Investigating the Language of Uncertainty
experimental data, formal semantics
&
probabilistic pragmatics

Dissertation

vorgelegt von

Michele Herbstritt
aus Borgomanero, Italien

zur
Erlangung des akademischen Grades Doktor der Philosophie
in der Philosophischen Fakultät der

Eberhard Karls Universität Tübingen.

2020

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Gedruckt mit Genehmigung der Philosophischen Fakultät der Eberhard Karls Universität Tübingen

Dekan: Prof. Dr. Jürgen Leonhardt

Hauptberichterstatter: Prof. Dr. Gerhard Jäger
Mitberichterstatter: Prof. Dr. Michael Franke, Prof. Dr. Friedrich Hamm

Tag der mündlichen Prüfung: 13. Mai 2020

Universitätsbibliothek Tübingen Online-Bibliotheksinformations und Ausleihsystem, TOBIAS-lib

Abstract

This dissertation reports a series of studies about the language of uncertainty in English. We investigate what we call “uncertainty expressions”, which include both verbal probabilities such as *probable* and *likely* and epistemic modals such as *might* and *possible*. Moreover, we look at complex, or nested, uncertainty expressions such as *certainly likely* and *might be possible*. The issues investigated in this work lie at the interface between semantics and pragmatics, and we attempt to answer questions such as: How do speakers communicate under uncertainty and about uncertainty? And what is the uncertainty intuitively expressed by uncertainty expressions? What do uncertainty expressions actually mean? And what is the role of context when we use them in a conversation?

Almost as frequent as logical connectives and quantifiers, uncertainty expressions are ubiquitous in everyday conversations. Unsurprisingly, uncertainty expressions have been extensively investigated by philosophers, logicians, linguists and cognitive scientists. One of the goals of this dissertation is to bring closer together the different traditional approaches to the study of uncertainty expressions. In our investigation, we strive for interdisciplinarity. In doing so, we integrate methods coming from formal semantics and pragmatics with experimental data and computational modeling. The focal point of the dissertation is a novel data-driven probabilistic model of the use and the interpretation of simple and complex uncertainty expressions.

Acknowledgements

My PhD is officially over, so I can finally tell the story of how I had applied for it on a (probably rainy) night in the Spring of 2014. At 23:55, exactly 5 minutes before the deadline. From my phone. Sitting in a bar somewhere in Amsterdam East. Sipping Gin & Tonic.

Whether such a peculiar beginning has been the inauguratory blessing of a successful endeavor or the foreshadowing omen of a doomed enterprise, it is not for me to say. What I will say is that if I'm not overly disappointed by how this dissertation eventually came together, that's certainly thanks to my supervisor, Michael Franke. The reason is simple enough: It would be hard to overstate the quantity and quality (and manner!) of things that I've learned from Michael, and much of what I've learned from him made its way into this dissertation. Among the many people which I'd like to thank, Michael surely deserves the biggest *thank you* of them all. He's been an amazing PhD supervisor. He was able to quickly realize how Michele-the-rookie-researcher functioned (certainly sooner than I did). He was always extremely supportive, pushing me hard enough to explore my half-baked ideas (to which he contributed a great deal!) while always giving me enough freedom to procrastinate, which, we all know that, is essential to academic research.

A very big *thank you* goes to Gerhard Jäger as well: He kindly agreed to be one of the reviewers of my dissertation (together with Michael Franke and Fritz Hamm) and he organized my PhD defense (*Promotionskolloquium* to its friends). Gerhard, Michael, and Fritz, together with Sigrid Beck and Tilman Berger, were the members of the defense committee, which I would also like to thank. They managed to make my defense a very pleasant experience, despite it being (so they told me) the first PhD defense ever happening on-line since the foundation of the Faculty of Philosophy of the University of Tübingen. Something to be proud of for sure, although I have to admit that during the first couple of centuries it was probably pretty hard to get a good enough Internet connection. I'm also grateful to Stefan Zauner for his substantial help with the dissertation submission and defense procedure.

Virtually all the work reported in this dissertation has greatly benefited from discussions I had with (and feedback I received from) many people during the past 5 years. Additionally (and perhaps more importantly?) my life has greatly benefited from the time I spent with many people during the past 5 years. Among those people, many will I be able to thank in the following lines; only a few, I hope, will I forget.

First of all, I'd like to thank Judith Degen, Noah Goodman and Dan Lassiter for hosting me at the Computation and Cognition Lab and the Department of Linguis-

tics of Stanford University in the Spring of 2015. Judith, Noah and Dan made me feel truly welcome during my two-month visit at Stanford, and they provided invaluable insights on my research during our discussions. I'm also grateful to Leon Bergen, Lorenza D'Angelo, Justine Kao, Ciyang Qing, Greg Scontras, Shane Steinert-Threlkeld, Michael Henry Tessler, Steve Woodworth for their insightful comments on my work and/or for the good times spent together outside working hours. Finally, among the Stanford crowd, a special *thank you* goes to Chris Mierzewski and Francesca Zaffora Blando.

Next, back to Tübingen: Thanks to Fabian Dablander for his help with the implementation of my very first experiment on Mechanical Turk, thanks to my office mates Anthea Schöller and Vinicius Macuch Silva and thanks to my housemate Ramon Ziai. Not too far from Tübingen is Aidlingen, where Giacomo Lini and Vicky Pedrazzoli live, together with their dog (my favorite dog) Zeno. I can't recall exactly how many times I've slept on their couch when I didn't have a place in Baden-Württemberg yet/anymore, and it was great to have them as (best!) friends throughout the years of my PhD: Thank you!

Many people don't fit in either of the two somewhat geographically inspired groups just mentioned, but they nonetheless contributed to this dissertation with their feedback and comments. Many thanks to Giovanni Cinà, Diana Mazzarella, Barbora Siposova, Giorgio Sbardolini and to the many people who asked insightful questions about my work at ESSLLI, XPRAG, and CogSci.

I'm grateful to my friends Alberto Gallo, Maurizio Ghione and Gabriele Vargiu for our *rendez-vous* around Europe during the years of my PhD (and before and after them as well): we've been to Rome, Isola del Giglio, Dublin, Budapest, Provence, Berlin, Amsterdam (and I might be forgetting a few), each and every time having a great time together despite the distances lying between us during the rest of the year.

Finally, even the most heartfelt *thank you* wouldn't come close to be enough for Margherita Isella. She has always been the #1 subject of my experiments, the #1 audience of my talks, the #1 reader of my drafts. And the #1 source of my happiness.

Per concludere, un grandissimo grazie va ai miei genitori, Chiara ed Edgardo, per il loro supporto incondizionato durante questi anni.

Contents

1 Introduction **1**
1.1 Chance will tear us apart 1
1.2 Translating uncertainty expressions 4
1.3 Formal and probabilistic semantics 5
1.4 Context sensitivity 7
1.5 Adverbs and adjectives 10
1.6 Higher-order uncertainty 12
1.7 Complex uncertainty expressions 15

2 Context dependency & QUD sensitivity **18**
2.1 Introduction 18
2.2 Background 19
2.3 Experiment 1 22
2.3.1 Setup and procedure 24
2.3.2 Results 27
2.4 Conclusion 35

3 Uncertainty adverbs & adjectives **38**
3.1 Introduction 38
3.2 Background: the Speech Act Modifier theory 39
3.3 Proposal: context-dependent semantics 45
3.3.1 Agreement, at-issueness and QUD-sensitivity 45
3.3.2 Back to the data 49
3.4 Conclusion 53

4 Higher-order uncertainty and rational beliefs **55**
4.1 Introduction 55
4.2 Experiment 2: measuring uncertain beliefs 59
4.3 Model: rational beliefs about the urn 60
4.4 Conclusion 64

5 Simple uncertainty expressions: data **66**
5.1 Introduction 66
5.2 Experiment 3 67

5.2.1 Experiment 3a: production 68
5.2.2 Experiment 3b: interpretation 73
5.3 Conclusion 76

6 Simple uncertainty expressions: model 78
6.1 Introduction 78
6.2 Rational Speech Act model 79
6.3 A model of uncertainty expressions 82
6.4 Model evaluation and criticism 88
6.5 Conclusion 94

