SEMANTIC-PRAGMATIC ADAPTATION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF LINGUISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Sebastian Schuster
July 13, 2020

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

(Judith Degen)   Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

(Christopher G. Potts)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

(Daniel Lassiter)

Approved for the Stanford University Committee on Graduate Studies

_____

Abstract

Acknowledgements

Contents

List of Tables

List of Figures

xiii

Chapter 1

Introduction

Speakers vary in their language use at all linguistic levels. This is most obvious at the phonetic level, since we constantly encounter speakers with different accents than our own and who consequently pronounce words differently from us. But this variation also exists at other levels. For example, at the level of syntax, different speakers show different preferences for syntactic alternations such as the frequency with which they use passive constructions [Weiner1983]. Or at the word level, different speakers use words differently and–to take a stereotypical example–the comment *"The movie was good."* by an overly enthusiastic American person is generally intended to convey a worse opinion than the same utterance by a British person.

This variability is at odds with the observation that successful communication is nevertheless possible most of the time. Unless a speaker has a very strong accent which a listener had not been previously exposed to, or a speaker uses many words in a very unexpected manner, listeners tend to be able to comprehend the utterances of our interlocutors. The ease of communication with interlocutors whose language use differs from their own suggests that listeners are equipped with a very dynamic comprehension system that can easily be adjusted to novel speakers, rather than processing utterances according to their own egocentric linguistic system.

In theory there are multiple possibilities of how such a dynamic comprehension system could work. For example, it could be that listeners *normalize* the linguistic input prior to interpreting utterances such that any variability is removed before interpretation [e.g,NewmanSawusch1996]. It could also be that listeners constantly *fine-tune* their linguistic representations such that they match the representations of their interlocutors [e.g, PickeringGarrod2004]. However, an increasing body of research suggests that listeners *adapt* to specific speakers and learn-speaker specific language models that enable comprehension of utterances by speakers who vary in their productions [Norris2003,Kraljic2007,Bradlow2008, Kamide2012,Kleinschmidt2015,Fine2016,Roettger2018]. The learning is driven by the statiscal input – by tracking speaker-speci?c statistics, listeners can estimate accurate generative models of a speaker, i.e., models predicting how a speaker would pronounce a word or what a speaker would say in different contexts, which in return can lead to more accurate comprehension processes.

As I discuss in more detail in Chapter 2, there have been many experiments and considerable modeling work investigating the properties of adaptation processes in phonetics and syntax. At higher levels such as semantics and pragmatics, however, we still know a lot less about the extent to which listeners adapt or the associated cognitive processes. In this dissertation, I therefore investigate the extent of semantic-pragmatic adaptation through multiple experiments as well as the adaptation processes through computational modeling experiments.

1.1   Why adapt?

Before, I turn to the specific research questions and experiments, let me explain why it is beneficial for listeners to adapt. We interact with many different speakers in our daily lives – either truly interactively in conversation or more passively when watching TV, listening to audio or video recordings, or consuming other types of media. Thus, if as listeners, we constantly adapt to all the speakers we encounter and update speaker-specific expectations, we have to keep track of considerable amounts of information. This process clearly incurs some cost, which raises the question what the benefits of semantic-pragmatic adaptation are, and whether the benefits outweigh the cost. I will not be able to answer the latter question since I can neither quantify the the cost associated with adaptation nor quantify the utility of adaptation. However, to answer the first question, there exist clear benefits of semantic and pragmatic adaptation, including the following.

First, at the semantic level, listeners will be able to better infer the intended speaker meaning if they know the speaker's mapping between words and world states. For example, if I know that a speaker only uses *some* to refer to quantities greater than 3, I will be better able to narrow down the state of the world after hearing *"I ate some of the cookies"* than I would have been able to if I had assumed that the speaker uses *some* exactly the same way as I do, which for the sake of the example, let's say is to refer to quantities greater than 0. Similarly, differences in speaker and listener meaning become even more striking if my meaning of *some* is narrower than the speaker's *some*: If a speaker uses *some* to refer to a quantity of 4 but my meaning of *some* is limited to quantities greater than 5, I will infer a state of the world that is incompatible with the actual world. [**TODO: make figure?**].

At the pragmatic level, adaptation can further help listeners to infer the speaker's intended meaning. To see this, note that one of the key assumptions about pragmatic reasoning is that listeners reason about alternative utterances when interpreting a speaker's utterance [Grice1975, Horn1984]. For example, consider the following sentence that gives rise to a scalar implicature.

(1)   Sue: It might snow tomorrow.

(2)   ⇝ It is not certain that it will snow tomorrow.

According to Gricean pragmatic theories, listeners assume that a speaker is cooperative and arrive at the inference in (2) through a counterfactual reasoning process: they reason that if Sue had wanted to communicate that it is certain that it will snow tomorrow, Sue would have uttered the more informative statement *It is certain that it will snow tomorrow* (or simply the bare assertion *"It will snow tomorrow"*). Assuming that Sue knew the truth regarding the more informative sentence, it must be
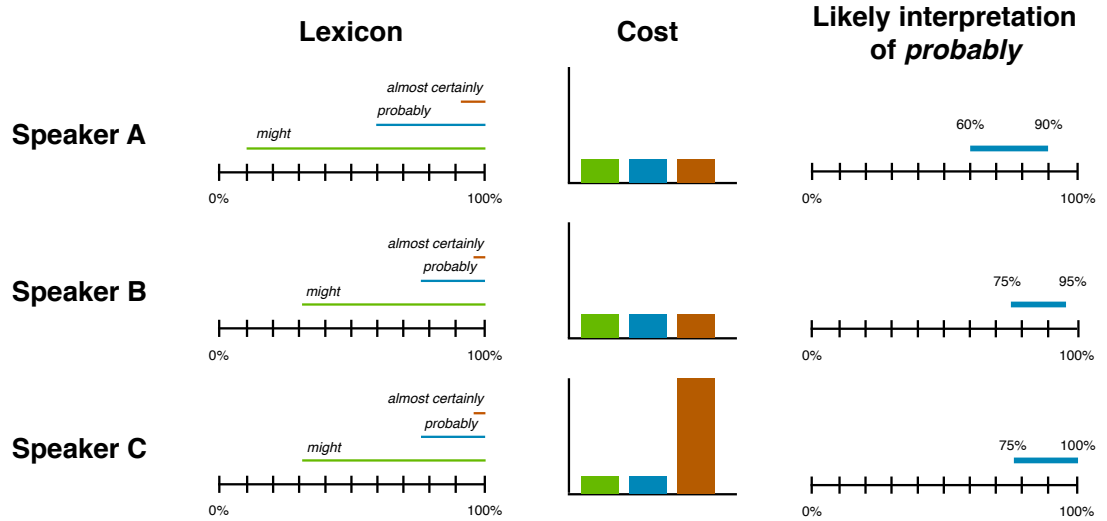
Figure 1.1: Lexica, utterance preferences and likely interpretation of *probably* for three different hypothetical speakers. The region of the probability scale covered by each line in the Lexicon panel indicates the corresponding expression's literal semantics. Height of bars in the Cost panel indicates the speaker's cost (dispreference) for each expression.

that the more informative statement is not true, which leads the listener to conclude (2).

Accounts of pragmatic reasoning share the implicit assumption that listeners have precise expectations about the speaker's language use – specifically, which utterance alternatives were available to the speaker that they didn't use – in different situations. Listeners can only draw correct pragmatic inferences if they know what a speaker would have said to communicate alternative world states. In large parts these expectations are guided by the meaning of words. as I illustrated with the cookies example above. However, speaker expectations also depend on other factors such as the speaker's preference for different lexical items or the set of alternative utterances from which they choose.

To illustrate how different beliefs about the meaning of words and utterance preferences can lead to different interpretations, consider the interpretation of the uncertainty expression *probably* produced by three different hypothetical speakers. For the sake of this example, let us assume the only three expressions that a speaker can choose from are *might*, *probably*, and *almost certainly*. A listener's beliefs about the three speakers' meanings and preferences are schematically illustrated in Figure 1.1.

First, consider speaker **A**, for whom *might* is true if the described event probability

(e.g., of snowing) exceeds 10%, *probably* if the event probability exceeds 60% and *almost certainly* if the event probability exceeds 90%. If a listener has accurate beliefs about **A**'s mapping between expressions and event probabilities and observes **A** produce the sentence *It will probably snow*, they will be likely to infer a probability of snowing between 60 and 90%. As illustrated above, the reasoning follows the schema of a standard scalar implicature [Grice1975, Horn1984]: if **A** had intended to communicate a probability above 90%, they could have said *It will almost certainly snow*, which would have been more informative and equally relevant. Assuming the speaker knows the actual event probability and is cooperative, it is therefore likely that the intended probability is not above 90%.[1]

Now, consider speaker **B**, for whom *might* is true if the event probability exceeds 30%, *probably* if the event probability exceeds 75% and *almost certainly* if the event probability exceeds 95%. If a listener has accurate beliefs about **B**'s mappings, they will be likely to infer, via the same reasoning as above, a chance of snow between 75% and 95% when they hear **B** produce the same sentence, *It will probably snow.*

Finally, consider speaker **C**. **C** uses the same mapping between expressions and event probabilities as **B**. However, **C** has a strong preference against producing *almost certainly*. If a listener has accurate beliefs about **C**'s lexicon and production preferences, they will be likely to infer a chance of snow between 75% and 100% when they hear **C** produce *It will probably snow* since they will not consider *almost certainly* a likely alternative. That is, the scalar inference will be blocked by the additional knowledge of the speaker's production preferences.

As this example shows, a listener who tracks the variability in these hypothetical speakers' meanings and production preferences will draw on average more accurate inferences about the world than a listener who relies on their own meanings and preferences for interpreting utterances.

The third advantage of adaptation is related to online language processing. The last several decades in psycholinguistic research produced a lot of evidence that listeners constantly engage in the prediction of the upcoming input [e.g.KuperbergJaeger2016].[2] On the one hand, this form of predictive processing makes language comprehension more robust. If due to noise, a listener is unable to perceive part of the signal, they can often fill in the blanks using their predictive model. On the other hand, predictive processing makes comprehension more efficient. If listeners constantly predict the upcoming signal, the early stages of comprehending an utterance reduce to comparing predictions about the upcoming linguistic input to the perceived input and, at these

---

[1]Under a standard Gricean view, the negation of the stronger alternative is inferred categorically. However, I adopt probabilistic language here in keeping with recent results that scalar inferences are more aptly viewed as probabilistic inference under uncertainty [Goodman2013].

[2]This is more generally true for many perceptual processes including vision (see e.g., [Clark2013; Friston2010]).

stages, listeners only have to process this difference, i.e., the error signal. However, according to such an account, rapid and accurate processing is only possible if listeners are able to make reliable predictions about the upcoming input. Accurate predictions in a variable and constantly changing environment, in return, are only possible through adaptation which highlights another advantage of constant adaptation.

1.2   Defining the scope

In this dissertation, I investigate the extent of semantic and pragmatic adaptation as well as the associated cognitive processes. That is, to what extent do listeners learn speaker-specific meanings of words; to what extent do listeners learn speaker-specific expectations of words, and to what extent does this speaker-specific knowledge affect interpretations of utterances? And what are the cognitive processes that lead to this behavior and what is the nature of the representations that are updated as a result of adaptation?

In this enterprise, I focus on uncertainty expressions such as *might* and *probably*. Thus, all findings will only directly apply to adaptation to variable use of uncertainty expressions. However, uncertainty expressions belong to the much larger class of context-sensitive expressions, a class of expressions for which it is generally assumed that their interpretation crucially depends on contextually specified parameters which–as I will show in subsequent chapters–are also tied to the speaker's identity. Given the extensive research on the parallels between uncertainty expressions and other context-sensitive expressions such as quantifiers and gradable adjectives [Lassiter2016, SchoellerFranke2017], the results in this dissertation should therefore also apply to any other types of context-sensitive expressions, and all the presented models could be easily extended to other classes of expressions.

1.3   Why uncertainty expressions?

Uncertainty expressions have several properties that make them a good testing ground for studying semantic and pragmatic adaptation. First, there is no consistent mapping between uncertainty expressions and event probabilities [e.g., Clark1990,Pepper1974], which suggests that listeners have to rely on additional contextual information (such as speaker identity) if they want to infer an event probability that a speaker intended to communicate using an uncertainty expression. Second, there is considerable inter-speaker variability in the use of these expressions [Wallsten1986] and therefore it is likely that listeners expect different speakers to use these expressions differently. Lastly, interpreting uncertainty expressions plays an important role in many everyday situations from the banal – such as talking about the weather – to the serious – such

as communicating about health risks [Berry2004, Lipkus2007, Politi2007] or making
financial decisions [Doupnik2003]. Thus, listeners would benefit from tracking how a
given speaker uses these expressions.

1.4   Structure of this dissertation

In Chapter 2, I provide background on three topics that I repeatedly touch upon in
the main part of the dissertation: partner-specific linguistic behavior, the semantics
of uncertainty expressions, and game-theoretic models of pragmatic reasoning. In
Chapter 3, I present experiments that investigate how English language users expect
a generic speaker to use uncertainty expressions, and I use this data to estimate the
parameters of a computational model of production expectations of a generic speaker.
In Chapter 4, I then turn to several research questions concerning semantic-pragmatic
adaptation. I first establish in experiments that listeners adapt to variable use of the
uncertainty expressions *might* and *probably*. I then present a computational model
of the adaptation process which is based on the generic speaker expectation model.
This model allows me to run different adaptation simulations and to investigate the
nature of representations that are updated during adaptation. I further discuss the
predictions of the model concerning the interpretation of uncertainty expressions after
adaptation, and I validate these predictions in another experiment. In Chapter 5, I
show that listeners can adapt to multiple speakers who use uncertainty expressions
differently, and I discuss the implications for different models of generalization. In
Chapter 6, I investigate to what extent the adaptation process can be modulated
by non-linguistic factors and show that information about a speaker's mood both
influences listeners' expectations before adaptation as well as their propensity to
adapt. In Chapter 7, I discuss what the findings in this dissertation taken together tell
us about semantic-pragmatic adaptation, and I outline promising future directions.

This dissertation is primarily focused on adaptation and therefore of primary
interest to researchers studying linguistic adaptation. However, given that I also
conduct many experiments probing the use of several epistemic modals, I also provide
novel data points in Chapter 3 for researchers interested in epistemic modality.

Chapter 2

Background

My investigation of the extent and properties of the semantic-pragmatic adaptation processes through the example of adaptation to uncertainty expressions falls within the broader topic of partner-specific linguistic behavior, which I review in the next section. Since I will also be making use of recent semantic theories of epistemic modals and game-theoretic pragmatic models, I also provide background on these two topics in this chapter.

2.1 Partner-specific linguistic behavior

Partner-specific linguistic behavior constitutes a large class of behaviors in which language users either change their comprehension behavior or their production behavior or both to better align with the idiosyncrasies of their conversational partners. Both the observed phenomena and the underlying cognitive processes and mechanisms have been discussed under several different names in the literature, including *alignment* (e.g., [PickeringGarrod2004]), *accommodation* (e.g., [Goldinger1998]), *convergence* (e.g., [Pardo2006]), *lexical entrainment* (e.g., [ClarkWilkesGibbs1986]), and–the topic of this dissertation–*adaptation* (e.g., [KleinschmidtJaeger2015]). The multitude of terms stems in part from different scientific communities working on different linguistic domains. For example, the terms *convergence* and *alignment* are both generally used to refer to the process of speakers and listeners converging in their production and comprehension behavior in interaction such that after several rounds of interaction, their phonetic productions and syntactic preferences are more similar to each other than at the beginning of the interaction. Researchers in phonetics, tend to describe this process as *convergence* whereas the sentence processing community prefers the term *alignment*. In part, however, these terms actually describe different phenomena. In particular, research in *adaptation* generally focuses exclusively on the comprehension side of language processing whereas research on *alignment*, *accommodation*, *convergence*, and *lexical entrainment* generally discusses partner-specific production as well as comprehension behavior, although the focus of most of this research lies on production behavior with the assumption that changes in production behavior follow from changes in comprehension behavior.

Experimental evidence for partner-specific behavior comes mostly from three different forms of studies: *exposure-test* studies, *continuous adaptation* studies, and *interactive conversation* studies. In exposure-test studies, participants first listen to some linguistic material, e.g., words in isolation or sentences describing an event depicted on a card. After this passive exposure, there is generally some form of test phase that probes whether participants updated their production and/or comprehension behavior. While frequently there is only one exposure and one test phase, there can also be multiple exposure-test blocks within one experiment.

In continuous adaptation studies, exposure and test trials are combined and thus

participants' behavior is probed after each exposure. The behavioral data in these studies often comes from online processing measures such as self-paced reading studies in which participants read a sentence word-by-word and one measures how long it takes them to read each word, or visual-world eye-tracking [Tanenhaus] or mouse-tracking [e.g., roettgerFranke] in which participants are instructed to point or click on an object and one tracks the movement of their gaze or their cursor.

In interactive conversation studies, either one participant and one confederate or two participants work on a task requiring language. For example, in the director-matcher paradigm, one participant (or a confederate) takes the role of the director who describes the order of different pictures in a display in front of them that is occluded from the other participant's vision, and the other participant (the matcher) has to put the same set of pictures into the same order as the director's set of pictures. The dependent measure in these experiments are usually either the recordings or transcripts which can be analyzed in terms of how similar the production behavior of the two interlocutors becomes over time, or one can also employ online measures such as eye-tracking to draw inferences about the comprehension behavior of the matcher.

Such experiments have been used in hundreds, if not thousands, of studies to investigate partner-specific linguistic behavior. I will not be able to review all of them or even the majority of them, but in what follows, I will present several foundational studies across linguistic domains and then discuss theories and computational models of partner-specific behavior. Finally, I will discuss several of the core questions regarding partner-specific behavior and what different accounts predict along with results from studies that try to adjudicate between different explanations.

2.1.1 Foundational studies

Phonetics. [Goldinger1998] conducted a series of shadowing experiments to investigate properties of the lexicon. In these exposure-test studies, he asked participants to first produce a set of words (or non-words) that were written on a screen (the baseline recordings). Then participants listened to recordings of these stimuli in an exposure phase. Finally, participants had to repeat ("shadow") words in a test phase which were recorded by the experimenter. The main dependent measure for his experiments came from an AXB task. In an AXB task, participants subsequently listen to three recordings A, X, and B and have to decide whether A sounds more like X than B does, or B more than A. A and B were always the baseline recording and the shadowing recording (in counterbalanced order) and X was the stimulus token that the shadower had heard in the shadowing experiment. [Goldinger1998] found that participants in the AXB task were consistently more likely to select the shadowing recording than the baseline recording, with modulation effects from overall word frequency, the number of repetitions of each stimulus, and whether participants immediately shadowed the

word or did so after a brief delay. While his main conclusion from this experiment was that language users store detailed episodes of perceived words in memory, this study was also among the first that provided evidence for phonetic accommodation: listeners changed their production behavior such that it matched more closely the behavior of the speaker whose productions they had heard during the exposure phase.

[Norris2003] conducted a series of exposure-test phonetic adaptation experiments investigating whether listeners could shift their perceptual boundaries between phonemes. Their exposure phase consisted of a lexical decision task; on each trial, Dutch participants listened to a recording and were asked to respond whether what they heard constituted a Dutch word or not. The recordings varied depending on the condition. In the /f/-biased condition, participants heard /f?s/, a sound that was ambiguous between the fricatives /f/ and /s/, embedded in words that always end in /f/ in Dutch; and in the /s/-biased condition, participants heard /f?s/ embedded in words that always end in /s/. Thus, the ambiguous sound was always disambiguated by the lexical context. During the test phase, participants categorized sounds on a continuum from /f/ to /s/ as either /f/ or /s/. [Norris2003] found that participants in the /f/-biased condition categorized more sounds on the continuum as /f/ than participants in the /s/-biased condition, suggesting that listeners shifted their perceptual boundary in response to the speech in the exposure phase.

[Clayards2008] studied another aspect of adaptation, namely whether listeners also update their beliefs about the intra-speaker variability of phonetic features. Specifically, they looked at the variance in voice-onset times (VOT) of the bilabial stops /p/ and /b/. Subjects participated in a visual-world eye-tracking study in which they saw pictures of four different objects on a screen. On each trial, participants listened to the recording of a single word and were asked to click on the object that had been mentioned. On critical trials, there were two objects whose names only differed in the word-initial stop, e.g., a beach and a peach. Across conditions, participants either listened to recordings of a speaker with very consistent VOTs (i.e., the variance in VOTs was very low; *narrow* condition) or to a speaker with highly varying VOTs (*wide* condition). [Clayards2008] found that in the wide condition, participants looked more at the competitor object (e.g., at the beach when hearing *peach*) than in the narrow condition, suggesting that participants exhibited uncertainty about which word the speaker produced and that over the course of the experiment, they learned the VOT distributions of the speaker.

Syntax.   [Bock1986] conducted one of the first studies systematically investigating partner-specific syntactic behavior through a repeated exposure-test experiment. Participants interacted with an experimenter who showed participants cards depicting an event. Half of the trials were priming trials on which participants listened to the experimenter describing the depicted event, and participants were asked to repeat

the description of the experimenter. Each priming trial was followed by a picture description trial on which participants were asked to describe a different event depicted on another card without any input from the experimenter. The critical priming trials varied in terms of the syntactic structure that was used to describe the event. Half of the events could be either described with a prepositional prepositional dative construction (e.g., *"A rock star sold some cocaine to an undercover agent"*) or a double object construction (*"A rock star sold an undercover agent some cocaine"*), and the other half could be described using an active sentence (e.g., *"One of the fans punched the referee"*) or a passive construction (*"The referee was punched by one of the fans"*). [Bock1986] found that the syntactic structure descriptions that participants produced on the picture description trials were influenced by the syntactic structure of the description of the priming trial; for example, when participants described an event after hearing a prepositional dative description on the preceding priming trial, they were more likely to also use a prepositional dative description than when they were primed with a double object construction. These findings, usually referred to as syntactic priming, have been replicated many times, including in interactive conversations between two participants [BrenniganPickering1998] and in comprehension (e.g., [Traxler2008]), and priming effects have also been observed in corpora of naturalistic speech (e.g., [Gries2005]).

[Kamide2012] investigated online comprehension of sentences with temporally ambiguous syntactic structures using a visual-world eye-tracking exposure-test experiment. Participants in her experiment saw displays with two persons (e.g., a man and a girl) and two objects (e.g., a motorbike and a carousel) and listened to a speaker producing a sentence with an ambiguously attached relative clause such as *"The uncle of the girl who will ride the motorbike is from France."* The relative clause in this sentence could either attach high as a modifier of *uncle* or low as a modifier of *girl* but given that children rarely ride motorbikes, the sentence is pragmatically disambiguated and the relative clause is likely attached high in this sentence. During the exposure phase, participants heard one of two speakers on each trial; speaker A always produced sentences with a high-attaching relative clause and speaker B always produced a sentence with a low-attaching relative clause. In the test phase, participants completed the same kind of trials with novel utterances produced by the two speakers. [Kamide2012] found that participants exhibited more anticipatory looks towards the pragmatically plausible object that would attach high (e.g., the motorbike in the example sentence) when the sentence was produced by speaker A as compared to when the sentence was produced by speaker B, and the opposite pattern when the sentence was produced by speaker B, suggesting that listeners learned speaker-specific preferences for syntactic structures that guided their online parsing behavior.

[FineJaeger2013] also studied the comprehension of temporally ambiguous syntactic structures. In a continuous adaptation experiment using a self-paced reading

paradigm, they asked participants to read sentences with a verb (e.g., *warned*) that could either be the main verb of a sentence (3a) or the verb of reduced relative clause (3b).

(3) a. The experienced soldiers **warned** about the dangers before the midnight raid.

     b. The experienced soldiers **warned** about the dangers conducted the midnight raid.

Given that reduced relative clauses are a lot less frequent than regular clause, readers usually experience a garden-path effect when reading the second verb (*conducted*) in (3b). In a self-paced reading experiment, in which participants read one word (or one segment) of a sentence after another with the time spent on each individual word being recorded, this garden-path effect typically manifests in longer reading times on the second verb in sentences like (3b). However, [FineJaeger2013] found that throughout the experiment in which participants were both exposed to sentences in which the ambiguous verb was a main verb and sentences in which it was a reduced relative clause, the garden-path effect slowly disappeared and at the end of the experiment, participants were as fast reading sentences with reduced relative clauses as they were reading sentences with a canonical main verb structure. This again provides evidence for listeners updating their expectations about syntactic parses in an environment (in this case in the experimental context) and integrating these updated expectations in online sentence processing.

While there have been several studies about cumulative syntactic adaptation similar to [Kamide2012] and [FineJaeger2013], it is also noteworthy that these effects appear to less reliable as compared to other partner-specific phenomena. [Liu2017] failed to replicate the experiment by [Kamide2012], [HarringtonStack2018] failed to replicate one experiment by [FineJaeger2013], and [PrasadLinzen2020] argued that one needs more than 1,200 participants to achieve sufficient power to detect syntactic adaptation effects in self-paced reading experiments.

Semantics. At the level of semantics, [ClarkWilkesGibbs1986] demonstrated in an interactive conversation study that interlocutors implicitly negotiate referring expressions in interaction. They asked pairs of participants to participate in a matcher-director task. One participant took the role of director and was asked to describe the order of 12 tangram shapes in a display to the other participant, the matcher. The matcher's task was to arrange the shapes in the same order as the director's display. The director and matcher could verbally communicate but they could not see each other or their interlocutor's display. Each pair of participants conducted six rounds of describing and arranging tangram figures. [ClarkWilkesGibbs1986] found that while in initial rounds directors used very long referring expressions describing

many details, over the course of the experiment, the expressions became shorter and shorter and participants started to associate the individual figures with short noun phrases such as *"the ice skater"* or *"chair"* which clearly identified the referents for the two interlocutors with the same conversational history but would be very opaque for listeners who were not part of the previous exchanges, a phenomenon usually referred to as *lexical entrainment*.

[Yildirim2016] investigated to what extent listeners learn speaker-specific expectations of vague lexical items through a series of exposure-test experiments. Specifically, they investigated how participants' expectations about a specific speaker's productions of the quantifiers *some* and *many* changed as a result of observing that speaker's use of quantifiers. On exposure trials, participants saw short video clips along with a bowl of blue and green candies. In each clip, the speaker produced an utterances such as *"Some of the candies are blue"* or *"Many of the candies are green"* along with different proportions of blue and green candies. On test trials, participants were asked to rate how likely they though it was that the speaker they just saw would use *some* or *many* (or something else) to describe different candy proportions by distributing 100 points across three utterance choices. Depending on the condition, the exposure speaker would either always use *some* (*some-biased* condition) to describe a bowl with approximately equal amounts of blue and green candy, or they would always use *many* (*some-biased* condition). This manipulation had an effect on participants' production expectations: Participants who were exposed to the *some-biased* speaker, rated *some* to be a more likely utterance choice for a larger range of proportions than participants in the *many-biased* condition, and the opposite was true for *many*. This was true both for between-participant manipulations with only one exposure speaker per participant, and in within-participant manipulations with two different exposure speakers per experiment, suggesting that participants learned speaker-specific production expectations of the use of quantifiers after brief exposure to a specific speaker.

Intonation and prosody. Lastly, there has also been work investigating the dynamicity of prosodic cues. [Kurumada2012] investigated two aspects of the interpretation of contrastive focus in multiple exposure-test experiments. In one experiment, they exposed participants either to a speaker who used contrastive focus reliably or a speaker with unreliable uses of contrastive focus. On each trial, participants had to select an image that they thought the speaker was referring to. The images always included an object that the speaker mentioned (e.g., a zebra) and a competitor that was similar to the target image (e.g., an okapi, a zebra-like animal which only has strips on its legs). During the exposure phase, the speaker produced utterances of the form *It looks like a zebra*, either with focus on *looks* (verb-focus) or on the noun (noun focus) followed by a continuation that clearly described the target (affirmative

continuation; e.g., *... because it has black and white strips all over its body*), or a continuation that described the competitor (negative continuation; e.g., *... but it's not; it has stripes only on its legs*). Depending on the condition, the speaker used different focus patterns. In the reliable condition, the speaker always used noun-focus with affirmative continuations, and verb focus with negative continuations; in the unreliable condition, the speaker used both focus patterns for both continuations equally often. [Kurumada2012] found that listeners adapted to the different uses and in a test phase without continuations, participants in the reliable speaker condition relied on prosodic cues more often than in the unreliable condition. This effect has also been replicated in German by [RoettgerFranke2019], and they have shown that prosodic adaptation affects incremental online processing of utterances and that learning the associations between speakers and their use of prosodic cues is an incremental learning process. Lastly, [Kurumada2012] further showed that listeners can also recalibrate their perceptual boundary between verb-focus and noun-focused versions of the same utterance.

2.1.2 Accounts and models of partner-specific behavior

Episodic memory. One account of partner-specific behavior, especially for phonetic partner-specific behavior, is based on the theory that linguistic representations consist of rich episodic memory traces from individual perceptive events. Episodic memory accounts generally focus on the processes involved in word recognition and word production (e.g., [Goldinger1998,Johnson1996,Pierrehumbert2001]). In this context, a rich episodic memory trace consists of information about the word identity, the phonetic properties of the produced word as well as other contextual factors such as the situational context. In comprehension, word recognition then operates through activation of memory traces as suggested by cue-based memory retrieval models (e.g., [Ratcliff1979]). For example, if a listener hears a speaker produce the word "dog", all memory traces of past productions of the word dog will be activated and once activation exceeds a certain threshold, the listener infers the word identity. Crucially, however, not only the word identity but also contextual factors such as the situational context and the phonetic play a role in activation and activation of traces will be stronger if it closely matches contextual factors. Thus memory traces of previous productions of a specific speaker are activated more than traces of the same word produced by other speakers, which means that the representations of words, i.e., the aggregate activation of memory traces, are different depending on the speaker.

On the one hand, if one makes the assumption that representations for production and comprehension are shared and that activation of memory traces persists for at least short periods of time, such accounts predict that language users accommodate to their interlocutors in interaction. If a listener hears words produced by a speaker,

the activation of memory traces that match the phonetic properties of the speaker will be stronger than other memory traces of the same word. If the listener then produces the same word shortly after hearing it, there will be residual activation of the memory traces that are specific to their conversational partner and the listener's production will be more similar to the production of the original speaker than it would have been if they had produced the word without previously hearing it [Goldinger1998].

On the other hand, in comprehension, episodic memory accounts predict that comprehension is facilitated when one hears a word produced by a familiar speaker and, with additional stipulations, that listeners may interpret words differently depending on the speaker. The reason for facilitated comprehension is that memory traces that match the input both in word identity and speaker identity (or more broadly phonetic properties) receive more activation and the word representation therefore reaches the activation threshold faster than when the input only matches the word identity, as it is the case when a word is produced by a new speaker with different phonetic properties of their productions. The reason for speaker-specific interpretations are also increased activation of speaker-specific representations. If one makes the assumption that interpretation of words are a result of activating memory traces of contexts in which the word had been previously produced, and if a specific speaker uses a word only in certain contexts, contextual representations associated with the word representations are going to be strongly activated for the contexts in which the speaker produced the words in the past but less so for contexts in which other speakers produced the word in the past, ultimately leading to interpretations that may differ depending on the speaker identity.

[TODO: make a figure visually explaining episodic memory accounts]

While episodic memory accounts are capable of explaining a lot of empirical behavior in the domain of word recognition and word production, it remains an open question how exactly such accounts explain processing at higher linguistic levels such as syntax and semantics. Te explanation of speaker-specific interpretations that I provided here already goes beyond the original discussions of episodic memory accounts (though see, for example, [HortonGerrig2005,HortonGerrig2016], for episodic memory accounts of partner-specific productions and comprehension of referring expressions), and as also discussed by [Goldinger1998], it remains unclear how episodic memory accounts can explain parsing and composition of utterances, both in a partner-independent and partner-specific way.

Alignment. [PickeringGarrod2004] proposed an account for partner-specific linguistic behavior based on the idea that in interaction, conversational partners *align* their linguistic representations and representations of context, i.e., gradually converge to very similar or identical representations.This account has been primarily

inspired by the findings from structural priming in dialog (e.g., [BranniganPickering-Cleland2000]) and findings that the production of sentences is shared across conversational partners (e.g., [GarrodAnderson1987]), and [PickeringGarrod2004] argue that alignment and the joint productions in dialog happen as a result of a low-level priming mechanism. Similarly as according to episodic memory accounts, representations are shared by the production and comprehension systems and listeners and speakers activate representations at different linguistic levels as well as representations of the context in comprehension and production. This activation then makes certain lexical items, syntactic structures or pronunciations easier to produce and subsequently easier to comprehend, and conversational partners therefore automatically converge to similarly activated representations which leads to similar productions and facilitates comprehension. Thus, according to this account, partner-specific linguistic behavior happens automatically and effortlessly as a by-product of how the language processing system works.

[TODO: include Figure 2 from P&G2004]

[PickeringGarrod2004] acknowledge that this account is too simplistic to explain partner-specific behavior in some cases. For example, when speakers and listeners' visual common ground differs because some objects are occluded from the visual field of the listener (e.g., as in experiments by [Keysar2000,Heller2008]) it is impossible for the conversational partners to align their contextual representations. For this reason, [PickeringGarrod2004] argue for a two-step process: If a listener can interpret an utterance according to their own (potentially aligned) representations, they will do so; and only if this egocentric interpretation fails, they will actively reason about the common ground which requires additional resources and may be constrained by time or other pressures on resources. However, what it exactly means at a mechanistic level to fail interpreting an utterance remains an open question.

The alignment account readily predicts effects from structural priming experiments: when listeners hear an utterance with a certain syntactic structure, the syntactic representation of that structure is activated and the residual activation in subsequent productions primes the listener to produce utterances with the same syntactic structure as in the perceived utterance. Alignment accounts also predict the results from phonetic accommodation: again, phonetic representations are activated in comprehension and the residual activation can lead to productions of words that are phonetically more similar to the productions of the conversational partner. The account also predicts some properties of the behavior observed in lexical entrainment experiments: when two interlocutors are repeatedly using certain lexical items in referring expressions, certain lexical representation will be activated and the residual activation leads speakers and listeners to reuse the same referring expressions. However, the priming mechanism does not predict why speakers and listeners progressively

shorten their utterances in repeated reference games, and without additional stipulations one would expect the opposite to happen, since according to the alignment account, speakers and listeners should be using identical referring expressions. Lastly, if we make the assumption that the lexical representations also prime the contexts in which words had been recently used, this account does predict some of the semantic adaptation behavior that has been observed in the context of quantifiers. When listeners hear a speaker produce a quantifier such as *some* with specific proportions, activation of the representation of *some* and the context of the recently observed proportions will be higher and therefore listeners will expect that *some* is more likely to be used with these proportions. However, without additional stipulations, alignment fails to account for several behaviors in semantic adaptation experiments, and in particular, the fact that listeners form speaker-specific expectations when exposed to multiple speakers. I will discuss this issue and the shortcomings of the alignment account to explain speaker-specific behavior in more detail in Section 2.1.3 below.

Connectionist implicit learning accounts. [Cheng2006] developed a connectionist model of syntactic learning and processing and showed that such a model can also predict several empirical findings from the structural priming literature. Their model is a recurrent neural network (RNN) model [Elman1990] that predicts morphemes one-by-one. Each recurrent unit consists of two parts, a meaning system and a sequencing system, which together allow the model to simulate production and comprehension. In production, the model bases its predictions on a global sentence meaning, represented by abstract semantic roles and lemmas, and the previous word. in comprehension, the model bases its predictions only on the previous context since the message has to be inferred from the perceived utterance (a *messageless* model input). The model is trained on meaning-utterance pairs (an utterance in this model is an order list of morphemes) such that the weights of the RNN are updated whenever the predicted next morpheme is different from the actual morpheme in the training utterance. This process is intended to simulate the acquisition process. Importantly, the model keeps learning when comprehending utterances. When the model receives a messageless input, it compares its morpheme predictions to the actual morphemes in the input utterance and if there is a mismatch between predictions and input, the model updates its weights.

Like the episodic memory and alignment accounts, this model assumes shared representations, i.e., weights of the RNN, for comprehension and production. This leads to the model predicting structural priming. For example, when the model is input an utterance with a prepositional dative it will compare its predictions to the actual utterance. If it correctly predicts that *to* follows the object in the sentence, the model already appears to favor a prepositional dative structure and it will not update its weights significantly. On the other hand, if the model predicted a double

object construction and therefore did not predict the *to*, there will be a mismatch between predicted morphemes and actual morphemes and the model will updates it weights accordingly. The effect of priming can then be probed by having the model predict the production of utterances given a message. [Cheng2006] found that the predicted proportions of syntactic structures closely matched human behavior and that the model also exhibited priming behavior independent of specific lexical items (e.g., priming from a prepositional construction with *to* to a prepositional construction with *for*).

