Linguistic Variability and Adaptation in Quantifier Meanings

Ilker Yildirim, Judith Degen, Michael K. Tanenhaus, T. Florian Jaeger
Department of Brain & Cognitive Sciences, University of Rochester, Rochester, NY 14627
{iyildirim, jdegen, mtan, fjaeger}@bcs.rochester.edu

Abstract

People's representations of most and arguably all linguistic and non-linguistic categories are probabilistic. However, in linguistic theory, quantifier meanings have traditionally been defined set-theoretically in terms of categorical evaluation functions. In 4 "adaptation" experiments, we provide evidence for the alternative hypothesis that quantifiers are represented as probability distributions over scales (e.g., Zadeh, 1965). We manipulate exposure to different distributions of "some" and "many" and find that listeners adapt to those distributions, as predicted. Our results suggest that the interpretation of quantifiers is best modeled as a process involving rich, probabilistic representations.

Keywords: Quantifiers; Semantics; Language processing; Adaptation; Generalization

Introduction

In linguistic theory, quantifier meanings have traditionally been defined set-theoretically in terms of categorical evaluation functions (Barwise & Cooper, 1981) yielding either truth or falsity of a sentence containing a quantifier. Quantifiers are understood as relations between sets:

(1) some(A, B) is true iff $||A|| \cap ||B|| \neq \varnothing$

(2) many(A, B) is true iff $||A|| \cap ||B|| > n$, where n is some large number

For example, the sentence *Some candies are green* is true just in case the intersection of the candies and the green things is not empty. Similarly, *Many candies are green* is true just in case the cardinality of the intersection of the candies and the green things is larger than some contextual norm n. This points to a notable feature of some quantifiers: they exhibit both vagueness and context-dependence (Solt, 2009).

A class of alternative views tries to incorporate this feature by representing quantifiers probabilistically. For example, fuzzy logic (Zadeh, 1965) approaches to meaning consider quantifiers such as "some" as probability distributions over scales (e.g., Moxey & Sanford, 1993). Probabilistic quantifier semantics are at the heart of recent models of both syllogistic reasoning (Chater & Oaksford, 1999) and scalar implicature (Goodman & Stuhlmüller, 2013). Here we provide further evidence that quantifiers are indeed interpreted in a probabilistic, graded manner. The novel empirical contribution lies in addressing the adaptability of these distributions to variable language environments.

The probabilistic view on quantifier meaning is illustrated in Figure 1a: "some" and "many" form graded distributions over a contextually determined scale. [1] Previous work has implicitly assumed that these distributions are invariant across linguistic environments, in that the distribution corresponding to, for example, "some" is stationary across different dialects, speakers, genres, and so on.

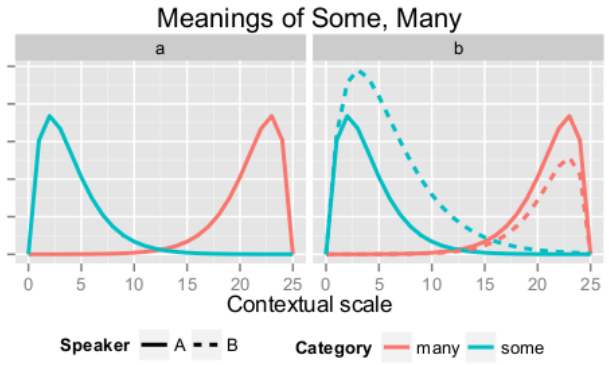

Figure 1: Illustration of across speakers variability in meanings of quantifiers.

However, variability in language use is the norm. Speakers differ in their realization of phonemes (cf. Allen, Miller, & DeSteno, 2003), lexical preferences (e.g., couch vs. sofa), as well as syntactic preferences (e.g., some speakers use passives more often than others, Weiner & Labov, 1983). Such linguistic variability is a challenge for comprehenders that must be overcome to achieve successful communication. One solution for dealing with variable linguistic environments is to track and adapt to the joint statistics of linguistic categories (e.g. phonemes, words, syntactic structures) and contextual cues, including the speaker.