7 Complex uncertainty expressions: data & model 96
7.1 Introduction 96
7.2 Model 97
7.3 Experiment 4: complex uncertainty expressions 103
7.4 Inference and model evaluation 108
7.5 Conclusion 113

Conclusion 115

References 118
References 118

List of Figures

1.1 Probability assigned to each phrase, expressed as percentage. 2

1.2 Two possible ways to partition the event space: on the left, the event p with 45% chances to happen is compared with several events $q_1, \dots, q_n$ each with small chances; on the right, the same event p is compared with a single alternative event q with higher chances. 8

1.3 Rational belief distributions about the contents of an urn containing 100 balls after drawing 8 balls and observing that 5 are red (left panel) and after drawing 80 balls and observing that 50 are red (right panel). 14

2.1 Examples of urns with different likelihood of drawing a red ball: from left to right 3/10, 5/10, 10/10. 23

2.2 Examples of urns with same likelihood of drawing red but different scenarios: dual on the left, plural on the right. 23

2.3 Example of a training trial. 25

2.4 Example of a critical trial. 26

2.5 Question choice percentage for each betting option. Black bars represent bootstrapped 95% confidence intervals. 27

2.6 Question choice percentage for each betting option, split by trial. Black bars represent bootstrapped 95% confidence intervals. 28

2.7 Expression choice percentage for each value. Black bars represent bootstrapped 95% confidence intervals. 30

2.8 Expression choice percentage for each value, scenario. Black bars represent bootstrapped 95% confidence intervals. 30

2.9 Expression choice percentage for each value, QUD. Black bars represent bootstrapped 95% confidence intervals. 31

2.10 Expression choice percentage for each value, scenario and qud. Black bars represent bootstrapped 95% confidence intervals. 37

4.1 Examples of partial observations of the contents of the urn. Left hand side: $a = 4$, $o = 3$; right hand side: $a = 8$, $o = 6$. 57

4.2 Examples of rational belief distributions given two partial observations of the urn and assuming a uniform prior over states. 58

4.3 Sample stimulus and input slider bins. 59

4.4 Measured belief distributions (red lines) in 15 observation conditions together with 95% bootstrapped confidence intervals (ribbons). 61
4.5 The data-generating model as a probabilistic graphical model. White nodes represent latent variables, shaded nodes represent observed variables. Single-bordered nodes represent stochastic dependence, double-bordered nodes represent deterministic dependence. Boxes indicate scope of indices. 62
4.6 Posterior predictive distributions (black lines) in each of 15 observation conditions together with Bayesian 95% highest density intervals (grey ribbons). Red lines and ribbons display mean values and 90% CIs of observed data. 64

5.1 Example of picture displayed to participants in the production trails. From left to right, it represents the unknown urn, a partial observation of 3 red balls out of 4, and the random draw of a new ball whose color needs to be predicted by the participant. 69
5.2 Percentages of expression choices in each observation condition, together with bootstrapped 95% confidence intervals. 70
5.3 Input sliders in the interpretation tasks, observation trials. The picture on the right provided immediate and interactive visual feedback, dynamically displaying the current slider selection (see the midpoint of the slider). 73
5.4 Counts of state, access and observation value choices in each expression condition, together with bootstrapped 95% confidence intervals. . 74

6.1 The universe of discourse in a simple reference game. 80
6.2 Examples of rational belief distributions given two partial observations of the urn. 83
6.3 Examples of literal belief distributions over states as a function of the received message, assuming flat priors and setting $\theta_{certainly} = 0.99$, $\theta_{probably} = 0.5$, $\theta_{possibly} = 0.01$. 85
6.4 Examples of speaker's distributions over messages given two partial observations of the urn. 87
6.5 The data-generating model. White nodes represent latent variables, shaded nodes represent observed variables. Single-bordered nodes represent stochastic dependence, double-bordered nodes represent deterministic dependence. 89
6.6 Posterior cumulative density distributions for the semantic threshold parameters. 91
6.7 Percentages of expression choices in each observation condition compared to posterior predictive distributions. Rectangular “confidence areas" are bootstrapped 95% confidence intervals of the data against Bayesian 95% HDIs of the posterior predictive. 93

6.8 Counts of state, access and observation values choices in each expression condition compared to posterior predictive distributions. Rectangular “confidence areas" are bootstrapped 95% confidence intervals of the data against Bayesian 95% HDIs of the posterior predictive. . . . 94

7.1 Rational belief distributions given observations 3/4 and 6/8. States which make *likely* true have been highlighted in red. Red numbers near the bars indicate the (rounded) probability mass of the corresponding state, assuming flat prior beliefs. 101

7.2 Input menus in the production task. 104

7.3 Percentages of expression choices in each observation condition, together with bootstrapped 95% confidence intervals (black bars). . . . 105

7.4 Counts of state, access and observation value choices in each expression condition, together with bootstrapped 95% confidence intervals (black bars). 107

7.5 Means of choices in the state and access interpretation conditions, with bootstapped 95% confidence intervals. 108

7.6 Mean predicted percentages of expression choices in each observation condition, together with Bayesian 95% HDIs (black bars). Red crosses mark the empirically observed counts for comparison (see also Figure 7.3). 110

7.7 Means of the predicted counts of state, access and observation value choices in each expression condition, together with Bayesian 95% HDIs (black bars). Red crosses mark the empirically observed counts for comparison (see also Figure 7.4). 111

7.8 Posterior predictive of the means of choices in the state and access interpretation conditions. Error bars are 95% HDIs. Red crosses are the empirically observed means (see also Figure 7.5). 113

List of Tables

2.1 Mean values and HDIs for model coefficients for the main effects of `value`, `scenario` and `qud` and their interactions. Coefficients credibily different from zero are marked with *. 32

2.2 LOO-IC scores with standard errors (SE). Lower is better. 33

2.3 Mean values and HDIs for model coefficients for the main effects of `value` and `scenario` and their interactions. Coefficients credibily different from zero are marked with *. 34

2.4 Mean values and HDIs for model coefficients for the main effects of `value`, `scenario` and `qud` and their interactions, excluding wrong responses in control conditions. Coefficients credibily different from zero are marked with *. 34

2.5 Mean values and HDIs for model coefficients for the main effects of `value`, `scenario` and `qud` and their interactions, excluding all data from control trials. Coefficients credibily different from zero are marked with *. 35

3.1 Association between QUD types and typical at-issue contents and produced answers. 46

5.1 Experimental conditions in Experiment 3a. The fractions represent observations of the urn. The labels on the left refer to levels of higher-order uncertainty. 68

5.2 Mean values and HDIs for model coefficients for the main effects of `observation` and `access`. Coefficients credibly different from zero are marked with *. 72

5.3 LOO-IC scores and standard errors (SE) of the models with `obs.+acc.`, `mode` and `ev` as predictors —lower is better. 72

5.4 Mean values and HDIs for model coefficients for the main effects of `expression` on `state`. Coefficients credibly different from zero are marked with *. 75

5.5 Mean values and HDIs for model coefficients for the main effects of `expression` on `access`. Coefficients credibly different from zero are marked with *. 76

6.1 Examples of EU of each message given two partial observations of the urn, rounded to two decimal places. 86
6.2 Mean inferred values and HDIs of the semantic threshold parameters free in the model, given experimental data collected in Experiment 1. . 91
6.3 Mean Pearson's correlation scores and HDIs between model posterior predictive distributions and exerimental data collected in Experiment 1. 92

7.1 Complex uncertainty expressions. 98
7.2 Experimental conditions in Experiment 4a. The fractions represent observations of the urn. The labels on the left refer to levels of higher-order uncertainty. 104
7.3 Mean inferred values and HDIs of the semantic threshold parameters free in the model given experimental data collected in Experiment 2. . 109
7.4 Mean Pearson's correlation scores and HDIs between model posterior predictive distributions and exerimental data collected in Experiment 4. 109