[Cheng2006]'s model was trained on data from a relatively small grammar and therefore cannot be used to simulate production and comprehension on a rather small set of utterances. However, recently, similar connectionist models that make use of large-scale RNN models developed for human language technology tasks have been shown to predict reading times from syntactic adaptation experiments [vanSchijndelLinzen2018] and to learn conventions in producing and comprehending referring expressions in repeated reference games [Hawkins2019].

Further, implicit learning accounts make the more general prediction that the magnitude of learning depends on how unexpected (or surprising) the input is. For example, the verb *sell* rarely appears in double object constructions and therefore *"The painter sold the art dealer a new work"* is considerably more surprising than *"The painter sold a new work to the art dealer."* Thus upon hearing a very unexpected utterance, implicit learning accounts predict that listeners will exhibit stronger learning effects, which in fact has been demonstrated, for example, in a re-analysis of a visual-world eye-tracking experiment to investigate priming effects in comprehension [ThothathiriSnedeker2008,SniderJaeger2013].

This expectation-based learning behavior is predicted by the connectionist models that I discussed here. Connectionist models, however, are not the only models predicting expectation-based learning, and I will discuss in the next section a class of models based on Bayesian belief updating that make similar predictions as the RNN-based models and have been used to model a wider range of partner-specific language behavior.

Bayesian belief updating adaptation models. A separate line of work focused on modeling several partner-specific linguistic behaviors using computational models within the framework of rational analysis [Marr1982,Anderson1990]. Rational analysis models are less concerned with the exact properties of representations as compared to the above discussed episodic memory and alignment accounts, but insetad are intended to formalize the optimal behavior of a cognitive agent to achieve precisely specified goals. In this vein, one successful method of formalizing cognitive processes has been to implement them as probabilistic programs and assume that learning and inferences happen as a result of Bayesian inference (e.g., [Tenenbaum2011]).

Within this framework, [Clayards2008] proposed the *ideal observer* model of phoneme recognition. According to this model, a listener represents phonemes as distributions over relevant phonetic cues. For example, if we consider again the main difference between the bilabial stops /b/ and /p/, namely the voice-onset time (VOT), an ideal observer represents /b/ and /p/ as distributions over VOTs, $P(\text{VOT} \mid /\text{b}/)$ and $P(\text{VOT} \mid /\text{p}/)$. When perceiving a sound with a specific VOT, assuming that all other cues unambiguously lead the agent to infer that the sound is either /b/ or /p/, the ideal observer then computes the probability of having perceived a specific sound such as /b/, $P(/\text{b}/ \mid \text{VOT})$, using Bayesian inference:

$$P(/\text{b}/ \mid \text{VOT}) = P(/\text{b}/) \times \frac{P(\text{VOT} \mid /\text{b}/)}{P(\text{VOT} \mid /\text{b}/) + P(\text{VOT} \mid /\text{p}/)}$$

According to this model, listeners take into account their prior beliefs about the phoneme in question (here $P(/b)$ and the relative likelihood of the observed VOT for the phoneme in question (the numerator) as compared to the sum of the likelihoods of the given VOT for all considered phonemes (the denominator). The model closely predicts looking patterns in the experiment by [Clayards2008], which exposed participants either to speakers with very narrow or very wide VOT distributions. The model predicts that participants should exhibit more uncertainty in the wide condition than in the narrow condition because in the wide condition, the VOT distributions for the two phonemes overlap more and therefore the probability of /b/ and the probability of /p/ after hearing a phoneme with a specific VOT are both closer to .5 in the wide condition as compared to the narrow condition. As I discussed above, this predicted behavior closely matches the behavior of participants in the experiment who indeed exhibited more uncertainty in the wide condition.

While the ideal observer model closely predicts participants' behavior if we stipulate distributions over VOTs, it does not explain how participants may learn these distributions and how individual observations change beliefs about these distributions. For this reason, [KleinschmidtJaeger2015] presented the *ideal adapter* framework, a probabilistic model that jointly predicts speech perception and learning as part of phonetic adaptation. The central idea of the ideal adapter is that rather than assuming that listeners have fixed beliefs about distributions over phonetic cues for different phonemes, listeners have higher-order beliefs about these distributions represented by another set of distributions. For example, if we assume that the distributions $P(\text{VOT} \mid /\text{b}/)$ and $P(\text{VOT} \mid /\text{p}/)$ can be approximated with two normal distributions with mean and variance parameters (μ_b, σ_b) and (μ_p, σ_p), respectively, then according to the ideal adapter model, listeners have beliefs about these parameters in the form of distributions $P(\mu_x)$ and $P(\sigma_x)$. When listening to speech, listeners then infer the phoneme according to the ideal observer model through marginalizing over their uncertainty over the relevant distributions. Thus, for example, the probability

of perceiving a /b/ as compared to a /p/ sound is then:

$$P(/b/ \mid \text{VOT}) = P(/b/)$$
$$\times \int_0^1 P(\mu_b, \sigma_b, \mu_p, \sigma_p)$$
$$\times \frac{P(\text{VOT} \mid \mu_b, \sigma_p)}{P(\text{VOT} \mid \mu_b, \sigma_b) + P(\text{VOT} \mid \mu_b, \sigma_p)} d(\mu_b, \sigma_b, \mu_p, \sigma_p)$$

Here the distributions over voice-onset times are guided by the mean and variance parameters for each phoneme (e.g., $P(\text{VOT} \mid \mu_b, \sigma_p)$), and listeners have probabilistic beliefs about the values of all the mean and variance parameters ($P(\mu_b, \sigma_b, \mu_p, \sigma_p)$). When trying to infer the probability of /b/, listeners then average over the uncertainty about the VOT distributions (hence the integral over the mean and variance parameters in the formula above).

Apart from introducing uncertainty about the cue distributions, the second crucial addition of the ideal adapter framework is an account of updating beliefs about cue distributions. For simplicity, let us assume that in a given context, a listener unambiguously infers the identity of a phoneme.[1] This could be, for example, because of visual disambiguation as when a listener hears the word /b?p/each[2] while looking a screen which contains a peach but crucially not a beach, or this could be because of lexical disambiguation as when a listener hears the word /b?p/rother, in which case the fact that *prother* is not a word makes it very unlikely that the word-initial phoneme is a /p/. Thus given a perceived phonetic cue (e.g., the VOT) and the identity of a phoneme, listeners update their beliefs about the mean and variance parameters using Bayesian belief updating:

$$\underbrace{P(\mu_b, \sigma_b \mid \text{VOT}, /b/)}_{\text{posterior}} \propto \underbrace{P(\mu_b, \mu_p)}_{\text{prior}} \times \underbrace{P(\text{VOT} \mid /b/, \mu_b, \sigma_b)}_{\text{likelihood}}$$

According to this model, when observing a /b/ with a specific VOT, a listener reweights their *prior* beliefs about the values of the parameters influencing the VOT-distribution based on the *likelihood* of observing that VOT given different parameterizations of the VOT distribution, resulting in *posterior* beliefs about parameter values. Thus, with every observation, a listener updates their beliefs about the mappings between phonetic cues and phonemes.

---

[1]This is not a critical assumption but it simplifies the formalization of the model because we have to account for less uncertainty.

[2]As above, I am using the notation /A?B/ to indicate a sound that is at least to some extent ambiguous between /A/ and /B/ when heard in isolation.

The ideal adapter model in its most basic form already captures the behavior from numerous phonetic adaptation experiments, including the study by [Norris2003]. However, as I will discuss in more detail below, listeners can learn *speaker-specific* distributions over phonetic cues and they can generalize from previously encountered speakers to novel speakers, and if we assumed that listeners held only one set of beliefs about phonetic distributions independent of the speaker or other contextual factors, the model would make incorrect predictions. In its full form, the ideal adapter model therefore assumes that listeners have speaker-specific beliefs about VOTs and that the priors for newly encountered speakers are highly structured based on regularities across categories, and may differ, for example, depending on the speaker's gender or information about where the speaker grew up [Kleinschmidt2019].

Models based on Bayesian belief updating have also been proposed to explain adaptation behavior in other linguistic domains. [Kleinschmidtetal2012] proposed a model based on Bayesian belief updating that predicts syntactic adaptation as in the experiments by [FineEtAl2013]. Similarly, [Hawkins2018] proposed a model of lexical entrainment in repeated reference games that assumes that listeners and speakers update their beliefs about different possible lexica in the course of the experiment. [Qing2014] proposed a model that qualitatively predicts semantic adaptation to different uses of quantifiers. Finally, [RoettgerFranke2018] proposed a model according to which listeners learn associations between prosodic cues and interpretations, and [DelaneyBuschEtAl2019] showed that a model based on Bayesian belief updating can also predict (to some extent) adaptation observed in event related potential (ERP) experiments investigating the extent of semantic priming.

Further, Bayesian belief updating models are not limited to linguistic phenomena. Hierarchical probabilistic models in combination with Bayesian learning had already earlier been proposed for many other cognitive phenomena such as visual perception (e.g., [Clark2013]) and categorization (e.g., [Tenenbaum2011]) and given the success of this family of models to closely model human behavior across many different cognitive tasks, some form of learning processes that can be effectively described with Bayesian models may be a fundamental property of human cognition (e.g., [Clark2013], [Friston2010]).

As I mentioned in the previous section, Bayesian belief updating models also predict expectation-based learning. The reason for this is that belief updating is based on two components, the prior and the likelihood. If the input is highly expected and therefore prior beliefs about model parameters make the input very likely, i..e, the likelihood is very high for parameterizations that have a high probability according to the prior, then very little reweighting of the prior happens and posterior is very similar to the prior. On the other hand, if the input is very surprising and is only likely for parameterizations that have a very low probability according to the prior, then considerable reweighting of the prior happens and the posterior differs considerably

from the prior, i.e., expectations have been updated.

One important difference from the other accounts that I presented in this section is that belief updating accounts do not make the assumption that representations for comprehension and production are identical, which explains why, for example, listeners can learn to understand a strong Scottish accent without actually starting to sound Scottish themselves. However, since these accounts are primarily concerned with comprehension, it remains an open question how the representations between comprehension and production are linked and how exactly adaptation and partner-specific production behaviors such as structural alignment or phonetic accommodation are connected.

2.1.3 Core questions

Numerous studies have shown that language processing is a highly dynamic system and that listeners and speakers exhibit change in their linguistic behavior in interaction. Thus there is an abundance of evidence for the question *whether* language users exhibit partner-specific behavior. However, when it comes to more detailed questions regarding behavior and processes, there are many important questions that have only been partially answered so far. Here, I discuss a selection of the core questions, experiments that tried to address these questions, and how the question and the results relate to the accounts and models of partner-specific behavior that I discussed above.

Partner-specific vs. local context adaptation. One of the most important questions concerning partner-specific behavior has been whether these processes are in fact speaker-specific or whether language users simply update their behavior in a local context and then abandon their partner-specific behavior once they interact with another interlocutor. The question of speaker-specificity has been most extensively studied in the domain of phonetic adaptation. [EisnerMcQueen2005] extended the research by [Norris2003] and investigated whether perceptual boundaries between the fricatives /f/ and /s/ were speaker-specific. They found no adaptation effects when the speaker producing the fricatives in the test phase (a categorical perception task) was different from the speaker in the exposure phase (a lexical decision task with Dutch words). [KraljicSamuel2006], on the other hand, found that updated perceptual boundaries between the stops /d/ and /t/ persisted even when the exposure and test speaker differed, suggesting that phonetic adaptation to stop consonants is not speaker-specific. Based on these findings, [KraljicSamuel2006] hypothesized that speaker-specific adaptation only happens for some phonetic contrasts, and they confirmed this hypothesis in another study replicating the speaker-specific adaptation effects for fricatives and the speaker-independent adaptation effects for stop consonants [KraljicSamuel2007]. They argue that this sensitivity to different phonemes

might either stem from the fact that the acoustic signal of fricatives provides more information about the speaker identity than stops do, or similarly, stem from the fact that intra-speaker variability for fricatives is much lower than inter-speaker variability whereas intra-speaker variability for stops is almost as high as inter-speaker variability. A rational agent thus would only retain speaker information for fricatives, which appears to closely match behavior observed in adaptation experiments. Further evidence for this comes from [TrudeBrownSchmidt2012] who also found speaker specific adaptation to vowels which again provide a lot of information about the talker and vary considerably across talkers.

At the syntactic level, [Kamide2012] found speaker-specific adaptation to attachment preferences but as mentioned above, these effects failed to replicate [Liu2017]. Further, [OstrandFerreira2019] systematically investigated whether structural alignment is bound to the speaker identity and while they were able to replicate many known structural alignment effects, they found no speaker-specific alignment in any of her five studies. [Krozcek2017], on the other hand, found speaker-specific effects in an exposure-test experiment in which participants learned word order preferences of two German speakers. When asked to interpret sentences in which the determiners had been replaced by white noise leading to utterances without case information for the subject and the object noun phrases, participants were more likely to interpret the sentence according to an SVO parse if produced by a speaker who showed such a preference during exposure, and more likely to interpret the sentence according to an OSV parse if produced by another speaker who showed the opposite preference during exposure.Taken together, it therefore seems that partner-specific syntactic behavior is rarely partner-specific but can be in some instances.

At the semantic/pragmatic level, [BrennanClark1996] found some evidence for speaker-specific behavior: in repeated reference games, speakers were more likely to change the referring expression when their interlocutor changed during the experiment than when the interlocutor remained the same. [MetzingBrennan2003] further found that listeners were slower to resolve referring expressions when a confederate started referring to an object with a new expression halfway through the experiment, but did not find such a slowdown when a new confederate was using a different referring expression than the original confederate. These findings were near replicated in a very similar experiment by [BrownSchmidt2009] though she also found that partner-specific effects in online processing are limited to truly interactive settings in which two interlocutors communicate and do not occur when participants listen to recordings of referring expressions. Finally, in the domain of expectations about the use of quantifiers, [YildirimEtAl2016] found that listeners are able to form-speaker specific expectations when they are exposed to multiple speakers who differ in their use of quantifiers.

In summary, the experimental evidence suggests that adaptation and other partner-specific behaviors are indeed speaker-specific for many, but not all, phonetic contrasts as well as for phenomena related to the interpretation of words or utterances. At the same time, at the level of syntax, there is only very little evidence for speaker-specific adaptation and it remains an open question in which cases listeners reliably learn speaker-specific preferences for specific syntactic structures. It further remains an open question why we find this selective adaptation behavior. [OstrandFerreira2019] argue that adaptation might be guided by utility: in cases in which adaptation is useful for successful communication–such as adapting to certain pronunciations of words or to the meaning of referring expressions or quantifiers–listeners readily adapt, but in cases in which adaptation is not required for useful communication–such as learning speaker-specific syntactic attachment preferences– listeners do not expend resources on adaptation. Further, as the results of selective phonetic adaptation suggest, listeners might also be aware of the structure in the variability, e.g., whether the speaker's identity explains some of the variance, and it could be that listeners only adapt to specific speakers if the average inter-speaker variability exceeds the intra-speaker variability.

With regards to processing accounts and models, speaker-specific behavior is predicted by episodic memory accounts and by probabilistic models with speaker-specific beliefs. The alignment account, on the other hand, does not predict speaker-specific behavior. According to the alignment account, partner-specific behavior is a result of residual activation of representations and this account does not stipulate any mechanisms that would link representations to specific speakers. Thus according to this account, language users should always behave most similar to the most recent speaker they interacted with, which is incompatible with the empirical results from many studies. Without additional stipulations, connectionist accounts also do not predict speaker-specific behavior but one could easily imagine to extend the recurrent units of a model such as the one by [Cheng2006] to be able to link the speaker's identity to linguistic representations.

Long-term vs. short-term effects. A second important question, in particular in the structural priming literature, has been whether the observed effects constitute long-term learning or only persist for a short period of time. [BockGriffin2000] and [BockEtAl2007] found that structural priming effects persists over many intervening filler trials, suggesting that priming involves at least to some extent long term memory. At the same time, however, there are also priming effect that seem to diminish rapidly. For example, [WheltonSmith2003] found in a production experiment, that participants were faster to produce a target utterance if they had been primed with an utterance having the same syntactic structure if the prime occurred immediately before the production trial, but they found no such speedup if the prime and test

trial had intervening fillers.

In the domain of syntactic adaptation, [Krojcek2017] found that adaptation effects persisted for at least 24 hours. Participants who returned to a second experimental session on the next day had retained the associations between syntactic preferences and speakers that they learned in the first session. However, in a third session 9 months later, the adaptation effects had disappeared. Similarly, [KaschakKuttaSchatschneider2010] found that cumulative structural priming effects persisted between two experimental sessions that had been one week apart.

In phonetic adaptation, there is more of a consensus that adaptation effects persist for longer period of times. For example, the adaptation study by [BradlowBent2007] involved two sessions and listeners retained what they learned in the exposure phase until the second session. Similarly, [XieEarleMyers2019] also found that adaptation effects persisted across two sessions separated by 12 hours.

This question has been of great interest because different accounts of partner-specific linguistic behavior make different predictions about long-term effects. Episodic memory accounts predict that effects should persist over long periods of time, though as with any memory system there may be eventual memory decay. Connectionist implicit learning models also predict long-term behavior and in fact, [Chang2006] demonstrated that their model closely predicts human behavior from experiments with filler trials in between priming and test trials. Bayesian adaptation models qua purely computational models [Marr1982] do not make any specific predictions about the structure and properties of memory but they do generally stipulate that speaker-specific beliefs are maintained indefinitely and thus also predict long-term effects.

Alignment accounts, however, do not predict long-term effects. This is because one core assumption of these accounts is that partner-specific behavior is a result of *recently* activated representations and that this activation rapidly decreases, especially once a language user processes additional utterances or encounters new speakers, which means that any partner-specific behavior should also be short-lived. Considering the considerable evidence for long-term priming and adaptation, one therefore may be inclined to dismiss the alignment account altogether. However, as [FerreiraBock2006] argue, there are effects of structural priming such as facilitation of production [WheltonSmith2003], which rapidly disappear and therefore priming and other partner-specific behavior may be a result of both long-term learning as well as short-term activation (see also [ReitterEtAl2011] for a concrete model within the ACT-R framework that supports this hypothesis).

Mechanistic vs. context-sensitive learning. Several studies also investigated closer the claim that partner-specific behavior happens automatically and effortlessly as a by-product of the language processing system, as argued by [PickeringGarrod2004]. According to this view, contextual factors should not matter and language

users should exhibit partner-specific behavior after interacting with a speaker independent of other factors such as visual cues or social information. In other words, partner-specific behavior should only be affected by language use and not other contextual factors.

There are several studies challenging this prediction. [KraljicEtAl2008] found that phonetic adaptation to a shifted perceptual boundary between /s/ and /sh/ was suppressed if participants were shown a speaker with a pencil in their mouth which explained why they pronounced /s/ more like /sh/. In the control condition in which participants saw a picture of a speaker without a pencil in their mouth, participants shifted the perceptual boundary, suggesting that contextual factors such as information from visual cues can modulate whether listeners adapt or not and that listeners are not mechanistically adapting exclusively based on linguistic input.

In the domain of accommodation, [Babel2012] found that social factors modulate accommodation in intricate ways. For example, in a shadowing experiment, heterosexual female participants aligned their productions closer to a male speaker if they considered the speaker attractive, and heterosexual male participants exhibited the opposite behavior and aligned their productions closer to a male speaker if they considered him less attractive. These findings again suggest that listeners are not mechanistically engaging in partner-specific behavior but instead also consider non-linguistic factors.

All these findings challenge that idea of a purely mechanistic account such as the alignment account. Probabilistic adaptation accounts, on the other hand, assume that the prior beliefs can be influenced by many contextual factors and therefore, this family of accounts readily predicts the context-sensitive adaptation behavior. Episodic memory accounts and connectionist accounts make no specific predictions about context-sensitive learning but they generally appear to be compatible with context modulating the learning behavior, and for example, [SumnerKataoka2013, Sumner2014] proposed an episodic memory account of speech perception in which they argue that social factors – such as attractiveness in the [Babel2012] study – modulate how strongly individual events are encoded in memory and that differences based on social or other contextual factors may be due to the strength of encoding of individual episodes in memory.

Nature of representations. Another core question, again in particular in the structural priming literature, concerns the nature of the representations involved in partner-specific language processing, and in particular whether priming involves abstract syntactic representations or whether priming is tied to specific lexical representations and therefore limited to utterances with overlapping lexical items. Overall,

there are many studies providing evidence that structural priming for production happens independent of lexical overlap and there is agreement that priming involves abstract syntactic representations (e.g., [Bock1986,PickeringBranigan1998,HartsuikerEtAl2008]).

In comprehension, there have been more conflicting results about whether priming happens independent of lexical overlap or not. [TooleyTraxler2010] reviewed numerous studies of priming effects in comprehension and concluded that priming effects in comprehension "have been less frequently observed in instances where only the syntactic structure is repeated across the prime and target sentences" [TooleyTraxler2010], but subsequently there have been several studies investigating syntactic adaptation (or cumulative structural priming) which found that lexical overlap is not necessary for facilitated comprehension (e.g., [FineJaeger2016]). Further, there have been several studies suggesting that priming is "boosted" by lexical overlap, i.e., priming effects are stronger when some of the lexical items (e.g., the main verb) are shared between the prime and the test sentence (e.g., [SegertEtAl2013,TraxlerTooleyPickering2014,ClelandPickering2006,CorleyScheepers2002,BraniganPickering The partially observed dependence on lexical overlap in comprehension, and the lexical boost effect both suggest that some partner-specific syntactic behavior also depends on lexical representations, which again suggests that multiple mechanisms are involved in partner-specific language processing. This hypothesis is further corroborated by experimental evidence that suggests that the lexical boost effect is short-lived (e.g., [HartsuikerEtAl2008]), and all of this behavior is predicted by the hybrid model by [ReitterEtAl2011] which includes both a long-term syntactic adaptation and a short-term lexical activation component.

The question of the nature of representation has gained less attention at other linguistic levels. At the phonetic level, it is generally assumed that speaker-specific representations are stored at the phoneme-level (e.g., [EisnerMcQueen2005]). At the semantic level, [Yildirim2016] hypothesized that semantic adaptation to variable uses of quantifiers involves both learning speaker-specific semantic representations and learning speakers' productions preferences but they did not systematically investigate this issue.

Generalization to novel speakers. Another important question in the domain of phonetic adaptation has been to what extent listeners generalize and their learned representations of the productions of one speaker or a group of speakers with similar characteristics transfers to novel speakers. [BradlowBent2008] investigated this question in the context of adaptation to Chinese-accented and Slovakian-accented speakers of English. They exposed participants either to a single accented talker or five accented talkers and then tested transcription performance on noisy versions of the speech of a novel accented talker. They found that participants in the single talker condition did not show better transcription performance than participants in a

control condition who were exposed to a non-accented speaker. However, participants in the five-talker condition were more accurate at transcribing the noisy speech of the novel talker, suggesting that exposure to multiple speakers of an accent is necessary to generalize to novel speakers with a similar accent. [XieEarleMyers2018], on the other hand, provided weak evidence of generalization from one Mandarin-accented talker to another and found that generalization was improved if participants slept in between the exposure and test phase. So the question of the amount of exposure (both in terms of number of distinct speakers and experience) has not been fully settled but there is consensus that listeners can generalize to novel talkers.

This behavior is predicted by episodic memory accounts. If a listener hear speech of a novel speaker that is very similar to the speech of a speaker the listener previously interacted with, memory traces of the previous speaker will receive strong activation (due to the similarity) and therefore listeners should behave similarly in response to the novel accented speaker as to the familiar accented speaker(s). The ideal adapter framework also predicts this behavior. [KleinschmidtJaeger2015,Kleinschmidt2018] argue that listeners have structured hierarchical beliefs about the distributions mapping phonemes to phonetic cues, and for example, when hearing a Chinese-accented talker, the beliefs about the current talker as well as higher-level beliefs affecting the beliefs of all Chinese-accented talkers will be updated. Connectionist account make no specific predictions about phonetic adaptation but it seems likely that neural networks would be capable to generalize across talkers. The alignment account does not predict systematic generalization to novel talkers since it does not assume speaker-specific representations.

Generalization to novel speakers at other levels of linguistic representation has not been systematically investigated.

2.1.4 Summary

To summarize this long section, I discussed a large number of studies that show that partner-specific behavior occurs at all linguistic levels and both in production and comprehension. Speakers align their pronunciations, choice of syntactic structures, and choice of referring expressions to their interlocutors; and listeners can learn speaker-specific phonetic representations; can learn in some cases, speaker-specific syntactic preferences; and they can learn speaker-specific interpretations of referring expressions and other lexical items.

What is still less clear is the extent of speaker-specificity, the permanency of alignment and adaptation, the exact properties of the learning process, the nature of the representations, and the extent of generalization across speakers, especially at higher linguistic levels such as semantics and pragmatics.

In terms of models, probabilistic computational models appear to be best at

capturing the majority of comprehension behavior, especially because they predict that listeners form long-term speaker-specific representations and that these speaker-specific representations are systematically structured. However, the shortcoming of these models is that they do not make specific predictions about production, and since they are not concerned with algorithmic or implementational concerns [Marr1982] such as short-term activation of memory or memory decay, they do not make any predictions about transient partner-specific behavior.

At the other end of the spectrum, alignment accounts that are based on the idea of short term activations cannot explain many empirical phenomena, such as adaptation to multiple speakers, long-term priming, or contextually modulated adaptation. At the same time, such accounts can readily explain transient partner-specific behavior such as the lexical boost in structural priming.

As argued, for example, by [FerreiraBock2006] and [ReitterEtAl2011], the advantages and shortcomings of models of persistent learning and transient activation suggest that both processes likely play a role in partner-specific behavior, and therefore hybrid models with a long-term learning component and a short-term activation component are likely best suited for simulating the full range of partner-specific behaviors.

In the subsequent chapters of this dissertation, however, I will only consider comprehension behavior, and I will only focus on cumulative adaptation rather than any short-lived priming effects. I will therefore primarily make use of probabilistic computational models, though I will also discuss the implications of my empirical findings for other types of models. On the empirical side of things, I will investigate the nature of representation in semantic adaptation (Chapter 4); the extent of speaker-specific semantic adaptation (Chapter 5); and whether the learning process is context-sensitive or purely mechanistic (Chapter 7). Before turning to these issues, I will provide more background on the semantics of uncertainty expressions and a computational framework to model comprehension and production of utterances.

2.2 The semantics of uncertainty expressions

While this is not a dissertation about modality, many of the utterances that I consider in my experiments contain an epistemic modal. Since I discuss the implications of my experimental results for popular theories of the semantic of epistemic modals in Section 3.3, and since my computational model is inspired by recent accounts of epistemic modality, I provide a brief introduction to several theories of modality in this section.

Here, and throughout this dissertation, I adapt the broad notion of modality by [Portner2009] and [Kratzer2012Ch2], which not only includes modal auxiliaries (e.g., *might*, *could*) but also other evidential devices such as probability operators (e.g.,

probably) and attitude verbs (e.g., *think*). At the same time, however, to limit the scope of this discussion, I will only cover epistemic modality, and therefore omit any discussion of deontic modals, i.e., modals to express how the world should be according to laws, societal norms, etc., which are frequently discussed together with epistemic modals. I will also omit discussions of the important connections between modals and conditionals [see e.g., Lewis1973?,Kratzer1978,Kratzer1979,Kratzer2012] and discussions of the extent to which different semantic theories validate desired and undesired logical inferences [see e.g., Yalcin2010].

2.2.1   Background: Possible world semantics

Classical modal logic and most other semantic theories of modals are based on the concept of possible worlds [kripke1963]. A possible world is a world which differs in one or multiple properties from the actual world. For example, while the proposition ϕ_{brown} expressed by the sentence "I have brown hair" is true in the actual world w, one possible world w_1 is identical in every regard to the actual world except that the proposition ϕ_{blond} encoded by "I have blond hair" is true and the one encoded by "I have brown hair" (ϕ_{brown}) is false.

According to a possible world semantics, all sentences have to be evaluated relative to a possible world w and propositions ϕ can be represented as a set of worlds in which ϕ is true. If we consider the worlds w and w_1 as described here, the propositions expressed by the sentences "I have brown hair" and "I have blond hair" evaluate to different truth conditions, depending on the possible world.

$$\llbracket \phi_{brown} \rrbracket^{w} = 1 \text{ iff } w \in \phi_{brown} = 1$$

$$\llbracket \phi_{brown} \rrbracket^{w_1} = 1 \text{ iff } w_1 \in \phi_{brown} = 0$$

$$\llbracket \phi_{blond} \rrbracket^{w} = 1 \text{ iff } w \in \phi_{blond} = 0$$

$$\llbracket \phi_{blond} \rrbracket^{w_1} = 1 \text{ iff } w_1 \in \phi_{blond} = 1$$

2.2.2   Modal logic

In classical modal logic, the truth conditions of sentences with epistemic modals depend on an accessibility relation R. R determines which worlds w' are epistemically accessible from the actual world w, i.e., which worlds are epistemically consistent with the actual world. For example, consider rolling two six-sided dice, one after another. Before you roll the first die, all worlds in which the sum of the two dice is between 2 and 12 (all possible combinations of two dice) are epistemically accessible since they are compatible with the actual world. Now, if you roll one of the dice and it comes

up 4, only worlds in which the sum of the two dice is between 5 and 10 (all possible sums of 4 and a number between 1 and 6) are epistemically accessible.

Formally, if wRw' is true then w' is epistemically accessible from w. A proposition ϕ embedded under an epistemic modal is then true if either ϕ is true in all epistemically accessible worlds (for necessity modals such as *must*) or ϕ is true in at least one epistemically accessible world (for possibility modals such as *might*).

(4) $[\![\text{must } \phi]\!]^w = 1$ iff $\forall w' \in W : wRw' \rightarrow [\![\phi]\!]^{w'} = 1$

(5) $[\![\text{might } \phi]\!]^w = 1$ iff $\exists w' \in W : wRw' \rightarrow [\![\phi]\!]^{w'} = 1$

If we again use the example of rolling two dice and assume that the world w_x corresponds to the sum of the two dice being x, then wRw' is true iff $w' \in \{w_2, w_3, ..., w_{11}, w_{12}\}$. Therefore, for example,

$[\![\text{must roll a number between 2 and 12}]\!]^w = 1$

(since *roll a number between 2 and 12* is true in all epistemically accessible worlds)

$[\![\text{must roll a 7}]\!]^w = 0$

(since *roll a 7* is only true in some epistemically accessible worlds)

$[\![\text{might roll a 7}]\!]^w = 1$

(since *roll a 7* is true in the epistemically accessible world w_7)

$[\![\text{might roll a 1}]\!]^w = 0$

(since *roll a 1* is false in all epistemically accessible worlds).

While this approach seems intuitively correct for scenarios like rolling two dice, it is very challenging to represent utterances that convey more fine-grained meanings than mere possibility or necessity. As [Lassiter2017] points out, one could extend this proposal to modal expressions such as *probably* and *likely* by assuming that *probably* ϕ is true if ϕ is true in more epistemically accessible worlds than epistemically accessible worlds in which ϕ is false:

(6) $[\![\text{probably}\phi]\!]^w = 1$
      iff $|\{w' \in W \mid wRw' = 1 \wedge [\![\phi]\!]^{w'} = 1\}| > |\{w' \in W \mid wRw' = 1 \wedge [\![\phi]\!]^{w'} = 0\}|$

However, this proposal comes with at least two shortcomings if one wants to consider it as a complete theory of epistemic modals. First, one has to make the limit assumption [Lewis1981], i.e., one has to assume that W contains a finite number of possible worlds. Second, this proposal does not provide a theory of interpretation for any type

of graded epistemic modal expressions such as *It is 60% likely that...* or *It is highly probable that...* or modal expressions in comparative constructions such as *It is twice as likely that X than Y* [Lassiter2017]. Third, there is no connection between event probabilities and the use of different modals except that this account would predict that *might ϕ* is true when the probability of ϕ is greater than 0, and *must ϕ* is true if the probability of ϕ is 1.

2.2.3   Double relativity of modals

The most prominent semantic theory of epistemic modals (and all other flavors of modals) is the account by Kratzer [Kratzer1981,1991], later revised in [Kratzer2012]. Building on [Lewis1973], she developed a unifying account of all modal flavors, which assumes that there are several core meanings of modals that can be expressed by various linguistic devices (e.g., *could* and *might* are both possibility modals), and that the interpretation of a sentence with a modal depends on two *conversational backgrounds*, that is, contextually specified functions from possible worlds to sets of propositions: the *modal base $f(w)$* and the *ordering source $g(w)$*.

   The intuition behind the modal base $f(w)$ is that one can explicitly state which worlds the modal quantifies over using an "In the view of ..." adverbial clause. For example, for epistemic modals, the modal base may be a function that returns the set of propositions that are compatible with what is knowns in the current world, which can be explicitly expressed in a sentence with an epistemic modal through the adverbial clause "In the view of what is known":

   (7)   In the view of what is known, it could rain tomorrow.

Kratzer argues that utterances without explicit mention of a conversational background are interpreted by contextually resolving the relevant modal base. If we ignore the ordering source for a second, this leads to the following definitions of f-necessity and f-possibility.

   f-necessity:
   ϕ is a necessity with respect to a modal base f iff $\phi \subseteq \cap f(w)$.

   f-possibility:
   ϕ is a possibility with respect to a modal base f iff $\cap\{\{\phi\} \cup f(w)\} \neq \emptyset$.

   The semantics of sentences with necessity modals such as *must* and possibility modals such as *might* can then be expressed in terms of f-necessity and -possibility:

   (8) $[\![\text{must } \phi]\!]^{w,f} = 1$ iff ϕ is an f-necessity in w
   (9) $[\![\text{might } \phi]\!]^{w,f} = 1$ iff ϕ is an f-possibility in w

The semantics of these expressions is equivalent to the classical modal logic semantics presented in (4) and (5), since the accessibility relation R can be defined as $wRw' = w' \in \cap f(w)$. For this reason, this account suffers from the same issues as the classical modal logic account: it cannot be used to derive interpretations for graded modals or comparatives.

Kratzer partially resolves these issues by introducing a second conversational background, the ordering source $g(w)$. $g(w)$ defines a partial preorder $\leq_{g(w)}$ over the set of possible worlds W such that

$$u \leq_{g(w)} v \text{ iff } \{p \in g(w) \mid u \in p\} \subseteq \{p \in g(w) \mid v \in p\}.$$

That means, v is at least as close to an ideal as u iff all propositions in the set of ideal propositions $g(w)$ that are true in u are also true in v. In the case of epistemic modals, the ideal as defined by $g(w)$, is usually assumed to contain propositions corresponding to a normal course of events.

Necessity and possibility then be defined as follows.

> *Necessity*:
> ϕ is a necessity with respect to a modal base f and an ordering source g iff for all $u \in \cap f(w)$, there is a $v \in \cap f(w)$ such that $u \geq_{g(w)} v$ and for all $z \in \cap f(w)$: if $v \geq_{g(w)} z$, then $z \in \phi$.

> *Possibility*:
> ϕ is a possibility with respect to a modal base f and an ordering source g iff $\neg\phi$ is not a necessity with respect to f and g.

Intuitively, the definition of necessity can be seen as further restricting the modal source such that something must be true if it is true in all epistemically accessible worlds that come closest to the ideal defined by the ordering source.

To illustrate how the modal base and the ordering source work together, imagine a murder case in a small town in which there are four suspects A, B, C, and D who all have a motive.[3] Further, there was a tourist T from Iceland in town when the murder happened. Since random tourists rarely murder somebody without a motive, the set of normal propositions $g(w)$ in this example could be $\{a, b, c, d\}$, where each proposition corresponds to A, B, C, and D committing the murder, respectively. However, the conversational background of what is known $f(w)$ is compatible with A, B, C, D, or T being the murderer, i.e., $\cap f(w) = \cup\{a, b, c, d, t\}$. Now, it seems natural for a police officer to utter (10) or (11) but unlikely for the officer to utter (12).

(10)   A might have committed the murder.

(11)   A or B or C or D must have committed the murder.

---

[3]Example adapted from [Kratzer2012Ch2].

(12)   T might have committed the murder.

Kratzer's account makes exactly these predictions. It predicts that (10) is true because $\neg a$ is not a necessity and therefore a is a possbility; it predicts that (11) since $a \vee b \vee c \vee d$ is a necessity; and it predicts that (12) is false since $\neg t = a \vee b \vee c \vee d$ is a necessity and therefore t is not a possibility.