A powerful way to test whether listeners adapt to the statistics of the input is to determine whether categorization functions shift with exposure. If listeners adapt to new environments in which the statistics diverge from their prior beliefs, this would suggest that linguistic representations are sensitive to and adapt to such sources of variability. This reasoning has been successfully applied to phonetic categories (e.g., Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Vroomen, Linden, Gelder, & Bertelson, 2007; Kraljic & Samuel, 2006), prosodic categories (Kurumada, Brown, & Tanenhaus, 2012), and syntactic categories (Fine, Jaeger, Farmer, & Qian, under-review; Kamide, 2012).

Here we ask whether listeners' representations of the quantifiers "some" and "many" are probabilistic and sensitive to environmental variability. Figure 1b depicts hypothetical *some* and *many* distributions over cardinalities for two speakers whose use of the quantifiers differs.

In four adaptation experiments, we provide evidence that quantifiers are represented as probability distributions. More-

[1] For example, it is not as plausible to quantify 18 out of 1000 as "many" as to quantify 18 out of 20.

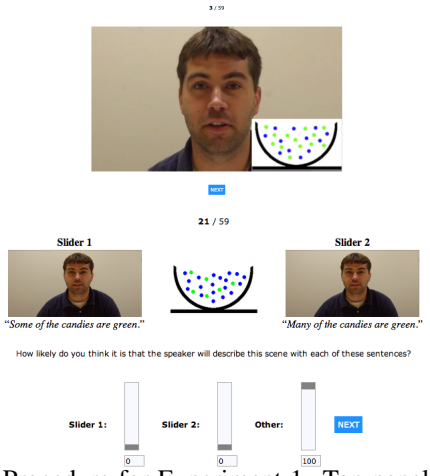

Figure 2: Procedure for Experiment 1. Top panel illustrates an exposure phase trial. Bottom panel illustrates a test phase trial.

over, we present evidence that listeners' interpretations of quantifiers rapidly adapt to the statistics of the local linguistic environment represented by a novel speaker. Furthermore, we provide evidence suggesting that listeners' adaptation might be taking place across multiple levels (or types) of representations. We argue that the rapid adaptation that we observe involves both speaker-specific and quantity level representations enabling transfer of adaptation across visual object types.

Experiment 1

Behavioral evidence strongly suggests that listeners dynamically adapt to the phonetic and syntactic variability in their language environment (Vroomen et al., 2007; Bertelson, Vroomen, & Gelder, 2003; Kamide, 2012; Fine et al., underreview). But do such adaptive processes also occur at the level of *meaning*? We addressed this question in an experiment by investigating whether listeners adapt their interpretations of the two English quantifiers "some" and "many" based on experience with a speaker who uses these quantifiers in a way that deviates from the listener's prior expectations.

Our experimental logic followed that of previous adaptation experiments (e.g., Bertelson et al., 2003). The experiment employed a by-2 between-participant design. One group of participants was exposed to a novel speaker's use of the word "some" (*some-biased group*). Another group of participants was exposed to a novel speaker's use of "many" (*many-biased group*). Participants in both groups were then tested on how they interpreted that speaker's utterances.

Participants

80 participants were recruited over Amazon's crowd-sourcing service Mechanical Turk. All participants were self-reported native speakers of English. Each experimental session took about 15 minutes and participants were paid $2.

Procedure and Materials

Figure 2 illustrates the materials and procedure for this experiment. The experiment proceeded in two phases, the *exposure phase* and the *test phase*.

In the exposure phase, participants watched videos as in the top panel of Figure 2. The video showed a bowl of 25 candies in the bottom right of the screen. The bowl always contained a mixture of green and blue candies, but the number and spatial configuration of the candies differed between trials. Importantly, the video showed a speaker describing the scene in a single sentence. The videos played automatically at the start of the trial and the scene — the candy bowl — remained visible even when the video had finished playing (as shown in Figure 2, top). Two different speakers were employed between participants to ensure that effects were not due to a particular speaker.