Chapter 1

Introduction

1.1 Chance will tear us apart

On August 25th, 2017, Reddit user `zonination` posted a very nice looking picture on the community `dataisbeautiful`. The plot displayed in the picture was made using the R package `ggjoy`, inspired by the cover art of Joy Division's famous album Unknown Pleasures. The result was indeed beautiful. As it happens on the internet sometimes, the post went somewhat viral and reached an audience even beyond the (admittedly already pretty big) community of data-nerds (myself included) where it was originally posted.[1]

We reproduced the plot here in Figure 1.1. The title given to the plot by its author is "Perceptions of probability". On the y-axis several English phrases are listed, such as *almost certainly*, *very good chance*, *about even*, *highly unlikely* and so on. The x-axis, labeled "Assigned Probability", shows a scale of percentages, from 0% to 100%. For each phrase on the y-axis, the plot displays the distribution of probability values, expressed as percentage, assigned to the phrase by a group of 46 Reddit users who participated in a survey informally run by `zonination` on `samplesize` (another Reddit community). We do not know much about the procedure followed in the survey. What we know is that the author collected users' responses to questions of the form

> *What probability would you assign to the phrase "X"?*

where *X* was replaced by each of the phrases listed in the plot. Participants' were instructed to reply with a percentage value between 0% and 100%. The plot displays distributions of participants' responses for each phrase.

Let us look at some examples. Glossing over the presence of a few puzzling outliers, there are a number of intuitive results. For example, look at the distribution assigned to *almost certainly*, at the very top of the plot: it is a reasonably peaked distribution, its peak somewhere between 95% and 100%. This seems intuitive: if something will *certainly* happen, then its chances are 100%; adding *almost* makes it just a little more vague.

[1] The dataset and R script used to produce the plot are available under MIT License at `https://github.com/zonination/perceptions`.

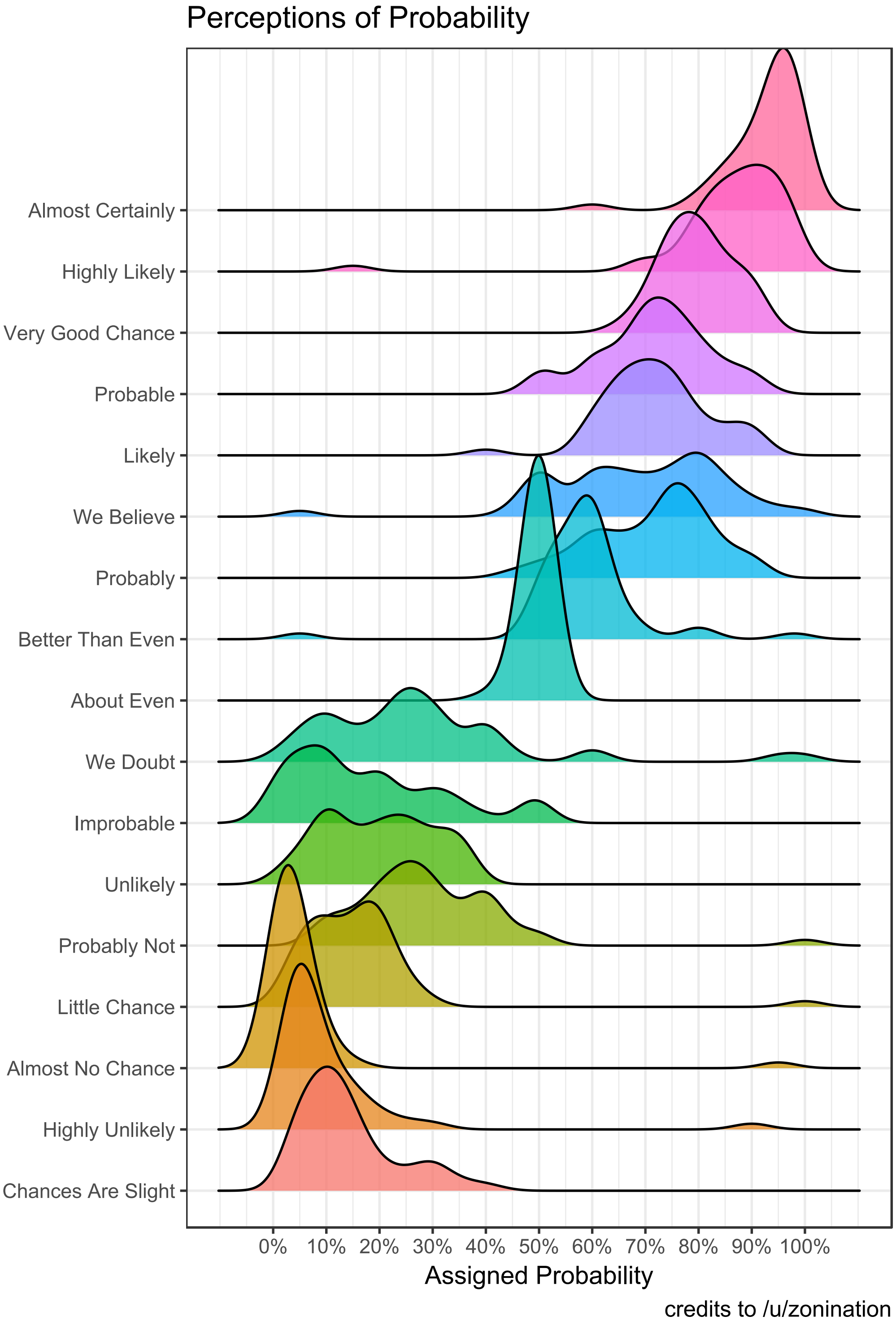


Figure 1.1: Probability assigned to each phrase, expressed as percentage.

Immediately below *almost certainly* on the y-axis we find *highly likely*. The distribution is not too dissimilar, but evidently less peaked, its peak somewhere between 90% and 95%. So *highly likely* appears to be somewhat weaker than *almost certainly*. Again, pretty intuitive. Next, let us go almost all the way to the bottom of the plot. Look at the distribution obtained for *almost no chance*: it is essentially a symmetric version of the one obtained for *almost certainly*, rather peaked between 0% and 5%. What about *highly unlikely*? It appears to behave symmetrically to *highly likely*. Next, look at *probable* and *likely*, very similar to each other, both peaked around $70\% - 75\%$ and both with seemingly quite high variance. And both behaving symmetrically to *improbable* and *unlikely*. Finally, what is the central axis of all these symmetries? Clearly, *about even*, with its distribution noticeably peaked around 50%.

The plot in Figure 1.1 is undoubtedly beautiful, very easy to read and its symmetrical features resonate nicely with our naive intuitions, all of which surely played a role in its viral diffusion. However, we believe that the plot is (and has been for many Reddit users) also very thought-provoking. In fact, imagine being one of the participants to `zonination`'s survey. The request is to produce a number expressing the probability intuitively assigned to the expression *likely*. Which number would you choose? Exactly 72%, which is the mean of the 46 participants' responses? Or some value smaller than 70%, as 15 participants seem to think, six of them going as low as 60%? On the other hand, eleven participants chose a number greater than 75%, four of them going as high as 90%. In a nutshell, as much as the mean probability value associated with *likely* appear to be intuitive, the variance between participants cannot be ignored. A quick look at the plot is enough to realize that analogous observations hold for essentially all the expressions.

A small terminological excursus is in order here. Phrases such as *(almost) certainly*, *(un)likely*, *(im)probable* etc. are often collectively referred to as *verbal probabilities* or *probability expressions*. In this dissertation we often use the slightly more general term *uncertainty expressions*, which we take to include also auxiliaries such as *might* and adverbs/adjectives such as *possibly* and *possible*, i.e. those expressions best known in philosophy, logic and linguistics as "epistemic modals".