Apart from introducing this distinction between normal courses of events and theoretically possible courses of events, Kratzer's account also provides a semantics for comparatives using the notion of comparative possibility.[4]

> *Comparative possibility*:
> ϕ is at least as good a possibility as ψ in w with respect to a modal base f and an ordering source g iff
>
> $$\neg \exists u (u \in \cap f(w) \wedge u \in \phi - \psi \wedge \forall v ((v \in \cap f(w) \wedge v \in \psi - \phi) \to v <_{g(w)} u))$$

Further, ϕ is a better possibility than ψ iff ϕ is at least as good a possibility as ψ and ψ is not at least as good as possibility as ϕ. Using this definition, [Kratzer1991] defines the semantics of *probably* as

(13) $[\![$probably $\phi]\!]^{w,f,g} = 1$ iff ϕ is a better possibility than $\neg\phi$.

Similarly, the semantics of ϕ *is more likely than* ψ can be defined as

(14) $[\![\phi$ is more likely than $\psi]\!]^{w,f,g} = 1$ iff ϕ is a better possibility than ψ.

As compared to classical modal logic, this proposal has the advantage of providing a semantics for comparative constructions and a semantics for *probably*. However, this account still does not provide a compositional account for modal expressions and therefore does not provide a semantics for expressions such as *very likely*. Second, this account also does not make predictions about the use of epistemic modals to communicate and infer event probabilities. [Kratzer2012] briefly discusses event probabilities and shows that one can come up with probability measures on the set of sets of possible worlds $\mathcal{P}(W)$ such that $\phi \geq_{g(w)} \psi$ implies $P(\phi) \geq P(\psi)$. However, her discussion does not go beyond showing that a connection between event probabilities and possible worlds is possible and her semantic account of modals leaves it open how speakers and listeners map modals to event probabilities.

2.2.4   Threshold semantics

in recent years, there have been several proposals for an alternative semantics of epistemic modals based on the idea that the meaning of epistemic modals is determined

---

[4]This is the revised definition of comparative possibility from in [Kratzer2012Ch2] which is slightly different from the original notion of comparative possibility in [Kratzer1981].

by the position of the probability of the embedded proposition on a probability scale [e.g., Swanson2006,Yalcin2010,Lassiter2017]. The truth condition of utterances with epistemic modals is then determined by whether the probability of the embedded proposition exceeds some threshold θ_x associated with the modal x:

(15) $[\![\text{must } \phi]\!]^w = 1$ iff $P(\phi) > \theta_{must}$ in w

(16) $[\![\text{might } \phi]\!]^w = 1$ iff $P(\phi) > \theta_{might}$ in w

(17) $[\![\text{probably } \phi]\!]^w = 1$ iff $P(\phi) > \theta_{probably}$ in w

This account is inspired by accounts of gradable adjectives [e.g., Kennedy2007] and the observation that many epistemic modals are likely gradable adjectives. For example, [Lassiter2017] argues that *possible*, *probable*, *likely*, and even *certain* all behave in many regards like gradable adjectives: Similarly to gradable adjectives like *tall*, they can be used with degree modifiers (18), and measure phrases (19), and they can be part of comparative clauses (20).

(18)   a. Joan is **very** tall.

        b. It is **very** probable that you'll win the lottery.

(19)   a. Joan is **6ft** tall.

        b. It is **70%** likely that you'll win the lottery.

(20)   a. Joan is tall**er** than Bob.

        b. That you'll get hit by lightning is **more** likely than that you'll win the lottery.

Because of this parallelism, threshold accounts, which have been successfully used to represent the meaning of gradable adjectives, can also be straightforwardly applied to graded or comparative epistemic modals. For example, the utterances in 18b), 19b), and (20b) can be represented as follows.

(21) $[\![\text{very probable } \phi]\!]^w = 1$ iff $P(\phi) \gg \theta_{probable}$ in w

(22) $[\![70\% \text{ likely } \phi]\!]^w = 1$ iff $P(\phi) = .7$ in w

(23) $[\![\phi \text{ more likely than } \psi]\!]^w = 1$ iff $P(\phi) > P(\psi)$ in w

As the choice of examples in this section indicates, the arguments for this account are primarily based on observations concerning epistemic adjectives. Whereas the modal logic account and Kratzer's account focus on the meaning of epistemic auxiliaries and much less on observations concerning epistemic adjectives and adverbs, threshold accounts put epistemic adjectives and adverbs front and center and treat auxiliaries more as an afterthought. Given that auxiliaries are generally not gradable, it might appear surprising to assume a threshold semantics for auxiliaries. However,

beyond the fact that a unified account leads to a simpler theory, there is also the argument that epistemic adjectives and auxiliaries can be used in the same utterance to strengthen the meaning.

(24)   You might win the lottery but you probably won't.

It would be strange to assume that *probably* in this utterance expresses that the probability of not winning the lottery exceeds some threshold but that *might* does not make any reference to an event probability.

One other question this account raises, is how to set the thresholds. For modals such as *probably* that describe event probabilities that are clearly not at either end of the probability space, there is agreement that the threshold is to at least a certain extent context-dependent. For example, in the case of a coin flip with two possible outcomes the threshold will most likely be above .5.

(25)   The coin will probably land on tails.

It would be unexpected for a speaker to produce this utterance if the probability of the coin landing on tails was actually below .5 and it would be more likely for the coin to land on heads. However, when there are more outcomes, the threshold can be lower. For example, [Teigen1988] asked participants 10 days before the 1986 Eurovision Song Contest finals to estimate the chances of winning for 20 contestants using the responses *probable*, *not probable*, and *neither probable nor improbable*. He found that no subject used *probable* only once, which – given that participants could have only assigned a winning probability of more than .5 to at most one of the contestants – indicates that participants also endorsed the statement for event probabilities less than .5. In part, this result was likely driven by the very limited response options in this forced choice production experiment but later work has also found that participants use different uncertainty expressions to describe an event depending on the alternative outcomes and that, for example, likely is sometimes also used to describe event probabilities lower than .5 [e.g., WindshitlandWells1998].

For auxiliaries, on the other hand, there is less agreement about whether the threshold should be context-sensitive. [Yalcin2010] argues for a probabilistic version of existential and universal quantifications such that the thresholds for *might* and *must* are 0 and 1, respectively. [Swanson2006], on the other hand, argues for *might* and *must* being duals, that is that the threshold for *might*, θ_{might}, and the threshold for *must*, θ_{must}, both depend on a single parameter μ such that $\theta_{might} = \mu$ and $\theta_{must} = 1 - \mu$. He further argues for a weak interpretation of epistemic must that implies a threshold below 1, and consequently a $\mu > 0$, which in return also implies that *might* is only true for probabilities greater than $\theta_{might} = \mu > 0$. Based on data from truth value judgement tasks, [Lassiter2017] more generally argues that the thresholds for *might* and *must* are context-sensitive and generally greater than 0

and smaller than 1, respectively. He also considers *might* and *must* as being duals and takes the result that participants sometimes rejected statements with *might* for very low event probabilities as evidence for the possibility of *might* using a threshold greater than 0 and for *must* having a weak interpetation. In summary, there exist three proposals for the thresholds of epistemic auxiliaries: a) treating *might* and *must* as duals and assuming a strong interpretation of *must* [Yalcin2010]; b) treating them as duals and assuming a weak interpretation of *must* [Swanson2006], and c) treating *might* and *must* as duals and assuming context-sensitive thresholds [Lassiter2017]. I will discuss this issue in much more detail in the subsequent chapters.

Finally, we can again ask whether this representation can predict how language users map event probabilities to uncertainty expressions. This account is based on probabilities and therefore does make predictions about this mapping. However, given that uncertainty expressions only have a lower bound, this theory by itself does not predict why a speaker would produce *probably* instead of *might* when there is a high event probability, or why a speaker would produce *certainly* instead of *possibly* when the event probability is 1, since in either case both expressions are true. As has been demonstrated by [HerbstrittandFranke2019] and as I will also show in the next chapter, this issue can be easily reconciled by adding a pragmatic machinery on top of a threshold semantics that predicts that speakers will prefer more informative utterances.

To conclude this section, while the question of how thresholds should be set has not been fully settled, this account has – unlike the two previous accounts that I presented – the advantage of directly linking event probabilities to uncertainty expressions and being able to make predictions about ranges of event probabilities for which different uncertainty expressions can be used.

2.3 Verbal probability expressions

In a line of work that has largely operated in parallel to modality research, there has also been considerable research into how uncertainty expressions or – as they are referred to in this literature – verbal probability expressions are used to communicate probabiklities of future events. The overarching goals of this line of work have been to determine how language users map uncertainty expressions to probabilities and how probabilities communicated through uncertainty expressions shape decision behavior [BudescuWallsten1995].

Initially, this field operated under the assumption that there exists a conventional mapping between uncertainty expressions and probabilities and several works tried to develop methods intended for probing this mapping. For example, [BethMarom1982] tried to infer single point probabilities associated with various uncertainty expressions; [Hamm1991] tried to elicit which ranges of probabilities uncertainty expressions map

to; and [WallstenBudescuRapoportetal.1986] tried to infer membership functions, i.e., functions that assign a value from 0 to 1 to each probability indicating how exemplary for an uncertainty expression a given event probability is. Membership functions are a concept from fuzzy logic and are therefore distinct from probability distributions but membership functions and probability density functions are homomorphous, so one can think of them as representing how likely a probability is communicated by a given uncertainty expression.

While [WallstenBudescuRapoportetal1986] successfully managed to infer membership functions of uncertainty expressions that largely remained constant for individual participants across multiple experimental sessions, they also found that there is considerable inter-subject variability, suggesting that the mapping varies considerably across individual language users and challenging the assumption of a conventional mapping. [WallstenBudescuRapoportetal1986] only speculated about the reasons for this variability but they assumed that this variability was caused by individual differences such as different educational backgrounds or different experience levels with uncertainty expressions.

The finding that there exists considerable variability inspired a lot of follow-up work to identify contextual factors affecting interpretations and productions of uncertainty expressions. These factors include base probability events (e.g., *probable* describing the likelihood of snow in the North Carolina mountains in December is interpreted to describe higher event probabilities than when describing the probability of snow in October [WallstenFillenbaumandCox1986]); outcome severity (more severe consequences lead to inferring higher probabilities [WeberHilton1990]); outcome valence (positive events lead to higher inferred probabilities [MulletRivet1991]); and the alternatives presented in the experiment [Filenbaumetal1991].

Despite this progress in identifying contextual factors, there has not been progress in modeling how contextual factors affect the interpretation of uncertainty expressions nor has there been much work on identifying systematicity in inter-speaker variability. Instead, [Budescuetal2009] declared defeat and argued that uncertainty expressions lead to an "illusion of communication" because they are so vague that speakers and listeners employ different meanings without being aware of these differences (see also [AmerHackenbrackNelson1994,BrunTeigen1998,TeigenBrun1999]). [Budescuetal2009] therefore changed their focus to studying how to best present probabilities in texts as to avoid the illusion of communication. I will argue in subsequent chapters that declaring complete defeat has been premature. While I do not question that there are communicative scenarios in which speakers and listeners employ different meanings and therefore fail to communicate event probabilities properly, I will argue that there are scenarios in which listeners readily update speaker expectations and interpretations of uncertainty expressions, suggesting that accurate communication can be achieved in many scenarios through adaptation.

Setting aside the issue of variability for a moment, we can also ask whether membership functions could be a suitable meaning representation for uncertainty expressions. Membership functions – unlike accounts based on possible worlds – are capable of making predictions about the use of uncertainty expressions to communicate event probabilities, so prima facie they appear to be a good candidate for a meaning representation. However, one challenge for this account is the fact that according to most experimentally inferred membership functions, the meaning of uncertainty expressions is both lower-bounded and upper-bounded, i.e., according to this account an uncertainty expression can only be used if the probability that one wants to communicate is above a certain lower bound α and below a certain upper bound β. For example, if we assume that the membership function of *might* assigns non-zero values to probabilities in the interval $[0.1, 0.5]$, then a speaker can only use *might* to communicate event probabilities above $\alpha = .1$ and below $\beta = .5$. This upper-boundedness then leads to contradictions if the scalar inference from *might* to a stronger alternative such as *certainly* is explicitly canceled, as in (26), and the membership functions of the two uncertainty expressions do not overlap.

(26)   You might get a scholarship, in fact you certainly will.

This utterance communicates an event probability p such that $\alpha_{\text{might}} \leq p \leq \beta_{\text{might}}$ and $\alpha_{\text{certainly}} \leq p \leq \beta_{\text{certainly}}$ which does not exist if the membership functions of these two expressions do not overlap and β_{might} is lower than $\alpha_{\text{certaintly}}$, leading to a contradiction. This indicates that while membership functions are possibly a good representation for the use of uncertainty expressions in many contexts, they are not well-suited as a proper semantic representation.

2.4   The Rational Speech Act framework

The Rational Speech Act (RSA; [GoodmanFrank2016], see also [FrankeJaeger2016]) framework provides a computational-level theory of how speakers choose utterances from a set of alternative utterances and how listeners interpret utterances. The framework is a game-theoretic probabilistic formalization of Gricean [Grice1975] pragmatic reasoning and casts interpretation of utterances as probabilistic Bayesian inference, following recent successes in Bayesian modeling of many cognitive processes (e.g., [TenenbaumGrowingAMind]).

According to an RSA model, two types of agents engage in a communicate game: speaker agents S and listener agents L. The goal of a speaker agent is to communicate a world state w from a set of possible worlds W to a listener by choosing an utterance u from a set of alternative utterances U; the goal of the listener agent is to infer the world state w after hearing the utterance u. Speaker agents and listener agents reason about each other when choosing and interpreting utterances, leading to

Figure 2.1: : Referents in ad-hoc implicature reference game.

a recursive probabilistic reasoning process. Generally, as a base case of this recursive process, the lowest level is represented by a so-called literal listener L_0.[5] The literal listener interprets an utterance probabilistically according to a semantic interpretation function $[\![\cdot]\!]$ of the utterance, resulting in a distribution over world states w given an utterance u.[6]

$$L_0(w \mid u) \propto P(w) \times \begin{cases} 1 & \text{if } w \in [\![u]\!] \\ 0 & \text{otherwise} \end{cases}$$

The literal listener L_0 thus chooses a world state at random from all the worlds in which u is true (as returned by $[\![u]\!]$).

A pragmatic speaker agent S_1 then reasons about this literal listener when choosing their utterance. The formal goal of the pragmatic speaker is to soft-maximize [Luce1959,Sutton1998] the speaker utility $U(w, u)$, which balances how informative the utterances is to L_0 and how costly u is to produce, as indicated by the cost function $c(u)$. Informativity is defined as the negative log surprisal of L_0, resulting in the following utility function U and speaker distribution S_1.

$$U(w, u) = \log L_0(w \mid u) - c(u)$$

$$S_1(u \mid w) \propto \exp\left(\lambda U(w, u)\right)$$

λ is a rationality (or temperature) parameter guiding how likely a speaker is to choose the optimal utterance. Speaker choices are rational if $\lambda > 0$ and choices become increasingly optimal as λ approaches ∞.

At the next level, pragmatic listeners L_1 try to infer the world state that the speaker intended to communicate through Bayesian inference: L_1 integrates their prior beliefs about the state of the world $P(w)$ with the likelihood of the speaker S_1

---

[5] [**TODO: mention cases in which S_0 is the base case.**]

[6] To keep the definitions compact, I omit the normalization terms from all definitions here. Note, however, that all listener and speaker distributions are probability distributions and therefore always have to be normalized to sum to 1.

choosing the observed utterance to communicate w:

$$L_1(w \mid u) \propto P(w) \times S_1(u \mid w)$$

This recursive reasoning process can theoretically be continued ad infinitum by adding additional speaker and listener agents that reason about their respective counterpart agents one level below $i - 1$ the current level i:

$$U_i(w, u) = \log L_i(w \mid u) - c(u)$$

$$S_i \propto \exp\left(\lambda U_{i-1}(w, u)\right)$$

$$L_i(u \mid w) \propto P(w) \times S_i(u \mid w)$$

[**TODO: mention IBR**]

In practice, however, this recursive process is usually capped at the L_1 and L_2 level ([**TODO: add reference. Franke and Degen?**]) and the models that I will consider in this dissertation are also all capped at the L_1 level.

This recursive reasoning process implicitly models a Gricean counterfactual reasoning process in which listeners reason about alternative utterances that a speaker could have produced but didn't to interrpret To illustrate how this works, consider a simple reference game as shown in Figure 2.1.[7] In this game, there are three faces: one without any accessories ($\mathbf{n}$), one with glasses ($\mathbf{g}$), and one with glasses and a hat ($\mathbf{gh}$). A speaker chooses one of the three faces and tries to communicate her choice to the listener through an utterance u. In this contrived example, let us assume that the only three possible utterances that the speaker can choose from are $U = \{$FACE: My friend has a face, GLASSES: My friend has glasses, HAT: My friend has a hat$\}$, and that the listener is aware that the speaker can only choose from these three utterances. Speakers tend to show the following behavior in this game. If they refer to $\mathbf{n}$, they produce FACE, if they refer to $\mathbf{g}$ they produce GLASSES, and if they refer to $\mathbf{gh}$ they produce HAT. Conversely, listeners infer that the speaker likely referred to $\mathbf{n}$ after hearing FACE, to $\mathbf{g}$ after hearing GLASSES, and to $\mathbf{gh}$ after hearing HAT.

An RSA model rooted at a literal listener L_0 with a pragmatic speaker S_1 and a pragmatic listener L_1 captures this behavior. In this game, the semantic interpretation function $\llbracket \cdot \rrbracket$ is defined as:

u	$\llbracket u \rrbracket$
FACE	$\{\mathbf{n}, \mathbf{g}, \mathbf{gh}\}$
GLASSES	$\{\mathbf{g}, \mathbf{gh}\}$
HAT	$\{\mathbf{gh}\}$

[7]Example adapted from [GoodmanFrank2016].

Computing the distributions for $L_0(w \mid u)$ – if we assume uniform priors over the three referents – then results in:

u	$L_0(\mathbf{n} \mid u)$	$L_0(\mathbf{g} \mid u)$	$L_0(\mathbf{gh} \mid u)$
FACE	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
GLASSES	0	$\frac{1}{2}$	$\frac{1}{2}$
HAT	0	0	1

At the two extremes, L_0 thus chooses a face at random after hearing FACE, which is true about all three faces but it will always choose $\mathbf{gh}$ after hearing HAT, which is only true for $\mathbf{gh}$. As I mentioned above, the pragmatic speaker S_1 then chooses an utterance by soft-maximizing their utility, which depends on the informativity of u to communicate w to L_0 and the cost $c(u)$, which I assume to be 0 for all utterances in this example. If we further assume that the rationality parameter $\lambda = 1$, then $S_1(u \mid w)$ is:

w	$S_1(\text{FACE} \mid w)$	$S_1(\text{GLASSES} \mid w)$	$S_1(\text{HAT} \mid w)$
$\mathbf{n}$	1	0	0
$\mathbf{g}$	$\frac{2}{5}$	$\frac{3}{5}$	0
$\mathbf{gh}$	$\frac{2}{11}$	$\frac{3}{11}$	$\frac{6}{11}$

This speaker model qualitatively predicts the behavioral data from forced choice production experiments [GoodmanFrank2016]: for all three referents the most likely utterance choice according to S_1 matches the most likely choice in the experimental data. However, this behavior could have also been predicted by purely symbolic counterfactual reasoning processes as outlined in [Grice1975, Hirschberg1989]. The real advantage of this model comes from the fact that it is a quantitative model and if one fits the rationality parameter λ to the experimental data, one can also quantitatively predict utterance choices in a population of participants.

Finally, the pragmatic listener $L_1(w \mid u)$ can be used to predict the interpretation of utterances in this context. If we again assume that the prior over referents $P(w)$ is uniform, L_1 is:

u	$L_1(\mathbf{n} \mid u)$	$L_1(\mathbf{g} \mid u)$	$L_1(\mathbf{gh} \mid u)$
FACE	$\frac{55}{87}$	$\frac{22}{87}$	$\frac{10}{87}$
GLASSES	0	$\frac{11}{16}$	$\frac{5}{16}$
HAT	0	0	1

This listener model qualitatively predicts the behavioral data from the interpretation experiments, and once again one can quantitatively predict participants' proportions

of choosing the three referents after hearing a given utterance by fitting the rationality parameter λ.

Predicting utterance choices and interpretations in this contrived example of an ad-hoc implicature may not seem particularly impressive. However, many subsequent works following the original RSA model have extended this basic model to predict a wide range of pragmatic phenomena, including predicting scalar inferences [GoodmanStuhlmuehller2013], embedded implicatures [Potts2016], M-implicatures [Bergen2016], metaphors and hyperbole [Kao2013,Kao2014,Kao2015], irony [Kao,CohnGordon2019], the use of gradable adjectives [LassiterGoddman2015, QingFranke2014], generics [Tessler2019], vague quantifiers [SchoellerFranke2017], politeness [Yoon2019], social meaning[Burnett2017], and – as I will discuss in more detail in the next chapter – to predict the use and interpretations of epistemic modals [HerbstrittFranke2019]. While almost all of these models introduce additional extensions to the model as presented here, they crucially all rely on the recursive Bayesian reasoning process outlined in this section, highlighting the versatility of models that cast pragmatic behavior as an instance of Bayesian reasoning.

Chapter 3

Production expectations

In order to study to what extent listeners adapt to specific speakers and the associated cognitive processes, I first had to establish listeners' prior expectations about a generic speaker and to determine to what extent these expectations can be modeled. I therefore conducted a norming study, which served the following theoretical and methodological purposes. First, it served as a methodological check on whether the paradigm is suited for manipulating fine-grained event probabilities. Second, it addressed the theoretical question of whether listeners vary in their expectations about a generic speaker's use of uncertainty expressions, by collecting participants' judgments about uncertainty expressions they expected speakers to use for varying probabilities of receiving gumballs of a particular color from a gumball machine. Third, the results from this study informed the experimental design of the adaptation experiments reported in later chapters, by allowing me to both choose which pair of uncertainty expressions to test adaptation on, and to determine the particular event probability for which participants had roughly equi-probable expectations about which expression of uncertainty a generic speaker would use to report an event with that probability. Lastly, I used the data collected in this study to estimate population-level beliefs for the generic production model reported in Section 3.2, which served as the prior belief model of the adaptation model reported in the next chapter.

Moreover, the work in this chapter not only forms the basis for the experimental and modeling work in subsequent chapters, but independently, it also touches on some of the questions concerning semantic theories of epistemic modals that I discussed in Chapter 2. At the end of this chapter, I therefore also discuss to what extent the results are compatible with existing semantic theories of epistemic modals and what they can tell us about contextually specified parameters.

3.1 Experiment 1: Pre-exposure ratings

3.1.1 Method

Participants

I recruited a total of 420 participants (20 per condition) on Amazon Mechanical Turk. I required participants to have a US-based IP address and a minimal approval rating of 95%. Participants were paid $1.80 (condition 1), $1.50 (conditions 2-15), or $2.00 (conditions 16-21), depending on the number of trials, which amounted to an hourly wage of approximately $12–$15.

Materials and Procedure

This study was a production expectation experiment intended to probe listeners' expectations about a generic speaker's language use. Participants were instructed

Figure 3.1: Example trial in Experiment 1.

that over the course of the experiment, they would see several scenes with an adult man, a young girl, and a gumball machine on a table and that the gumball machine is too high up on the table for the girl to see (see Figure 3.1 for an example scene). After completing an attention check which asked participants whether the girl could see the gumball machine,[1] participants saw a series of scenes and were asked to rate how likely they thought it was that the adult would produce two given responses by distributing 100 points across the two given utterances and the blanket *something else* option (OTHER). Sliders automatically jumped back if participants tried to distribute more than 100 points. In each scene, the child uttered *"I want a blue one"* (target color: blue) or *"I want an orange one"* (target color: orange), randomized across participants.[2] The gumballs in the machines were tossed around continuously to

---

[1]Participants had to go back to the instructions in case they responded incorrectly. This was the case for 41 participants.

[2]In condition 1 (*bare-might*), as well as conditions 16-21 (all conditions with *bare not*), the target color was randomized across trials. While randomization of the target color across trials increased the correlation between the ratings for the two colors, the average ratings for each condition independent of the target color were not affected by this choice. See Appendix A or a detailed discussion of the

prevent participants from counting the gumballs and to make sure that participants did not consider it more likely to get one of the gumballs at the bottom of the machine. In each of the 21 conditions, participants saw only two of the following seven possible adult utterances with different uncertainty expressions:[3]

- You'll get a blue/orange one. (BARE[4])

- You might get a blue/orange one. (MIGHT)

- You'll probably get a blue/orange one. (PROBABLY)

- I think you'll get a blue/orange one. (THINK)

- It looks like you'll get a blue/orange one. (LOOKS LIKE)

- You could get a blue/orange one. (COULD)

- You won't get a blue/orange one. (BARE NOT)

Within each condition, I manipulated the percentage of target color gumballs across trials, which I take as proxy for the objective probability of receiving a gumball of the target color. Each participant saw 3 trials[5] for each of the following percentages: 0%, 10%, 25%, 40%, 50%, 60%, 75%, 90%, 100%. I randomized the order of expressions across participants and trials were presented in randomized order.

3.1.2 Results and Discussion

Figure 3.2 shows participants' ratings for different gumball proportions for 3 of the 21 conditions, namely all combinations of the conditions with the utterances BARE,

---

effect of this manipulation on the ratings.

[3]In the choice of the investigated expressions, I follow recent work on the interpretation of uncertainty expressions [Pogue2018] and aim to use naturalistic utterances. I therefore decided against using a frame such as *"It is UNCERTAINTY-ADJ that"* that would have fixed the syntactic structure across items because such a frame would have resulted in less naturalistic expressions like *"It is possible that"* or *"It is probable that"*, which are much less common than the expressions I considered (e.g., *might* appears more than 30 times more often than *possible that* in the spoken portion of the Contemporary American English [Davies2009]). A speaker's use of these rare expressions thus could have triggered additional pragmatic inferences due to violations of the maxim of manner [Grice1975] that are unrelated to my research questions.

[4]As a notational convention, I refer to utterances with uncertainty expressions in SMALL CAPS and to the uncertainty expression itself in *italics*.

[5]In condition 1 (*bare-might*), participants saw each gumball machine 6 times: 3 times when being asked to produce a statement about orange gumballs and 3 times when being asked to produce a statement about blue gumballs. In conditions 15-20 (all conditions with *bare not*), participants saw each machine 4 times: 2 times for each color.
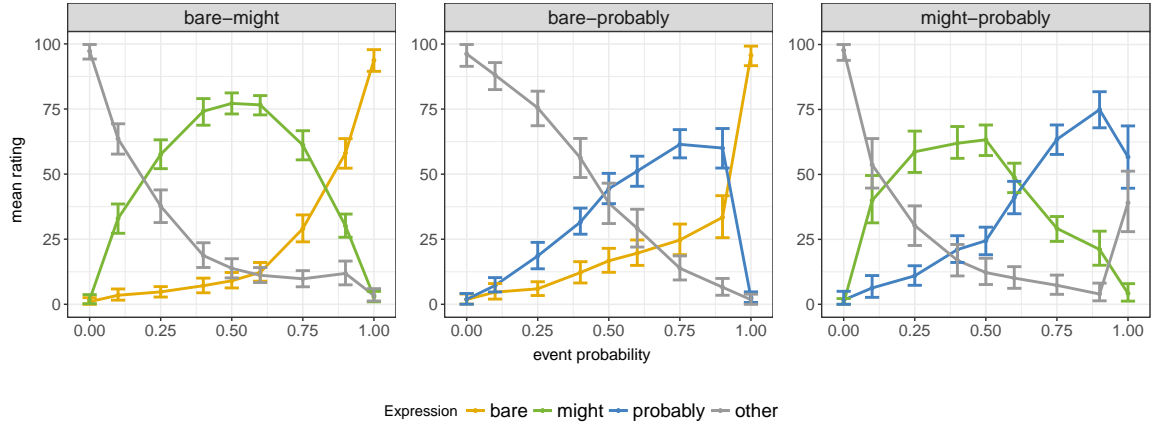
Figure 3.2: Results from 3 conditions of Experiment 1. Error bars correspond to bootstrapped 95%-confidence intervals.

PROBABLY, and MIGHT (see Appendix B for the results from the other 18 conditions). The results from these three conditions highlight several important properties of participants' behavior in this experiment that generalize to all conditions. First, the ratings for individual utterances are influenced by the utterance choices presented to participants. If we compare the ratings for MIGHT in the *bare-might* and the *might-probably* condition, we see that MIGHT received high ratings for a larger range of event probabilities when it is paired with BARE than when it is paired with PROBABLY. We observe similar effects for the other two utterances. This suggests that participants are cued towards using the utterances provided in the experiment and that their ratings depend on the presented alternatives – an effect that has also been observed for frequency expressions [Chase1969] and quantifiers [Degen2016].

Second, the results suggest that participants are sensitive to the different event probabilities and that this paradigm is well suited to study the mapping between event probabilities and uncertainty expressions. For example, in the *might-probably* condition, participants provided considerably different ratings when they were presented with a gumball machine with 50% target color gumballs than when they were presented with 60% target color gumballs.

Third, in all conditions, the mean ratings are graded and except for the 0% and 100% target color gumball trials, the average rating for none of the utterances is close to 100. There are two potential explanations for this observation. It could be that participants provided categorical ratings, i.e., generally assigned 100 points to one of the three options but the category boundaries vary across participants which leads to the graded average ratings. It could also be that participants' individual ratings are graded which could reflect participants' uncertainty about which utterance a speaker
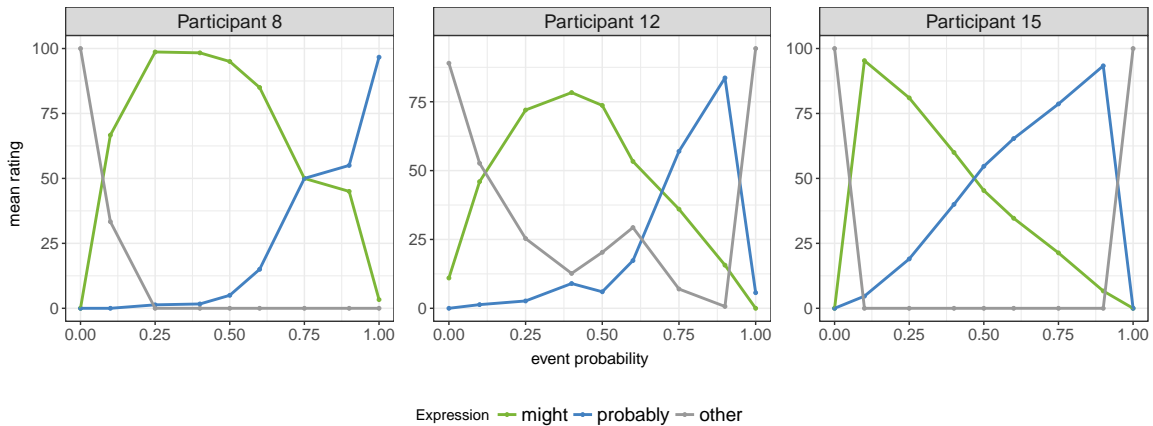
Figure 3.3: Results of three individual participants in the *might-probably* condition of the Experiment 1.

would use and that these individual graded ratings drive the graded average ratings. If we look at individual participants' ratings, it appears to be a combination of both. Figure 3.3 shows the responses of three individual participants in the *might-probably* condition. These figures show that there is a range of gumball proportions for each participant for which they assigned similar ratings to two utterances, which suggests uncertainty about the speaker's utterance choice. At the same time, however, this range also differed across participants: Participant #8, who considered the experimental speaker a "*cautious*" speaker, thought that the speaker would only be likely to use PROBABLY when the objective probability of getting a target color gumball was greater than 0.75, whereas participant #15, who considered the experimental speaker a ' '*confident*" speaker, thought that PROBABLY was a better utterance choice than MIGHT when the objective probability of getting a target color gumball was just greater than 0.5. These observations suggest that for some event probabilities, participants have uncertainty about a speaker's choice of uncertainty expression and that participants have a priori different expectations about how a generic speaker would use these expressions.

This uncertainty and variability seems to be particularly borne out in the *might-probably* condition. For this reason, I chose this pair of expressions to study listeners' adaptation to variable uses of uncertainty expressions.

Aside: The effect of the question under discussion

One potential issue with the experimental design is that participants could consider additional speaker goals beyond being informative. For example, given that the

speaker is talking to a child, it could be that participants assume that he wants to be particularly encouraging and the the choice of his expressions is shaped by both informative and social goals [Yoon2019]. To investigate whether this is the case, I therefore conducted additional exploratory experiments with the *might-probably* expression pair in which I manipulated the question under discussion [QUD; Robers1996] by manipulating the utterance produced by the girl. In the PREFERENCE condition, the speech bubble read *"I want a blue/an orange one"* as in the experiment above. In the DISPREFERENCE condition, the speech bubble instead read *"I **don't** want a blue/an orange one"*, indicating a dispreference by the girl for the color under discussion. In the DISJUNCTION condition, the speech bubble read *"Will I get a blue or an orange one?"*, indicating no preference for one or the other color.

If participants only consider informativity as a goal, there should be no effect of the QUD and the ratings in the DISPREFERENCE and DIJSUNCTION conditions should be identical to the ratings in the PREFERENCE condition. However, if participants expect the speaker to be particularly encouraging in the conditions in which the girl expresses a preference or dispreference for one color, we would expect different ratings depending on the QUD.[6] If the child's preference affects ratings, the ratings in the DISJUNCTION condition should lie between the other two versions of the experiment.

Methods. I recruited 40 participants (20 per version) on Amazon Mechanical Turk. The requirements were the same as in the original experiment and none of the participants participated in the original experiment. Participants were paid $2.50.

The procedure was identical as in conditions 16-21 in the previous experiment and participants completed 36 trials. Across conditions, I manipulated the QUD by changing the utterance in the speech bubble above the child:

- I want a blue/an orange one (PREFERENCE condition)

- I don't want a blue/an orange one (DISPREFERENCE condition)

- Will I get a blue or an orange one? (DISJUNCTION condition)

Results and discussion. Figure 3.4 shows the ratings for the three conditions grouped by uncertainty expression. Overall, the ratings are very similar across conditions and there are no significant differences between conditions, except that the

---

[6]One could imagine the effect going in either direction. It could be that participants expect the speaker to be particularly encouraging and therefore expect the speaker to use the stronger uncertainty expression *probably* for lower event probabilities in the PREFERENCE condition than in the DISPREFERENCE condition. On the other hand, it could be that participants expect the speaker to be scared of disappointing the child and therefore expect him to use *probably* for higher event probabilities in the PREFERENCE condition than in the DISPREFERENCE condition.
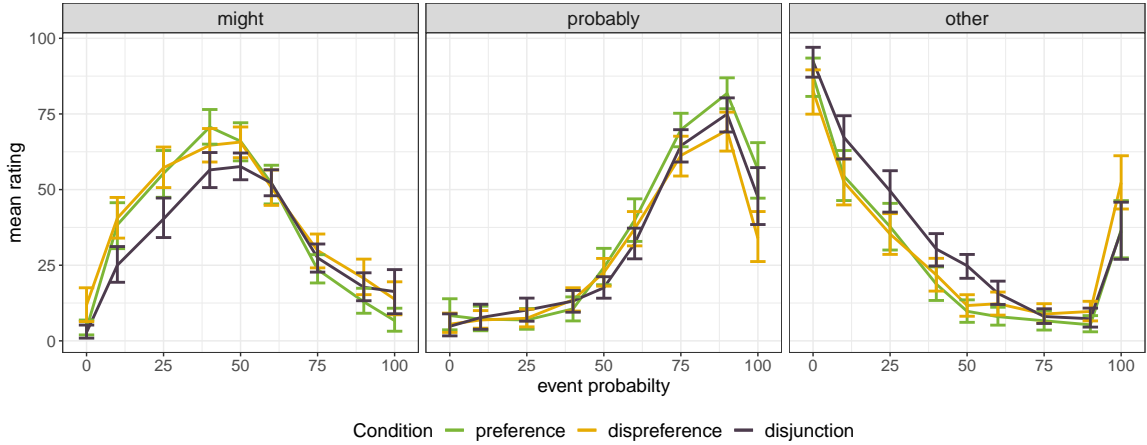
Figure 3.4: Mean ratings for each uncertainty expression with different questions under discussion. Error bars correspond to bootstrapped 95%-confidence intervals.

rating for *probably* is significantly higher in the PREFERENCE condition than in the DISPREFERENCE condition when there is a 100% event probability, and the reverse is true for the *other* utterance. The main takeaway from this experiment is therefore that the QUD has at most very little impact on speaker expectations in this paradigm.

