.72$, $t(139) = 1.30$, *n.s.*; speaker assignment: $\beta = 1.21$, $t(139) = 0.28$, *n.s.*; test block order: $\beta = 2.28$, $t(139) = 0.52$, *n.s.*). Further, a model that includes these four predictors does not explain significantly more variance than a model that only includes speaker type as a predictor ($F(3,139) = 0.67$, *n.s.*).

The results of this experiment suggest that listeners form speaker-specific expectations of how different speakers use uncertainty expressions after brief exposure. At the same time, the results provide concrete evidence against two other accounts. First, they provide evidence against an account according to which participants only adapt to the experimental situation: If participants had only updated their expectations of what a generic speaker would say in the scenes presented in the experiment, we would not have expected to see differences in ratings between speakers. Second, they also provide evidence against a pure priming account according to which listeners update their expectations to the most recent exposure. Note that the adaptation effect was independent of the order of presentation and the order of test blocks. If participants had been primed by the most recent exposure speaker, we would have expected that participants' post-exposure ratings were primarily influenced by the behavior of the second exposure speaker.

The results of this experiment also replicate the finding by Yildirim et al. (2016) of differing effect sizes between the single-speaker and two-speaker experiments: The adaptation effect was considerably smaller in this two-speaker experiment (Cohen's d: 0.486) than in the single-speaker adaptation experiment by Schuster and Degen (2018) (Cohen's d: 1.263).

As suggested by a reviewer, one reason for the smaller effect size in the two-speaker experiment could be some form of self-priming and that participants' responses in the first test block influenced their responses in the second block. We evaluated this hypothesis in a post-hoc analysis of the responses from the first test block. We compared the responses of participants who were first tested on the *cautious* speaker to the responses of participants who were first tested on the *confident* speaker. If responses in the first test block influenced responses in the second test block, we would expect a larger effect size if we only consider the data from the first block. We did indeed find a larger effect size in the first block (Cohen's d: 0.723), which suggests that participants exhibited some form of self-priming.

However, even if we only consider the first block of responses, the adaptation effect remains smaller in the two-speaker experiment (Cohen's d: 0.723) than in the one-speaker experiment (Cohen's d: 1.263). This could be either a result of memory limitations or a result of listeners jointly tracking the statistics of each speaker as well as of the overall experimental situation (situation-specific statistics). We further investigate these possibilities in the next experiment.

Experiment 2: Identical speaker types

In Exp. 1, we found that the adaptation effect was smaller than it was in the single-speaker version of the experiment, which could have either been a result of memory limitations or joint speaker-specific and situation-specific adaptation. In this experiment, we investigate whether there is evidence for one of these two accounts. We exposed listeners to two speakers of the same type.[1] If the smaller effect in Exp. 1 was caused by listeners' inability to separate their experiences with the two speakers in memory, i.e, some experiences might have been attributed to the incorrect speaker, we would expect the adaptation effect in this experiment to be on average the same as in the one-speaker experiment. This is because even if listeners cannot perfectly separate their experiences with each speaker, they would on average still have the same number of experiences with each of the two speakers as listeners had with the one speaker in the single-speaker experiment. If, on the other hand, the smaller effect in the previous experiment was a result of listeners jointly tracking speaker-specific and situation-specific statistics, we would expect the adaptation effect to be larger here than in the single-speaker experiment. This is based on the assumption that more exposures lead to a larger adaptation effect and thus listeners' should adapt more to the situation if they are exposed to two

[1] In the spirit of open science, we note that the data from this experiment comes from a faulty version of Experiment 1. A scripting error led to participants always being exposed to the same speaker type instead of two different speaker types. Because of this error, the pre-registered analysis (https://osf.io/3cw79) deviates from the analysis that we report here. The reported analyses here are the only additional analyses we performed on the data. The reason for not discarding the data from this experiment but rather including it here is that it provides an informative data point for the question of whether listeners track situation-specific expectations.

speakers and hence also twice the number of interactions.

Methods

Participants We recruited 104 participants on Amazon Mechanical Turk. Participants had to have a US-based IP address and a minimal approval rating of 95%, and they were paid $5 (approximately $12–$15/hr).