Introspective intuitions about uncertainty expressions seem to vary a lot among Reddit users (and arguably among other people too). And yet, we use them all the time to reason, communicate and coordinate successfully with our peers. We say that a coin flip is *equally probable* to land heads or tails. We say that buying a lottery ticket is *probably* a waste of money, because winning the lottery is *extremely unlikely*. We look out of the window and decide that it *might* rain soon; on the other hand, the weather forecast informs us that precipitation will only be *likely* in the evening. *Maybe* we can still have the pic-nic we planned. We know it is *possible*, but we predict that Trump will *almost certainly* not win the next presidential election. Therefore, colonizing Mars and moving there *might* not be necessary (yet).

If logic connectives (*and*, *or*, *if* ... *then* ..., etc.), negation (*no*, *not*, etc.) and quantifiers (*some*, *all*, *every*, *any*, *many*, etc.) come up virtually in every real-life conversation, uncertainty expressions are almost as frequent. For this reason, it is not surprising that uncertainty expressions have received a tremendous amount of attention by researchers in psychology and linguistics.

The amount of relevant scientific literature is gigantic. Rather then attempt a thorough review of the literature, which we believe would be an unrealistic and not-so-insightful task, this chapter has the much less ambitious goal of highlighting some of the key ideas found in previous works on uncertainty expressions, focusing on a limited number of issues which are especially important either from a historical or a theoretical point of view (or both). The remainder of this chapter is devoted to introducing four of these issues (which we sometimes call puzzles), around which almost all of this dissertation revolves.

While discussing these puzzles, we pay attention to the background against which they were developed. With only minimal simplification, this background can be described as characterized by the divide between psychological approaches to uncertainty expressions and linguistic ones. In a nutshell, psychological approaches have traditionally taken the form of what we can call "translation studies", i.e. experimental studies attempting to provide a one-to-one correspondence between uncertainty expressions and numeric chance levels or intervals. On the other hand, linguistic approaches mostly in the logico-philosphical tradition have focused on the formal semantics of uncertainty expressions, i.e. abstract descriptions of the logical properties of uncertainty expression and their contribution to a compositional theory of meaning.

We discuss these two traditions in some more detail and we provide relevant references to the literature in the next paragraphs of this introduction.

1.2 Translating uncertainty expressions

Even if we do not know much about the details of `zonination`'s Reddit survey, it is easy to see how their work can be categorized as a translation study: participants had to provide numbers expressing the probability that they would assign to each phrase. Assigning numbers to uncertainty expressions corresponds to only one "direction" of the translation. From a linguistic point of view we can think of this direction as *interpretation* of uncertainty expressions, where a listener grasps the meaning communicated by a speaker using the language of uncertainty. The other direction is just as important: given a certain likelihood (or interval of likelihoods) that an event will happen, what phrase would we use to describe its chances? From a linguistic point of view we can think of this direction as *production* of uncertainty expressions, where a speaker uses the language of uncertainty to describe a state of affair.[2]

Clark (1990) reviews approximately twenty years of psychological research which can fall under the category of the translation approach, starting from the seminal work by Lichtenstein and Newman (1967). The methodology followed in their work appears to be quite simple: the authors asked 188 participants to assign a number between 0.01 and 0.99 to 41 verbal expressions, ranging from *highly probable* and *very likely* to *very unlikely* and *highly improbable*, but including also phrases such as *rather likely*, *rare*, *seldom*, *usually* and *inconclusive*. The conclusion drawn by the authors is that even though the responses were reasonably consistent, the asymmetries in responses found between pairs of symmetrical phrases (e.g. *quite likely* and *quite unlikely*) justify some

[2] Notice that by pointing out that both directions are crucial we do not intend to criticize `zonination`'s approach by any means. After all, it is not uncommon to find psychological studies which tackle only one or another of the two directions.

skepticism towards the endeavor of assigning verbal labels to numerical probabilities. Clark (1990) appears to be more cautious in commenting the work by Lichtenstein and Newman, his main criticism being that the choice of some of the 41 stimuli may be questionable, because context-dependent modifiers were included (e.g. *quite*, *very*) as well as expressions with a "distinctly frequentistic meaning" (e.g. *usually*, *rare*).

Be that as it may, many subsequent studies followed the same approach of Lichtenstein and Newman, i.e. the translation from verbal expressions to numerical values (while some went the other direction, i.e. assigning verbal labels to frequency or chance values). We refer to the review by Clark (1990) for the relevant references. In general, Clark's conclusion is that there seems to be a reasonable consistency between the findings of several different studies. However, the within-subject variability in many studies appears to be generally high. Related to this, a somewhat more recent trend of research has attested several sources of contextual effects on the production and interpretation of uncertainty expressions. For example, Wallsten, Fillenbaum, and Cox (1986) found that subjects' translations are sensitive to the prior base rate of the events under discussion: for example, the perceived probability of an event described as *likely* will be higher if the event was initially perceived to be likely *a priori*. Next, it has been found that when the expressions are evaluated in context rather than in isolation the variability in the translations provided by experimental subjects is significantly higher (Beyth-Marom, 1982; Brun & Teigen, 1988). Finally, Teigen (1988) and Windschitl and Wells (1998) found that the way in which the set of possible alternative events is presented to, or conceptualized by, the subjects, has an effect on subjects' use and interpretation of uncertainty expressions. The studies by Teigen and Windschitl and Wells are especially relevant for us and we will return to them momentarily. Before doing that, we need to introduce the other tradition of research about uncertainty expressions, i.e. the logico-philosophical approach.

1.3 Formal and probabilistic semantics

As already mentioned, theoretical linguists and philosophers have traditionally approached uncertainty expressions from a more formal and abstract perspective, focusing more on the logical properties of such expressions and their role within a general compositional theory of meaning. In this tradition the focus has been largely on epistemic modals (e.g. *might*, *possible*, *possibly*) rather than probability expressions, at least since the seminal works by Carnap (1947), Hintikka (1961), Kratzer (1977) and Kripke (1980), which laid the foundations of so-called *possible world semantics*. Some exceptions are the works by Hamblin (1959) and Kratzer (1991), to which we will return momentarily.

There is an impressive amount of linguistic literature about epistemic modals. The debate is multifaceted and interdisciplinary, spanning from semantics and pragmatics to syntax and historical linguistics, and touching on related issues in mathematical logic, epistemology and metaphysics. The scope of this dissertation is much more limited, as we focus solely on some issues at the interface between semantics and pragmatics. For this reason, we take as our starting point the work by Kratzer (1991) mentioned above, because it contains, to the best of our knowledge, the first attempt to develop a unified formal account of the meaning of both epistemic modals and probability expressions.

Kratzer's approach is purely qualitative, in the sense that it does not make any reference to probabilities, chance levels, or any other quantitative measures. Uncertainty expressions are analyzed as operators (e.g. quantifiers) on a set S of possible worlds (or states of affairs, intuitively "ways the world could be like"), which typically represents the domain of the discourse, often referred to as *modal base* or *state space*. Let us briefly look at Kratzer's semantics for *likely*, adopting the notation introduced by Seth Yalcin in his influential paper *Probability Operators* (Yalcin, 2010). First, a preorder $\geq_O$ is defined on the worlds in the modal base.[3] The preorder represents the normality or stereotypicality of each world: it is defined on the basis of a set of propositions O, which is assumed to contain the propositions normally or stereotypically true, in an intuitive sense, in the given situation. Propositions, denoted as $\varphi, \psi, \ldots$, are defined as sets of possible worlds. For any pair of worlds $w, v \in S$, the preorder $\geq_O$ is defined as follows:

$$w \geq_O v \text{ iff } \{\varphi \in O \mid w \in \varphi\} \supseteq \{\varphi \in O \mid v \in \varphi\}$$

Intuitively, a world w is at least as stereotypical as another world v just in case w is contained in a higher quantity of stereotypical propositions than v. Next, the preorder $\geq_O$ is lifted to propositions:

$$\varphi \succsim_O \psi \text{ iff } \forall w \in \psi\ \exists v \in \varphi \text{ such that } w \geq_O v$$

The preorder $\succsim_O$ is taken to capture the semantic contribution of comparative expressions such as *at least as likely*: a proposition φ is at least as likely as another proposition ψ just in case for each world w in ψ there is a world v in φ which is at least as stereotypical as w. Based on this, the strict comparatives *more likely* can be easily defined:

$$\varphi \succ_O \psi \text{ iff } \varphi \succsim_O \psi \wedge \neg(\psi \succsim_O \varphi)$$