That being said, it is noteworthy that the numerical differences are very systematic and suggest subtle effects of the QUD on speaker expectations. First, unrelated to any social goals, I found that in the 10-50% event probability range, participants in the DISJUNCTION condition consistently rated MIGHT lower than in the other two conditions, and they rated the *other* utterances higher. A likely explanation for this observation is that participants find it more natural to respond with a statement involving the more frequent color if they child asks *"Will I get a blue or an orange one?"*. Therefore participants consider it less likely that the speaker would respond with *"You might get a blue one"* if there is a very low probability of getting a blue gumball. (Recall that the response options always included only one color term; in the PREFERENCE and DISPREFERENCE conditions, the color in the response options was the same as in the child's utterance but in the DISJUNCTION condition, the color in the response options was counterbalanced across trials.) In the other two conditions, the color term was part of the QUD and therefore participants find it more likely that the speaker would respond with MIGHT.

The second noteworthy difference is that on the 100% event probability trials, PROBABLY is rated lower in the DISPREFERENCE condition than in the other two conditions. A possible explanation for this observation is that participants believe that the speaker would try harder to avoid the scalar inference from *probably* to *not certainly* when the outcome is less desired based on the following reasoning: If the

child infers from the use of *probably* that it is not certain that she will get the undesired gumball, she might be somewhat hopeful that she could still get the desired gumball and will be more disappointed when she inevitably gets the undesired gumball. While this explanation is highly speculative, it is consistent with findings that social goals influence interpretations and expectations about productions of expressions that may give rise to scalar inferences [Bonnefon, Yoon2019].

To sum up this aside, these exploratory experimental results suggest that the QUD affects production expectations but importantly only to a very small extent. For this reason, I continue to use the child's request *"I want a blue/an orange one"* in the subsequent adaptation studies, assuming that in this paradigm, social goals have a negligible impact on participants' expectations.

3.2   Modeling expectations about uncertainty expression productions

In this section, I propose a computational model of expectations about uncertainty expression production that is informed by the data from Experiment 1. This model will serve as proxy for listeners' baseline generative model of a generic speaker and will be used as the basis for investigating adaptation processes in the next chapter. Experiment 1 confirmed previous findings that participants' expectations about how a generic speaker would use uncertainty expressions depend on the set of utterances that participants can choose from. I further found that ratings were graded in part because participants had uncertainty about how a generic speaker would use uncertainty expressions. Hence, a model that accounts for participants' beliefs about a speaker's production of uncertainty expressions should (a) be able to capture differences in ratings depending on the availability of alternative utterances; (b) provide graded predictions about utterance probabilities; and (c) be able to capture within-participant uncertainty about probability of use.

Computational game-theoretic models such as the Rational Speech Act framework (RSA; [Goodman2016]) are uniquely suited to satisfy the above desiderata. As I introduced in Chapter 2, RSA models are a probabilistic formalization of Gricean pragmatics which model comprehension as Bayesian probabilistic inference. They consist of *listener* and *speaker* agents which recursively reason about each other to derive interpretations and choose utterances. For my purpose of modeling production expectations, I focus on a model of listener beliefs about a speaker's production model. As common in RSA models, the speaker's utterance utility is determined by trading off the informativity of the utterance to a *literal listener* on the one hand and the cost of the utterance on the other.

In defining the informativity of an utterance with an uncertainty expression, I

follow previous RSA models of uncertainty expressions ([Herbstritt2019]) and assume
that uncertainty expressions have a threshold semantics [Swanson2006,Yalcin2010,Lassiter2016],
as introduced in Chapter 2.[7] Thus, for each uncertainty expression e in the model,
there exists some threshold $\theta_e \in [0, 1]$ such that an utterance u_e with e is true if the
probability ϕ of the proposition embedded under e exceeds θ_e. I base the computation
of informativity on the *literal listener* L_0, a probability distribution from utterances
to event probabilities ϕ.

$$L_0\left(\phi \mid u_e, \theta_e\right) \propto P(\phi) \times \begin{cases} 1 & \text{if } \phi > \theta_e \\ 0 & \text{otherwise} \end{cases} \qquad \text{(for positive embedded propositions)}$$

$$L_0\left(\phi \mid u_e, \theta_e'\right) \propto P(\phi) \times \begin{cases} 1 & \text{if } \phi < \theta_e' \\ 0 & \text{otherwise} \end{cases} \qquad \text{(for negated embedded propositions)}$$

According to this formalization, the literal listener randomly draws an event prob-
ability ϕ from all values above the threshold θ_e associated with the uncertainty ex-
pression e (or, in the case of negated propositions, from all values below the threshold
θ_e'). $P(\phi)$ is a prior distribution over event probabilities, which is independent of the
observed utterance.

The *pragmatic speaker* S_1 that intends to communicate an event probability ϕ
then chooses an utterance u_e with uncertainty expression e from a set of utterances
U such that the speaker's utility is soft-maximized:

$$S_1\left(u_e \mid \phi, \theta, c\right) \propto exp\left(\lambda\left(\log L_0\left(\phi \mid u_e, \theta_e\right) - c(u_e)\right)\right)$$

Recall that λ is a rationality parameter which governs how likely the speaker is to
choose the utterance that maximizes utility.

S_1 crucially depends on a vector of thresholds θ which contains a threshold for
each uncertainty expression in the utterances in U, as well as a cost function $c(u)$.
The values that participants expect a *pragmatic speaker* to assign to these variables
are unknown a priori; we infer these values from the data collected in the above
reported experiment. In Experiment 1, I found that both at the population level
and at the individual level, participants' ratings of the different expressions gradually
increased and decreased with changing event probabilities (as, for example, shown in
Figure 3.3). This is expected if participants have probabilistic beliefs about thresholds
θ (as illustrated in the right panels of Figure 3.5) but not if they reason based on

---

[7]As discussed in Chapter 2, unlike other accounts of epistemic modals, threshold semantics
accounts can directly map uncertainty expressions to probabilities and therefore I did not evaluate
any other semantic accounts.
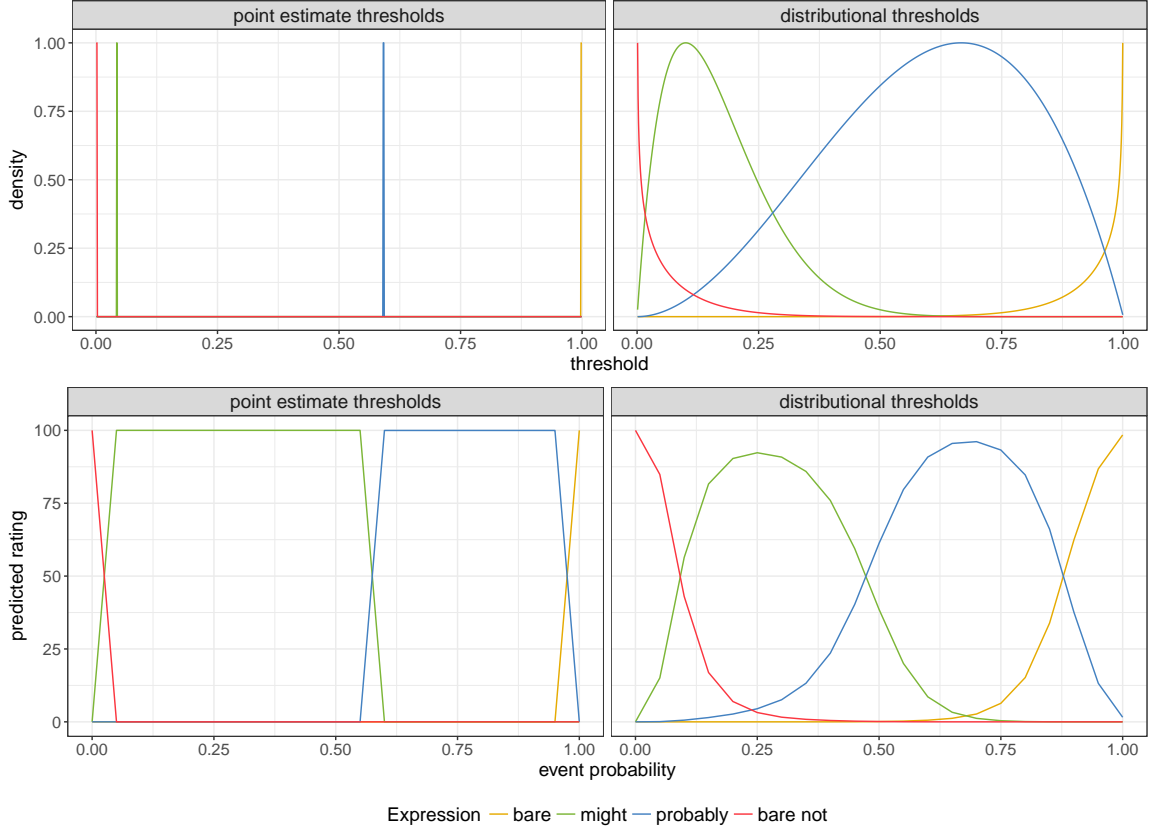
Figure 3.5: Example threshold distributions (upper panels) and corresponding model predictions for the *expected pragmatic speaker* model (lower panels). In this example, the set of possible utterances is $U = \{\text{BARE, MIGHT, PROBABLY, BARE NOT}\}$, all utterances have equal costs, the rationality parameter λ is set to 10, and the prior probability over event probabilities $P(\phi)$ is a uniform distribution. As the panels on the left show, point estimates of thresholds lead to sharp categorical boundaries in the model predictions, whereas distributions over thresholds, as in the panels on the right, lead to gradually increasing and decreasing predicted utterance ratings.

point estimates of θ (as illustrated in the left panels of Figure 3.5). Considering these observations, I assume that listeners hold beliefs about a speaker's thresholds in the form of a distribution $P(\theta_e)$.[8] Analogously, I assume that listeners also have beliefs

---

[8]I leave it open whether a speaker samples from a distribution over thresholds when producing utterances (as suggested by [Qing2015]) or always uses the same values for thresholds. In the former case, listeners could have higher-order beliefs $P(\eta)$ about different speakers' threshold distributions instead of having direct beliefs about the thresholds that different speakers use. For my purposes, this distinction does not matter since I assume that listeners would marginalize over higher-order

$P(c)$ about the speaker model's cost function. Using these two distributions, we can define the *expected pragmatic speaker* model $ES_1 (u_e \mid \phi)$ as follows:

$$ES_1 (u_e \mid \phi) = \int P(c) \int_0^1 P(\theta) S_1 (u_e \mid \phi, \theta, c) \, d\theta \, dc$$

This model predicts which utterance listeners with uncertainty about a speaker's thresholds and cost function would expect that speaker to use to describe different event probabilities. Intuitively, this model can be seen as a weighted average of different *pragmatic speaker* models, where individual models differ in terms of thresholds and cost functions. The weights of this average are determined by listeners' belief distributions over thresholds and costs. For example, if listeners believe that it is likely that a speaker uses a threshold $\theta_{probably}$ of 0.5, they will assign higher weight to speaker models using this threshold than models using other thresholds.

3.2.1   Linking function

I assume that in Experiment 1, participants, when asked to provide ratings for utterances, reasoned about a generic speaker's likely descriptions of varying event probabilities. I assume that this reasoning was guided by participants' beliefs about the speaker's thresholds and costs, and that participants averaged over their uncertainty. For this reason, I assume that the population-level average ratings of what participants expect the speaker to say in different situations correspond to the probabilities predicted by the *expected pragmatic speaker* model (with the *something else* option being predicted by the sum of the probabilities of all utterances not present in a condition; see below for more details). Further, given the forced choice nature of the experiment and that we estimate model parameters from limited and potentially noisy data, I make the following additional linking assumptions for which I provide a rationale and an assessment of their importance in turn.

- **Set of utterances**: Across all conditions, I assume that the set of utterances that participants are considering is the set of all utterances that I used in Experiment 1, i.e., $U = \{$ BARE, MIGHT, PROBABLY, THINK, LOOKS LIKE, COULD, BARE NOT$\}$. I include all utterances instead of only the utterances that are presented in a given condition since I assume that participants' general knowledge of English uncertainty expressions also influences their ratings. Ideally, I would include even more utterances in this set of alternatives but since I can only estimate parameters for uncertainty expressions for which I collected ratings, we are limited to the utterances in U.

beliefs $P(\eta)$ such that $P(\theta_e) = \int P(\eta) P(\theta_e \mid \eta) \, d\eta$ and I therefore take the simplest approach and directly model $P(\theta_e)$.

The exact set of utterances has only a small impact on the model's predictions as long as the set includes the BARE and BARE NOT utterances as well as at least one utterance with a weaker (e.g., *might*) and one with a stronger (e.g., *probably*) uncertainty expression. If this requirement is met, the model makes the same qualitative predictions, and closely matches the quantitative predictions of a model that includes all 7 utterances choices.

- ***something else* option**: Participants in condition $\mathscr{C} = \{u_a, u_b\}$ could only choose between the three utterances $U' = \{u_a, u_b, \text{*something else*}\}$. For modeling data from condition $\mathscr{C}$, I therefore need a function to predict the ratings for the utterances in U'. For u_a and u_b, this is straightforward: I assume the probability of a participant choosing u_a or u_b is proportional to $ES_1(u_a \mid \phi)$ and $ES_1(u_b \mid \phi)$, respectively. I model the probability of a participant choosing the *something else* option as the sum of the probability of all utterances that were not part of the condition as well as a constant O, which accounts for probability mass assigned to utterances that participants might be considering but which are not contained in U. This gives us the following condition-specific function $ES_1^{(\mathscr{C})}$ for predicting participants' ratings.

$$ES_1^{(\mathscr{C})}(u \mid \phi) \propto \begin{cases} ES_1(u \mid \phi) & \text{if } u \in \mathscr{C} \\ O + \sum_{u \notin \mathscr{C}} ES_1(u \mid \phi) & \text{if } u \text{ is *something else*} \end{cases}$$

This summation over alternative utterances is crucial for fitting the data since we need to capture the ratings for *something else*. The only viable alternative would be to fit individual curves for *something else* for each condition, which would require the estimation of considerably more parameters and would not explain the ratings for the *something else* option. The inclusion of the constant O is less important but it still improves model fit.

- **Cost function**: I assume that the cost function represents participants' beliefs about the speaker's preferences for different utterances. Lower costs of an utterance indicate higher speaker preferences. I further assume that we are cueing participants to believe that the speaker would be likely to use the two utterances, u_a and u_b, that are provided in condition $\mathscr{C} = \{u_a, u_b\}$ and that participants therefore primarily use the *something else* option when both of the two utterances are semantically infelicitous or otherwise highly unexpected. I model this cueing effect in my choice of the cost function $c(u)$, which depends on the condition. For the two utterances that are presented to the participants, I set the cost to 1 and for all the other utterances, I set the cost to a constant

$\gamma > 1$:

$$c(u, \mathscr{C}) = \begin{cases} 1 & \text{if } u \in \mathscr{C} \\ \gamma & \text{otherwise} \end{cases}$$

Theoretically, I could have also used a different constant γ_u for each utterance. The data from Experiment 1, however, suggests that participants generally did not prefer one utterance over the other. To limit the number of free model parameters and to prevent overfitting, I therefore use a single constant γ for all utterances. I will, however, relax this assumption in our adaptation model in the next chapter, which I use to investigate whether listeners update their beliefs about preferences during adaptation.

This condition-specific cost function is important for the model fit. If I didn't use such a cost function, the model would assign much higher ratings to the *something else* option than participants did.

- **Noise**: Finally, to account for participants not paying attention or making mistakes, I also include a noise term that models participants providing random ratings. The amount of noise is captured by the noise strength parameter δ. This parameter indicates the proportion of random responses, that is, the proportion of responses drawn from a uniform distribution over the three condition-specific responses U'.

  The inclusion of the noise term is not crucial for fitting the data but it does improve model fit and is common practice in RSA models whose parameters are estimated from experimental data [[]see][]Herbstritt2019,Tessler2019.

Incorporating all of these assumptions, we end up with the following noisy, condition-specific expected pragmatic speaker model $ES_1^{(\mathscr{C})'}(u \mid \phi)$, which I use to predict participants' ratings:

$$ES_1^{(\mathscr{C})'}(u \mid \phi) = \delta \times \frac{1}{|U'|} + (1 - \delta) \times ES_1^{(\mathscr{C})}(u \mid \phi)$$

For the prior distribution over event probabilities $P(\phi)$, which is used in the literal listener model L_0, I use a uniform distribution over the interval $[0, 1]$.[9] For the distributions over thresholds $P(\theta_e)$, I use a Beta distribution parametrized by α_e and β_e. The choice of Beta distributions is motivated by two of its properties. First, the

---

[9]To verify the assumption that the prior on event probabilities is uniform, I conducted a separate norming study in which participants rated how likely they thought it was that a speaker described different gumball machines containing different proportions of blue and orange gumballs after hearing an unintelligible utterance. I found that on average participants rated all gumball machines equally likely which suggests that the prior over event probabilities is indeed uniform.

support of a Beta distribution is the interval $[0, 1]$ which corresponds to the exact range of possible values for θ_e.

The second reason for using Beta distributions is that, depending on the parameterization, Beta distributions can take on very different shapes. This property is important because I am making the simplifying assumption that all utterances in my experiments have a threshold semantics. Such a semantics is commonly assumed for uncertainty expressions such as *probably* [[]e.g.,[]]Yalcin2010,Lassiter2016, but it is unconventional for bare assertions such as *"You'll get a blue one"*, which are generally assumed to be true only if the event is certain to happen, i.e., it has an event probability of 1. However, since Beta distributions can have a shape like the distribution for BARE in the upper right panel in Figure 3.5, the model has the capability to infer a semantics for the bare form that is almost equivalent to a traditional semantics of bare assertions. In this parameterization of the Beta distribution, most probability mass is assigned to values of θ close to 1, which is mathematically almost equivalent to a traditional semantics.[10] Therefore, using Beta distributions for the threshold distributions has the desirable effect of allowing me a unified treatment of all expressions included in the model.

3.2.2 Parameter estimation

Given all the assumptions outlined above, the model has 18 parameters in total: A cost parameter γ, a rationality parameter λ, a noise strength parameter δ, a constant corresponding to other utterances O, and for each utterance, Beta distribution parameters α_e and β_e. I estimated these parameters jointly from all 21 conditions of Experiment 1 using Bayesian data analysis [[]BDA; see, e.g.,[]]Kruschke2015. To construct the dataset, I treated the ratings by each participant as a probability distribution from which I sampled 10 utterances. I used highly uninformative uniform priors over the interval $[0, 15]$ for the Beta distribution and cost parameters, uniform priors over the interval $[0, 7]$ for the rationality parameter, and uniform priors over the interval $[0, 0.5]$ for O and the noise strength parameter. I estimated the vector of parameters Θ using MCMC with a Metropolis Hastings sampler. To decrease autocorrelation of the chain, I collected a sample only at every 10th iteration (i.e., I use thinning of 10). I discarded the first 10,000 burn-in samples and then collected 50,000 samples. I ran four MCMC chains and confirmed convergence by computing the $\hat{R}$-statistic [Gelman2003]. More details on the implementation of the model can

---

[10]Alternatively, one can also see the threshold distribution for the bare form as a distribution over a verification parameter η that governs how certain a speaker has to be to utter a bare assertion [[]see, e.g.,[]]Lewis1976,Potts2007,Davis2007,Moss2018. Mathematically, my assumption of bare forms having a threshold semantics is equivalent to assuming that bare assertions are only true when a speaker's credence of the proposition exceeds the verification threshold η.

be found in Appendix C.

3.2.3   Model evaluation
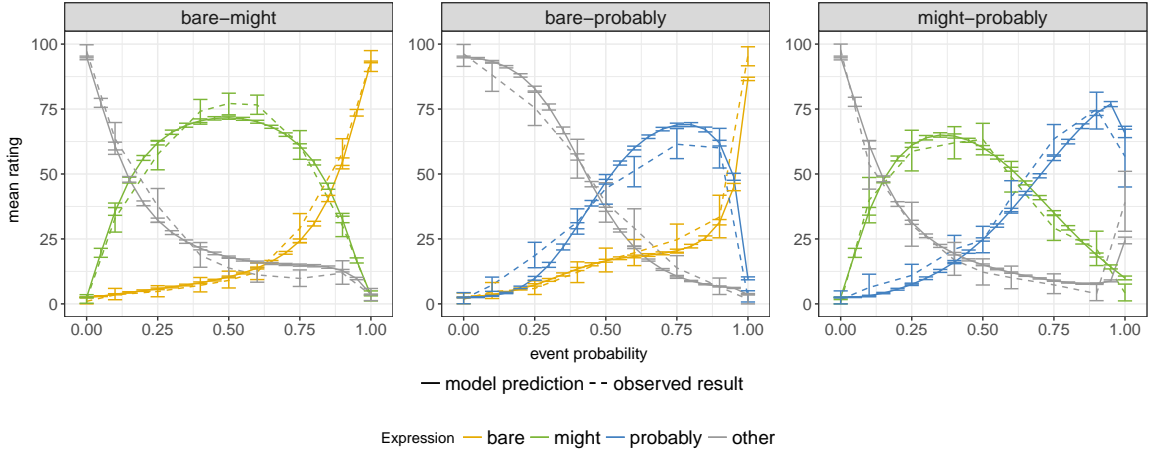
Figure 3.6: Model predictions and results from Experiment 1. Error bars correspond
to 95% high density intervals (model predictions) and bootstrapped 95%-confidence
intervals (observed results).

The result of the parameter estimation procedure is a posterior distribution over
parameters given the observed data $P(\Theta \mid D_{obs})$. I evaluated model fit by perform-
ing a posterior predictive check [[]PPC;][]Kruschke2015. To this end, I took 10,000
samples of parameters Θ from the posterior distribution. For each sample, I com-
puted the model predictions $ES_1^{(\mathscr{C})'}(u \mid \phi)$ parameterized by Θ. I then compared
the average model predictions to the mean ratings that participants had provided
in the pre-exposure experiments. I further computed the 95% high density interval
[[]HDI;][]Kruschke2015 which reflects the certainty of the model about its predictions.

Figure 3.6 shows the model predictions and the experimental data for three con-
ditions (see Table 3.1 and Appendix D for modeling results for all 21 conditions). As
these plots show, the model is able to capture almost the entire variance in partic-
ipants' average ratings. Further, the 95% HDIs are very small, which suggests that
the model is certain about its predictions. Both of these observations are also true
for the model's predictions for all the other conditions. For 19 of the 21 conditions,
the R^2 value between the model predictions and the experimental data exceeds 0.9,
and for the remaining 2 conditions, the R^2 value exceeds 0.88.

Most cases in which the model predictions and the experimental data deviate
concern the ratings at the two extremes of the event probability space. The model
often underpredicts ratings for the *something else* option when there is either a 0%

or a 100% chance of getting a target color gumball. In these situations, participants presumably thought that BARE and BARE NOT are the most appropriate utterances and therefore rate *something else* highly unless we provide them with the BARE or BARE NOT options. The model predicts this behavior to some extent but seems to assume that participants were cued more heavily towards the presented utterance options than they actually were. This could be an indicator that I should revisit my unconventional approach of treating the bare forms like uncertainty expressions with a threshold semantics, since the model would predict higher ratings for *something else* at both ends of the scale if I assumed that the bare form and its negation were only true in the cases of 100% and 0% event probabilities, respectively. However, for the purposes in this dissertation, the exact predictions about production choices for objectively certain events are not as important and hence I decided against revising the assumption that all utterances in the model have a threshold semantics.

The comparably low R^2 for the *probably-think* condition ($R^2 = 0.88$) stems from the fact that participants disagreed on the ordering of these two expressions. This led to a bimodal distribution of average ratings (see the figure in Appendix B) for THINK, which my population-level model cannot fully capture. This suggests that if I wanted to perfectly model participants' production expectations, I should additionally model participant-level differences. However, since the ordering of *probably* and *think* is not of great relevance for my investigation of adaptation to different uses of *might* and *probably*, I did not attempt to fit a more complex model that can account for individual listener differences.

One potential concern given the flexibility of the model is that it could be overfitting the data. This is unlikely considering that all parameters are shared across all conditions and thus I am estimating only 18 parameters to predict in total 567 data points (27 data points for each one of the 21 conditions). To nevertheless rule out the possibility of overfitting, I performed a leave-one-out cross-validation of the model. For each condition x, I estimated a distribution over parameters Θ_x using the data from all conditions but x. I then compared the model predictions of the model parametrized by Θ_x to participants' ratings in condition x. This way, the model has to predict participant behavior which it has not observed during parameter estimation. Table 3.1 shows the R^2 values for participants's ratings and model predictions for the model estimated from all conditions and the leave-one-out models.

As this table shows, the R^2 values remain high[11] even if I exclude the data on

---

[11]I only observed slightly bigger drops in correlations for the *bare not-bare* and the *bare not-could* conditions. In these conditions, I paired expressions that are a poor description for a large range of event likelihoods: BARE NOT and BARE are only expected to be good descriptions if the event probability is very low or very high, respectively, and the *bare not-could* condition includes expressions of which neither are well suited to describe high event probabilities. As a result, participants in these conditions seemed to be more uncertain what to do which was reflected in several participants commenting in a post-experiment survey that they would have answered differently had there

Condition	R^2 (all data)	R^2 (leave-one-out)
bare-might	0.992	0.988
bare-probably	0.978	0.976
bare-could	0.978	0.976
bare-looks like	0.927	0.896
bare-think	0.968	0.964
might-probably	0.964	0.954
might-could	0.921	0.910
might-looks like	0.934	0.918
might-think	0.946	0.934
probably-could	0.961	0.959
probably-looks like	0.944	0.931
probably-think	0.888	0.860
could-looks like	0.924	0.910
could-think	0.931	0.920
looks like-think	0.970	0.960
bare not-bare	0.894	0.848
bare not-might	0.968	0.958
bare not-probably	0.910	0.893
bare not-could	0.910	0.840
bare not-looks like	0.927	0.903
bare not-think	0.933	0.920

Table 3.1: R^2 values for experimental data and model predictions for model estimated from all data and for models estimated from all conditions except the predicted condition.

which the model is evaluated from the model's training data, which suggests that my proposed model indeed explains participants' expectations of a generic speaker's uncertainty expressions.

Lastly, one of the advantages of Bayesian cognitive models is that their parameters are interpretable. Figure 3.7 shows the maximum likelihood estimates of the inferred threshold distributions $P(\theta)$ for the seven uncertainty expressions that we included in my experiments. The first observation is that most threshold distributions have considerable variance rather than being peaked at a particular value. This suggests that listeners have probabilistic beliefs – i.e., uncertainty – about the semantic threshold for each utterance.

---

been better response options. Participants' ratings were therefore presumably more idiosyncratic, which made it harder for the model to predict ratings without having access to the data from these conditions.
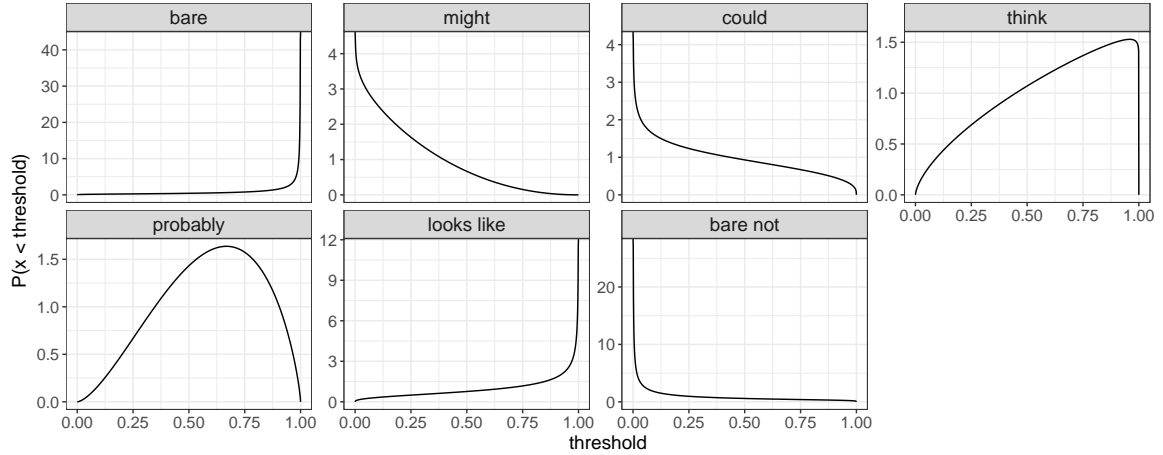
Figure 3.7: Inferred threshold distributions. For the negative bare utterance (BARE NOT), the distribution is over an upper threshold, i.e., a bare statement embedded under negation is true if the probability of the event is lower than the threshold. For all other utterances, the distribution is over a lower threshold.

Further, the inferred threshold distributions are broadly in line with qualitative accounts of the meaning of uncertainty expressions. As discussed above, for the bare form and its negation, we expected the model to infer threshold distributions whose probability mass is concentrated around $\theta = 1$ and $\theta = 0$, respectively.[12] As Figure 3.7 shows, this is indeed what the inferred threshold distributions look like.

To what extent are the other inferred distributions in line with previous formal accounts of uncertainty expressions? The probability mass of the threshold distribution for *might* is concentrated at values slightly above 0. This is–to a large extent– in line with non-probabilistic accounts of the interpretation of epistemic modals, such as the account by [Kratzer2012] that I discussed in Chapter 2. As discussed in the previous chapter, these accounts generally assume that *might p* is true if there exists a world w in a set of (contextually restricted) epistemically accessible worlds W such that p is true in w [[]e.g.,][]Kratzer1991,Swanson2008,Hacquard2011. This logical condition can be translated into a probabilistic framework by assuming that in my gumball machine context, there exists an epistemically accessible world w for each gumball g and that in world w, one gets gumball g. Under this assumption, *"You might get a blue gumball"* is true if there exists an epistemically accessible world w in which one gets a gumball g that is blue. At the same time, if such a world exists, then P(blue gumball) is greater than 0, which approximately corresponds to the threshold semantics with

---

[12]Note that since the negation of the bare form is a negative form, θ is an upper threshold. For the bare form as well as all the other utterances that we are are considering in this paper, θ is a lower threshold.

the inferred threshold distribution of the model. The inferred threshold distribution for *could* is very similar, but not identical, to the one of *might*.

The threshold distribution for *probably* has most of its probability mass concentrated at thresholds above 0.5. This again largely reflects the predictions of existing accounts that assume that *probably p* is true if *p* is more likely than the negation of *p* [[]e.g.,][]Kratzer2012.

The threshold distributions for the remaining expressions, *looks like* and *think* also match intuitions. The distribution for *looks like* has most of its probability mass near threshold values of 1 but is overall slightly weaker, i.e., assigns higher probabilities to lower thresholds, than the bare form. The distribution for *think* assigns most probability mass to high thresholds, which is compatible with the intuition that speakers use *think* when they strongly believe the embedded proposition but are not entirely certain that it is true.

In summary, the inferred threshold distributions are to a large extent compatible with previous accounts of uncertainty expression. However, both the modeling results and the experimental results suggest subtle but theoretically important deviations from predictions of some modality accounts, which I discuss in more detail in Section 3.3.1.

	λ (rationality)	γ (cost)	δ (noise)	O (other utterances)
MAP	2.21	3.03	0.074	3.64×10^{-5}
CI	[2.16, 2.26]	[2.98, 3.08]	[0.069, 0.079]	$[2.89 \times 10^{-5}, 4.50 \times 10^{-5}]$

Table 3.2: Estimated maximum a posteriori estimates (MAP) and 95% credible intervals (CI) for model parameters.

Table 3.2 shows the MAP values and credible intervals for the remaining parameters. The model inferred that speakers are relatively likely to choose an optimal utterance (reflected in the λ parameter being greater than 1); that utterances that are not included in the experiment incur a considerable cost (reflected in the γ parameter being greater than 1); that about 7.4% of the data should be treated as noise (reflected in the δ parameter); and that the production probability of utterances not included in my set of utterances is low.

3.3 General Discussion

As mentioned in the introduction of this chapter, the purpose of the experiments and models reported here was two-fold. On the one hand, I will use the empirical findings as well as the generic speaker expectation model in subsequent chapters to study adaptation. On the other hand, the results touch on several issues that are

relevant for accounts of epistemic modals, and they provide a new perspective on the "illusion of communication".

3.3.1 Implications for semantic theories of modals

The results from the experiments and the models provide several interesting new data points about the use of epistemic modals that can be used to evaluate different theories of epistemic modality. First, the inferred threshold distributions for *might* and *could* differ such that the distribution for *could* assigns higher probabilities to higher values than the distribution for *might*, and both of these distributions assign some probability mass to thresholds greater than 0. Second, the inferred threshold distribution for *probably* assigns some probability to values below .5.[13] Third, both the data and the inferred model parameters suggest that expectations about the use of epistemic modals are graded; for some event probabilities certain modals are more expected to be produced by a generic speaker than for other probabilities, and in only few cases, participants assigned all 100 points to a single utterance Fourth, the expressed expectations varied depending on the presented alternatives. For example, *might* was expected to be used to describe higher event probabilities when it was paired with the bare form than when it was paired with *probably*.

To what extent are these observations compatible with the two most prominent accounts of epistemic modals – Kratzer's [Kratzer2012] account and a threshold semantics account [Lassiter2017,Swanson2006,Yalcin2010] – and to what extent are these findings compatible with the proposal to use membership functions to predict the use of uncertainty expressions [Wallssten]?

Kratzer's account. Kratzer's account makes several assumptions and predictions that are challenged by these findings. Recall that she assumes that epistemic *could* and *might* are semantically equivalent and that utterances with these modals are true if there is a contextually restricted epistemically accessible world in which the embedded proposition is true. Given that the inferred threshold distributions suggest that *could* is more informative than *might*, the assumption of semantic equivalence does not necessarily hold. In her defense, however, if *could* was clearly more informative than *might*, we would not expect that (27) sounds like a contradiction since the same sentence is acceptable if we replace *could* with a considerably more informative expression such as *probably* (28).

---

[13]Since I did not investigate whether participants always correctly perceived the event likelihood to be less than 0.5 for gumball machines with less than 50% target color gumballs, an alternative explanation for this observation would be that participants sometimes overestimated the event likelihood and for this reason, they expected the generic speaker to use *probably* despite an objective event likelihood of less than 0.5.

(27) # You might get a blue gumball but it is not the case that you could get a blue one.

(28) You might get a blue gumball but it is not the case that you'll probably get a blue one.

One hypothesis for the unacceptability of (27) that is compatible with my findings is that the unacceptability is caused by the large overlaps between the expectations about thresholds for *might* and *could*. Since the expectations about threshold for these two expression are very similar and therefore there is a high expectation of them being the same or even the threshold for *could* being lower than *might*, this sentence indeed constitutes a contradiction according to many speaker models and because of that it could seem generally unacceptable. This issue merits more investigation but as it stands, the inferred distributions as well as the experimental data from all conditions that probed expectations about *might*, *could*, or both all suggest that *could* is more informative than *might*, which is not predicted by this account.

The second challenge for Kratzer's account is that *might* and *could* have expected thresholds greater than 0. It seems likely that all worlds in which the girl gets one of the gumballs are among best epistemically accessible worlds and therefore one would expect that as long as the probability of getting a gumball in desired color is greater than 0, *might* and *could* should be true. However, the threshold distributions for both of these utterances suggest that participants only expect these two expressions to be true when the event probability is considerably above 0. Under Kratzer's account it remains unclear why participants have generally very low expectations (or even no expectations) of the speaker to use *might* and *could* when there is a low but non-zero event probability.[14]

Third, the fact that participants assigned non-zero ratings for *probably* to be used to describe event probabilities smaller than .5 and that the inferred threshold distributions consequently also assign probability mass to values less than .5 is also not compatible with the assumption made by Kratzer's account that *probably* ϕ means that ϕ is a better possibility than not ϕ – if the probability of getting a blue gumball is less than .5, then it is more likely to not get a blue gumball than to get a blue gumball.

---

[14]It could be that this is an artifact of the provided alternatives and that participants gave low ratings to *might* and *could* for low event probabilities because they assumed that the speaker would produce more informative utterances involving negation such as *It is unlikely that you'll get a blue/orange one* or *You probably won't get a blue orange one*. However, this would still not explain the asymmetry between *might* and *could*. Further, participants were providing high ratings to utterances that were true but better described by alternative utterances not included in the response set in many other cases, so it seems unlikely that they would have provided low ratings for *might* for very low event probabilities if they had considered it to be likely true utterances in these situations.