The exposure phase consisted of 10 critical and 10 filler trials. In critical trials the speaker produced the sentence *Some of the candies are green* (some-biased group) or *Many of the candies are green* (many-biased group). On a critical trial, the bowl always contained 13 green candies and 12 blue candies. This scene was identified as the Most Ambiguous Quantity (MAQ) scene in a preceding norming study in which participants rated how well descriptions containing different quantifiers matched scenes sampled from a continuum of quantities.

The remaining 10 trials in the exposure phase were filler trials. On a filler trial, participants observed the speaker correctly describing a scene with no green candies in it as *None of the candies are green* (5 trials) and a scene with no blue candies in it as *All of the candies are green* (5 trials). The purpose of the filler trials was two-fold. First, it made our manipulation less obvious. Second, including clearly true descriptions of unambiguous scenes encouraged participants to believe that the speaker was indeed intending to accurately describe the scene. The order of the critical and the filler trials was randomized.

Following the exposure phase, participants entered the test phase. The test phase was intended to assess participants' beliefs about the speaker's use of both "some" and "many". On test trials, participants saw a candy scene in the center of the display and two identical still images of the speaker from the exposure phase on either side of the scene (see Figure 2, bottom).

The two images of the speaker were paired with one of the two alternative descriptions *Some of the candies are green* and *Many of the candies are green* each. The participants' were asked to rate how likely they thought the speaker would be to describe the scene using each of the alternative descriptions. They performed this task by distributing a total of 100 points across the two alternatives (the first and the second slider bars; see Figure 2, bottom panel) and a third alternative – namely "Other" – to reflect how much they thought that neither of the two alternatives fit the scene (the third slider bar). As in the exposure phase, scenes always consisted of a bowl of 25 candies with differing numbers of green candies. To assess participants' beliefs about the speaker's use of "some" and "many", we sampled scenes from the entire scale. Specifically, scenes contained one of

$\{1,3,6,9,11,12,13,14,15,17,20,23\}$ green candies out of 25 candies. Over 39 test trials, participants rated each scene 3 times. Different instances of the same scenes differed in the spatial configuration of the blue and green candies. The order of the scenes and the mapping from alternative descriptions to slider bars were randomized and counterbalanced.

To ensure that participants were attending to the task, we placed catch trials after about every six trials. On some of these trials, a gray cross appeared at a random location in the scene. Before the next trial began, participants were asked if they had seen a gray cross in the previous scene.

Data Analysis

We did not analyze the "Other" responses. The top row in Figure 3a shows the distribution of "some" and "many" in the test phase separately for the two groups of participants. The distributions were obtained by averaging participants ratings for the different scenes along the scale. We first averaged across the three instances of each scene within a speaker and then averaged those ratings across speakers (separately for each point on the scale). Those average ratings were then fit with a generalized linear model with cubic splines, which gave us the continuous curve for each of the two alternative descriptions shown in Figure 3a, top row. Participants in the *some-biased* group adapted in the opposite (and predicted) direction from participants in the *many-biased* group. That is, the distributions for participants in the *some-biased* group were updated such that they were more likely to rate a wider range of scenes as more likely with respect to the "Some" description. Such high ratings of the "Some" description came at the expense of the alternative description. Similarly, the distributions for participants in the *many-biased* group reflected that these participants were more likely to rate the "Many" description as more likely at the expense of the alternative description.

In order to quantify the shift in interpretations between the two groups of participants, we derived two measures. First, for each participant, we estimated the MAQ as the point where the two curves were closest to each other (excluding the extremes of the scene continuum).

Similar in logic to the phonetic adaptation experiments, we reasoned that participants in the *many-biased* group would come to interpret a "many" as applying to a wider range of scenes (and hence quantities). Because participants had to share a total of 100 points between the alternatives, this adaptation in favor of "many" would be at the expense of "some" ratings. Therefore, the MAQ scene should shift to the lower end of the continuum of set sizes compared to 13 (the MAQ scene from the norming study). In contrast, for participants in the *some-biased* group, if they were to adapt to the statistics of the speaker during the exposure phase, they should rate a wider range of scenes more likely to be described using the quantifier "some." These high ratings for "some" would come at the expense of "many." Therefore, the MAQ should shift to the higher end of the continuum of set sizes compared to the MAQ scene from the norming experiment.