Materials and procedure The materials and procedures were the same as in Exp. 1 except for the following two modifications. First, the speaker types for each participant were identical across the two exposure blocks: both speakers were either *confident* or *cautious* speakers. Second, the number of trials with PROBABLY and the number of trials with MIGHT were the same (10 trials per utterance and block) whereas in Experiment 1, the *confident* speaker produced only 5 instances of MIGHT and the *cautious* speaker produced only 5 instances of PROBABLY.[2] Assignment of speaker types was counterbalanced across participants, which means this experiment had a between-subjects manipulation.

As in Experiment 1, we excluded participants who provided correct responses to less than 11 of the attention checks as well as participants who seemed to provide random responses as defined above. In total, we excluded 11 participants because of the attention check criterion and 1 more participant because of random responses.

Analysis and predictions As the primary analysis, we compare the AUC differences between the splines for MIGHT and of the splines for PROBABLY between participants in the two conditions. Analogous to Experiment 1, we predict that the mean AUC difference will be bigger in the *cautious* speaker condition than in the *confident* speaker condition.

We again also investigate whether the assignment of speaker to speaker type or the order of the test blocks have an effect on the AUC difference. We do not expect either of these factors to affect adaptation.

Lastly, we compute the effect size measured by Cohen's d. As explained above, we expect the effect size either to be the same as in the single-speaker experiment or to be larger.

Results and discussion

Figure 3 shows the mean utterance ratings of participants for the two conditions. We again observe listener adaptation, resulting in a greater AUC difference in the *cautious* speaker condition than in the *confident* speaker condition ($t(89) = 8.01, p < 0.001$). Further, no factors other than speaker type are significant predictors of the AUC difference (speaker assignment: $\beta = -1.32$, $t(87) = -0.398$, *n.s.*; test block order: $\beta = 4.28$, $t(139) = 1.30$, *n.s.*).

Lastly, the effect size (Cohen's d: 1.68) was larger in this experiment than in Experiment 1 and the single-speaker experiment by Schuster and Degen (2018). While it would be premature to definitively conclude from these three experiments that listeners' expectations are jointly influenced by individual speaker's productions as well as all the productions in the experiment, our results point in this direction.

[2]The reason for the second modification is the above mentioned scripting error. See below for a discussion of potential implications.

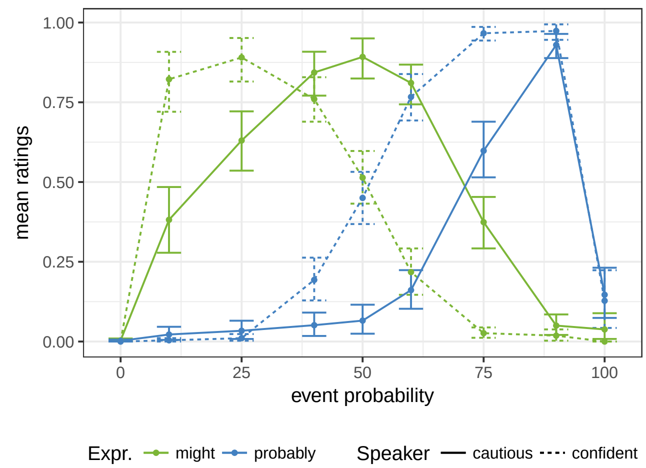

Figure 3: Mean utterance ratings for scenes with different event probabilities in Experiment 2. Error bars indicate bootstrapped 95% confidence intervals.

There is a potential confound in this experiment because participants saw 5 additional filler trials during each exposure block which could have led to the larger effect size as compared to the single-speaker experiment. However, this explanation seems unlikely considering previous work.[3]

General discussion and conclusion

In two experiments, we found that listeners form speaker-specific production expectations of uncertainty expressions after brief exposure to two speakers. This shows that the results by Yildirim et al. (2016) also extend to lexical items other than quantifiers.