Without additional stipulations, Kratzer's account also does not explain the graded production expectations. However, this issue could be salvaged pretty easily by assuming that listeners are uncertain and have probabilistic beliefs about the relevant conversational background of the speaker. Kratzer herself argues for the possibility of listeners considering multiple conversational backgrounds ("More often than not, conversational backgrounds for modals remain genuinely underdetermined and what speakers intend to convey is compatible with several choices of conversational backgrounds." [Kratzer2012,Ch2,p32]) and analogous to the beliefs about thresholds in the model above, one could also imagine that listeners have beliefs about relevant conversational backgrounds.

Finally, if combined with a model of pragmatic reasoning such as RSA, this theory is compatible with the varying ratings across alternatives. One could defined the literal listener L_0 based on the truth-conditions predicted by Kratzer's account and then have the pragmatic machinery as implemented by S_1 predict the dependence on salient alternatives.

Threshold account. Considering that the computational model above is based on a threshold semantics, it is unsurprising that a threshold account is considerably more compatible with the empirical findings. It is out of the box compatible with different and non-zero thresholds for *might* and *could* and it is compatible with a threshold below .5 for *probably*.

As I mentioned in Chapter 2, there has been disagreement about the thresholds for *might* with [Yalcin2010] arguing in accordance with [Kratzer1991] that *might* should have a threshold of 0. My findings here, however, provide additional evidence for [Lassiter2017]'s claim that *might*'s threshold is greater than 0 and context-dependent.

The experimental and modeling results also corroborate the theoretical arguments that *probably* can have a threshold lower than .5. Surprisingly, I found this in a setting in which there are only two possible outcomes unlike the Eurovision Song Contest scenario and lottery scenarios discussed by [Lassiter2017, Yacin2010, Teigen1988], suggesting that multiple outcomes are not necessary for the threshold being lower than .5.

Further, concerning the graded ratings, I had to make the additional stipulation that listeners have probabilistic beliefs about a speaker's thresholds. This further challenges [Yalcin2010]'s assumption of a fixed threshold for *might* but it is compatible with the assumption of context-dependent thresholds that [Lassiter2017] argues for. While he did not spell out a complete pragmatic model of epistemic modals, he does argue that the computation of thresholds could be akin to the RSA-based model of graded adjectives by [LassiterGoodman] which also assumes that there are

probabilistic beliefs about threshold distributions.[15]

Finally, as shown in the experimental results above, this account readily predicts the dependence of the ratings on the available alternatives.

Membership functions.   Lastly, let me turn to the proposal to represent the mapping between uncertainty expressions and event probabilities as membership functions over the interval $[0; 1]$ is compatible with the findings from each condition of the experiment, if we consider each condition separately. One could estimate functions based on the 9 ratings for each uncertainty expression through, for example, fitting splines or polynomial functions. This way one could represent any lower and upper thresholds, and membership functions are out of the box compatible with the graded ratings that I observed in Experiment 1.

However, one challenge for representing the mapping is that membership functions do not depend on the set of alternative utterances. Therefore, this account does not predict why the ratings for the same expression differ across conditions, e.g., why *might* is rated higher for higher event probabilities in the *bare-might* than in the *bare-probably* condition. This issue would also persist if we combined the concept of membership functions with an RSA model such that the membership function would define the literal listener L_0 and production expectations and interprations in context would be predicted by the pragmatic speaker S_1 and the pragmatic listener L_1, respectively. Note that membership functions also have an upper bound and if we set the upper bound for the uncertainty expression e, β_e, to a value less than 1, then the model will predict that e cannot be used to describe event probabilities greater than β_e. Ruling out the production of e for some interval $[\beta_e; 1]$, however, is in tension with the experimental results that suggest that for all the positive utterances that I considered, participants considered the likelihood of a speaker producing the utterance when there was an event probability of 1 to be greater than 0. To account for these results, we would therefore have to set the upper thresholds of all expressions to 1, which reduces the membership function to a fuzzy threshold semantics. While [Wallsten1986]'s proposal does not rule out membership functions of that form, these functions would no longer represent what [Wallsten1986] had intended as they would no longer represent how stereotypical an uncertainty expression is to express a certain event probability.

In conclusion, an evaluation of various semantic theories suggests that the choice of representing the meaning of uncertainty expressions with a threshold semantics

---

[15]I am deviating from the [LassiterGoodman2015] model, though, by assuming that each uncertainty expression as an associated non-flat threshold distribution whereas [LassiterGoodman2015] assume that the threshold distribution is inferred through joint reasoning about the informativity of the utterance and the prior beliefs about the relevant scale.

is compatible with all the empirical findings from Experiment 1, suggesting that when it comes to representing the mapping between uncertainty expressions and event probabilities, a threshold account combined with an RSA model and several additional stipulations makes a number of correct predictions about expected use of uncertainty expression.

3.3.2 Variability and the "illusion of communication"

The results from Experiment 1 also challenge the idea of the "illusion of communication" [Budescu2009] in the use of uncertainty expressions to communicate event probabilities. [AmerHackenbrackNelson1994, BrunTeigen1988, TeigenBrun1999] all argued that listeners are oblivious to the variability of use and are therefore not even considering the possibility that a speaker's use of uncertainty expressions could be different from their own. However, in my experiments, I found that participants exhibited uncertainty and that they considered a lot uncertainty expressions to be possibly used for a large range of event probabilities, suggesting that participants are aware that an unknown generic speaker could use uncertainty expressions differently from how they would use them.

I can only speculate about why my findings suggest the opposite than previous findings but I consider it likely that the experimental task had an effect on participants' behavior. In previous experiments, participants indicated their beliefs about variability by selecting a lower and upper bound. These bounds were supposed to represent the lowest and highest event probability for which any speaker would use a given uncertainty expression and the general finding is that the ranges provided by individual participants are much narrower than the variability observed in interpretations across participants. However, by forcing participants to choose fixed-point lower and upper bounds, participants inevitably had to express their likely probabilistic beliefs about the use of uncertainty expression in a deterministic way and it could be for example, that participants' ranges excluded event probabilities that were unlikely but still possible according to participants' beliefs. In the experiment above, on the other hand, participants had an explicit way of expressing uncertainty and therefore did not have to transform beliefs into a deterministic range, which allowed them to indicate their full range of beliefs.

Looking ahead, while adaptation does not require that listeners are actively aware of existing variability, the finding in Experiment 1 that they are aware of it makes it more likely that listeners would also try to adapt to variable uses in some way. Whether and how listeners do this will be the focus of the next chapter.

3.4 Chapter summary

In this section, I described a computational model of production expectations of uncertainty expressions. This model is couched within the RSA framework and assumes that listeners hold beliefs about a speaker's lexicon (in the form of utterance-specific threshold distributions) and about speaker preferences (in the form of utterance-specific costs). I estimated the free parameters of this model from the results of Experiment 1, which resulted in a model that is able to accurately predict participants' utterance ratings – i.e., their expectations of use – in Experiment 1 across all conditions with a shared set of parameters.

In the following chapters, I will use this model as the basis for modeling adaptation. Since this model is able to capture different beliefs about thresholds and preferences, it provides me with the opportunity to simulate the adaptation process as a result of updating beliefs about these model parameters.

Chapter 4

Adaptation

4.1 Experiment 2: Adaptation of speaker expectations

We now turn to our main research questions of whether and how listeners adapt to variable uses of uncertainty expressions. In Experiment 1, we found that participants show uncertainty in their expectations about a generic speaker's use of *might* and *probably*. Based on these results, we investigate whether participants form speaker-specific expectations about the use of *might* and *probably* after observing a specific speaker's use of uncertainty expressions for a short period of time. The procedure, materials and analyses were pre-registered at `https://osf.io/w926x/`.[1]

4.1.1 Method

Participants

We recruited a total of 80 participants (40 per condition) on Amazon Mechanical Turk. We required participants to have a US-based IP address and a minimal approval rating of 95%. Participants were paid $2.20 which amounted to an hourly wage of approximately $12–$15. None of the participants participated in Experiment 1.

Materials and procedure

Exposure trials: In the first part of the experiment, participants saw 25 exposure trials. These trials had a similar setup as the trials in Experiment 1: they also showed a child requesting a blue or orange gumball and a gumball machine with blue and orange gumballs. However, instead of the cartoon adult, they showed a video of an adult male or female speaker (counterbalanced across participants) producing 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 two conditions (see Table 4.1 for an overview). In the *confident speaker* condition, participants saw 10 critical trials with 60% target color gumballs and the speaker producing an utterance with *probably* (target color was randomized

---

[1]This experiment is a follow-up to a previous experiment with a potential confound due to different number of exposures across conditions. See Appendix E for a discussion of the previous experiment. The qualitative results of both experiments are identical.

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

Table 4.1: Number of exposure trials (n) per utterance (MIGHT, PROBABLY, BARE) and associated proportion of target color gumballs (ϕ) in the *cautious* vs. *confident* speaker conditions in Experiment 2. Critical trials bolded.

across trials), 5 filler trials with 100% target color gumballs and the speaker producing BARE, and 10 filler trials with 25% target color gumballs and the speaker producing MIGHT. In the *cautious speaker* condition, participants saw 10 critical trials with 60% target color gumballs and the speaker producing an utterance with *might*, 5 filler trials with 100% target color gumballs and the speaker producing BARE, and 10 filler trials with 90% target color gumballs and the speaker producing PROBABLY. The filler trials contained utterance-event probability pairs that were rated very highly in the *might-probably* condition of Experiment 1 (see Figure 3.2) and were intended to boost confidence in the speaker.

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 to the next scene, participants had to press a button which was disabled until the video clip had finished.

Test trials: The test phase was almost identical to the *might-probably* condition of Experiment 1 except that the cartoon figure of the man was replaced with a picture of the speaker that participants saw on the exposure trials. Participants were presented with scenes containing gumball machines with 9 different proportions of blue and orange gumballs (identical as in Experiment 1) and they were asked to provide ratings for the utterances MIGHT and PROBABLY by distributing 100 points across these two utterances and the blanket *something else* option. Participants provided two ratings for each of the 18 color-gumball machine combinations resulting in a total of 36 trials.

Both speakers were from the East Coast and Native Speakers of North American English. They were instructed to produce the utterances in a normal voice without any special prosody. The speakers were naïve to the purpose of the experiment.

Attention checks: In order to verify that participants were paying attention to the video and the scenes, we included 15 attention checks (6 during exposure and 9 during test trials), which were randomly positioned within the two experimental phases. Trials that contained an attention check either displayed or did not display

(pseudo-randomized) a small grey X somewhere around the gumball machine. After completing a trial with an attention check, participants were asked whether they had seen a grey X in the previous scenes or not.

Exclusions

We excluded participants who provided incorrect responses to more than 3 of the attention checks. Based on this criterion, we excluded 8 participants in the *cautious speaker* condition, and 7 participants in the *confident speaker* condition. None of the results reported below depend on these exclusions.

Analysis and predictions

Intuitively, we expect a more confident speaker to use lower thresholds for *probably* and *might* than a more cautious speaker. Therefore, if participants track these different uses, we expect their ratings to depend on how the speaker used uncertainty expressions during the exposure phase. Concretely, in our forced choice production paradigm, we expect participants in the *confident speaker* condition to rate PROBABLY highly for a larger range of event probabilities than participants in the *cautious speaker* condition. Following [Yildirim2016], we quantified 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 and participant. The area under the curve is proportional to how highly and for how large of event probabilities participants rate an utterance. If an utterance is rated highly for a larger range of event probabilities, the AUC will also be larger. We therefore tested whether listeners updated their expectations according to these intuitions by computing the difference between the AUC of the spline for MIGHT and of the spline for PROBABLY for each participant. We predicted that the mean AUC difference would be larger in the *cautious speaker* condition than in the *confident speaker* condition.

4.1.2   Results and discussion

The center and right panels of Figure 4.1 show the mean ratings for the three options in the two conditions. As these plots show, participants updated their expectations about the speaker's language use, and therefore made different predictions about how the speaker would use uncertainty expressions as compared to prior expectations elicited in Experiment 1 (left panel). In the *cautious speaker* condition, participants gave high ratings for MIGHT for a larger range of event probabilities than in the *confident speaker* condition. On the other hand, participants gave high ratings for PROBABLY for a larger range of gumball proportions in the *confident speaker* condition than in the *cautious speaker* condition. These differences result in a significantly
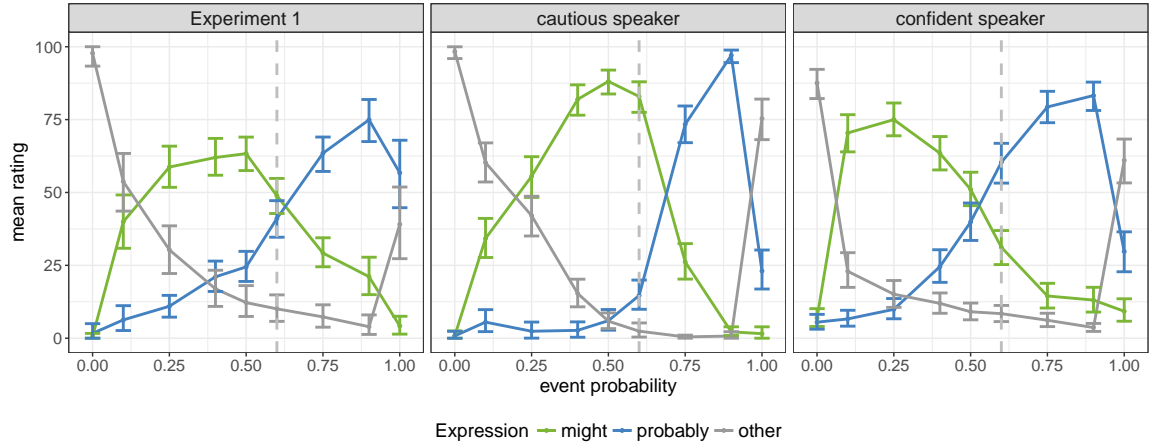
Figure 4.1: Mean ratings for the *might-probably* condition from Experiment 1 (repeated from Figure 3.2) and mean post-exposure ratings from Experiment 2. Error bars correspond to bootstrapped 95%-confidence intervals. The grey dotted line highlights the ratings for the 60% event probability ratings.
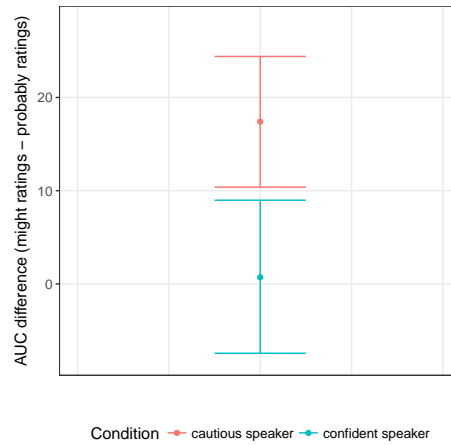

Figure 4.2: Area under the curve (AUC) differences from Experiment 2. Error bars correspond to bootstrapped 95%-confidence intervals.

larger AUC difference in the *cautious speaker* condition than in the *confident speaker* condition ($t(63) = 2.99$, $p < 0.01$, see also Figure 4.2).

As Figure 4.1 shows, participants also differed in their ratings of the two utterances when they were presented with a scene with 60% target color gumballs. In Experiment 1, participants assigned approximately equal ratings to both expressions; in the *cautious speaker* condition, participants rated MIGHT higher than PROBABLY; in the *confident speaker* condition, the pattern was reversed and participants rated PROBABLY higher than MIGHT. These expectations mirror the speaker behavior during the exposure phase and provide additional evidence that participants tracked the speaker's usage of uncertainty expressions.

The results further suggest that participants updated their mappings between uncertainty expressions and event probabilities: In the *confident speaker* condition, MIGHT and PROBABLY were rated highly for lower event probabilities than in the *cautious speaker* condition. This experiment further provides evidence against an account according to which participants only learn that the *cautious speaker* prefers to use MIGHT and the *confident speaker* prefers PROBABLY. Since the frequency of both expressions was the same in this experiment, participants could not have inferred a preference for one of these two utterances.

In sum, the results from this experiment provide evidence for listener adaptation to a specific speaker's use of uncertainty expressions after a brief exposure phase. Further, the results suggest that listeners' expectations about a speaker's language use are at least not exclusively driven by tracking speakers' preferences for different utterances. We investigate the nature of the updated expectations in the next section.

4.2 Adaptation model

The experimental results presented in the previous section suggest that listeners update *some* expectations about language use when they interact with a speaker. However, the nature of the updated representations is unclear. As mentioned in the introduction, there are three likely candidates: first, it is possible that listeners update their expectations about the speaker's lexicon (i.e., the mapping between event probabilities and uncertainty expressions); second, listeners might update their expectations about the speaker's preferences; and third, they might update both their expectations about the speaker's lexicon and about the speaker's preferences. The experimental results above suggest that it is unlikely that listeners track only speaker preferences, but considering that beliefs about preferences and beliefs about the lexicon can interact in complex ways (as illustrated in Figure 1.1), we investigate all three options.

The production expectation model in Section 3.2 provides us with the opportunity to formally evaluate these three hypotheses. Through a series of simulations of the

adaptation process, we can compare models in which different types of parameters are updated during adaptation. Following work in modeling adaptation in other linguistic domains [[]e.g.,[]]Kleinschmidt2012,Kleinschmidt2015,Qing2014,Hawkins2017,Roettger2019, we assume that in interaction, listeners form beliefs about a set of speaker-specific parameters Θ_S.[2] We further assume that the formation of these beliefs is an instance of Bayesian belief updating: listeners start off with prior beliefs about Θ_S based on their general knowledge about language and subsequently update their beliefs about Θ_S with every utterance they hear. That is, after observing a series of productions $D = d_1, ..., d_n$ where each d_i is an utterance-event probability pair $d_i = (u_i, \phi_i)$, listeners' beliefs about Θ_S are the result of performing Bayesian inference:

$$P(\Theta_S \mid D) \propto P(\Theta_S)P(D \mid \Theta_S) = P(\Theta_S) \prod_{i=1}^{n} P(d_i \mid \Theta_S)$$

We assume that the likelihood function is the *expected pragmatic speaker* ES_1 parameterized by Θ_S:

$$P(\Theta_S \mid D) \propto P(\Theta_S) \prod_{i=1}^{n} ES_1(u_i \mid \phi_i, \Theta_S)$$

4.2.1 Simulations

In order to investigate which parameters are updated during adaptation, we ran simulations with varying prior structures, which correspond to different assumptions about which parameters may be updated. The adaptation model crucially relies on a prior over speaker-specific parameters $P(\Theta_S)$ which reflects listeners' prior beliefs about the use of uncertainty expressions. For our simulations, we assumed that the means of this prior are given by the estimates of the model parameters that we obtained from fitting the model to the norming data. The variances reflect whether or not the parameter can be updated in response to exposure. In particular, we used delta distributions, i.e., distributions with zero variance, to model a parameter that cannot be updated. We ran simulations on models with the following three prior

---

[2]Since the manipulation in our experiments was between subjects, our results do not provide direct evidence that listeners are indeed adapting to speakers (as compared, for example, to the general experimental situation). For now, we assume that listeners are adapting to specific speakers and we return to this issue in the general discussion. Further, the speaker-specific parameters might be correlated and listeners might form beliefs about bundles of correlated parameters instead of forming beliefs about individual parameters. For simplicity, we assume here that individual parameters are independent but it would be interesting to investigate whether, for example, listeners who expect a speaker to use lower thresholds for *probably* also expect the same speaker to have lower thresholds for *might* (see also Section 4.2.3).

structures:

- **Costs**: The first prior structure corresponds to an adaptation process according to which participants only learn speaker preferences during adaptation. We modeled the prior over cost parameters as a log-normal distribution centered at the mean value inferred from the norming data. Because we were interested in whether listeners update their beliefs about speaker preferences, we relaxed the assumption from the norming data model that all utterances have the same cost and assumed that each expression has its own cost parameter indicating beliefs about the speaker's preferences.[3] Use of the log-normal distribution was motivated by two reasons: First, cost must be greater than zero, and the support of log-normal distributions is limited to numbers greater than 0. Second, since the cost term is part of an exponent in the expected pragmatic speaker model, differences on a logarithmic scale correspond to linear differences in the model's utterance probabilities. For the priors over all other parameters, we used a delta distribution.

- **Threshold distributions**: This prior structure corresponds to an adaptation process according to which participants only learn speaker-specific threshold distributions during adaptation. We parameterized threshold Beta distributions $P(\theta_e)$ with their mean μ_e and population parameter ν_e [Kruschke2015]. Since the threshold and therefore also the mean of the threshold distribution have to lie within the interval [0,1], we used a truncated normal distribution $\mathscr{N}_{[0,1]}$, which we centered at the mean value from the norming data. For the population parameters ν_e, which indicate how peaked a threshold distribution is, we assumed that distributions can only become more peaked when listeners are exposed to a speaker with very consistent language use and therefore modeled the prior as an exponential distribution shifted to the mean population parameter that we estimated from the norming data. We used a delta distribution for the priors over all other parameters.

- **Threshold distributions and costs**: This prior structure corresponds to an adaptation process according to which participants learn both speaker-specific threshold distributions and speaker preferences during adaptation. We used the log-normal distributions as priors over the cost parameters and the truncated normal and exponential distributions as priors over the threshold distribution parameters, as described above. This means that both the threshold

---

[3]Note that since all updates of parameters happen as part of the belief updating simulations and we are not fitting model parameters to post-exposure ratings, we do not have to be concerned about overfitting due to too many parameters.

	Range	Step size	MAP value
Variance for μ	[0.025,0.25]	0.025	0.175
Scale for ν	[0.5,4.5]	0.5	3.5
Variance for cost	[0.1,1.5]	0.2	0.7

Table 4.2: Explored hyperparameter ranges for variance parameters, and inferred MAP values, which were used in the adaptation simulations.

distributions and the cost parameters could be updated during the adaptation simulations.

Each of these prior structures corresponds to a different hypothesis about which expectations listeners update during adaptation. For comparison, we also considered a baseline in which none of the parameters are updated during adaptation (the *fixed* prior structure). To adjudicate between these three hypotheses, we ran simulations of the adaptation process for both (*cautious speaker* and *confident speaker*) conditions with different prior structures and compared the models in terms of their likelihood of generating the experimental data. During each simulation, we performed Bayesian inference to infer the posterior parameter distribution after observing the 25 data points that participants observed in the exposure phase (see Table 4.1 for an overview of the 25 utterances in the two conditions). We performed inference using MCMC with a Metropolis-Hastings sampler. We used thinning of 10, discarded the first 2,000 burn-in samples and collected 10,000 samples from each of the two chains.

The prior distributions over the different parameters that may be updated during the adaptation simulations are all parameterized by two constants which govern their mean and their variance. The first set of parameters (the mean of the log-normal and truncated normal distributions; the location parameter of the exponential distributions) was given by the estimates from fitting the model to the norming data. The second set of parameters (the variance of the log-normal and truncated normal distributions; the scale parameter of the exponential distributions) was treated as hyperparameters of the simulations. To keep the model as simple as possible, we only used three hyperparameters in total: a variance parameter for the cost for all expressions; a variance parameter for the mean of the threshold distributions for all utterances; and a scale parameter for the prior over population parameters for all utterances. We optimized these three parameters through a Bayesian hyperparameter search on the adaptation data, which provided us with a distribution over hyperparameter values. Considering that each simulation is computationally expensive, we could only test a few hundred hyperparameter combinations, which are listed in Table 4.2. We found that the resulting distributions were highly peaked and therefore, we used only the maximum a posteriori estimates of the hyperparameters (also shown

in Table 4.2) for the model comparisons below.

4.2.2 Model comparisons

We compared model fits via the likelihood of the model generating the post-exposure data. To compute this metric, we constructed a dataset D_{obs} of utterance-event probability pairs by treating each post-exposure rating as a probability distribution and sampling 10 utterances from it. We then computed the posterior likelihood odds between Model 1 with posterior distribution over parameters $P(\Theta_S^{(1)})$ and Model 2 with posterior distribution $P(\Theta_S^{(2)})$.

$$\text{posterior likelihood odds} = \frac{\int_0^1 P\left(\Theta_S^{(1)}\right) P\left(D_{obs} \mid \Theta_S^{(1)}\right) d\Theta_S^{(1)}}{\int_0^1 P\left(\Theta_S^{(2)}\right) P\left(D_{obs} \mid \Theta_S^{(2)}\right) d\Theta_S^{(2)}}$$

The posterior likelihood odds indicate how much more likely it is that the data was generated by Model 1 than by Model 2. Note that we are marginalizing over distributions over parameter values (the integrals in the above formula) and the more parameters we allow to update, the more dispersed the distributions over parameters $P\left(\Theta_S^{(1)}\right)$ and $P\left(\Theta_S^{(2)}\right)$ will be. At the same time, the likelihood terms $P\left(D_{obs} \mid \Theta_S^{(1)}\right)$ and $P\left(D_{obs} \mid \Theta_S^{(2)}\right)$ will only be high for a small range of parameter assignments. Taken together, this means that the posterior odds metric will naturally favor simpler models, because the more dispersed the distributions over parameters are, the less weight is assigned to parameter configurations that yield a high likelihood of generating the data [[]a property often referred to as Bayesian Occam's razor; see, e.g.,]][]MacKay1992,Neal1995.

To evaluate how well the different models predict the empirical post-exposure ratings, we also compute the R^2 value between participants' average post-exposure ratings and the maximum a posteriori predictions of the post-exposure model. However, while R^2 values close to 1 always indicate that model predictions closely match empirical ratings, it is important to note that there are multiple caveats with using R^2 in our setup. First, R^2 is based on the assumption that all predictions are independent and that the residual error follows a normal distribution. These assumptions generally hold for regression models but they are violated for our model predictions since the model predicts a probability distribution over utterances for each event likelihood and therefore predicted utterance ratings are not independent and the residual errors do not follow a normal distribution. Further, unlike the posterior odds, R^2 does not take uncertainty of the model predictions into account but instead compares the mean participant behavior to the mean model predictions, and therefore is not a well-suited

Model	odds	R^2
fixed	10^{-1265}	0.673
cost	10^{-478}	0.817
threshold distributions	10^{-187}	0.885
cost & threshold distributions	1	0.901

Table 4.3: Model evaluation results on data from Experiment 2. *odds* are the posterior likelihood odds of the models compared to the *cost and threshold distributions* model. R^2 are computed between the mean post-exposure ratings and the mean model predictions.

measure for comparing models exhibiting uncertainty.[4] For these reasons, we present R^2 values only to provide intuitions about how well the model predictions match the empirical ratings. For the formal model comparisons we rely on the posterior odds.[5]

Table 4.3 shows the posterior likelihood odds and the R^2 values between the the models and the experimental data from Experiment 2. As the values in this table show, the model in which the cost as well as the threshold distributions are updated during adaptation is much more likely to generate the experimental data than the other two adaptation models as well as the prior model. Further, this model closely predicts participants post-exposure behavior, as indicated by the high R^2 value.

What do these modeling results tell us about the semantic/pragmatic adaptation process? We assumed that each of these simulations correspond to an adaptation process in which different types of expectations are updated. The modeling results corroborate the experimental results from Experiment 2: the models according to which no expectations are updated during adaptation (the *fixed* model) or according to which only preferences are updated (the *cost* model) provide poor predictions for the post-adaptation ratings. The results further clearly indicate that listeners update expectations about the threshold distributions: The two models according to which listeners update threshold distributions were best at predicting post-adaptation behavior. Overall, however, the model that allows updating of preferences and threshold distributions (the *cost & threshold distributions* model) provides the best predictions of post-adaptation ratings, which provides evidence that listeners update expectations

---

[4]If the assumptions of independence and normally distributed residual errors were met, we could address the issue of uncertain model predictions by using a recently proposed Bayesian R^2 metric [Gelman2019], which takes uncertainty of model predictions into account and returns a distribution over R^2 values. However, the computation of this metric crucially relies on estimates of the variance of the residual error σ^2, and it neither makes sense nor is possible in our current framework to estimate this quantity. Hence, we also cannot compute a meaningful Bayesian R^2 distribution.

[5]For all simulations discussed in this and following sections, we find that the R^2 and the posterior odds lead to the same conclusions but in an additional simulation experiment, reported in Appendix F, we encountered a case where the two metrics disagreed.
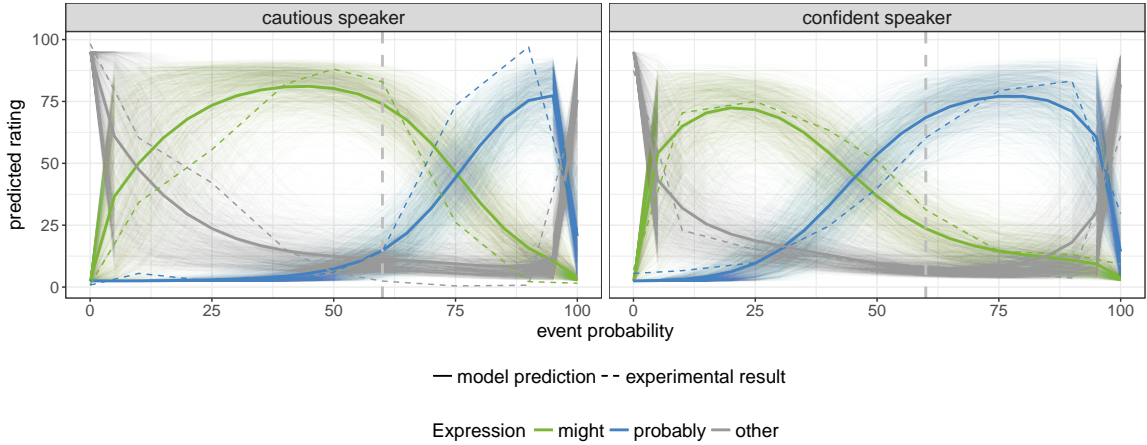
Figure 4.3: Post-adaptation model predictions from simulations for Experiment 2 and experimental results. The solid lines shows the mean model predictions and the thin lines around the mean show the distribution of model predictions.

about both threshold distributions and preferences.

4.2.3 Model evaluation

Apart from quantitatively assessing the fit of the model, it is informative to visually inspect the predictions of the model to verify that the model makes correct qualitative predictions. Figure 4.3 shows the post-exposure predictions of the *cost & threshold distributions* model compared to the average participant ratings for the two conditions from Experiment 2. Qualitatively, the model captures several important patterns in the post-adaptation behavior. The model correctly predicts that in the *cautious speaker* condition, ratings for MIGHT are higher than ratings for PROBABLY when there is an objective probability of 0.6. For the *confident speaker* condition, the model correctly predicts the opposite pattern. The model also predicts that in the *cautious speaker* condition, participants rate MIGHT highly for a larger range of event probabilities than in the *confident speaker* condition and the model predicts the reverse pattern for the PROBABLY ratings. Further, the model predicts that high ratings for MIGHT and PROBABLY are not limited to the utterance-event probability combinations that participants observed during the exposure phase. For example, the model correctly predicts high ratings of MIGHT for low event probabilities in the *cautious speaker* condition despite the fact that it never observed utterances for low event probabilities. Similarly, the model predicts high rating of PROBABLY for high event probabilities in the *confident speaker* condition – a combination which was again not part of the exposure trials of this condition.
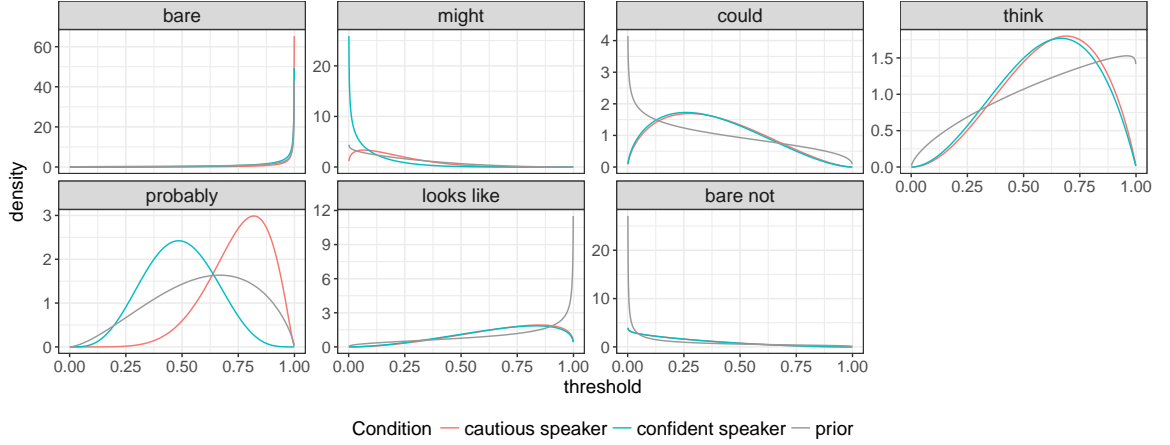
Figure 4.4: Post-adaptation threshold distributions from the simulations for Experiment 2.

Quantitatively, there are some differences between the model predictions and participant behavior. This is not surprising considering that the model predictions are a result of simulations and, with the exception of the three variance parameters of the prior distributions, we did not fit any model parameters to the behavioral data. One difference is that the model underpredicts the ratings of one of the PROBABLY utterances in the *cautious speaker* condition. One reason for this deviation could be the relatively simple prior structure. For the priors, we made the assumption that all model parameters are independent of each other and that the variance for the different parameter types is the same for all utterances. However, it could be that listeners have more structured prior beliefs such that priors over different parameters are correlated or variances of prior distributions differ. For example, it could be that listeners expect the thresholds for *might* and *probably* to be correlated such that higher thresholds for *might* are correlated with higher thresholds for *probably*. Or it could be that listeners expect more between-speaker variation for some expressions than for others. Considering that we only have data from two experiments to test model predictions and therefore would likely overfit to the data if we tried to fit more complex prior structures with more parameters, we leave the investigation of the exact structure of listeners' prior beliefs to future work.

The second noticeable deviation is that in the *confident speaker* condition, the model overpredicts the ratings of the OTHER utterance for event probabilities of 1 . This prediction is primarily driven by high values for the predicted ratings of BARE. However, we argue that the model predictions in this case are reasonable, and that the lower participant ratings are likely an artifact of the experimental task. After completing the experiment, several participants indicated in a feedback form that

they were confused by the lack of an option to rate the BARE utterance, which they had heard during the exposure phase. In light of this confusion, almost all individual participants chose among two strategies when there was an event probability of 1: they either provided a rating of 100 for OTHER or they provided a rating of 100 for PROBABLY, which on average leads to the ratings shown in Figure 4.3.

With the exception of these two deviations, the model makes not only correct qualitative, but also accurate quantitative predictions for the post-exposure ratings.

Lastly, we can also inspect how the model arrived at its predictions by looking at the inferred model parameters. Figures 4.4 and 4.5 show the inferred post-exposure threshold distributions and costs for the two conditions as well as the distributions inferred from the norming data. Figure 4.4 shows that the threshold distribution for *probably* changed considerably depending on the exposure phase: in the *cautious speaker* condition, its mean shifted to a higher value than inferred from the norming data; in the *confident speaker* condition the mean shifted to a lower value. To a lesser extent, we observe similar shifts in the mean of the threshold distributions for *might*. We further observe that for all expressions, the variance of the threshold distributions decreased as a result of adaptation. In the case of the expressions that were part of the exposure phase, this is expected, since the exposure speaker used these expressions very consistently; in the case of the other expressions, this decrease in variance is a result of the exponential prior over the population parameter, which biased the model towards lower variance. For some of the thresholds, this resulted in differently shaped distributions. However, note that the area under the curve of all threshold distributions except for *probably* is still very similar to the area under the curve of the norming data threshold distributions. And overall, except for the distributions for *might* and *probably*, the post-exposure threshold distributions are almost identical in both conditions. This suggests that the post-adaptation expectations are in part a result of updated threshold distributions for *might* and *probably*.