Finally, the semantic contribution of the bare uncertainty expression *likely* is equated to *more likely than not*. Intuitively, we can say that a proposition p is likely just in case the states of affairs described by p are strictly more stereotypical than those described by the negation of p. For any atomic proposition p:

$$\text{likely}(p) \text{ iff } p \succ_O \neg p$$

Kratzer's approach is reasonably intuitive, it is conservative (as it sticks to the possible world framework) and it validates many inference patterns involving uncertainty expressions together with logical connectives and other modal operators (see Yalcin, 2010). Nonetheless, Yalcin (2010) shows that there are intuitively valid inference patters which cannot be captured by Kratzer's approach and, conversely, that Kratzer's semantics validates inference patterns which appear to be intuitively invalid. On the basis of his analysis, Yalcin argues in favor of a quantitative approach, i.e. an analysis of uncertainty expressions within a semantic system which incorporates at least some reference to probability measures or similarly rich structures. Similar approaches can

[3] A preorder (or quasiorder) is a reflexive and transitive binary relation defined on the elements of a set.

be found in the works by Swanson (2006, 2016), Moss (2015), Lassiter (2011a, 2017). A simple (perhaps the simplest) way to implement such a quantitative semantics for uncertainty expressions is to extend a possible world model with probability measures. More precisely, given the domain of the discourse S, each world $w \in S$ is associated with a function from propositions (subsets of S) to real numbers: $P_w : \mathcal{P}(S) \to \mathbb{R}$, such that P_w is additive and it assigns 1 to tautologies.[4] With this machinery in place, the meaning of *probably* can be defined as follows:

probably(p) is true in w *iff* $P_w(p) > 0.5$

The idea behind this definition is quite simple: a proposition can be said to be *probable* in a world just in case the probability of the proposition in that world is greater than 0.5. In other words, the proposition is more likely than not, which resonates with the intuitive Kratzerian definition above, while adding a more precise and quantitative sense to it. Along the same lines, the meaning of *possibly* and *certainly* can be defined by setting reasonably low (e.g. 0.1) and reasonably high (e.g. 0.9) probabilistic thresholds, respectively:

possibly(p) is true in w *iff* $P_w(p) > 0.1$
certainly(p) is true in w *iff* $P_w(p) > 0.9$

It has to be noted that the specific threshold values used in these definitions are intended to be first approximations, capturing pre-theoretic intuitions about the meaning of the corresponding expressions. As we will see momentarily, however, things can get more complicated than this. In any case, we assume these definitions (or some variations thereof) as our starting point in basically every chapter of this work.

The psychological and the logico-philosophical traditions represent two approaches with different goals and methodology and, despite having pretty much the same topic of interest, have rarely interacted with each other in the literature. That is, translation studies have generally not engaged with the formal intricacies which arise if we try to integrate uncertainty expressions (or any kind of non-trivial expression, really) into a general compositional semantics.[5] On the other hand, the logico-philosphical approaches which *did* focus on formal intricacies have generally (with some exceptions) relied on armchair intuitions about the meaning and use of uncertainty expressions, rather than experimental data.[6] Both approaches have their own merits, and we believe that both would greatly benefit from interacting with each other, and so would our understanding of uncertainty expressions. For this reason, one of the goals of this dissertation is to bring the two traditions closer to each other.

1.4 Context sensitivity

There is at least one recent exception to the general observation that the psychological and the linguistic traditions have not engaged with each other's results. We have al-

[4] Additivity means that $P_w(p \cup q) = P_w(p) + P_w(q)$ for any $p, q \in \mathcal{P}(S)$ with $p \cap q \neq \emptyset$. Tautologies are propositions true in every possible world, i.e. they contains the entire state space S.

[5] The reader will be able to get a taste of such intricacies later on in this dissertation.

[6] While linguists have certainly performed countless empirical investigations about epistemic modals, linguistic data specifically about probability expressions is arguably sparse.

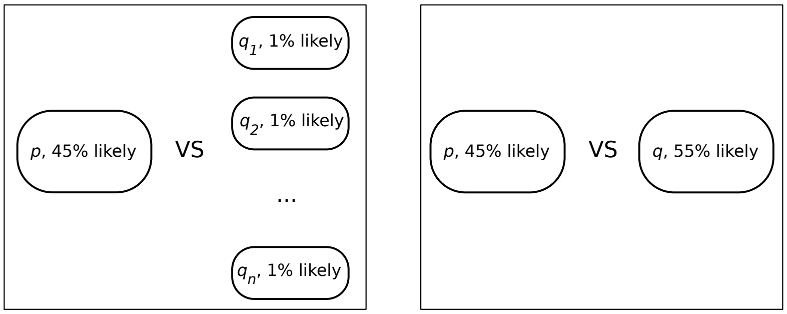


Figure 1.2: Two possible ways to partition the event space: on the left, the event p with 45% chances to happen is compared with several events $q_1, \dots, q_n$ each with small chances; on the right, the same event p is compared with a single alternative event q with higher chances.

ready mentioned that several studies in the translation approach established that many factors can influence the production and interpretation of uncertainty expressions, and in particular that the context in which these expressions are used seems to play a crucial role. Here, and in Chapter 2, we focus on the works by Teigen (1988) and Windschitl and Wells (1998), specifically because they pose a challenge for formal theories of the meaning of uncertainty expressions, a challenge which has been accepted by philosophers of language and linguists in recent years (Yalcin, 2010; Lassiter, 2011a).

In a nutshell, Teigen and Windschitl and Wells found that the way in which the possible alternative events to the event under discussion are presented to the subjects, or conceptualized in the context, has an effect on subjects' use and interpretation of uncertainty expressions. For example, given an event p which has 45% of chances to happen, subjects tend to say that *p is likely*, when p is presented alongside a set of many alternative events $q_1, \dots q_n$ each with very small chance to happen (e.g. 1%, see left panel of Figure 1.4 for an intutitve representation) significantly more often than when p, with the same chance of 45%, is presented alongside a unique alternative event q with 55% chance to happen (right panel of Figure 1.4).

It is not hard to see how this phenomenon, dubbed *Alternative Outcome Effect* by Windschitl and Wells, can be puzzling under the light of the aforementioned formal theories of uncertainty expressions. The issue is discussed in detail by Yalcin (2010) and Lassiter (2011a). Assuming that a sentence of the form *likely(p)* means that p *is more likely than not* (as in the qualitative approach *á la* Kratzer) or that $P(p) > 50\%$ (as in a simple probabilistic semantics), how is it possible that subjects' intuitions regarding utterances about the same event p, with the same chances of 45%, can vary depending on the distribution over the *other* events appearing in the context?

Yalcin (2010) and Lassiter (2011a) both recognize the importance of the issue and explore a number of approaches to reconcile the empirical data with a Kratzerian take on the semantics. This is the topic of the first part of Chapter 2, where we look in more detail at the Alternative Outcome Effect and at a thought-provoking intuition put forward by Lassiter about the sensitivity of uncertainty expressions to the Question Under Discussion (Roberts, 1996, 2012) in the conversation, sensitivity which –according to Lassiter– could help explaining the effects attested by Teigen and Windschitl and Wells.

The notion of Question Under Discussion (QUD) in a conversation is a formal concept that can be used to make sense of intuitive concepts such as the topic of a conversation or “what is relevant” in a conversation. In a nutshell, the QUD can be thought of as the question which the participants are (explicitly or implicitly) trying to answer with their contributions at any given point of the conversation (Roberts, 1996, 2012). The phenomenon linking QUD and uncertainty expressions put forward by Lassiter (2011a) is the following. Imagine a fair lottery with one winning ticket out of one million sold tickets. Imagine Mr. Burns has bought $300,000$ tickets, and $700,000$ other people have bought one ticket each. Therefore, Mr. Burns' odds to win are $^3/_{10}$. We know all this, and we make the prediction in (1):

(1) It's likely that Mr. Burns is the winner.