To ensure that our findings were not just an artifact of the way the analysis was conducted, we performed a separate set of analyses by computing the Area Under the Curve (AUC) for each of the two alternative descriptions. That is, again, we first fit a generalized linear model with cubic splines for each participant. Then we computed the AUC for each alternative description (by summing up the area under the fitted curve) and subtracted the AUC for the "Some" curve from the "Many" curve.

We reasoned that if participants adapted their quantifier interpretations in the predicted direction, then the AUC difference should be smaller (or negative) for participants in the *many-biased* group and larger (or positive) in the *some-biased* group.

All analyses were conducted using the R statistics software package (R Development Core Team, 2005).

Results

Middle row in Figure 3a presents the results for MAQ analysis. As predicted, for each speaker, the MAQ values were significantly smaller for the *many-biased* group than for the *some-biased* group ($p < 10^{-6}$).

Bottom row in Figure 3c shows re-evaluation of the same data using the AUC analysis. As predicted, for both speakers, the AUC difference for the *many-biased* group and the *some-biased* group grew in opposite directions ($p < 10^{-6}$).

These results suggest that listeners indeed track the joint statistics of quantities, speakers, and the quantifiers in their environment, and rapidly adapt their interpretations in response to the new input.

Experiment 2

One limitation of Experiment 1 is that effects might be speaker and/or scene specific. Experiments 2 and 3 were designed to test the hypothesis that the updating was more general. Experiment 2 examined adaptation when the emphasis is shifted away from the specific speaker by changing the instructions and by removing the speaker's face from the test phase trials. Experiment 3 used different objects in the test phase — Xs and Os instead of candies of different colors.

Participants

Participants were 80 Mechanical Turk workers. All participants were self-reported native speakers of English. Each experimental session took about 15 minutes, and participants were paid $1.5.

Procedure and Materials

The experimental stimuli were identical to those of Exp. 1.

The procedure was identical to that of Exp. 1 with the exception of the test trials. Unlike the previous experiment, participants did not see a cue to the speaker's identity. Instead, they saw only the two sentences providing the two alternative descriptions for the scene located at the center. The participants' task was to rate how likely that they thought that *a*

Exp	Pre-exposure	Exposure	Test (Post-exposure)	Groups
1	N/A	Candy scenes in videos	VS: Candies LS: Typed sentences + speaker images	Some-biased vs. Many-biased
2	N/A	Candy scenes in videos	VS: Candies LS: Typed sentences	Some-biased vs. Many-biased
3a	VS: Candies LS: Typed sentences	Candy scenes in videos	VS: Letters LS: Typed sentences	Some-biased vs. Many-biased
3b	VS: Candies LS: Typed sentences	Candy scenes in videos	VS: Candies LS: Typed sentences	Some-biased vs. Many-biased

Table 1: Summary of the experimental designs. VS: visual stimuli. LS: linguistic stimuli.

speaker would describe the scene with each of the alternative descriptions. They again distributed a total of 100 points across the two alternative descriptions and choice of "Other."

As in Exp. 1, 40 participants were assigned to each of the *some-biased* and *many-biased* groups. For each group, of the 40 participants, 20 were assigned to each of the speakers in the videos.

A summary of the procedures used in the different experiments is provided in Table 1.

Results

We excluded one of the participants from the analysis because they never adjusted the sliders on the test trials. Top row in Figure 3b plots the mean ratings by participants in each of the two groups. Participants adapted their interpretations of the quantifiers in accordance with the speaker-provided statistics, though less so than in Exp. 1.

We performed the same MAQ and AUC analysis as for Exp. 1. Middle row in Figure 3b illustrates that the MAQ for participants in the *many-biased* group was significantly smaller than the MAQ for participants in the *some-biased* group. This was true for both speakers ($p < 0.01$). The AUC analysis, bottom row in Figure 3b, also revealed significant adaptation ($p < 0.01$).