At the same time, however, we found that the adaptation effect size varied depending on whether the two speakers had the same or divergent bias during the exposure phase. When listeners were exposed to two different speaker types, the adaptation effect was smaller and their expectations seemed to have been shaped by their experiences with the two speakers as well as all the experiences encountered in the experiment. When both speakers behaved the same, on the other hand, the adaptation effect was much more pronounced and even greater than in the single-speaker experiment from previous work.

[3]Yildirim et al. (2016) used a very similar paradigm to study semantic adaptation to the use of the quantifiers *some* and *many*. Analagous to our *confident* and *cautious* speakers, they had a *some-biased* and a *many-biased* speaker. They report two versions of their experiment: one in which there were no filler trials with the other quantifier and another version in which there was a balanced number of exposure trials with both quantifiers in both conditions. They found that the adaptation effect was smaller when there were more filler trials, so we would expect that if the additional fillers affected the size of the adaptation effect, the effect would be even larger had we not presented the extra fillers to participants.

One likely explanation for these observations is that apart from tracking speaker-specific statistics, listeners also track the situation-specific statistics of all interactions in the experiment and their expectations are guided by both of these factors. In the case of speakers with different uses of uncertainty expressions, speaker-specific adaptation is attenuated since the overall statistics guide listeners towards an "average" speaker whose use falls somewhere in between the *cautious* and the *confident* speaker. When listeners are exposed to two speakers of the same type, on the other hand, the situation-specific statistics reinforce the speaker-specific statistics and hence listeners adapt more to the two speakers.

An account based on "faulty" memory, according to which listeners have trouble keeping the speaker-specific experiences separate, does not predict the larger adaptation effect when listeners are exposed to two speakers of the same type. If every experience is encoded as an episode in memory but some with the incorrect speaker information, on average, the number of experiences with each speaker should still be the same as in the one-speaker condition and therefore it is unclear why listeners adapt more in the two-speaker experiment than in the one-speaker experiment.

Our findings also have implications for current models of semantic adaptation. Following the recent successes in modeling phonetic adaptation as an instance of Bayesian belief updating (Kleinschmidt & Jaeger, 2015), Schuster and Degen (2018) propose a computational model of semantic adaptation. According to this model, when interacting with a speaker *Sp*, listeners update their beliefs about a set of speaker-specific parameters Θ_{Sp}, which govern the speaker's lexicon and preferences.[4] Their model predicted the results of the single-speaker experiment well, but without modifications, it does not predict the differences in effect size.

We consider two promising extensions of this model. First, the model could be cast as a hierarchical model. Hierarchical models have been argued to explain many cognitive and perceptual phenomena (see e.g., Clark, 2013, for a review), including phonetic adaptation (Kleinschmidt, 2019), and also seem applicable here. In a hierarchical version of the adaptation model, we would assume that the speaker-specific parameters Θ_{Sp} are not only shaped by the listener's prior beliefs and the observed interactions by a speaker *Sp* but rather also depend on a distribution reflecting the situation-specific expectations. Figure 4 shows a sketch of a potential hierarchical model. Such a model would explain the differences in effect size: When listeners are exposed to different speaker types, the situation-specific parameter distribution would be influenced by two speaker types that essentially cancel each other out, which in turn would lead to less extreme speaker-specific distributions. On the other hand, when both of the speakers are of the same type, the situation-specific parameter distribution would be more strongly shifted towards the observed distributions which in turn would lead to more extreme speaker-specific distributions.

[4] See also Hawkins, Frank, and Goodman (2017) for a similar model of the formation of conceptual pacts.

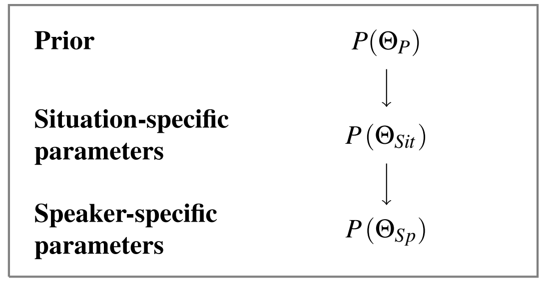

Figure 4: Hierarchical model of semantic adaptation. Situation-specific parameters $P(\Theta_{Sit})$ depend on prior beliefs $P(\Theta_P)$ and speaker-specific parameters $P(\Theta_{Sp})$ depend on the situation-specific parameters.