True or false? Lassiter's intuition (which he tested in informal surveys) is that in this context the evaluation of (1) can vary depending on the QUD in the conversation. In particular, (1) is not very likely to be evaluated as true if we are addressing a polar question such as (2-a):

(2) a. Is Mr. Burns the winner?
b. (Yes,) it's likely that Mr. Burns is **the winner**.

In (2-b), the boldface font indicates that the focus of the sentence is *the winner*, which strengthens the interpretation of the sentence as being intended to be about whether Mr. Burns is the winner or not (Roberts, 1996, 2012). Intuitively, if we compare Mr. Burns' odds to win ($^3/_{10}$) with the odds to lose ($^7/_{10}$), it seems hard to conclude that (2-b) can be true in the context. Consider now (2-b) repeated below as (3-b) but intended as an answer to a *who*-question such as (3-a):

(3) a. Who's the winner?
b. It's likely that **Mr. Burns** is the winner.

Lassiter's intuition and findings are such that (3-b) is more likely to be evaluated as true, because the *who*-question is about the specific outcome of the lottery and makes it so that Mr. Burns' odds to win ($^3/_{10}$) are not compared to his own odds to lose but to average winning odds across participants ($^1/_{700001}$). In other words, Mr. Burns *is* the most likely participant to win by a wide margin, which, under the question in (3-a) seems enough to make us conclude that it is likely that he will win.

Building on the discussion of Yalcin's and Lassiter's works, the main sections of Chapter 2 of the dissertation are experimental. The chapter allows us to introduce the reader to the urns-and-balls experimental setting which will return in many chapters of the dissertation. We report on design and results of an empirical study, which investigated the Alternative Outcome Effect and the QUD-Effect. Spoiler alert: we will be able to report only a half positive result, because while the Alternative Outcome Effect proved reasonably easy to replicate, the QUD-Effect was more elusive.[7]

1.5 Adverbs and adjectives

In the preceding discussion we have glossed over the fact that the category of uncertainty expressions, which we take to include epistemic modals and verbal probabilities, is general enough to include expressions belonging to several different syntactic categories (parts of speech), i.e. at least auxiliary verbs (*might*), adjectives (*possible*, *probable*) and adverbs (*possibly*, *probably*). Nowhere in the preceding paragraphs have we reflected on the possible syntactic differences between different uncertainty expressions. In particular, we have tacitly assumed that the adjectival and adverbial forms of any specific expression (e.g. *possible/possibly*, *probable/probably* and *certain/certainly*) can be treated essentially as synonyms. This was for good reasons, as the synonymity between corresponding pairs of uncertainty adverbs and adjectives appears to be intuitive enough, at least at first sight:

(4) a. It's probable that the election is rigged.
b. The election is probably rigged.

Moreover, the synonymity is generally assumed in all the literature cited so far.

However, a relevant number of authors have not failed to notice that *there are*, in fact, several differences between uncertainty adverbs and adjectives (e.g. Bellert, 1977; Nuyts, 2001a, 2001b; Nilsen, 2004; Piñón, 2006; Ernst, 2009). A series of recent works by Lavi Wolf and colleagues (Wolf, 2014; Wolf & Cohen, 2009; Wolf, 2015; Wolf, Cohen, & Simchon, 2015) have collected and summarized a significant body of literature on this topic. The observed phenomena seem to span across syntax, semantics and pragmatics. Let us look at some examples.

First, uncertainty adjectives seem to be much easier to embed than adverbs. For example, under negation:

(5) a. It's not probable that Trump will win the next election.
b. *Not probably, Trump will win the next election.

While (5-a) sounds perfectly fine, (5-b) seems hardly grammatical at all.

Next, uncertainty adverbs seem to be *speaker oriented* (c.f. Jackendoff, 1972) or in other words "more subjective" than adjectives:

(6) a. *Alice:* Trump will probably lose.
b. *Bob:* #Who says so? / #Whose opinion is this?

(7) a. *Alice:* It's probable that Trump will lose.
b. *Bob:* Who says so? / Whose opinion is this?

[7] The research reported in Chapter 2 was conducted as a follow-up to the study reported in Herbstritt (2015). The basic assumptions and the goal are essentially the same: we investigate the Alternative Outcome Effect and the QUD-Effect on uncertainty expressions. However, the study presented here is more refined, the experimental task is designed to be more interesting for the participants, and the analysis is much more thorough (and Bayesian).

Bob's follow-up question in (6-b) sounds out of place, as it seems clear from Alice's remark in (6-a) that she is communicating her own beliefs. On the other hand, Alice's use of the adjective *probable* in (7-a) does not appear to give rise to the same inference, and Bob's question sounds appropriate.

Next, adverbs and adjectives appear to give rise to different agreement/disagreement conversational patters:

(8) a. *Alice:* Trump will probably lose.
b. *Bob:* I agree.
c. Bob agrees that Trump will lose.

(9) a. *Alice:* It's probable that Trump will lose.
b. *Bob:* I agree.
c. Bob agrees that it's probable that Trump will lose.

Intuitively, it seems that agreeing to *Trump will probably lose* amounts to agreeing that Trump will lose: (8-c) is a good paraphrase of (8-b). On the other hand, agreeing to *It's probable than Trump will lose* amounts to agreeing to the full proposition that it is probable that Trump will lose: (9-c) is a good paraphrase of (9-b).

Several more examples like these are discussed in Chapter 3, which is entirely devoted to investigating the differences between uncertainty adverbs and adjectives. At the core of the chapter lies the issue, raised by the set of phenomena collected by Wolf and colleagues, of how to reconcile the traditional assumption of the synonymity between each pair of uncertainty adverbs/adjectives and the fact that they seem to exhibit appreciably different behaviors, not only regarding their syntactic distribution but also semantic and pragmatic effects.

Wolf and colleagues put forward their own theory of uncertainty expressions in order to solve the puzzle, the so-called *Speech Act Modifier theory* (SAM). In particular, they propose a rather unconventional take on uncertainty adverbs, analyzed as expressive devices modifying the assertive force at the speech-act level, as opposed to uncertainty adjectives, analyzed more traditionally as modals, contributing compositionally to the meaning of the sentence in which they appear. On the basis of this difference, according to the authors, all the observed differences between adverbs and adjectives can be explained. In Chapter 3 we look at SAM in more detail. As we will see, our goal is not so much to criticize this approach (although we will have critical remarks), but rather to explore how far we can go in explaining the data without assuming such an extreme and sharp distinction between adverbs and adjectives. Do we really *need* to suppose that uncertainty adverbs are expressives, that they do not contribute compositionally to the meaning of the sentence in which they appear and that they only apply at speech-act level?

Our answer is that such an extreme approach might not be necessary. In Chapter 3 we propose an alternative account of the phenomena collected by Wolf and colleagues. In a nutshell, we stick with the assumption of synonymity at the semantic level, and we try to explain the observed differences as pragmatic phenomena. The crucial idea underlying our approach is Robert's QUD-sensitive semantics for epistemic auxiliary *might* (Roberts, 2015, 2017), which we summarize in Chapter 3 and then adapt to the case of uncertainty adverbs and adjectives.

1.6 Higher-order uncertainty

Generally speaking, the translation attempts carried out in the psychological tradition seem to aim at establishing a one-to-one correspondence between expressions and levels or intervals of objective probability.

Recall the question asked by `zonination` to their participants, i.e. *What probability would you assign to the phrase "X"?*. When we ask such a judgment of probability, it seems natural to take the answer as reflecting the participant's intuition about the communicated probability in an objective sense. For example, imagine flipping a coin and saying that *there's an about even chance that the coin will land heads*. This seems to convey that the coin is fair, or in other words that the likelihood or chance that the coin will land heads (i.e. its bias) is equal to 50%. Similarly, imagine hearing that *it is almost certainly going to snow tomorrow* on the weather channel. This seems to convey that the chance of snow is pretty high, say something between 90% and 100%.