The results from Exp. 2 suggest that the adaptation observed in Exp. 1 is not a simple speaker-specific adaptation effect and suggest instead that listeners' adaptation to the statistics of the linguistic environment might occur at multiple levels of representations. Adaptation was stronger in Exp. 1 where a cue to the speaker was provided in the test phase. However, the fact that we also observe adaptation in Exp. 2 (when no such cue is available) suggests that this adaptation was to some extent generalized across speakers.

Experiment 3a

It is nevertheless possible that the adaptation effects found in Exps. 1 and 2 is object-specific, i.e. quantifier interpretations are only updated for quantities of *candies*. Exp. 3a tested this by replacing the candy scenes in the test phase trials with scenes containing letters (Xs and Os).

Participants

We recruited 40 participants over Mechanical Turk who were self-reported native speakers of English. Each experimental session took about 15 minutes. Participants were paid $1.5.

Materials and Procedure

The test stimuli differed from the previous experiments. On each test trial we presented 25 letters, each of which was either an X or an O. The letters in each scene were scattered within a circle (but there was no visible boundary). The descriptions that participants rated were *Some of the letters are Xs* and *Many of the letters are Xs*. Number of Xs in a scene could be any of the values that the number of green candies could be in a scene from Exps. 1 and 2. Participants' task was again to rate (by distributing 100 points) how likely that they thought a speaker would describe the scene with each of the alternative descriptions and the third choice of "Other."

The stimuli in the exposure phase were identical to Exp. 1 and 2 but speaker identity was not varied between participants. Half of the participants were assigned to the *some-biased* group and half to the *many-biased* group.

In order to establish that transfer occurred between the candy and the letter scenes, we included a pre-exposure test phase. The aim of these pre-exposure test trials was to measure participants' prior interpretations of quantifiers in candy scene descriptions and compare them to quantifiers in letter descriptions following exposure to candy scenes. That is, we analyzed participants' responses to descriptions of letter scenes in the post-exposure test trials and responses to descriptions of candy scenes in the pre-exposure test trials together to measure whether participants' interpretations changed with exposure.

Data Analysis

For each participant in the MAQ analysis, we determined the MAQ for the pre- and post-exposure test responses separately. Then we subtracted the pre-exposure MAQ from the post-exposure MAQ. A positive difference is expected for the *some-biased* group and a negative one for the *many-biased* group.

For the AUC analysis, we first calculated the AUC difference on pre-exposure test trials for each participant. Then we calculated the AUC difference on post-exposure test trials. The pre-exposure AUC difference was then subtracted from the post-exposure AUC difference. The expected patterns of results was the same as in the previous experiments.

Results

Top row in Figure 3c illustrates the group mean ratings for the post-exposure test trials. Participants' ratings in the *some-*