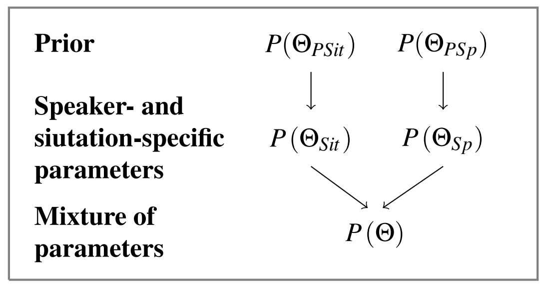

Figure 5: Mixture model of semantic adaptation. Overall production parameters $P(\Theta)$ are a weighted combination of situation-specific parameters $P(\Theta_{Sit})$ and speaker-specific parameters $P(\Theta_{Sp})$.

A second possibility would be to cast the model as a mixture model in which overall production parameters are a weighted combination of situation-specific and speaker-specific parameters (and potentially other factors). Figure 5 shows a sketch of a potential mixture model. According to such a model, listeners would form both situation-specific and speaker-specific expectations as a result of adaptation and then combine these expectations to their overall expectations. Such a model would also predict the smaller effect size in Experiment 1 since it would predict that the overall production expectations are influenced by the speaker-specific statistics as well as the situation-specific statistics and the latter drive the production expectations to be more similar to an "average" speaker. When listeners are exposed to two identical speakers, on the other hand, the situation-specific expectations (which are in line with the speaker type of both exposure speakers) would reinforce the speaker-specific expectations and therefore lead to a larger adaptation effect. Future experimental work should adjudicate between the hierarchical and the mixture model account.

In conclusion, we presented new experimental results from the domain of uncertainty expressions which suggest that speaker-specific semantic adaptation is a product of forming speaker-specific expectations and forming expectations about the situation independent of the speaker. These results raise a number of interesting questions, most pressingly regarding transfer effects to novel speakers, which have been observed in other linguistic domains (e.g., Bradlow & Bent, 2008; Xie, Earle, & Myers, 2018). In our experiments, the exposure and test speakers did not differ. This raises the question about whether and to what extent updated expectations transfer to novel speakers whose similarity to the exposure speaker(s) varies. Both models sketched above lend themselves well to capturing such transfer effects. In addition, participants saw very similar visual scenes on each trial. Another potential direction would be to study the extent of speaker-specific adaptation when listeners encounter more novel scenes during the test phase to investigate to what extent listeners form speaker-specific expectations independent of other contextual factors. Answering these questions will help disentangle the different adaptation processes and give us a better understanding of how listeners infer meanings in context.

References

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Fine, A. B., & Jaeger, T. F. (2016). The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning Memory and Cognition*, *42*(9), 1362–1376.

Finegan, E., & Biber, D. (2001). Register variation and social dialect variation: The register axiom. In *Style and Sociolinguistic Variation* (pp. 235–267).

Hawkins, R. X. D., Frank, M. C., & Goodman, N. D. (2017). Convention-formation in iterated reference games. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society (CogSci 2017)* (pp. 482–487).

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127–142.

Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition*, *124*(1), 66–71.

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience*, *34*(1), 43–68.

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), 148–203.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

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.

Roettger, T. B., & Franke, M. (submitted). Evidential strength of intonational cues and rational adaptation to (un-)reliable intonation. *Cognitive Science*.

Schuster, S., & Degen, J. (2018). Adaptation to variable use of uncertainty expressions. In *Proceedings of Architectures and Mechanisms for Language Processing 2018 (AMLaP 2018).*

Weiner, E. J., & Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics*, *19*(1), 29–58.

Xie, X., Earle, F. S., & Myers, E. B. (2018). Sleep facilitates generalisation of accent adaptation to a new talker. *Language, Cognition and Neuroscience*, *33*(2), 196-210.

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.