Needless to say, these intuitions and examples are not enough to conclude that in general uncertainty expressions "stand for" numeric values or intervals representing objective probabilities, and even `zonination` seems to be aware of the importance of other levels, including "we believe" among their expressions.

If a meteorologist runs her simulations and then announces that "snow is likely", it seems reasonable to assume that she is referring to a rather specific interval of objective chance values assigned to the event of snow, say something between 60% and 80%. Now imagine hearing the meteorologist claim and accepting it with the usual advisable moderate skepticism. A few moments later, a friend of yours confidently announces that "with a sky looking like that, it's definitely going to snow". Now, is *he* referring to a specific interval of objective chance values too? And what about the random guy at the grocery store assuring you that "it's been snowing for days, it can't possibly snow today as well."? And, again, after hearing all these different probabilistic claims about the event of snow, what interval of objective chance values will *you*, the listener, assign to the event of snow, if any at all?

These are real-life examples involving situations where uncertainty expressions seem to associate with increasingly vague, subjective, intuitions of uncertainty rather than to the more objectively sounding concepts of bias or chance. To drive the point home more precisely, we can look at an artificial example as well. Imagine an urn containing 100 balls of two different colors, e.g. red and blue. You do not know the exact contents of the urn and you cannot look inside. What you are allowed to do is drawing a number of balls at once and look exclusively at the drawn (or *accessed*) balls. For example, imagine drawing 80 balls from the urn and observing that 50 of the balls are red and the remaining 30 are blue. What do you know about the contents of the urn after this partial observation? How many red balls do you think there are, in total, in the urn? A reasonable approximation of the ratio of red balls in the urn is provided by the ratio of *observed* red balls over the number of *accessed* balls, i.e. $^{50}/_{80} = 0.625$ or equivalently 62.5%. This number represents the best guess available to a rational agent as to what the "bias towards red" might be for the given urn in the given situation or, in other words, the objective chance of randomly drawing a red ball from that urn.

On the basis of this number, if you are asked to draw a ball at random and make a prediction about its color using an uncertainty expression, you might say something along the lines of (10):

(10) The ball will probably be red.

Intuitively speaking, it seems to us that (10) would be a perfectly acceptable prediction in the given situation. Moreover, if we look again at Figure 1.1, we can see that a chance level around 62% is perfectly compatible with the phrase *likely* and, generally speaking, results from the translation studies would agree that the association holds. And finally, according to the simple probabilistic semantics sketched above, (10) is to be evaluated as true, because the (agent's best guess about the) chance of the event *the ball will be red* exceeds the threshold of 0.5.

So far, so good. Then what is the puzzle? Imagine the same urn as before, containing 100 red and blue balls. Once again, you do not know the ratio, but you can draw some balls from the urn in order to get an idea, except this time you are only allowed to draw exactly 8 balls from the urn (instead of 80). Imagine observing that 5 balls are red and 3 are blue. What do you know about the contents of the urn based on this observation? The best guess available to you is once again the ratio of observed red balls: $^5\!/\!_8 = 0.625 = 62.5\%$. This is exactly the same number as before. And yet, when asked to make a prediction in this situation, would you confidently say that a randomly drawn ball will probably be red, i.e. something along the lines of (10), which seemed appropriate before? It seems to us that looking at 8 balls out of 100 and observing that 5 of them are red is really not enough evidence to say something like (10), and that (11) or similar would be much more appropriate:

(11) The ball might be red.

What changed? If the appropriateness of (10) and (11) is intuitively very different in the two situations, despite the fact that (our best guesses about) the objective chance of the focal event "a randomly drawn ball will be red" is the same in both situations, then we should infer that (our best guess about) the objective chance of the event is not the (only) factor that matters in determining the truth value or (most likely) the degree of appropriateness of a sentence containing uncertainty expression such as (10) and (11).

What matters, then? A large part of this dissertation attempts to answer this question. Our main intuition, which we share with Moss (2015) and we assume as our working hypothesis, is that several (at least two) levels or layers of uncertainty play a crucial role in our use and interpretation of uncertainty expressions. We will return to this terminology in Chapter 4, where we introduce an experimental scenario, based on urns and colored balls, in which two levels of uncertainty can be easily manipulated in an intuitive way. For the purpose of this introduction, suffice it to say that we assume a first level of uncertainty, which corresponds to the objective chance of an event (think: the bias of a coin, the overall ratio of red balls inside the urn). Knowing the objective chance of a stochastic event A corresponds to having perfect information about A, but it is still not enough to known with certainty whether A will happen or not. Now, what if we are not even sure about the objective chance of A? This is the second level of uncertainty, which we will refer to as *higher-order uncertainty*, and it corresponds to having some subjective uncertainty (or, equivalently, some degree of confidence) about the objective chance of an event.

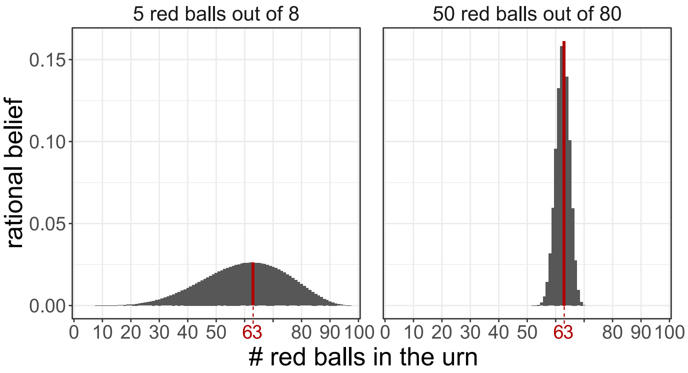


Figure 1.3: Rational belief distributions about the contents of an urn containing 100 balls after drawing 8 balls and observing that 5 are red (left panel) and after drawing 80 balls and observing that 50 are red (right panel).

To go back to the example above, in both situations (our best guess about) the objective chance of the event “red ball” is the same (i.e. 62.5%) but the degree of confidence with which we would make those guesses is intuitively very different: it is reasonably high if we have observed 80 balls out of the total 100 contained in the urn, much lower if we have observed only 8. This difference is apparent in Figure (11), where we display the belief distributions about the composition of the urn that an unbiased rational agent *should* hold after each of the partial observations in our example (according to a normative model of rational belief formation to which we will return in Chapter 4). Both distributions peak at 63, indicating that 63 is the single most credible amount of red balls contained in the urn. However, the distribution on the right is much more closely concentrated around the mode, indicating that the agent’s beliefs are much more precise. *This difference*, we claim, is behind the difference in the appropriateness of (10) and (11) in the two situations.

To the best of our knowledge, the distinction between levels of uncertainty has not been explicitly discussed in previous empirical work about uncertainty expressions, let alone actively manipulated in experimental studies. Among the formal linguists, Moss (2015) and Lassiter (2018) recognize the importance of such a distinction and explore it in some detail. As the reader might have inferred on the basis of the length of this section, entirely dedicated to introducing the distinction between levels of uncertainty, this distinction plays a very important role in our work, and four chapters of the dissertation are related to it in one way or another.

In particular, in Chapter 4 we discuss some intuitions about the distinctions between objective/subjective uncertainty and first-order/higher-order uncertainty, and we attempt to capture some of these intuitions in a way which is both precise and formal but also intuitive and easy to implement and manipulate in experimental settings. The basic idea was introduced above: we manipulate the number of balls which can be drawn from an urn and observed by an agent, in such a way that the ratio of, say, red balls is constant but the confidence associated to the observation varies. The chapter reports on an interesting result, namely that participants in our experiment can be taken to behave in an approximately rational way —i.e. according to a normative rationalistic model of belief formation— when they are asked to reason under (higher-order) uncertainty about the contents of an urn.

On the basis of this result we developed an experimental study to scrutinize our hypothesis that higher-order uncertainty plays a crucial role in participants' production and interpretation of uncertainty expressions. Materials, procedures and results of the experiment are detailed in Chapter 5. The goal of the chapter is to investigate *how* participants use and interpret simple uncertainty expressions (e.g. *probably*, *possibly*) in situations of higher-order uncertainty. Building on the answer to this question, the following Chapter 5 attempts to answer the question *why* participants behave the way they do. Our answer will take the form of a probabilistic pragmatic model of language use and interpretation, which is discussed in detail in the chapter and whose predictions are evaluated against experimental data. Finally, we apply the same data-driven modeling approach to complex uncertainty expressions, which we introduce in the next section and which are the topic of Chapter 7.