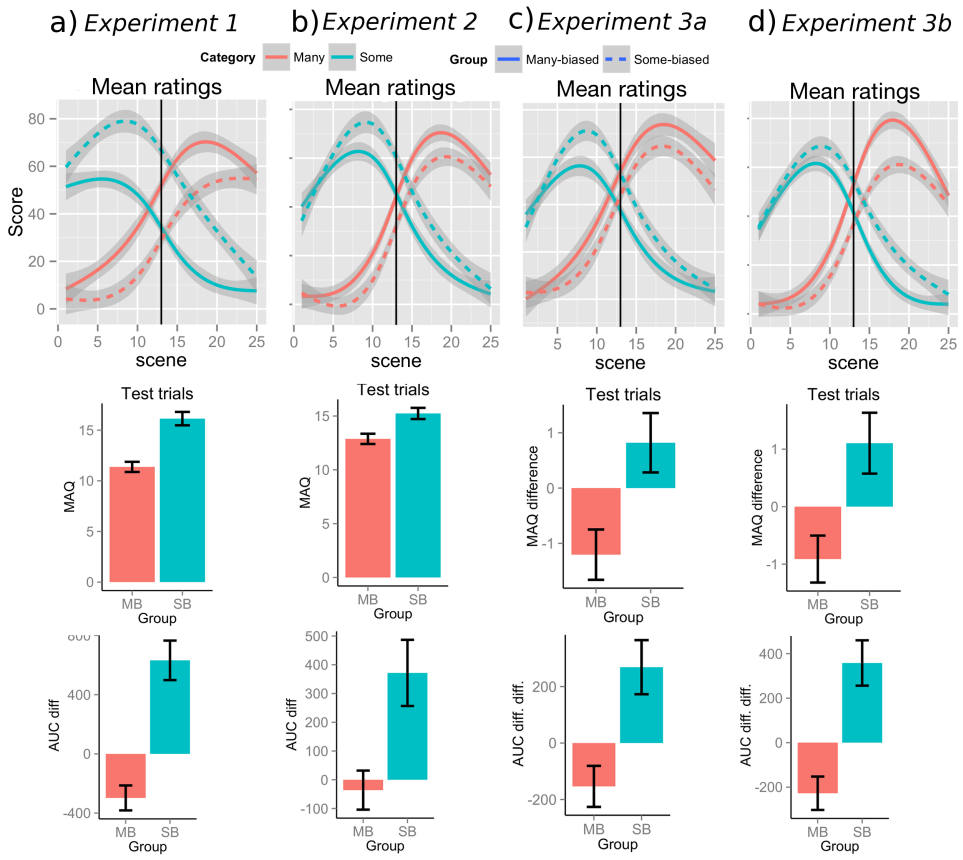

Figure 3: Each column shows data for experiment (e.g., left-most column is Experiment 1, right-most column is Experiment 3b). The vertical lines in the density panels at the top denote the MAQ scene (scene 13) determined based upon a preceding norming study. MB=Many-biased, SB=Some-biased.

biased group and the *many-biased* group did not differ before exposure. However, following the adaptation trials, participants' responses reflect that they adapted in the predicted directions:

The MAQ difference analysis in middle row in Figure 3c shows that indeed participants in the *some-biased* group rated "Some" descriptions as more likely across the whole continuum of scenes, whereas participants in the *many-biased* group favored "Many" descriptions at the expense of the alternative descriptions ($p < 0.01$). The difference in AUC difference analysis in Figure 3c, bottom row, reaffirmed our findings ($p < 0.01$).

The results from Exp. 3 suggest that participants' quantifier interpretations did not adapt candy-specifically - instead, quantifier adaptation transferred to a different visual environment. That is, the quantity level representation itself adapted.

Experiment 3b

To establish that the results we obtained in Exp. 3 were not due merely to the additional pre-exposure test trials, we re-ran Exp. 2 with pre-exposure test trials. The pre- and post-exposure test trials were identical and contained candy scenes.

We recruited 120 participants over Mechanical Turk who were self-reported native speakers of English. Each experimental session took about 15 minutes, and participants were paid $1.5.

60 participants were assigned to each of the the *some-biased* group and the *many-biased* group. 30 participants in each group were assigned to each of the speakers.

Top row in Figure 3d shows the mean post-exposure test trial responses (responses did not differ on pre-exposure test trials between groups). Following adaptation trials, there is a clear effect of group in the predicted direction, replicating the results from Exp. 2.

Middle row in Figure 3d shows the results of the MAQ difference analysis. The qualitative patterns of our results reflects the predicted pattern, such that the MAQ difference was positive in the *some-biased* group and negative in the *many-biased* group. This difference was significant ($p < 0.01$). In the difference in AUC difference analysis (Figure 3d, bottom row) the participants adapted to the speakers in the predicted directions ($p < 10^{-4}$).

We thus replicated the results from Exp. 2, again indicating that listeners' adaptation of quantifier meanings is broad. It also confirms that the inclusion of pre-exposure test trials

is most likely not the reason for the transfer effect found in Exp. 3.