Figure 4.5 shows that the costs of the MIGHT, PROBABLY and BARE utterances, i.e., the three utterances that participants observed during the exposure phase, all decreased while the costs of the other four utterances increased compared to the costs inferred from the norming data. Further, the post-exposure cost of *might* is lower than the cost of *probably* in the *cautious speaker* condition and the opposite relation between these costs holds in the *confident speaker* condition, which suggests that that the post-adaptation expectations are in part also a result of updated beliefs about preferences of MIGHT and PROBABLY. This finding also highlights the complex interplay between threshold distributions and preferences: The number of exposure trials with MIGHT and PROBABLY was identical in both conditions, so participants could not have inferred a preference based on exposure frequencies. Instead, participants seemed to indirectly infer that the *cautious speaker* prefers to use MIGHT from the speaker's uses of MIGHT to describe a larger range of event probabilities and that the
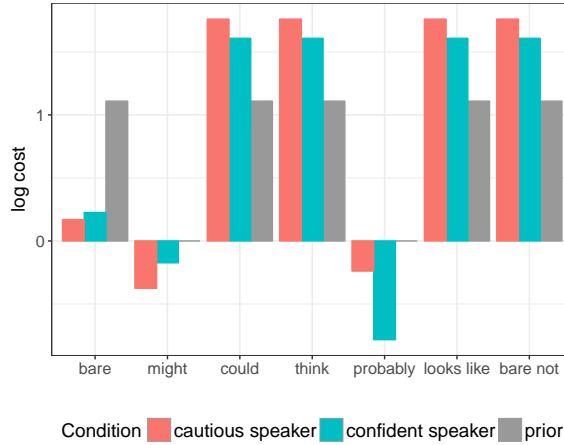
Figure 4.5: Post-adaptation *log* cost values from simulations for Experiment 2. Note that the cost of MIGHT and PROBABLY in the norming data model was 1 and therefore the *log* cost for these utterances is 0.

confident speaker prefers to use PROBABLY from the speaker's uses of PROBABLY for a larger range of event probabilities. However, frequency nevertheless has an effect on inferred preferences. As shown in Appendix F, the difference in inferred preferences was larger when participants in the *cautious speaker* condition were exposed to more trials with MIGHT than in the *confident speaker* condition and participants in the *confident speaker* condition were exposed to more trials with PROBABLY than in the *cautious speaker* condition.

4.2.4 Interim summary

In the previous two sections, we presented the results from an experiment that provides strong evidence for listeners updating expectations about a speaker's use of uncertainty expressions after brief exposure to that speaker. We further presented a computational adaptation model which models the adaptation process as an instance of Bayesian belief updating. Using different implementations of that model to investigate which kind of expectations listeners update during adaptation, we found strong evidence that listeners update beliefs both about threshold distributions and about speaker preferences.

4.3   Experiment 3: Effect of adaptation on interpretation

Up to this point, we focused on listeners' expectations about a speaker's use of uncertainty expressions. As we discussed in the introduction, we expect updated expectations to also have an effect on the *interpretation* of uncertainty expressions. This effect is also predicted by RSA models, which assume that a pragmatic listener L_1 tries to infer the state of the world (in our case, the event probability ϕ) by reasoning about their prior beliefs about the world state and their expectations about a speaker's language use (in our case, the expected pragmatic speaker ES_1) via Bayes' rule:

$$L_1(\phi \mid u_e) \propto P(\phi)ES_1(u_e \mid \phi).$$

According to such a model of interpretation, the shifts in expectations that we observed in the previous experiment should also lead to a shift in interpretations. If we assume a uniform prior over event probabilities,[6] then the model predicts that listeners who were exposed to a *cautious* speaker should infer higher event probabilities when hearing MIGHT or PROBABLY than listeners who were exposed to a *confident* speaker. Figure 4.6 shows the distribution over event probabilities after hearing three different utterances as predicted by L_1 parameterized by the inferred parameters from our adaptation simulations in the previous section. As these plots show, in the *cautious speaker* condition, the distribution over event probabilities after hearing *might* and *probably* is shifted towards higher values as compared to the distributions in the *confident speaker* condition.

In this experiment, we tested whether this prediction is correct and whether listeners' change in expectations transfers to a change in interpretations. The procedure, materials and analyses were pre-registered at `https://osf.io/ghnc3`.[7]

---

[6]To reiterate, this assumption was motivated by the study reported in Footnote 9, which suggested that participants on average assign equal probability to each gumball machine a priori. In this special case, we expect interpretations to be an exact mirror image of production expectations but in many other scenarios, listeners will have skewed prior beliefs about the event likelihood (e.g., consider beliefs about the likelihood of rain on a specific summer day in California, which will be highly skewed towards low values) and in such scenarios, listeners integrate prior beliefs and production expectations to arrive at interpretations.

[7]This experiment is a modified version of a previous experiment, which qualitatively yielded the same results but seemed to confuse some participants. See Appendix G for a discussion of the original experiment.
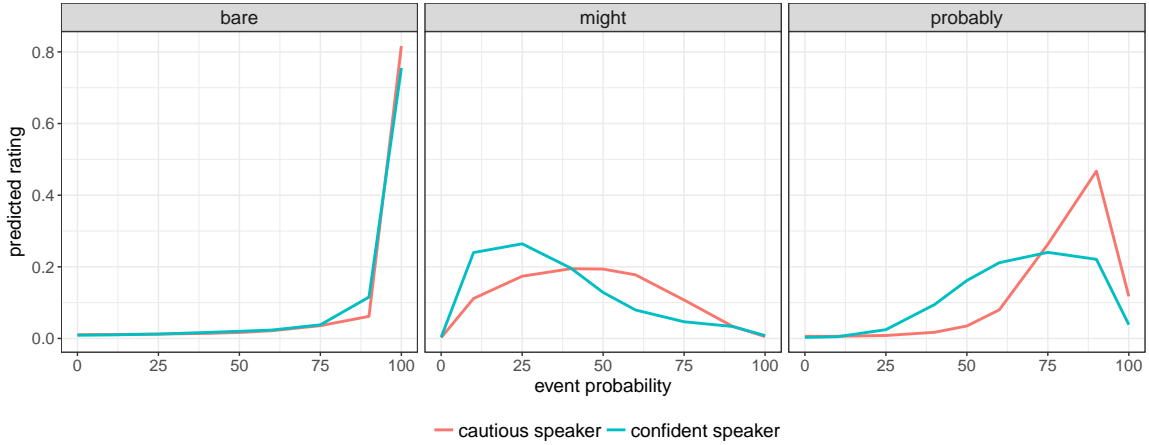
Figure 4.6: Post-adaptation interpretation distributions for the utterances BARE, MIGHT, and PROBABLY as predicted by the pragmatic listener L_1.

4.3.1 Method

Participants

We recruited a total of 80 participants (40 per condition) on Amazon Mechanical Turk. We required participants to have a US-based IP address and a minimal approval rating of 95%. Participants were paid $1.5 which amounted to an hourly wage of approximately $15. None of the participants had participated in any of the previous experiments.

Materials and Procedure

Participants completed a set of exposure trials followed by a set of test trials. The exposure trials were identical to the exposure trials in Experiment 2. The test trials probed participants' interpretations of the utterances MIGHT, PROBABLY and BARE. On each test trial, participants listened to a recording of the speaker from the exposure phase producing MIGHT, PROBABLY and BARE and then participants were asked to rate for 9 gumball machines with the same proportions of blue and orange gumballs as in the previous experiments how likely they thought it was that the speaker saw each of these gumball machines by distributing coins. Participants distributed 10 coins per trial and completed 6 trials in total – one for each expression-color pair. The exposure phase again contained 6 attention check as in the previous experiment. However, given the low attention check performance in the previous experiments, we modified the attention checks. Instead of asking participants whether they saw an X on the previous trial, we asked participants to choose the gumball machine that they
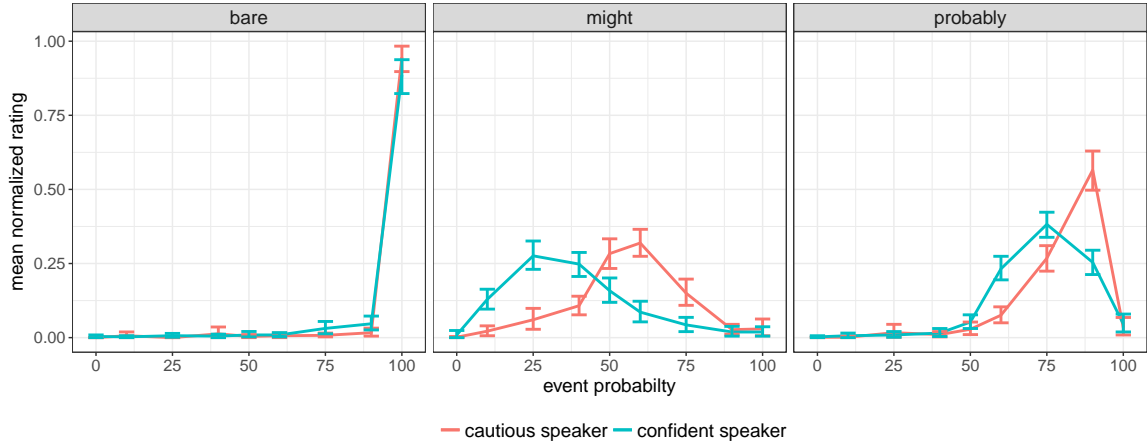
Figure 4.7: Aggregated post-exposure ratings from Experiment 3. Error bars correspond to bootstrapped 95%-confidence intervals.

had seen on the previous trial among two machines displayed in random order.

Exclusions

We excluded participants who failed more than 2 attention checks, which led to 1 exclusion in the *cautious speaker* condition and 1 exclusion in the *confident speaker* condition.

Analysis and Predictions

If participants update their expectations of a specific speaker's use of uncertainty expressions, we expect them to interpret a more confident speaker's utterance to communicate a lower event probability than a more cautious speaker's utterance. We tested this prediction by treating participant's distributions of coins of gumball machines as a probability distribution over gumball proportions (and consequently event probabilities). For each utterance, we normalized participants' coin distributions such that they summed up to 1, so that we could interpret the normalized scores as a categorical probability distribution over gumball machines given an utterance. We then compared the resulting distributions over target color gumball proportions for each utterance in terms of their expected value, using a *t*-test. We predicted that the expected values of MIGHT and PROBABLY would be larger in the *cautious speaker* condition than in the *confident speaker* condition.

Model	odds	R^2
fixed	10^{-452}	0.827
cost	10^{-231}	0.883
threshold distributions	10^{-129}	0.887
cost & threshold distributions	1	0.936

Table 4.4: Model evaluation results on data from Experiment 3. *odds* are the posterior likelihood odds of the models compared to the *cost and threshold distributions* model. R^2 are computed between the mean post-exposure ratings and the mean model predictions.

4.3.2 Results and Discussion

Figure 4.7 shows the aggregated and normalized ratings for the two conditions. As predicted, participants provided higher ratings for gumball machines with higher target color percentages after hearing MIGHT and PROBABLY in the *cautious speaker* condition than in the *confident speaker* condition. This also led to a significantly higher expected value for MIGHT ($t(76) = 5.84$, $p < 0.001$) and PROBABLY ($t(76) = 3.92$, $p < 0.001$) in the *cautious speaker* condition as compared to the *confident speaker* condition.

These results suggest that listeners not only update their expectations about a speaker's use of uncertainty expressions, but also use those updated expectations in interpretation. Here, we made the implicit linking assumption that the distribution of coins reflects participants' beliefs about event likelihoods after hearing an utterance. The choice for this paradigm, which is very similar to betting paradigms that have been used to study utterance interpretation in reference games [Frank2012,Goodman2013] as well as for probing subjective probabilities [[]e.g.,][]Hampton1973, was motivated by the assumption that listeners will have uncertainty about the exact event likelihood after hearing MIGHT or PROBABLY. Allowing participants to assign multiple coins to gumball machines with different proportions provided them with the ability to convey this uncertainty. The results from this experiment suggest that participants behaved as expected: they assigned almost all coins to the gumball machine with 100% target color gumballs after hearing BARE, an utterance about whose interpretation participants likely have very little uncertainty, whereas they assigned coins to multiple machines after hearing MIGHT or PROBABLY.

4.3.3 Model comparison

We return again to our main research question regarding which expectations are updated during adaptation. The production expectation experiments and model simulations provided strong evidence for listeners updating their beliefs about the threshold distributions and preferences. To determine the stability of these results, we also compared the pragmatic listener L_1 predictions from the simulations with different prior structures to the post-exposure ratings in Experiment 3. To this end, we computed the predictions of the L_1 model from the posterior distributions over model parameters that we obtained through the simulations in the previous section. Table 4.4 shows the model fit for the different types of simulations. As this table shows, the model according to which both threshold distributions and costs are updated provides the best fit according to both metrics. Considering that the posterior likelihood odds consistently favored this model in all model comparisons, we take these results together as strong evidence that listeners update their expectations about threshold distributions and costs.

4.3.4 Model evaluation

Figure 4.8 superimposes the model predictions and the experimental data. As these plots show, the model accurately captures most of the qualitative and quantitative patterns. First, the model makes both qualitatively and quantitatively accurate predictions for the interpretation of the BARE utterance in both conditions. Second, the model makes the crucial qualitative prediction that participants expect the speaker to communicate lower event probabilities in the *confident speaker* condition than in the *cautious speaker* condition, which we also observed in Experiment 3. Further, even though we used the parameters that we obtained in the simulations from the previous section and did not fit any parameters to the data from Experiment 3, the model also provides good quantitative predictions of participant's interpretation of MIGHT and PROBABLY, which provides further support for the hypothesis that semantic/pragmatic adaptation is an instance of Bayesian belief updating.

The main deviation between the model predictions and the experimental data lies in the interpretation of MIGHT in the *cautious speaker* condition. For this interpretation, the model predicts a less peaked distribution than the empirical distribution. One explanation for this deviation could be that participants are considering alternative uncertainty expressions (e.g., *very unlikely*) that we did not include in the model. However, since fine-tuning the set of alternative utterances would not change the qualitative predictions of the model and would not provide additional theoretical insights, we leave a more detailed exploration of this issue to future work.
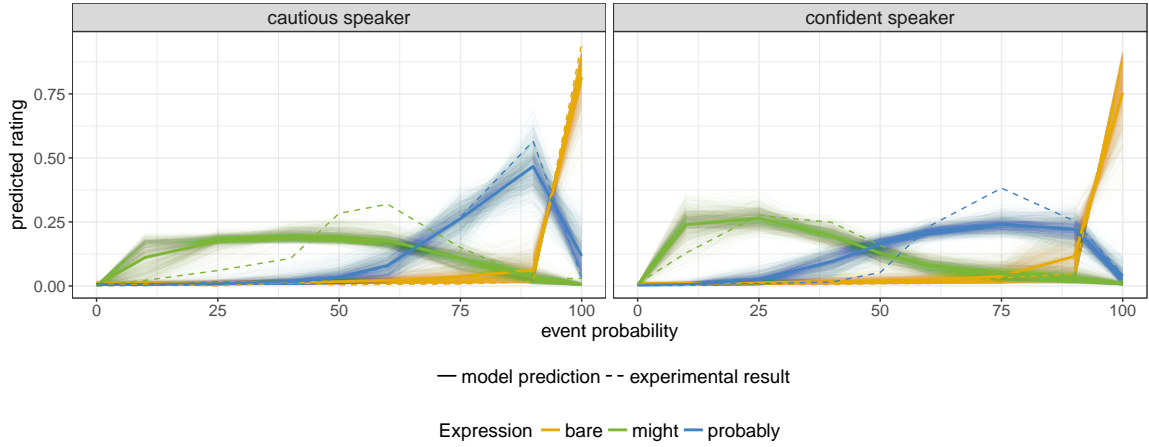
Figure 4.8: Predictions of *threshold distributions and costs* model and data from Experiment 3. The thin lines around the mean show the distribution of model predictions.

4.4 General Discussion

While adaptation in language is a widely attested phenomenon, the nature of the representations that are updated during semantic/pragmatic adaptation has largely remained a mystery. In this paper our contribution has been three-fold:

First, in a production expectation experiment (Experiment 2), we showed that listeners adapt to speakers who vary in their use of uncertainty expressions, demonstrating that that the findings from [Yildirim2016] extend to the class of uncertainty expressions. Second, we showed that updated *production* expectations also affect subsequent utterance *interpretation* (Experiment 3). Finally, in a series of model comparisons, we found strong evidence for listeners updating their beliefs about both the speaker's lexicon as well as the speaker's preferences, which suggests that semantic/pragmatic adaptation is a result of updating both of these representations. We further found that modeling the adaptation process as an instance of Bayesian belief updating explains participants' post-adaptation behavior in both the production expectation (Experiments 2) and comprehension (Experiment 3) experiments.

We next discuss the implications of this work for other accounts of adaptation and for semantic theories of uncertainty expressions, as well as its methodological implications. We then turn to limitations of the current results and account and discuss promising future research avenues this work opens.

4.4.1 Implications for and relation to other accounts of adaptation

The model in this paper is formulated at the computational level [Marr1982,Anderson1990] and is therefore directly only comparable to other computational models. However, we can still assess the compatibility of our findings with mechanistic accounts. We first discuss the relation to existing computational models of adaptation and then discuss what the results tell us about existing mechanistic accounts of adaptation.

The model presented above follows several other computational models of linguistic adaptation that are based on Bayesian belief updating, including models of phonetic adaptation [Kleinschmidt2015], syntactic adaptation [Kleinschmidt2012], adaptation in the interpretation of prosodic cues [Roettger2019], and adaptation to variable use of the quantifiers *some* and *many* [Qing2014]. All of these models are based on the same Bayesian belief updating procedure according to which listeners integrate their prior linguistic expectations and observed linguistic behavior to form updated expectations that facilitate comprehension. This updating procedure is in line with recent proposals of a "Bayesian brain" [[]e.g.,[]]Clark2013,Friston2010 which argue that many cognitive and perceptual processes can be seen as an instance of integrating prior beliefs with observed signals from the environment.

Similarly, the formation of conceptual pacts [Clark1986] – alignment in the formation of referring expressions – has been explained using a model similar to the one we have proposed here [Hawkins2017]. Hawkins et al.'s model is based on the assumption that speakers and listeners have uncertainty about the lexicon [[]see also[]]Bergen2016 and that in interaction, speakers and listeners update their beliefs about the shared lexicon, akin to the updating of threshold distributions in our model. This further suggests that belief updating plays an important role in partner-specific language processing.

In the space of mechanistic accounts, [Pickering2004] argued that a lot of partner-specific linguistic behavior can be explained as the result of priming, i.e., the automatic activation of linguistic representations when a speaker produces an utterance or a listener hears an utterance. Their account has the appeal of explaining why partner-specific language use often appears to happen automatically and effortlessly. With the additional stipulation that the activated representations include information about both language and the situational context and thus are able to represent variable semantics of uncertainty expressions, their account appears to be compatible with our results. However, considering that [Yildirim2016] and [Schuster2019] found that listeners form speaker-specific expectations rather than expecting speakers to behave like the most recent speaker they encountered, adaptation seems to be a more complex process than predicted by a priming account.

In a more recent proposal, [Pickering2013] agued that at least sometimes listeners perform *prediction-by-association* when processing an utterance, that is, listeners make predictions about what the speaker would say based on the context and their experience with the speaker. This appears to be compatible with our computational adaptation model but more details need to be worked out about how such predictions operate at the implementational level [Marr1982].

In a second line of work, [Horton2005,Horton2016] argued that partner-specific language use can be explained by an episodic memory account [Goldinger1998,Johnson1997,Pierrehu... According to this account, individual linguistic events are encoded together with speaker information and the world state in memory, which results in speaker-specific linguistic representations. This account is compatible with our findings, if we assume that individual utterance-world state pairs are stored in memory together with the speaker's identity, and that some additional inference mechanism gives rise to the more complex pragmatic behavior that we observed in our experiments.

4.4.2 Implications for the semantics of uncertainty expressions

Our results also have implications for semantic theories of uncertainty expressions. The finding that listeners rapidly update their beliefs about semantic thresholds of uncertainty expressions suggests that the semantics of these expressions is highly dynamic and context-sensitive. This is broadly compatible with theoretical accounts of probability operators [[]a subset of uncertainty expressions; e.g.,][]Yalcin2010 which state that the meanings of probability operators are highly dynamic and largely determined by the context. Our results suggest that the meaning of uncertainty expressions is even more dynamic than predicted by these accounts. First, we show that this dynamicity extends to a broader set of uncertainty expressions than is typically considered (e.g., *might*), as has been recently also argued by [Lassiter2016]. Second, while these accounts generally assume that the main source of variability in interpretation is the probability of the event embedded under the uncertainty expressions, we find that knowledge of speaker identity also importantly contributes variability.

Dynamic and context-sensitive semantics have also been proposed for many other types of expressions. For example, [Clark1983] argued that speakers and listeners are able to compute novel senses of nouns and verbs on the fly. Similarly, in the domain of gradable adjectives such as *tall*, [Kennedy2007] and many others have argued that the interpretation of these adjectives crucially depends on contextual parameters. Considering the prevalence of dynamic meanings for so many other types of expression, it is therefore not surprising that the interpretation of uncertainty expressions also appears to be highly context-sensitive.

4.4.3 Methodological implications

Our results also have implications for conducting psycholinguistic experiments. First, the finding that listeners adapt to the statistics of their environment within a short experiment suggests that experimenters should be cognizant of potential adaptation effects when probing production expectations or interpretations of uncertainty expressions [[]see also][]Jaeger2010.

Further, the results of Experiment 1, and in particular, the finding that participants' expectations about the use of utterances in the experiment strongly depended on the alternative utterances that we provided, highlights the need to be cautious about drawing general conclusions about expectations of use from single experiments. For example, had we only considered the results from the *bare-might* condition (see Figure 3.2), we might have concluded that "might" is an expected expression to communicate an event probability of 75%, whereas if we had only considered the results from the *might-probably* condition we might have instead concluded that it is *not* an expected expression to communicate an event probability of 75%. This is where explicit modeling of the sort we have engaged in here is hugely helpful: formulating a concrete linking function which models the effects of alternatives allows for inferring the latent meanings of utterances by combining data from different experiments [[]see also][for similar approaches]Franke2014,Peloquin2016.

4.4.4 Limitations and future directions

One limitation of the present research is that the experimental paradigm is not interactive and that participants likely engaged in meta-linguistic reasoning in providing production expectation and interpretation ratings. While we tried to make the communicative situation depicted in the experiments natural, the paradigm is clearly different from everyday dialog. This limitation was necessary for the tight coupling between the experimental work and the model simulations that allowed us to investigate what kind of representations listeners update during adaption; in a more naturalistic and unconstrained setting, we would not have been able to obtain information about listener's production expectations and about their uncertainty in both production expectations and interpretations. However, considering that our task was different from everyday interactions, investigating to what extent the results in the present research translate to less scripted and more interactive settings is an important area for future research. Employing measures like eye movements or mouse-tracking could provide insight into whether participants' updated beliefs affect online language processing, i.e. where meta-linguistic reasoning is unlikely to occur. In this vein, mouse-tracking has recently been employed to investigate the incremental nature of adaptation in the domain of prosodic cues [Roettger2019]. Both eye-tracking and mouse-tracking experiments allow for implementing more natural

interpretation tasks while still providing information about participants' uncertain beliefs via fixation patterns or cursor trajectories.

Throughout this paper, we made the assumption that listeners form *speaker-specific* production expectations. However, since all our experiments had a between-subjects design, it could be that participants were only adapting to the experimental situation, independent of the speaker. This seems unlikely given the results reported by [Yildirim2016], who found that participants formed speaker-specific production expectations after being exposed to multiple speakers whose use of quantifiers differed. Moreover, [Schuster2019] have provided evidence of speaker-specific adaptation to uncertainty expressions. However, exactly which aspects of a situation (e.g., the speaker, the topic of conversation, the visual context, etc.) listeners adapt to is an issue that merits further investigation.

One advantage of formalizing a theory as a computational model is that the model makes concrete predictions to test in future experiments. For example, the proposed model is able to make quantitative predictions about the relation between the number of exposure trials and the size of the adaptation effect. Qualitatively, the model predicts that more exposure should lead to more adaptation, for which some evidence is reported by [Schuster2019]. However, a systematic investigation of whether the model makes the correct quantitative predictions remains to be conducted.

Further, the presented adaptation model is built around the assumption that the utility of an utterance is exclusively determined by the informativeness and the cost of the utterance. However, it has been observed that other speaker goals such as being polite or convincing could also factor into the interpretation of uncertainty expressions [[]see e.g.,][]Pighin2011,Juanchich2013,Holtgraves2016. It could therefore be that, for example, listeners explain away the behavior of a "confident" speaker if the context suggests that the speaker has an incentive to be encouraging or has additional goals besides being informative [[]see also][]Yoon2016,Yoon2017. Investigating whether listeners draw such complex inferences could provide insight about which kind of potential speaker goals enter into listeners' pragmatic reasoning process.

4.4.5   Conclusion

We began with the puzzle of how to reconcile the assumption of stable utterance alternatives required for pragmatic reasoning with variability in speakers' language use. The work reported here, building on much previous work on adaptation, suggests that this apparent tension is easily resolved if listeners form speaker-specific utterance expectations that they can recruit when interpreting utterances produced by that speaker.

In a series of web-based experiments, we found that after exposure to a specific

speaker, listeners rapidly update their expectations about which uncertainty expressions that speaker is likely to produce to describe varying event probabilities. Moreover, these updated expectations also enter into subsequent utterance interpretation. We provided a formal account of semantic/pragmatic adaptation and modeled this behavior using a Bayesian cognitive model which assumes that (listeners reason about) speakers (who) efficiently trade off utterance informativeness and cost. Through a series of simulations we found strong evidence that semantic/pragmatic adaptation is the result of updating beliefs about both a specific speaker's underlying lexicon as well as the speaker's utterance preferences. These results both provide new insights into the cognitive processes involved in semantic/pragmatic adaptation and demonstrate that tight coupling of quantitative computational models and experimental results can shed light on unobservable representations.

Chapter 5

Speaker-specific adaptation

5.1   Introduction

Speakers exhibit considerable production variability at all levels of linguistic representation [¡]e.g.,¿Liberman1967, Weiner1983,Finegan2001. This includes variation in lexical choice to describe a world state. For example, [Yildirim2016] 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.

[Schuster2018] 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.,¿Norris2003,Kraljic2007,Bradlow2008, Kamide2012,Kleinschmidt2015,Fine2016,Roettger2018 and that in interaction, they learn how speakers choose among alternative utterances. In the domain of quantifiers, [Yildirim2016] 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, [Schuster2018] 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. [Metzing2003] 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, [Yildirim2016] 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, [Kraljic2007] 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, [Yildirim2016] 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 [Horton2005] 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 [Schuster2018] 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.

5.2 Experimental paradigm

In both of our experiments, we build upon the semantic adaptation paradigm used in [Schuster2018], 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.,¿Norris2003,Kleinschmidt2015,Yildirim2016. As shown in Figure 5.1, 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 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. 5.1), 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.

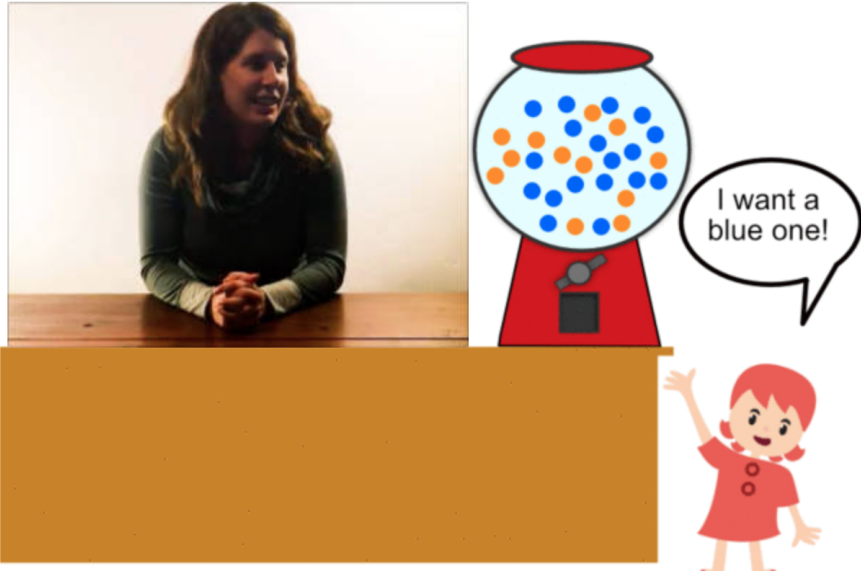## 5.3 Experiment 1: Different speaker types

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).

5.3.1 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).

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

Figure 5.1: 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.

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

Table 5.1: 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.

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 5.1 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 [Yildirim2016] and [Schuster2018], 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 [Yildirim2016], 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.

5.3.2   Results and discussion

Figure 5.2 shows the mean utterance ratings of participants grouped by the two post-exposure test blocks. As this plot shows, participants expected the *confident*
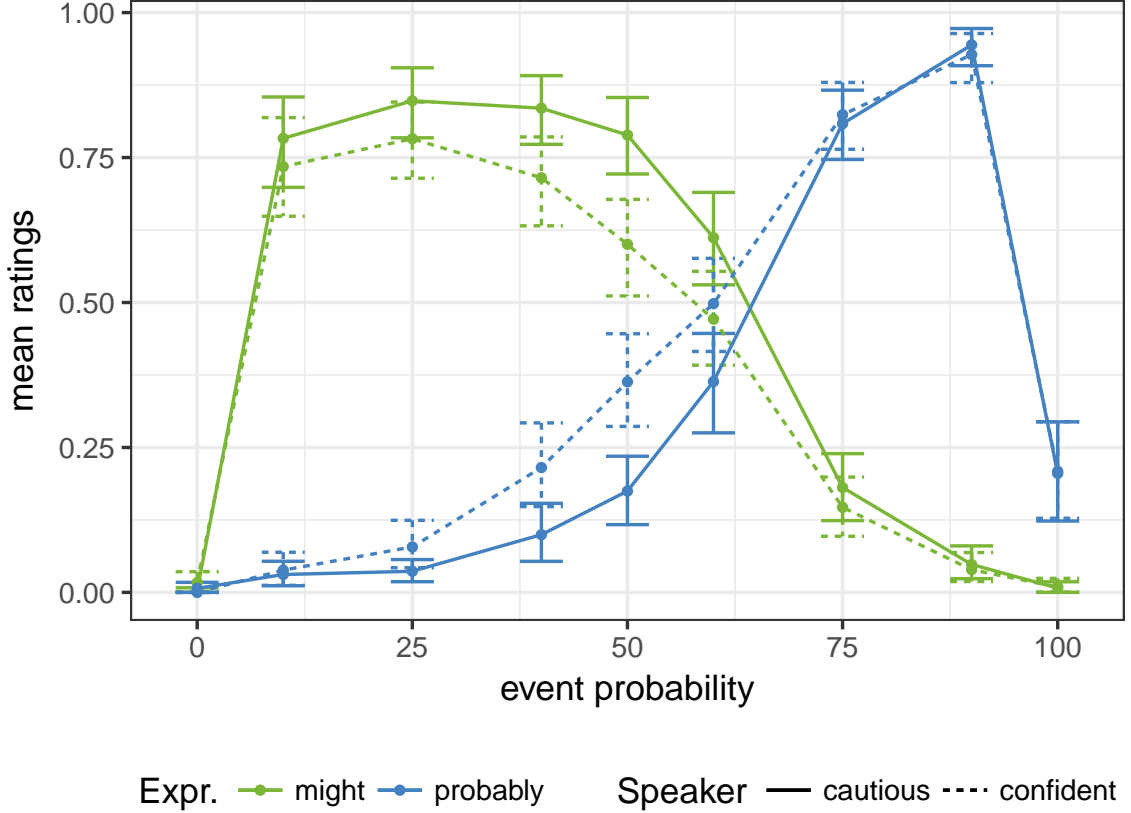
Figure 5.2: Mean utterance ratings for scenes with different event probabilities in Experiment 1. Error bars indicate bootstrapped 95% confidence intervals.

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 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 [Yildirim2016] 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 [Schuster2018] (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.

5.4 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 speakers and hence also twice the number of interactions.

5.4.1 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

---

[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.

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

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.

5.4.2 Results and discussion

Figure 5.3 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 [Schuster2018]. 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.

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]

---

[3][Yildirim2016] used a very similar paradigm to study semantic adaptation to the use of the quantifiers *some* and *many*. Analogous 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.
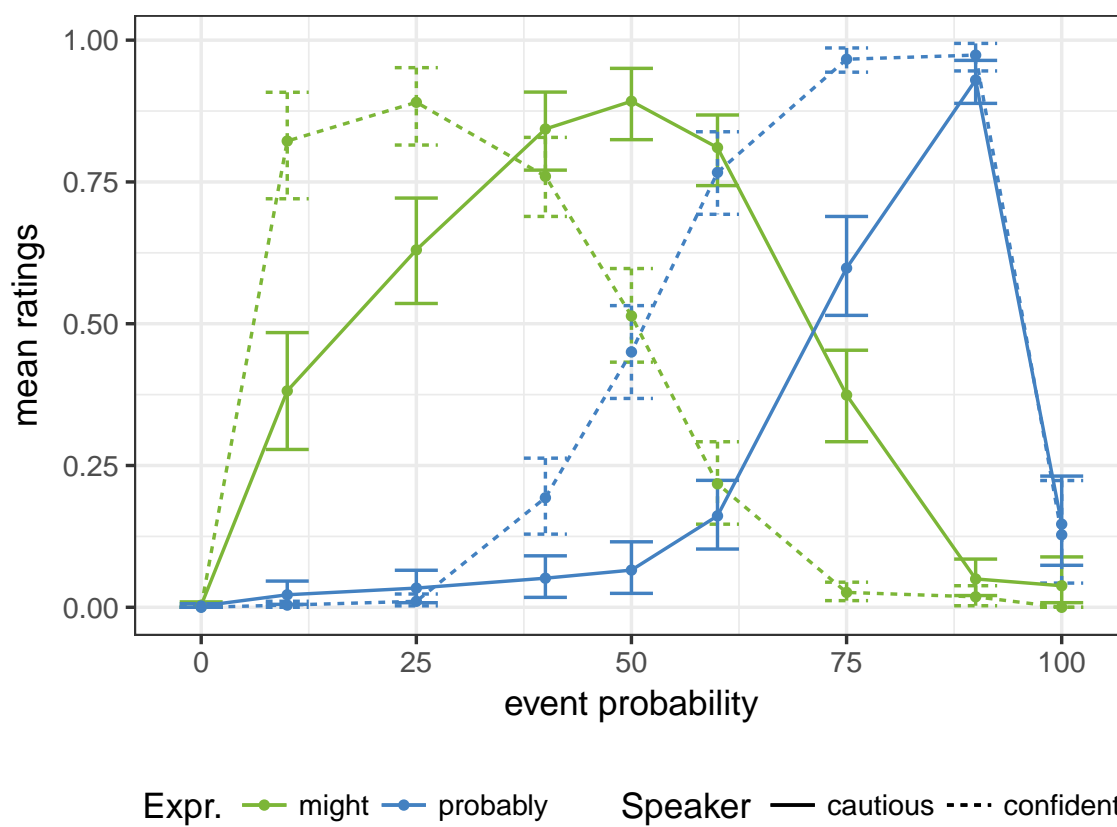
Figure 5.3: Mean utterance ratings for scenes with different event probabilities in Experiment 2. Error bars indicate bootstrapped 95% confidence intervals.

5.5 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 [Yildirim2016] 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.

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 [Kleinschmidt2015], [Schuster2018] 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

---

[4]See also [Hawkins2017] for a similar model of the formation of conceptual pacts.

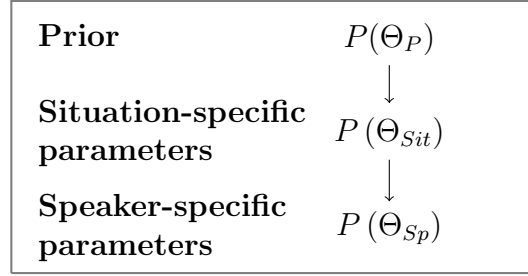|	
Prior	$P(\Theta_P)$
	$\downarrow$
Situation-specific parameters	$P(\Theta_{Sit})$
	$\downarrow$
Speaker-specific parameters	$P(\Theta_{Sp})$

Figure 5.4: 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.

cognitive and perceptual phenomena [¡]see e.g.,¿[for a review]Clark2013, including phonetic adaptation [Kleinschmidt2019], 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 5.4 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.

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 5.5 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.

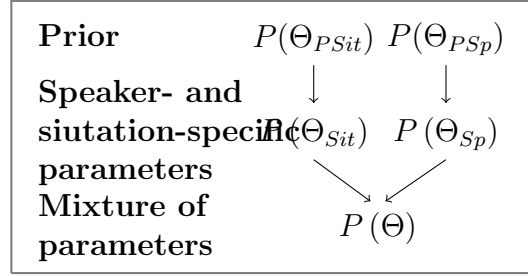|	
Prior	$P(\Theta_{PSit})$ $P(\Theta_{PSp})$
Speaker- and siutation-specific parameters	$\downarrow$ $\downarrow$ $P(\Theta_{Sit})$ $P(\Theta_{Sp})$
Mixture of parameters	$P(\Theta)$

Figure 5.5: 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 $P(\Theta_{Sp})$.

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.,¿Bradlow2008,Xie2018. 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.
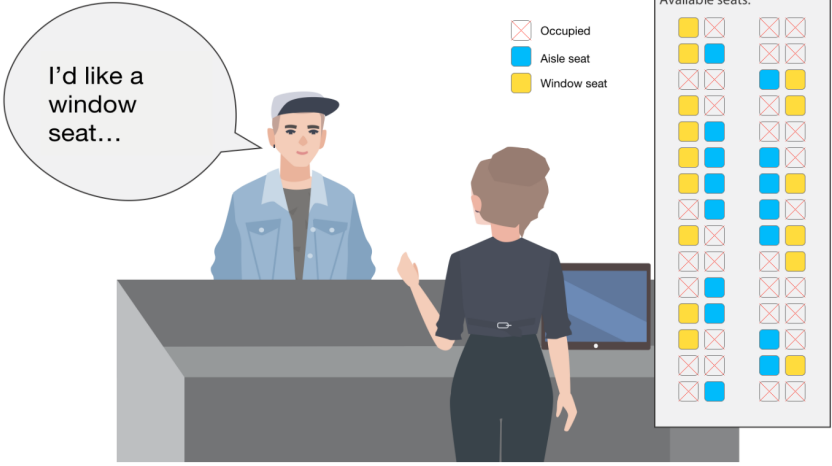
Chapter 6

Explaining away

6.1 Introduction

Speakers exhibit production variability at all levels of linguistic representation. Listeners deal with this variability by adapting to speakers and forming speaker-specific expectations about language use [¡e.g.,¿Norris2003,Kraljic2005,Bradlow2008,Kurumada2012,Kamide For example, at the lexical level, [Yildirim2016] found that participants were uncertain whether a generic speaker would use the utterance *Some of the candies are green* or the utterance *Many of the candies are green* to describe a bowl in which there were approximately equal numbers of green and blue candies. However, if participants briefly observed a speaker describing this scene either consistently using *some* or consistently using *many*, they updated their expectations about how that speaker would use quantifiers to describe different proportions. Similarly, [[]]henceforth S&D]Schuster2019 found that listeners update their expectations about a speaker's use of uncertainty expressions like *might* and *probably* after brief exposure to that speaker.

While previous work shows that listeners *can* adapt to variable language use, the contextual conditions and limits on adaptation are under-explored. In principle, it is possible that listeners might update their production expectations regardless of the context of utterance. Alternatively, listeners' production expectations, and consequently their adaptive behavior, might be modulated by non-linguistic contextual factors. To illustrate this, consider a speaker S who is in a very good mood and wants to be encouraging. If S tells a listener L "you'll *probably* win the sweepstake" when there is only a 60% chance of winning, L may consider S's use of *probably* instead of a weaker alternative such as *might* to be the result of S's mood. Consequently, L would not necessarily expect S to use *probably* to describe the same event probability when S is in a worse or more discouraging mood.

Recent computational models suggest that adaptation may be modulated by contextual factors. S&D proposed a computational model of adaptation to variable use of uncertainty expressions based on Bayesian belief updating. According to this model, when interacting with a speaker and observing their language use, listeners integrate their prior beliefs about the speaker's semantic representations and lexical preferences with the observed utterances to arrive at updated speaker-specific production expectations. Similar Bayesian belief-updating models have been proposed for adaptation in other linguistic domains, including in phonetic adaptation [Kleinschmidt2015], syntactic adaptation [Kleinschmidt2012], and prosodic adaptation [Roettger2019]. All of these models predict that the extent to which listeners adapt depends on how they initially expect a speaker to use language. Consequently, contextual factors that affect listeners' expectations about language use should also affect how much listeners adapt to specific speakers. If, given contextual information, a speaker's behavior matches prior expectations, there is no need to adapt.

Figure 6.1: Example trial from Experiment 1 and the post-exposure blocks from Experiments 2 and 3.

In addition to the model-predicted influence of contextual factors on adaptation, there is empirical evidence from phonetic adaptation: [Kraljic2008] used a lexical re-tuning paradigm to investigate adaptation to a speaker who produced a sound that was ambiguous between /s/ and /sh/. They found that without additional information listeners adapted, such that their perceptual boundary between /s/ and /sh/ shifted. However, when participants were shown a picture of the speaker with a pencil in their mouth, they explained away the observed signal as a pencil-distorted /sh/-sounding /s/ rather than as an intentionally produced /sh/-sounding /s/. Consequently, they did not adapt, i.e., their perceptual boundary between /s/ and /sh/ did not shift.

In this work, we investigate for the first time whether listeners explain away otherwise unexpected behavior at the lexical level if they are presented with contextual information that provides a reason for a speaker's productions. Concretely, we investigate whether one contextual factor – the speaker's mood – provides such a reason for otherwise less expected uses of the uncertainty expressions *might* and *probably* (EXPLAINING AWAY HYPOTHESIS). However, considering that previous experimental studies on semantic adaptation [Yildirim2016,Schuster2019] kept all aspects of the context constant between the exposure and test phase, it could also be that listeners simply learn associations between the use of uncertainty expressions and speakers, independent of other contextual information (ASSOCIATIVE HYPOTHESIS).

We investigate this issue as follows. We first establish that language users have different expectations about a generic speaker's use of uncertainty expressions depending on their beliefs about the speaker's mood (Exp. 1). We then investigate how much participants adapt when they are provided with information about the speaker's mood that makes their use of uncertainty expressions more expected, and compare participants' adaptation behavior to a neutral adaptation setting in which participants do not receive any information about the speaker's mood (Exp. 2). Finally, we investigate the relationship between adaptation and highly unexpected behavior by exposing participants to a speaker whose use of uncertainty expressions is incongruent with their mood (Exp. 3). We find that listeners adapt less when they are presented with a reason for the speaker's behavior. However, surprisingly, we also find that listeners do not adapt more when the behavior is highly unexpected given contextual information, potentially suggesting that there are limits on adaptation.

6.2   Experiment 1: Effect of speaker mood

In Exp. 1, we investigated how one contextual factor, the speaker's mood, affects listeners' expectations about a speaker's use of the uncertainty expressions *might* and *probably* for a range of event probabilities. The choice to manipulate the speaker's mood was guided by the intuition that listeners expect a speaker in a good mood

to use uncertainty expressions differently from a speaker in a bad mood. Moreover, mood is a non-inherent property of speakers that can change over time, which is important for Exps. 2 and 3.

Procedure, materials, analyses, exclusions and predictions were pre-registered on OSF (`http://osf.io/anonymized`).

6.2.1   Methods

Participants

We recruited 60 participants (20 per condition) from Amazon's Mechanical Turk.

Materials and procedure

At the beginning of the experiment, participants were introduced to an airline representative. Depending on the condition, the instructions explained that the representative was having a particularly bad day and feeling pessimistic and angry (*pessimist* condition); that she was having a particular great day and feeling optimistic and helpful (*optimist* condition); or that she was having a normal day (*neutral* condition). In addition to the textual mood information, the drawing of the representative showed her with an angry face (*pessimist*), a big smile (*optimist*), or a neutral facial expression (*neutral*).

Participants were then instructed that they would see scenes in which a customer of a cheap airline, who had the choice between getting a seat assigned at random or paying \$50 to pick their seat, would ask the representative about their possible seat assignment, to determine the likelihood of getting their preferred seat without paying. As shown in Fig. 6.1, participants could see the seat map and thus determine the number of available window and aisle seats and estimate the probability of getting the preferred seat. On each trial, participants had to indicate how likely they considered the representative to respond with one of the following two utterances:

- You might get a window seat/an aisle seat. (MIGHT)

- You'll probably get a window seat/an aisle seat. (PROBABLY)

Participants indicated their production expectations by distributing 100 points across these two utterances using a slider. If they thought that neither of the two utterances were likely responses, they could assign points to a blanket *something else* option. Participants completed 36 trials: they provided 4 ratings for each of 9 different probabilities of getting a preferred seat, ranging from 0% to 100% as indicated by the seat map. Trials were counterbalanced on whether the customer asked for a window or an aisle seat and trial order was randomized.

Analysis and exclusions

Following S&D, we quantified the production expectations for MIGHT and PROBABLY by fitting a spline with 4 knots for each participant and expression and computing the area under the curve (AUC) for each of these splines. A larger AUC indicates that an expression was rated highly for a larger range of event probabilities. To compare production expectations across conditions, we computed the difference in AUC between MIGHT and PROBABLY and compared the average difference across participants in the two conditions.

We excluded participants who provided random responses. Concretely, we excluded participants whose ratings for different event probabilities highly correlated ($r > 0.75$) with their average rating, suggesting that they always provided approximately the same rating independent of the event probability. Based on this criterion, we excluded 7 participants (*optimist*: 1, *pessimist*: 3, and *neutral*: 3).

Predictions

We predicted that participants expect the *optimistic* speaker to be encouraging and therefore to use PROBABLY for a wider range of event probabilities than the *pessimistic* speaker. Conversely, we predicted that participants expect the *pessimistic* speaker to use MIGHT for a wider range of probabilities than the *optimistic* speaker. This prediction should be reflected in larger AUC differences in the *pessimist* condition than in the *optimist* condition. Since it was unclear what mood participants attributed to the *neutral* speaker, we only predicted that the mean AUC difference for this third condition should lie between the mean AUC differences of the *optimist* and *pessimist* conditions, with the possibility of being equal to one of the two conditions.

6.2.2   Results and discussion

Fig. 6.2 shows participants' mean ratings for MIGHT and PROBABLY across the three conditions. As expected, we observe ratings close to 0 for both utterances when there is a 0% chance of getting the preferred seat (where there is a preference for the *something else* option, not shown), high ratings for MIGHT for low event probabilities, and high ratings for PROBABLY for high event probabilities. At the same time however, there are also differences across conditions. As the left panel shows, MIGHT was rated higher in the *pessimist* condition than in the *optimist* condition for a large range of event probabilities; as the right panel shows, the opposite was true for PROBABLY. Ratings in the *neutral* condition were almost identical to the ratings in the *optimist* condition. All these observations were also reflected in the AUC differences: The AUC differences in the *pessimist* condition were greater than in the *optimist* condition ($t(34) = 2.51$, $p < 0.05$). The AUC differences in the *neutral* condition – while
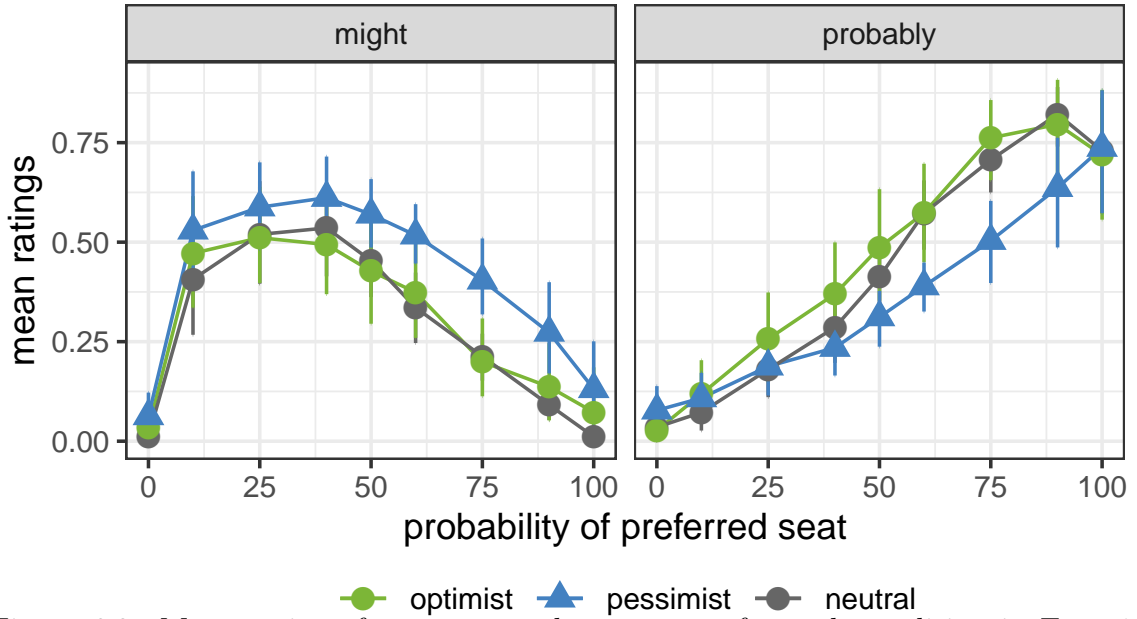
Figure 6.2: Mean ratings for MIGHT and PROBABLY for each condition in Exp. 1. Error bars correspond to bootstrapped 95%-confidence intervals.

numerically slightly larger – were not significantly different from the differences in the *optimist* condition ($t(34) = 0.35$, $p > 0.7$).

These results provide evidence that listeners have mood-dependent expectations about a generic speaker's use of uncertainty expressions. In Exp. 2 we investigate whether speaker mood affects the extent to which listeners adapt to that speaker's use of uncertainty expressions.

6.3 Experiment 2: Explaining away

In an exposure-and-test paradigm, we investigated whether adaptation to a specific speaker is modulated by knowledge about the speaker's mood. We follow S&D and either exposed participants to a speaker who always used MIGHT to describe an event probability of 60% (henceforth a "*cautious*" speaker) or a speaker who always used PROBABLY to describe an event probability of 60% (a "*confident*" speaker). Based on the results of Exp. 1, the behavior of a *cautious* speaker is more expected of a speaker who is having a bad day, and the behavior of a *confident* speaker is more expected of a speaker who is having a good day. We thus hypothesized that participants' beliefs about the speaker's mood influence how much they adapt: If the speaker's behavior is mood-congruent, we expected participants to experience a weaker expectation violation and adapt less than in the neutral conditions.

Condition	*pessimist*	*cautious*	*confident*	*optimist*
Mood	bad	neutral	neutral	good
$p = 25\%$	–		MIGHT x5	
$p = 60\%$	MIGHT x5		PROBABLY x5	
$p = 90\%$	PROBABLY x5		–	
$p = 100\%$	BARE x3		BARE x3	

Table 6.1: Overview of exposure utterances in Exp. 2. p indicates the proportion of preferred available seats shown on the seat map while the speaker produced the utterance. Critical trials are highlighted in gray.

Procedure, materials, analyses, exclusions and predictions were pre-registered on OSF (`http://osf.io/anonymized`).

6.3.1   Methods

Participants

We recruited 320 participants (80 per condition) from Amazon's Mechanical Turk.

Materials and procedure

The first block of the experiment consisted of an exposure phase with 13 trials (5 critical, 8 filler). On each trial, participants first saw a scene in which a customer asked about a specific seat and a seat map which indicated the number of available window and aisle seats (see top part of Fig. 6.1). To make sure participants paid attention to the seat map, they were then asked to rate how likely the customer was to get the preferred seat. They then listened to a pre-recorded response from the airline representative. Exposure trials were identical across the *pessimist* and *cautious speaker* conditions and identical across the *optimist* and *confident speaker* conditions but differed across these two pairs of conditions (see Table 6.1 for an overview): In the *pessimist* and *cautious speaker* condition, there were 5 critical trials in which the representative described a 60% probability of getting the preferred seat with "You might get one" (MIGHT); in the *optimist* and *confident speaker* conditions, the speaker responded with "You'll probably get one" (PROBABLY). 5 filler trials in the *pessimist/cautious speaker* conditions consisted of *probably* responses when there was a 90% preferred seat probability, and 5 filler trials in the *optimist/confident speaker* conditions combined *might* with a 25% preferred seat probability. Finally, 3 additional filler trials in all four conditions consisted of the response "You'll get one" (BARE) when it was 100% likely for the customer to get their preferred seat. Filler trials were
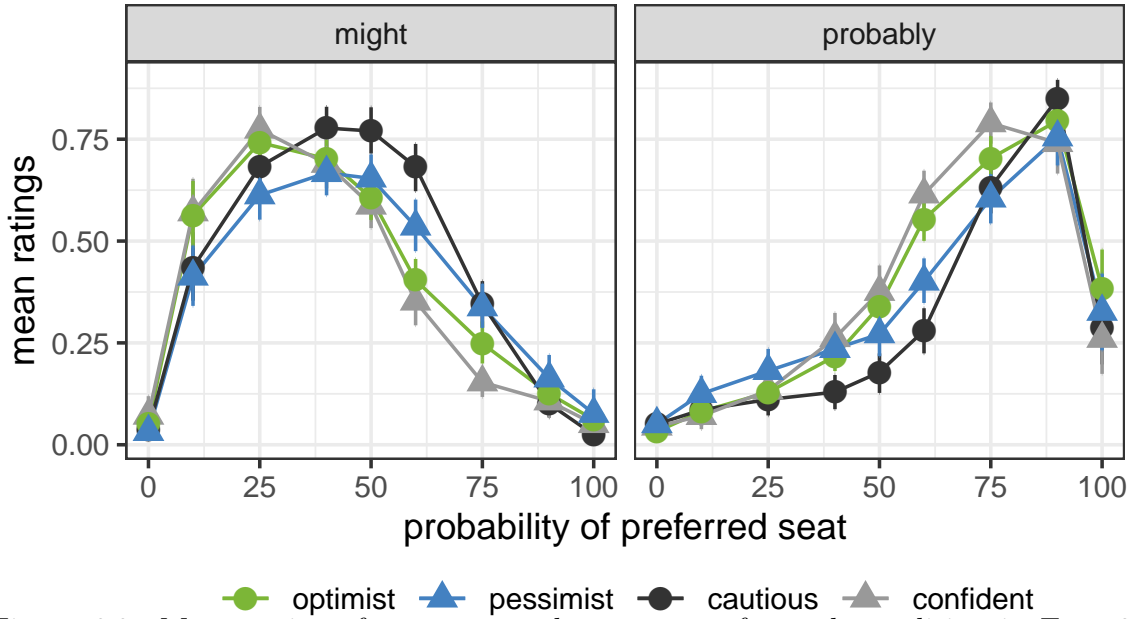
Figure 6.3: Mean ratings for MIGHT and PROBABLY for each condition in Exp. 2. Error bars correspond to bootstrapped 95%-confidence intervals.

intended to boost credibility of the speaker.

The exposure block was followed by another instruction, informing participants in all conditions that it was a week later and that the airline representative was having a normal day, followed by another manipulation check asking participants to rate how they thought the representative was feeling.

The last block of the experiment was identical to the trials in Exp. 1: participants completed 36 trials and rated how likely they thought it was that the speaker produced MIGHT, PROBABLY or *something else* for 9 different preferred seat probabilities.

Analysis and exclusions

We computed the AUC difference between the splines for MIGHT and PROBABLY for each participant as in Exp. 1. We again excluded data from participants providing random responses, which led to 52 exclusions (*pessimist*: 9, *cautious*: 14, *optimist*: 18, *confident*: 11).

Predictions

We expected that participants would adapt to the use of uncertainty expressions by the different speakers and update their expectations. Further, in line with the EX-PLAINING AWAY HYPOTHESIS, participants in the *pessimist* and *optimist* conditions,

who should experience less of an expectation violation, should adapt less than participants in the other two conditions. In terms of the AUC difference (AUC(might) - AUC(probably)), we therefore expected the following ordering: *cautious speaker < pessimist ≤ optimist < confident speaker.*

However, in Exp. 1, we also found that the ratings in the *neutral* condition did not significantly differ from the ratings in the *optimist* condition. This suggests that listeners' initial expectations about the speaker's mood and the associated production expectations only slightly differ across the *optimist* and the *confident speaker* conditions and therefore we also expected similar adaptation behavior in these two conditions.[1] We therefore, while expecting the numerical ordering described above, expected and pre-registered only significant differences between the *pessimist* and *cautious speaker* conditions, and between the *cautious speaker* and *confident speaker* conditions.

If listeners' adaptation behavior is not affected by contextual information, in accordance with the ASSOCIATIVE HYPOTHESIS, there should be no difference between the *cautious speaker* and *pessimist* conditions and no difference between the *confident speaker* and *optimist* conditions.

6.3.2 Results and discussion

Fig. 6.3 shows the mean ratings for MIGHT and PROBABLY for the four conditions. The results are consistent with the predictions according to the EXPLAINING AWAY HYPOTHESIS: First, participants in the *confident speaker* condition rated PROBABLY higher for a larger range of event probabilities than in the *cautious speaker* condition and the opposite was true for MIGHT. This pattern is also reflected in the mean AUC difference, which is larger in the *cautious speaker* condition than in the *confident speaker* condition ($t(133) = 5.18$, $p < 0.001$). This result replicates the adaptation effect found by S&D and suggests our seat map paradigm is suited for studying adaptation in the use of uncertainty expressions.

Second, we also observe differences between the *cautious speaker* and *pessimist* conditions. The mean AUC difference is larger in the *cautious speaker* condition than in the *pessimist* condition ($t(135) = 2.38$, $p < 0.02$).

Third, we also observe a numeric difference between the *confident speaker* and *optimist* conditions. Numerically, the AUC difference is larger in the *optimist* condition than in the *confident speaker* condition, but not significantly so ($t(129) = 1.61$,

---

[1]This intuition was further confirmed in a pilot study with 10 participants per condition, which we conducted prior to pre-registration. In the pilot, we found the expected ordering for the *cautious speaker*, *pessimist*, and *confident speaker* conditions but the difference between the *optimist* and *confident speaker* condition was so small that we would have needed more than 205 participants per condition to achieve power of $\beta = 0.8$.
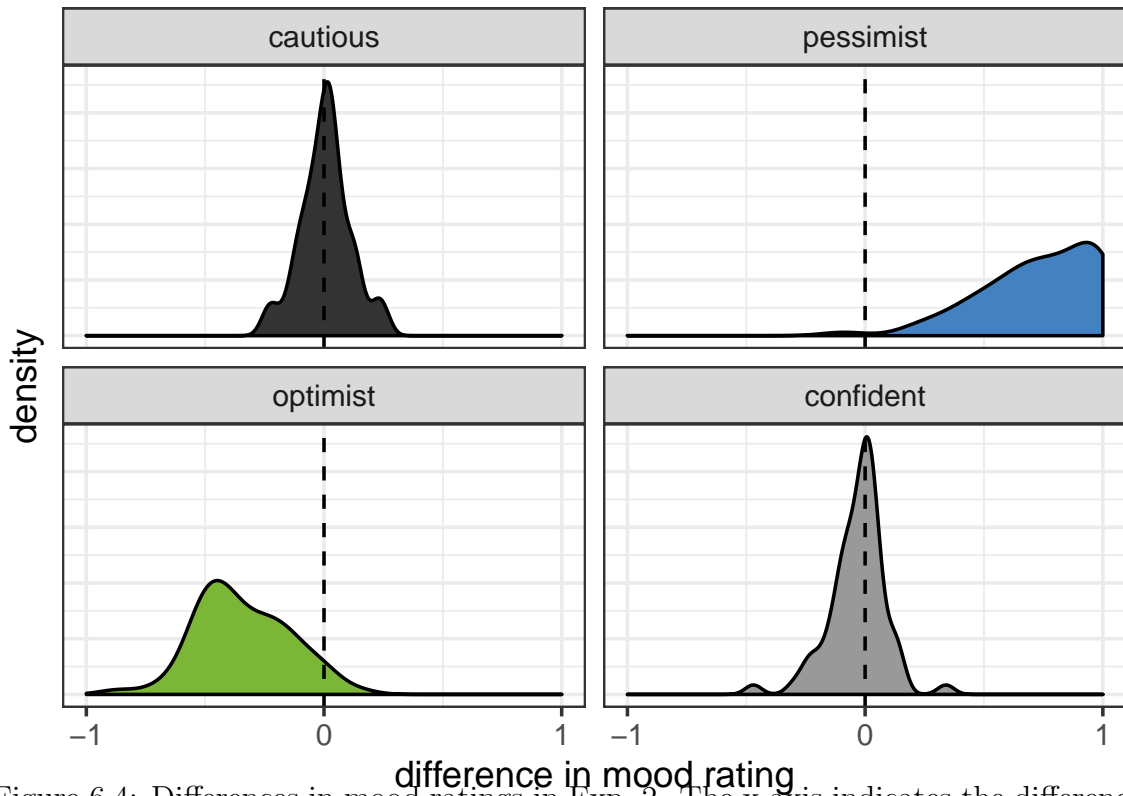
Figure 6.4: Differences in mood ratings in Exp. 2. The x-axis indicates the difference between the mood rating before the exposure block and the mood rating before the test block.

$p = 0.11$).

Lastly, as shown in Fig. 6.4, participants in the *optimist* and *pessimist* condition updated their beliefs about the mood after we instructed them that the speaker was now in a normal mood, suggesting that this instruction was sufficient to update participants' beliefs about the speaker's mood. As expected, participants in the two neutral conditions did not change their beliefs about the speaker's mood.

The results from this experiment provide evidence for listeners explaining away otherwise unexpected productions if they are presented with a reason for the speaker's behavior, and are predicted by the EXPLAINING AWAY account.

However, with additional stipulations, these results are also compatible with the ASSOCIATIVE account. One aspect of the context, the speaker's mood, changed between the exposure block and the test block in the *pessimist* and *optimist* conditions but not in the other two conditions. Therefore, it could be that this difference between the blocks leads to weaker associations between utterances and the context and therefore we observe less adaptation in the *pessimist* and *optimist* conditions. To

Condition Mood	*pessimist incongr.* bad	*optimist incongr.* good
$p = 25\%$	MIGHT x5	–
$p = 60\%$	PROBABLY x5	MIGHT x5
$p = 90\%$	–	PROBABLY
$p = 100\%$	BARE x3	BARE x3

Table 6.2: Overview of exposure utterances in Exp. 3. p indicates the proportion of preferred available seats shown on the seat map while the speaker produced the utterance. Critical trials highlighted in gray.

evaluate this possibility, we conducted Exp. 3.

6.4   Experiment 3: Incongruent conditions

In Exp. 3, we investigated participants' adaptation behavior when the speaker's use of uncertainty expressions was incongruent with the information about the speaker's mood, i.e., a speaker in a bad mood using uncertainty expressions like the *confident speaker* in Exp. 2, or a speaker in a good mood behaving like the *cautious speaker*. According to the EXPLAINING AWAY account, this should lead listeners to experience a stronger expectation violation than in the neutral and congruent conditions in the previous experiment and therefore listeners should adapt more. According to the ASSOCIATIVE account, on the other hand, listeners should adapt less than in the neutral conditions because according to this account, the smaller adaptation effect that we found in the *optimist* and *pessimist* conditions in the previous experiment was caused by a difference between the exposure phase and the test phase, which is still present in this experiment.

6.4.1   Methods

Participants

We recruited 160 participants (80 per condition) from Amazon's Mechanical Turk.

Materials and procedure

The procedure was identical as in Exp. 2. There were two conditions: *optimist incongruent* and *pessimist incongruent*. The *optimist incongruent* condition showed a speaker in a good mood who produced the same utterances as the *pessimist* and *cautious* speaker from the previous experiment. The *pessimist incongruent* condition

showed a speaker in a bad mood who produced the same utterances as the *optimist* and *confident* speakers in Exp. 2. See Table 6.2 for an overview of the exposure trials.

Analysis and exclusions

Analyses and exclusions were identical to the ones of Exp. 2. We excluded 27 participants (*optimist incongruent*: 15, *pessimist incongruent*: 12).

Predictions

We predicted that participants would adapt to different uses of uncertainty expressions: We expected the AUC difference in the *optimist incongruent* condition to be larger than in the *pessimist incongruent* condition. We further predicted that listeners experience a stronger expectation violation than in the neutral conditions in Exp. 2. We therefore also predicted the AUC difference in the *optimist incongruent* condition to be larger than in the *cautious speaker* condition, and the AUC difference in the *pessimist incongruent* condition to be smaller than in the *confident speaker* condition.

6.4.2 Results and discussion

Fig. 6.5 shows the mean ratings for MIGHT and PROBABLY for the two conditions in this experiment as well as the mean ratings from the neutral conditions from Exp. 2. As this plot shows, participants adapted to the different uses of uncertainty expressions. The mean AUC difference in the *optimist incongruent* condition was larger than in the *pessimist incongruent* condition ($t(131) = 5.90$, $p < 0.001$). However, unexpectedly, participants did not adapt more in the incongruent conditions than in the neutral conditions. The mean AUC difference in the *optimist incongruent* condition was not larger than in *cautious* condition ($t(129) = 0.004$, $p = 0.99$), and the mean AUC difference in the *pessimist incongruent* condition was not significantly smaller than in the *confident* condition ($t(135) = -1.18$, $p = 0.24$).

In this experiment, we again replicated the adaptation effect. However, we did not find a significantly stronger adaptation effect across the two incongruent conditions in this experiment as compared to the neutral conditions from the previous experiment, despite the fact that listeners should have experienced a stronger expectation violation.

What do these results imply for the EXPLAINING AWAY and ASSOCIATIVE accounts that we presented above? Together with the results from Exp. 2, the results from this experiment are unexpected under the ASSOCIATIVE account: if the reason for participants adapting less in the *pessimist* condition had been the difference in context
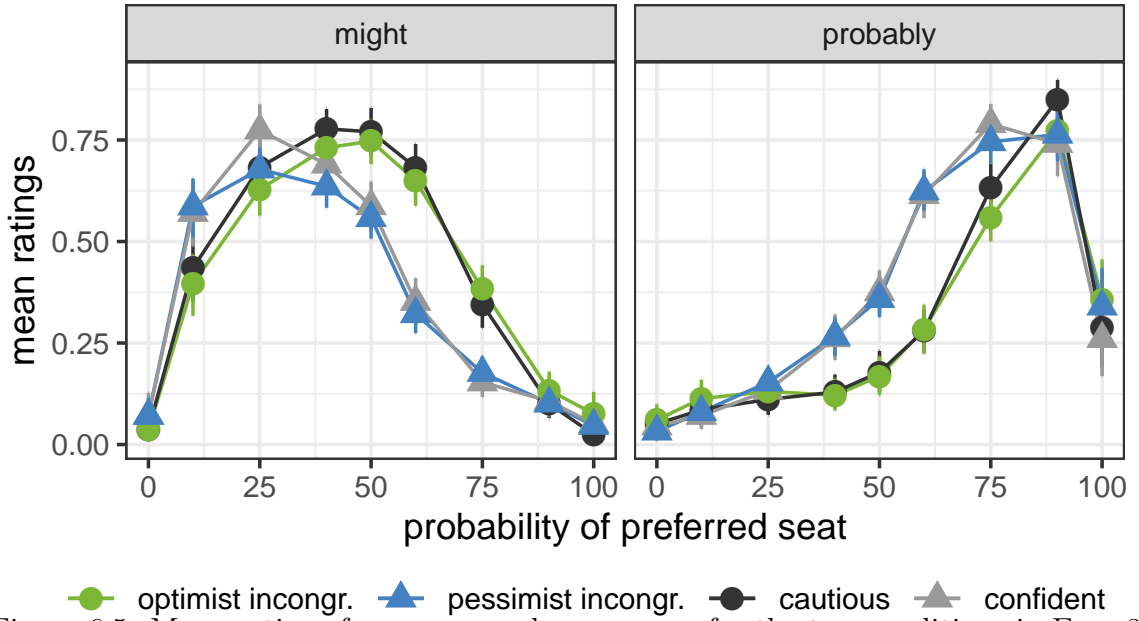
Figure 6.5: Mean ratings for MIGHT and PROBABLY for the two conditions in Exp. 3 as well as the neutral conditions in Exp. 2. Error bars correspond to bootstrapped 95%-confidence intervals.

between the exposure and test blocks, we would have expected less adaptation in this experiment as well.

However, we also did not find stronger adaptation, as we would have expected under the EXPLAINING AWAY account. We can only speculate about the reasons for this, but two explanations seem likely. First, given that there was a numerical difference between the *pessimist incongruent* and *confident* conditions in the expected direction, it could be that our experiment was underpowered to detect a potentially very small effect. However, a power analysis suggests we would need more than 500 participants per condition to achieve power of $\beta = 0.8$ and therefore we did not explore this option further.

Second, it could be that there is a limit on how much listeners can adapt and that this limit is already reached in the neutral conditions. If this was the case, listeners could still experience a stronger expectation violation when the behavior is incongruent with contextual factors but this stronger violation still does not lead to stronger adaptation.

6.5 General Discussion

In three experiments, we showed that language users have different expectations about the use of uncertainty expressions depending on their beliefs about the speaker's mood, and that this difference in expectations affects the extent of semantic adaptation. The results suggest listeners explain away otherwise unexpected behavior when they are presented with a cause, similarly as [Kraljic2008] found for phonetic adaptation.

Our results further largely confirm a prediction made by computational models of adaptation, namely that the extent of adaptation depends on how much observed behavior deviates from prior expectations. Similar results have been found for syntactic priming [Jaeger2013] and are predicted by connectionist models of syntactic learning [Chang2006].

However, the results from Exp. 3 suggest that expectation violation is not the only factor that influences adaptation and that there potentially exist limits to how much listeners can adapt. Investigating the limits of adaptation and developing quantitative linking functions between prior expectations and the extent of adaptation will therefore be important avenues for future work.

Chapter 7

Conclusions

Appendix A

Effect of color in Experiment 1

As mentioned in a footnote, I ran the norming studies in three batches using three slightly different procedures across conditions. I originally ran condition 0 (*bare-might*) as a pilot condition. In the results, I noted that participants did not differ in their ratings depending on whether the girl asked for a blue or an orange gumball ($R^2(27) = 0.997$ between mean ratings for blue and orange trials). To lower the number of trials, I therefore asked each participant to provide ratings for only one of the two colors (randomized across participants) for the next batch of conditions (conditions 1-14). I found that in some conditions, this led to small differences in ratings between participants who always rated utterances with *blue* and participants who always rated utterances with *orange* ($R^2(27)$ between 0.864 and 0.984). I hypothesize that this is a result of participants paying less attention if they were asked to do exactly the same task over and over again (in condition 0, the color and the associated utterances could change across trials). In order to verify the stability of our results, I replicated one of the conditions, condition 5 (*might-probably*), and asked participants to provide two ratings for each color and gumball proportion. I found that despite the lower correlation between average ratings for utterances with *blue* and utterances with *orange* in the original run ($R^2(27) = 0.929$), there was a very high correlation between the average ratings independent of the color of the original study and the average ratings of the replication ($R^2(27) = 0.975$), which suggests that the average ratings largely do not depend on whether I ask participants to provide ratings for both colors or just one color. Nevertheless, I used the modified procedure in which I asked participants to provide 2 ratings for each color and gumball proportion for the last batch of conditions (conditions 15-20). In all conditions in which I asked participants to provide ratings for utterances with both colors, the correlation between average ratings for utterances with *blue* and utterances with *orange* was almost perfect ($R^2(27) > 0.988$).

Appendix B

Additional results of Experiment 1.

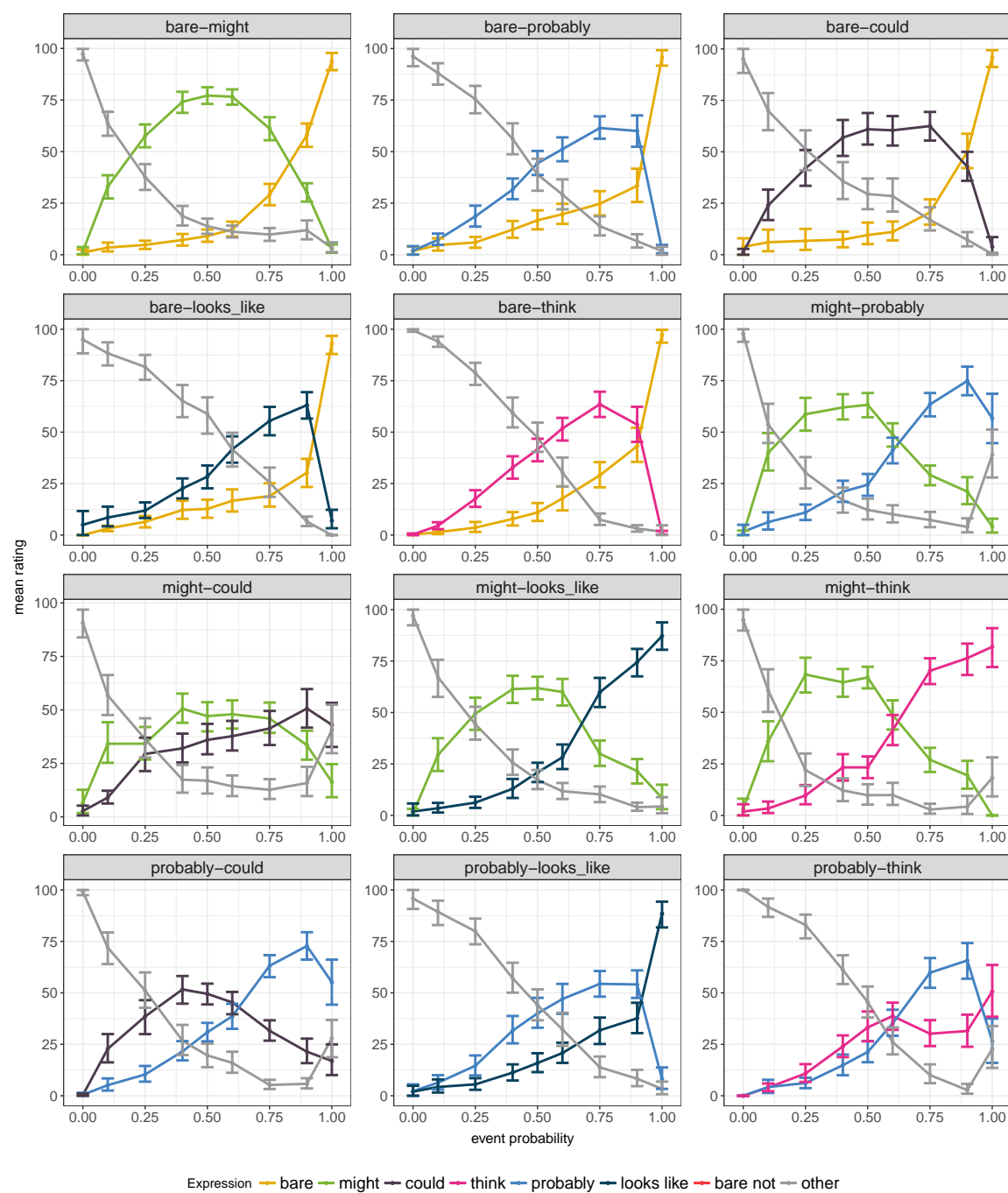Figures B.1 and B.2 show the results from all conditions in Experiment 1.