Discussion

Our results indicate that semantic representations can be adapted to new linguistic environments. At least in situations like the ones investigated here, this adaptation seems to be rapid, requiring only very limited exposure. Our observation that adaptation can be transferred across multiple linguistic and visual environments suggest that these adaptations are not limited to the specific nature of the scale, although it remains to be seen how such adaptation generalizes to scales of different ranges. Our experiments support probabilistic theories of quantifier meaning over set-theoretic ones. Our results are also compatible with a soft version of set-theoretic representations under which there are core logical representations that are enriched with probabilistic expectations about the use of quantifiers with different set sizes.

In this paper, we addressed the question of whether and how listeners adapt to speakers' use of the quantifiers "some" and "many." A recently emerging literature in other domains of language processing has provided evidence that listeners can rapidly adapt to speaker-specific variability in their language environment. Most of this line of work has focused on adaptation to phonological variability across speakers (e.g., Kraljic & Samuel, 2006; Clayards et al., 2008; Vroomen et al., 2007). To our knowledge, our work is the first to extend the logic of language adaptation experiments to semantic representations.

Future experimental work should address whether listeners can adapt to multiple speakers' quantifier use statistics simultaneously. While the relative magnitude of the shift in interpretations of "some" and "many" between Experiments 1 and 2 might be taken to provide preliminary evidence that listeners maintain both speaker-specific and speaker-general representations and that both of these are affected by recent experience with a specific speaker, future work is required to address more directly the nature of representations that are adapted by recent exposure. For example, it is possible that listeners maintain hierarchically structured representations over speakers, groups of speakers (based on their similarity), and so on (cf. modeling of phonetic adaptation; Kleinschmidt & Jaeger, 2011). Future research will also need to address how much of the adaptation comes from base-rate effects (e.g., changes in the prior probabilities of quantifiers) and how much of it comes from adaptation of the meaning of each quantifier (e.g., changes in the likelihood functions of quantifiers). In pursuing these questions, we believe it will be necessary to take a two-pronged approach, combining behavioral paradigms like the one introduced here with computational models that provide clear quantifiable predictions about how listeners adapt previous experience with other linguistic environments based on recent experience with a specific linguistic environment.

Acknowledgments

This work was supported by research grants from the National Institute of Health (NIH HD 27206) to MKT, and by a NSF CAREER award (IIS-1150028) as well as an Alfred P. Sloan Research Fellowship to TFJ.

References

Allen, J., Miller, J., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*, 544.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*(2), 159–219.

Bertelson, P., Vroomen, J., & Gelder, B. de. (2003). Visual recalibration of auditory speech identification a mcgurk aftereffect. *Psychological Science*, *14*(6), 592–597.

Chater, N., & Oaksford, M. (1999). The Probability Heuristics Model of Syllogistic Reasoning. *Cognitive Psychology*, *38*(2), 191–258.

Clayards, M., Tanenhaus, M., Aslin, R., & Jacobs, R. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804.

Fine, A. F., Jaeger, T. F., Farmer, T., & Qian, T. (underreview). Rapid adaptation of syntactic expectation.

Goodman, N. D., & Stuhlmüller, A. (2013, January). Knowledge and implicature: modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–84.

Kamide, Y. (2012). Learning individual talkers structural preferences. *Cognition*.

Kleinschmidt, D., & Jaeger, T. (2011). A Bayesian belief updating model of phonetic recalibration and selective adaptation. In *ACL CMCL 2011* (p. 10).

Kraljic, T., & Samuel, A. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268.

Kurumada, C., Brown, M., & Tanenhaus, M. (2012). Pragmatic interpretation of contrastive prosody: It looks like speech adaptation. In *CogSci 2012*.

Moxey, L. M., & Sanford, A. J. (1993). Prior expectation and the interpretation of natural language quantifiers. *European Journal of Cognitive Psychology*, *5*(1), 73-91.

Solt, S. (2009). *The semantics of adjectives of quantity*. Unpublished doctoral dissertation, CUNY.

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

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

Zadeh, L. (1965). Fuzzy sets. *Information and Control*.