Figure B.1: Results of Experiment 1 – Part 1. Error bars correspond to bootstrapped 95%-confidence intervals.
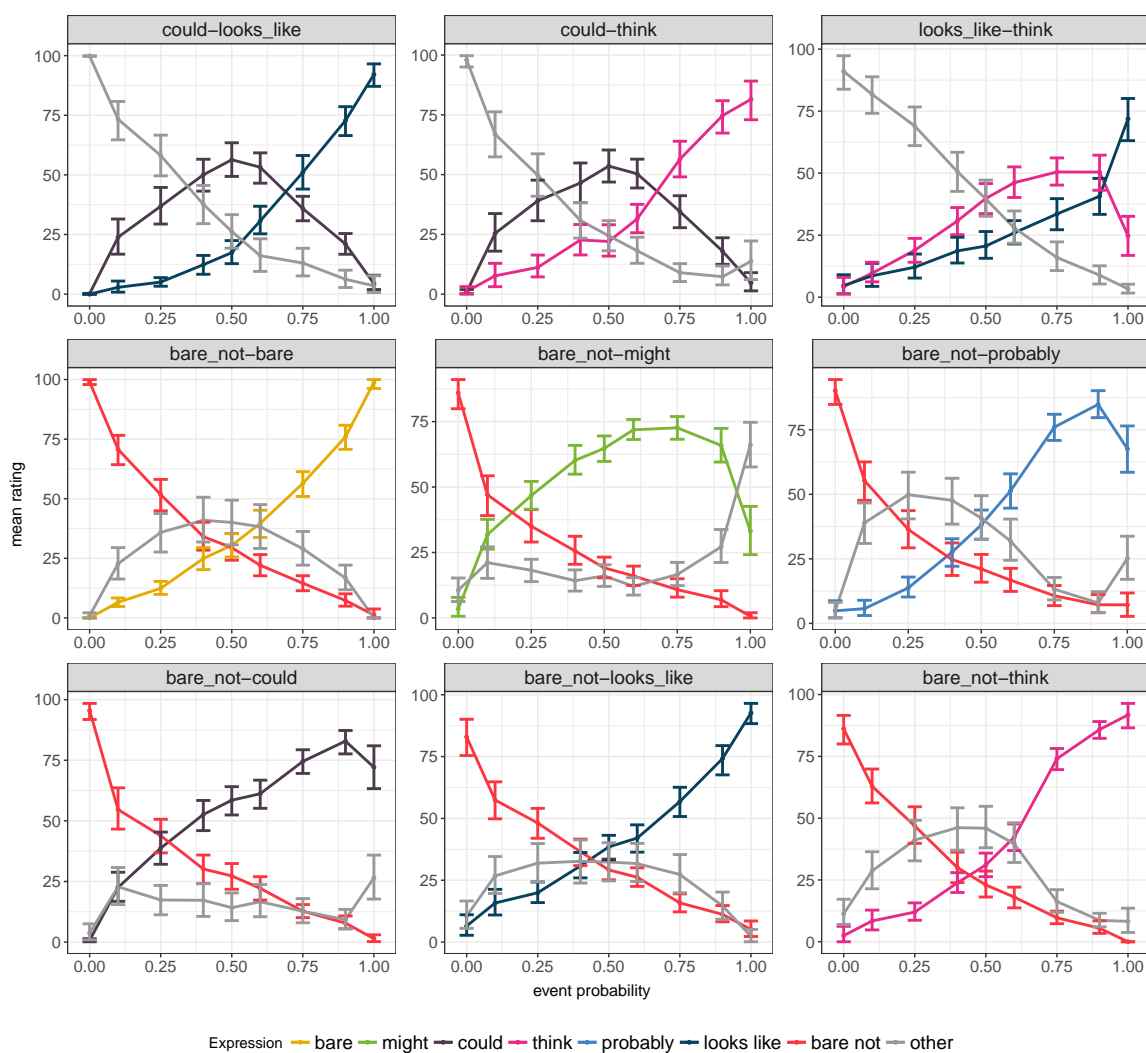
Figure B.2: Results of Experiment 1 – Part 2. Error bars correspond to bootstrapped 95%-confidence intervals.

Appendix C

Model implementation details

The model presented above poses some challenges for performing Bayesian data analysis with considerable amounts of data. Concretely, the integral over threshold distributions in the expected pragmatic speaker model ES_1 (repeated here) makes it hard to compute the distribution ES_1 given a set of parameters Θ.

$$ES_1\left(u_e \mid \phi\right) = \int P(c) \int_0^1 P(\theta) S_1\left(u_e \mid \phi, \theta, c\right) d\theta \; dc$$

The reason for this is two-fold: First, there is no analytical solution for this integral, and second, since S_1 depends on thresholds for all uncertainty expressions $P(\Theta)$ is a multidimensional distribution which cannot be easily approximated.

I solve this issue by introducing two approximations. First, I discretize the threshold distributions by distributing the probability mass of the Beta distributions across 20 equally-wide bins, resulting in a discrete probability distribution $P_d(\theta)$ (see [Tessler2019] for a similar approach). Since all event probabilities for which participants were asked to provide ratings in the experiments were multiples of 5%, we do not lose any accuracy and gain the advantage that we can now sum over a discrete probability space:[1]

$$ES_1\left(u_e \mid \phi\right) = \sum_\theta P_d(\theta) S_1\left(u_e \mid \phi, \theta, c\right)$$

While this approximation can in theory be computed exactly, its computation remains intractable even for the small number of utterances that I included in the

---

[1] In my data analysis procedure, I assumed that the distribution over cost functions, $P(c)$, is a delta distribution which assigns all probability mass to the condition-specific cost function $c(u, \mathscr{C})$ parameterized by the cost parameter γ. Since this implies that $P(c)$ is zero for all other cost functions, we can omit the integral and replace c with the condition-specific cost function, which I implicitly did here.

model presented above. Note that the discrete version of the vector of thresholds θ has one dimension with 20 possible values for each utterance, which implies there are $20^{|U|}$ possible assignments of θ. This means for estimating parameters for a model with 7 utterances, I would have to sum over $20^7 = 1.28 \times 10^9$ parameterizations of the pragmatic speaker model S_1 to compute the likelihood for one sample of parameters in the BDA.

I solve this problem through another approximation, which exploits the fact that $S_1(u_e \mid \phi, \theta, c)$ only depends on the thresholds for uncertainty expressions other than e in the normalization term. I approximate the normalization term by marginalizing over θ'_e and thus making S'_1 independent of all thresholds except θ_e:

$$\widetilde{S}_1(u_e \mid \phi, \theta_e, c) = \frac{exp\ \mathbb{U}(\phi, u_e, \theta_e, c)}{exp\ \mathbb{U}(\phi, u_e, \theta_e, c) + \sum_{u'_e \neq u_e} \sum_{\theta_{e'}} P_d(\theta_{e'})\ exp\ \mathbb{U}(\phi, u_{e'}, \theta_{e'}, c)},$$

where $\mathbb{U}(\phi, u_e, \theta_e, c) = \log L_0(\phi \mid u_e, \theta_e) - c(u)$ is the speaker utility as defined in Chapter 3.

This approximation allows me to define the following approximation of ES_1, which is tractable since we only have to sum over all values of one threshold instead of all combinations of thresholds:

$$\widetilde{ES}_1(u_e \mid \phi) \propto \sum_{\theta_e} P_d(\theta_e)\widetilde{S}_1(u_e \mid \phi, \theta_e, c)$$

This approximation leads to identical results as ES_1 if each threshold distributions assigns all probability mass to one value, i.e., if we have point estimates for thresholds. To assess how much ES_1 and its approximation, $\widetilde{ES}_1$ deviate when the threshold distributions have non-zero variance, I performed several simulations, with different threshold distributions. For these simulations, I assume that there are only two possible utterances, which makes the computation of ES_1 tractable.

Figure C.1 shows the results of these simulations. As these plots show, the approximate model $\widetilde{ES}_1$ is a very close approximation of the expected pragmatic speaker model ES_1, which suggests that this approximation should only minimally affect our modeling results.

The model is implemented in Python using the scikit-learn [Scikit2011] and numpy [vanderWalt2011] libraries.
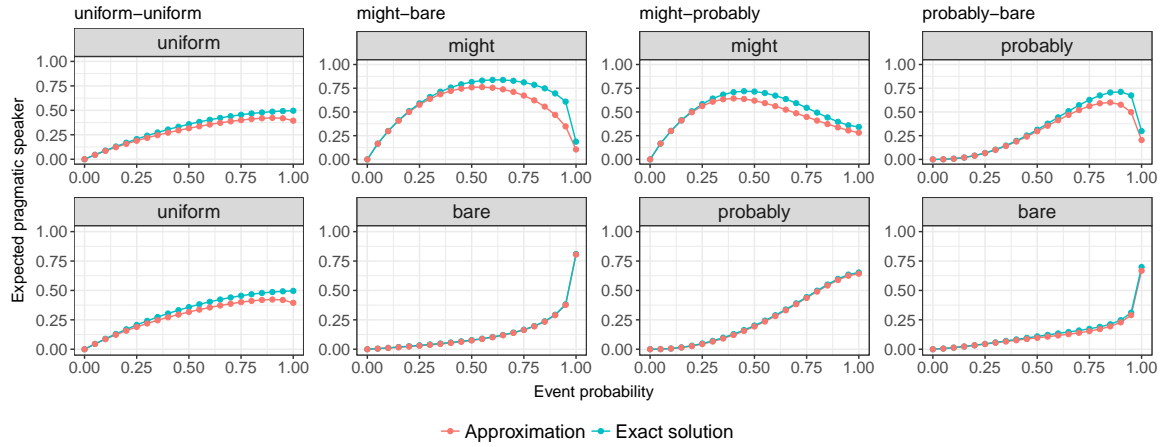
Figure C.1: Predictions of exact and approximate expected pragmatic speaker model for different combinations of thresholds. The leftmost panels (uniform) shows predictions of both models if both utterances have uniform threshold distributions, i.e., threshold distributions with very high variance. The other panels show model predictions under the assumption that the utterances have the threshold distributions that I inferred in Chapter 3.

Appendix D

Additional model predictions

Figures D.1 and D.2 show the model predictions and the results from all conditions in Experiment 1.
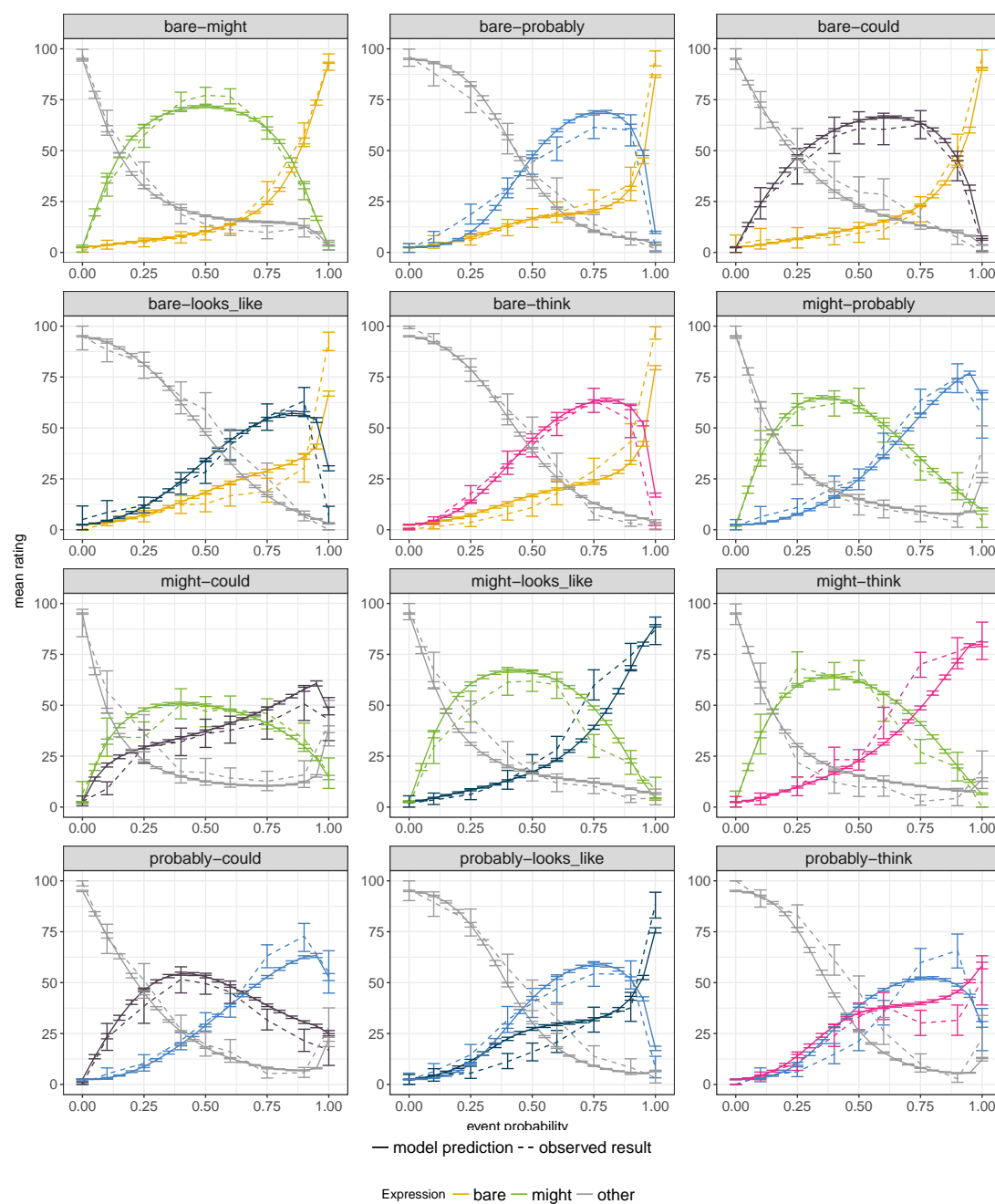
Figure D.1: Model predictions and results of Experiment 1 – Part 1. Error bars correspond to 95% high density intervals (model predictions) and bootstrapped 95%-confidence intervals (observed results).
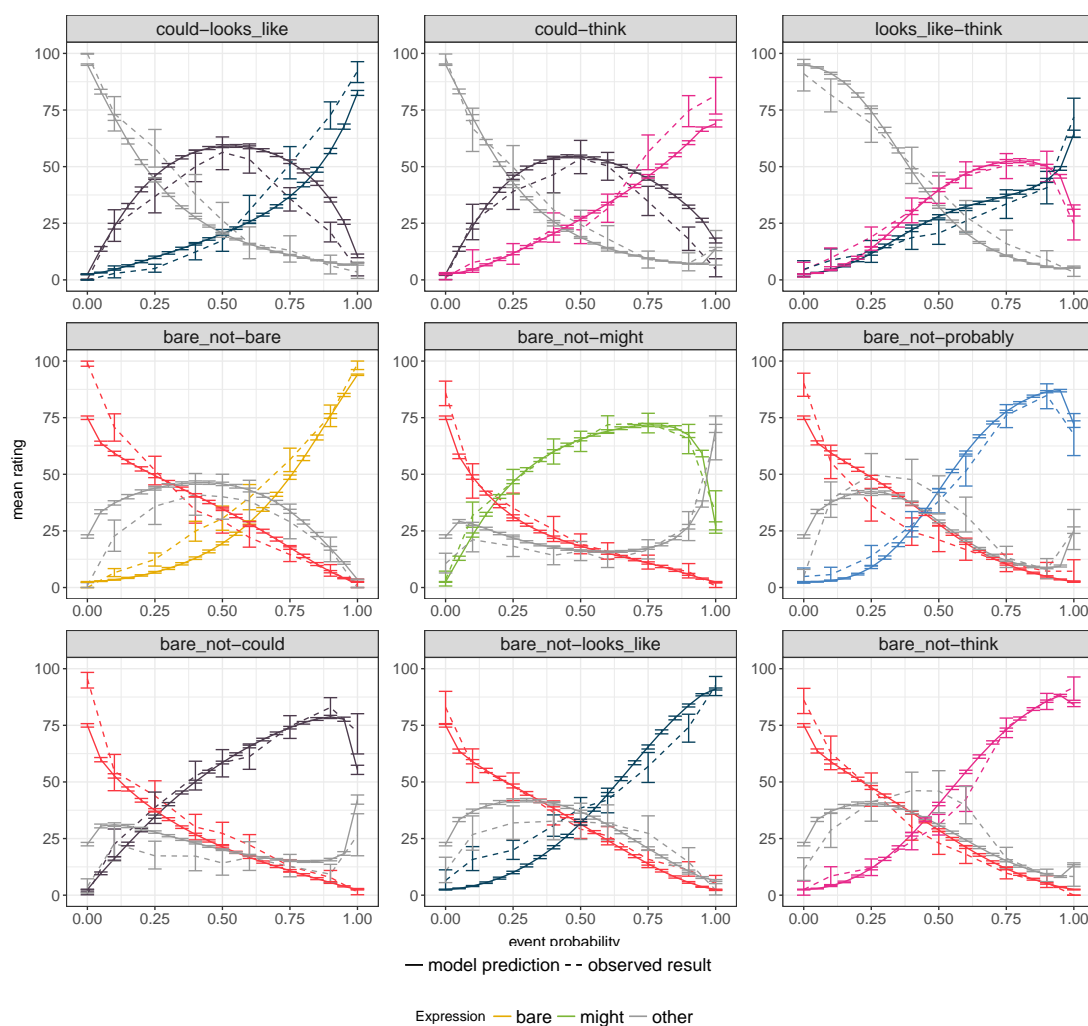
Figure D.2: Model predictions and results of Experiment 1 – Part 2. Error bars correspond to 95% high density intervals (model predictions) and bootstrapped 95%-confidence intervals (observed results).

Appendix E

Original production expectation adaptation experiment

I originally ran a different version of the production expectation experiment which included a potential confound because the number of utterances with each uncertainty expression was not matched across conditions. Qualitatively, this lead to the same results as Experiment 2 in Chapter 4 but from this experiment, it remained unclear whether the different post-exposure ratings were a result of the different number of exposure trials across conditions or a result of listeners updating the mapping between uncertainty expressions and event likelihoods.

For transparency, I report the procedure and the results of the original experiment here. The procedure, materials and analyses were pre-registered at `https://osf.io/w926x/`.

E.1 Method

E.1.1 Participants

We recruited a total of 80 participants (40 per condition) on Amazon Mechanical Turk. We required participants to have a US-based IP address and a minimal approval rating of 95%. Participants were paid $2 which amounted to an hourly wage of approximately $12–$15. None of the participants had previously participated in Experiment 1.

E.1.2 Materials and procedure

Materials and procedure were identical to Experiment 2. The only difference between Experiment 2 and this experiment were the number of filler trials with the other

	Original experiment						Experiment 2					
	MIGHT		PROBABLY		BARE		MIGHT		PROBABLY		BARE	
	n	ϕ	n	ϕ	n	ϕ	n	ϕ	n	ϕ	n	ϕ
cautious	**10**	**60%**	5	90%	5	100%	**10**	**60%**	10	90%	5	100%
confident	5	25%	**10**	**60%**	5	100%	10	25%	**10**	**60%**	5	100%

Table E.1: Number of exposure trials (n) per utterance (MIGHT, PROBABLY, BARE) and associated proportion of target color gumballs (ϕ) in the *cautious* vs. *confident* speaker conditions in this original experiment and Experiments 2. Critical trials bolded.

uncertainty expression, as shown in Table E.1.

E.1.3 Exclusions

I excluded participants who provided incorrect responses to more than 3 of the attention checks. Based on this criterion, I excluded 11 participants in the *confident speaker* condition and 8 participants in the *cautious speaker* condition. None of the results reported below depend on these exclusions.

E.2 Analysis and predictions

As in Experiment 2, I tested whether listeners updated their expectations after exposure by computing the difference between the AUC of the spline for MIGHT and of the spline for PROBABLY for each participant. I predicted that the mean AUC difference would be larger in the *cautious speaker* condition than in the *confident speaker* condition.

E.3 Results and discussion

This experiment yielded the same results as Experiment 2. As the panels in Figure E.1 show, participants updated their expectations about the speaker's language use and therefore made different predictions about how the speaker would use uncertainty expressions. In the *cautious speaker* condition, participants gave high ratings for MIGHT for a larger range of event probabilities than in the *confident speaker* condition. On the other hand, participants gave high ratings for PROBABLY for a larger range of gumball proportions in the *confident speaker* condition than in the *cautious speaker* condition. These differences result in a significantly larger AUC difference in the
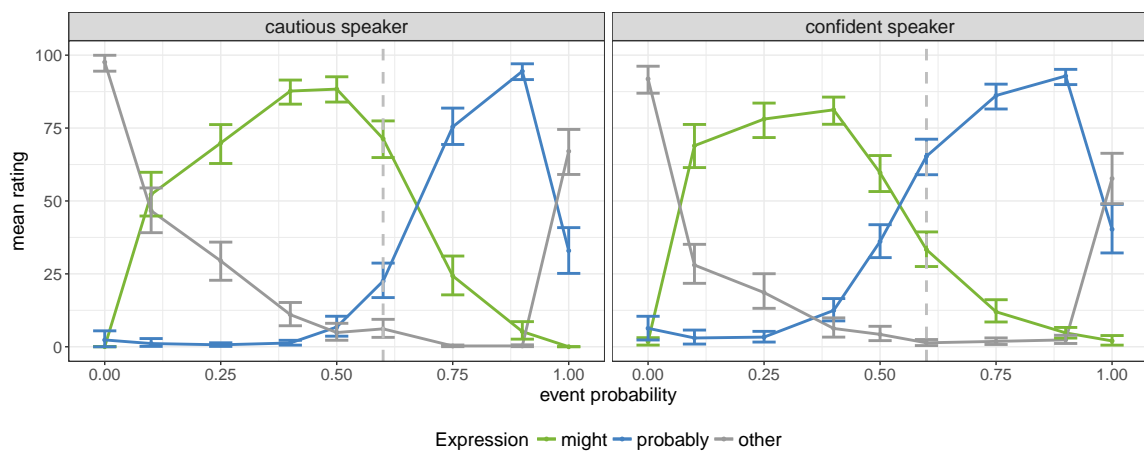
Figure E.1: Mean post-exposure ratings from original production expectation experiment. Error bars correspond to bootstrapped 95%-confidence intervals. The grey dotted line highlights the ratings for the 60% event probability ratings.
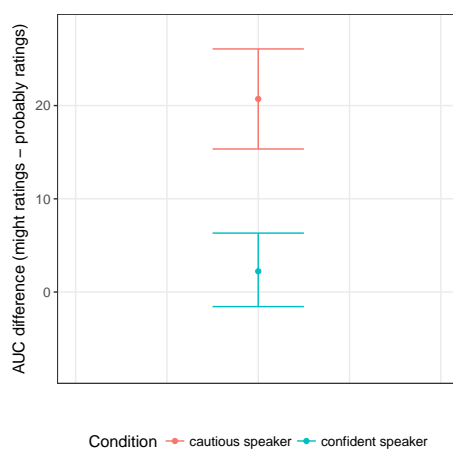

Figure E.2: Area under the curve (AUC) differences from original production expectation experiment. Error bars correspond to bootstrapped 95%-confidence intervals.

cautious speaker condition than in the *confident speaker* condition ($t(59) = 4.98$, $p < 0.001$, see also left panel of Figure E.2).

However, from these results it remains unclear whether listeners update their expectations about the mapping between uncertainty expressions and event likelihoods. In this experiment, the number of utterances with *might* and *probably* differed across conditions. It is therefore possible that participants only learned that the *cautious speaker* overall prefers to use *might* and the *confident speaker* prefers *probably*. To address this confound, I conducted Experiment 2 which is reported in Chapter 4.

Appendix F

Model simulations for original production expectation experiment

Table F.1 shows the results of the model comparison for the original production expectation experiment and Figures F.1, F.2, and F.3 show the posterior predictions of the model simulations, the post-adaptation threshold distributions, and the post-adaptation costs, respectively. For these simulations, we took the MAP variance parameters that I estimated for the data from Experiment 2, so I did not fit any parameters to the data from the original production expectation experiment. In each condition, the model was exposed to the 20 utterances that participants were exposed to in the experiment (see left part of Table E.1).

These modeling results further demonstrate the stability of the results reported in Chapter 4. I again find that the *cost & threshold distributions* model predicts the post-exposure data best according to the posterior odds metric, and the inferred post-exposure threshold distributions and cost values exhibit the same patterns as I found in the simulations for Experiment 2. Not surprisingly, the inferred cost differences between *might* and *probably* are bigger in the present simulations for the original

Model	odds	R^2
fixed	10^{-1137}	0.746
cost	10^{-386}	0.766
threshold distributions	10^{-207}	0.856
cost & threshold distributions	1	0.809

Table F.1: Model evaluation results on data from original production expectation experiment. *odds* are the posterior likelihood odds of the models compared to the *cost and threshold distributions* model. R^2 are computed between the mean post-exposure ratings and the mean model predictions.

experiment reflecting that the exposure across conditions was not balanced.

However, as shown in Table F.1, I also find that according to the R^2 metric, the model according to which listeners only update their beliefs about threshold distributions predicts the post-exposure behavior better than the *cost & threshold distributions* model. While we generally expect the ranking of models to be the same according to both metrics, I explained in Chapter 4 that in my setup, multiple assumptions of the R^2 metric are violated and therefore, I do not consider the ranking of models according to the R^2 metric as evidence that listeners only update beliefs about threshold distributions – especially given that in all other simulations, both metrics suggested that listeners update both types of representations.
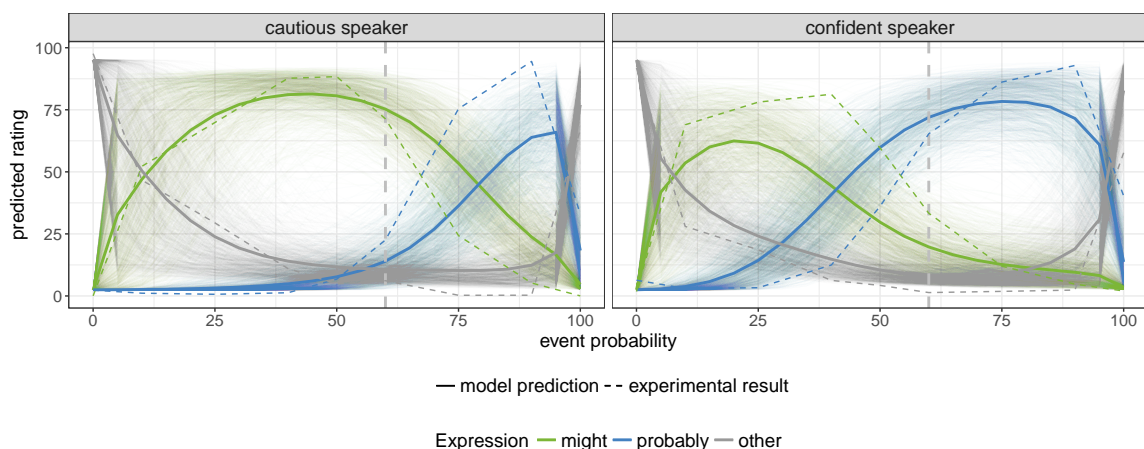

Figure F.1: Post-adaptation model predictions from simulations for original production expectation experiment and experimental results. The solid lines shows the mean model predictions and the thin lines around the mean show the distribution of model predictions.
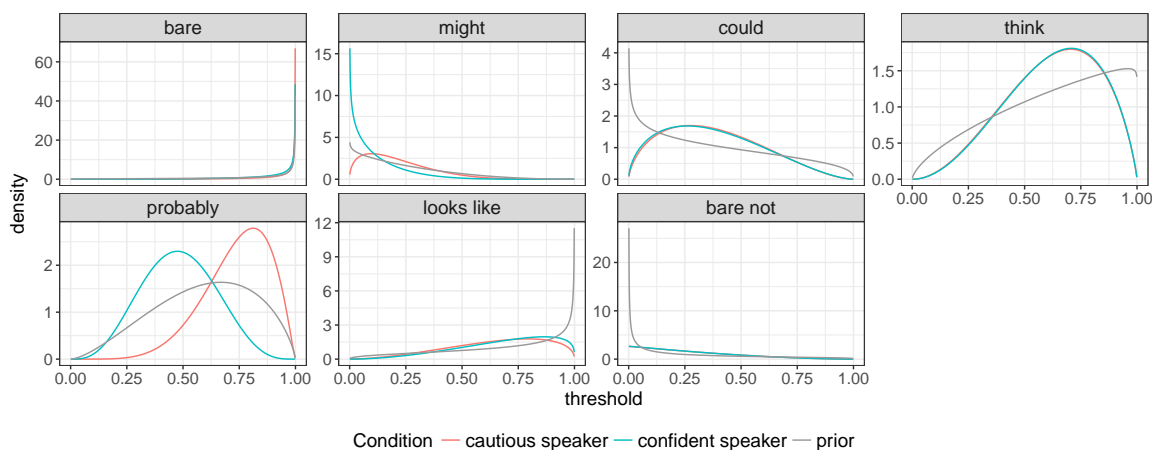
Figure F.2: Post-adaptation threshold distributions from the simulations for original production expectation experiment.
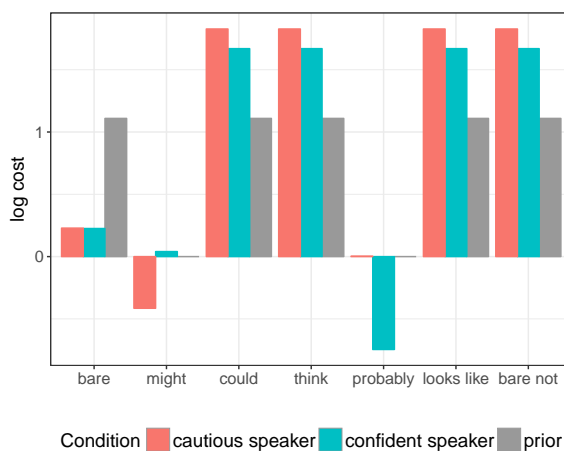

Figure F.3: Post-adaptation *log* cost values from simulations for for original production expectation experiment. Note that the cost of MIGHT and PROBABLY in the norming data model was 1 and therefore the *log* cost for these utterances is 0.

Appendix G

Original interpretation experiment

As I mentioned in Chapter 4, I originally ran a slightly different version of the comprehension experiment in which participants used sliders to rate which gumball machine they thought the speaker was describing. While the results were qualitatively the same as in the experiment reported above, the use of sliders seemed to confuse some participants (see details below) and therefore I changed the procedure such that participants provided ratings by distributing coins. For transparency, I report the procedure and the results of the original experiment here.

G.0.1 Method

Participants

I recruited a total of 80 participants (40 per condition) on Amazon Mechanical Turk. I required participants to have a US-based IP address and a minimal approval rating of 95%. Participants were paid $2 which amounted to an hourly wage of approximately $10–$12. None of the participants had participated in any of the previous experiments.

Materials and Procedure

The exposure phase was identical as in the other adaptation experiments: participants were either exposed to a *cautious* speaker or a *confident* speaker. Six of the exposure trials included attention checks in which participants had to indicate whether they saw a grey X on the previous trial or not.

Similar to Experiment 3, the test trials probed participants' interpretations of the utterances MIGHT, PROBABLY, and BARE. On test trials, participants listened to a recording of the speaker they encountered during the exposure phase and then rated how likely they thought it was that the speaker saw different gumball machines. On

each trial, like in Experiment 3, participants provided ratings for 9 gumball machines. However, unlike in Experiment 3, participants indicated their ratings by adjusting 9 sliders. Participants completed 6 test trials in total – one for each expression-color pair.

Exclusions

I excluded participants who failed more than 2 out of 6 attention checks, which led to 2 exclusions in the *cautious speaker* condition and 1 exclusion in the *confident speaker* condition.

G.0.2   Analysis and Predictions

As for Experiment 3, I expected that listeners interpret a more confident speaker's utterance to communicate a lower event probability than a more cautious speaker's utterance. I measured the interpretation of utterances by normalizing the ratings across the 9 gumball machines so that they sum to 1 and then computing the expected value for the proportion of blue and orange gumballs. I predicted that the expected values of target color gumball proportions after hearing MIGHT and PROBABLY were going to be larger in the *cautious speaker* condition than in the *confident speaker* condition.
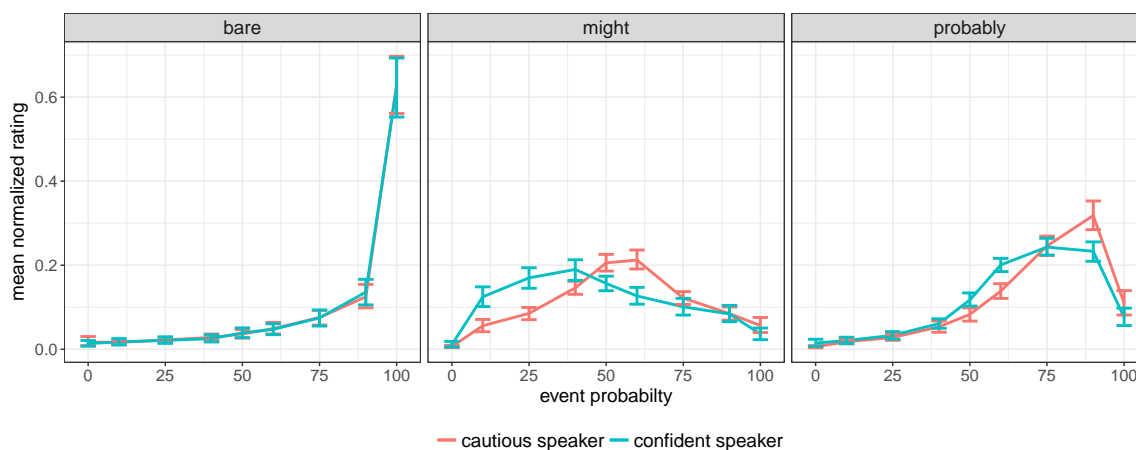
G.0.3   Results and Discussion

Figure G.1: Aggregated post-exposure ratings from the original interpretation experiment.

Figure G.1 shows the aggregated and normalized ratings for the two conditions. As predicted, participants provided higher ratings for gumball machines with higher target color percentages after hearing MIGHT and PROBABLY in the *cautious speaker* condition than in the *cautious speaker* condition. This also led to a significantly higher expected value for MIGHT ($t(75) = 3.05$, $p < 0.01$) and PROBABLY ($t(75) = 3.08$, $p < 0.01$) in the *cautious speaker* condition as compared to the *confident speaker* condition.

This means that qualitatively, the results are the same as in Experiment 3. However, since participants had the option to assign high ratings to all gumball machines (they could assign a maximum rating to each gumball machine if they wanted to), I noticed that many participants assigned very high ratings to most gumball machines and therefore did not indicate their interpretation of the utterance. Further, it seemed that some participants understood the instructions as rating the likelihood of getting a target color gumball and provided ratings proportional to the target color gumball proportion independent of the utterance. For these reasons, I revised the original paradigm as described in Chapter 4 and asked participants to indicate their interpretation using a limited set of coins, which appeared to be less confusing for participants.