1.7 Complex uncertainty expressions

So far in this introduction we have talked exclusively about simple uncertainty expressions. That is, one-word expressions such as *certainly*, *likely* and *possible*, or modified versions thereof such as *almost certainly* and *highly unlikely*, where the modifiers do not belong themselves to the language of uncertainty. Instead, we refer here to complex phrases obtained by combining two (and potentially more) uncertainty expressions, such as *might be likely*, *certainly possible*, *probably unlikely*. The vast majority of previous work on uncertainty expressions, both in psychology and linguistics, have focused on simple constructions.[8] However, complex uncertainty expressions are attested in English (and other languages) as well, and some authors have recently turned their attention to them, investigating their distribution, meaning and use (Moss, 2015; Lassiter & Goodman, 2015b; Lassiter, 2018).

Recall the two situations described above, i.e. drawing 80 balls from an urn and observing that 50 of them are red VS drawing 8 and observing that 5 are red. We said that (our best guess about) the objective chance of drawing a red ball from the urn should be the same after either observation, but at the same time our confidence is much lower in the latter case. If it turns out to be true that this difference affects our use and interpretation of simple uncertainty expressions, then we should expect it to have an effect on our use and interpretation of complex uncertainty expressions as well. This is a point originally made by Moss (2015), who probes our intuitions about two predictions such as (12) and (13) in two situations of uncertainty analogous to our urn observations:

[8] As pointed out by Lassiter (2018), the reason behind this fact might be found in the ties between the linguistic analysis of uncertainty expressions, especially modals such as *must* and *might*, and modal logic, i.e. the formal investigation of the logical properties of the concepts of possibility and necessity. In fact, the system of modal logic usually regarded as the one capturing the intuitive meaning of *might* and *must* allows us to prove that any number of stacked modal operators can be always reduced to the innermost operator. For example: *It's necessary that it's possible that P* is equivalent to *It's possible that P*.

(12) It's certainly likely that the ball is red.

(13) It might be likely that the ball is red.

Clearly, the point here is that despite the fact that the nested expression is the same in both sentences (*likely*) and despite the fact that the objective chance of drawing red is the same in both situations (62.5%), a sentence such as (12) sounds much more appropriate in the 50/80 situation (where we are indeed pretty confident that the odds of drawing red are high) than in the 5/8 situation (where we suspect that the odds are high, but we cannot really tell). In the latter situation, a more cautious sentence such as (13) seems preferable.

The puzzle raised by this example is clearly analogous to the one raised by our discussion in the previous section of this chapter, in particular the comparison between the sentences in (10) and (11). If the objective chance of red is the same in the two situations, what makes it so the sentences in (12) and (13) have different degrees of appropriateness in the two situations? As before, we hypothesize that what matters for evaluating sentences containing (simple and/or) complex uncertainty expressions is not only the objective chance of the focal event, but also (at least) one higher layer of uncertainty.

This is the topic of Chapter 7, in which we extend the pragmatic model developed in the preceding chapter in order to include complex uncertainty expressions. Moreover, we report on the results of two experimental studies, analogous to the ones described in Chapter 5, in which we investigate how participants use and interpret complex uncertainty expressions in situations of higher-order uncertainty. The data collected in the experiments is then used to train and evaluate our pragmatic model.

Structure of the dissertation. The succession of sections in this introduction mirrors the succession of chapters in the dissertation, almost in a recursive way. That is, Chapter 1 is the introduction itself, which the reader is reading right now and it introduces all the topics discussed in the dissertation. The first few sections of Chapter 1 introduce the remaining sections of the same chapter. In turn, these sections correspond, more or less, to the subsequent chapters of the dissertation. Luckily, the structure of the other chapters is more streamlined.

Chapter 2 is about the puzzle of context-sensitivity. First, we look at the Alternative Outcome Effect in more detail and at philosophers' and linguists' take on the consequences of the effect for standard semantic theories of uncertainty expressions. The focus of the Chapter is however experimental: we report on design and results of an empirical study collecting human data on the Alternative Outcome Effect and the closely related Question Under Discussion Effect hypothesized by Lassiter.

Chapter 3 turns to the uncertainty adjectives VS adverbs puzzle. First, we summarize the Speech Act Modifier Theory by Wolf and colleagues and the data they collected in order to support the theory. The focus then shifts to our own account of the semantic, pragmatic and conversational differences between uncertainty adverbs and adjectives, which is inspired by Robert's contextualist semantics and QUD-sensitivity of epistemic modality.

Chapter 4 is about higher-order uncertainty. First, we introduce a formal but hope-fully intuitive way to capture some of the intuitions behind the distinction between objective and subjective uncertainty. Subsequently, we report on design and results of an empirical study collecting human data about belief formation under uncertainty: we investigate whether a rationalistic model of belief formation introduced by Goodman and Stuhlmüller (2013) is good enough in our setting to be assumed in the linguistic model developed in subsequent chapters of the dissertation.

Chapter 5 is entirely dedicated to experimental data on the use and interpretation of simple uncertainty expressions. We report on design and results of two empirical studies investigating how English native speakers communicate with simple expressions such as *certainly*, *possibly* and *probably* in situations of higher-order uncertainty.

Chapter 6 is about our pragmatic model of uncertainty expressions. First, we introduce the framework of the Rational Speech Act model (RSA) in an intuitive manner. Second, we spell out the details of our version of RSA, specifically aimed at modeling pragmatic use and interpretation of simple uncertainty expressions in situations of higher-order uncertainty. We train the model on the data collected in the experiments reported in Chapter 5, and we evaluate model's performance from a Bayesian perspective.

Chapter 7 turns to complex uncertainty expressions. First, we extend the model developed in Chapter 6 in order to encompass messages containing nested occurrences of uncertainty expressions, such as *certainly possible* and *probably likely*. We then turn to empirical data. We report on design and results of two additional empirical studies, very similar to those reported in Chapter 5, investigating English native speakers' intuition about complex uncertainty expressions. The collected data are then used to train and evaluate the model.

Publications. The research reported in Chapters 2-3 is unpublished. Chapter 2 is based on a pre-registered project by Michele Herbstritt and Michael Franke (publicly available at `https://osf.io/5u3gw`), and it substantially expands on the idea already explored in Herbstritt (2015) (presented at ESSLLI Student Session 2015). Chapter 3 is based on a talk with the title *On the difference between modal adverbs and adjectives*, given by Michele Herbstritt at the workshop New Ideas in Semantics and Modeling (Paris, September 8, 2016). The research reported in Chapters 4-7 appeared with some differences and a slightly more condensed form as Herbstritt and Franke (2019) published in *Cognition*. Earlier versions of the work leading to the paper were presented at the 38th and 39th Annual Conference of the Cognitive Science Society. The conference papers appeared in the proceedings of the conference, respectively as Herbstritt and Franke (2016) and Herbstritt and Franke (2017).

Data and code. In the spirit of open science and reproducible research, the data collected in all the experiments reported in this dissertation together with the code used to run the experiments, analyze the data, implement the models, and generate the plots are publicly available at `https://github.com/mic-he/ProbExp-PhD`.

Chapter 2

Context dependency & QUD sensitivity

2.1 Introduction

In Chapter 1 we have seen that a relatively recent trend of research in formal linguistics convincingly argues in favor of an analysis of uncertainty expressions which goes beyond a purely qualitative account *à la* Kratzer and incorporates some reference to a quantitative measure of probability, likelihood or credence (e.g. Swanson, 2006; Yalcin, 2007, 2010; Lassiter, 2011a; Moss, 2015). As already mentioned, in this work we do not concern ourselves with the details of the arguments in favor or against such a quantitative or probabilistic approach, nor do we aim at evaluating the specific proposals available on the market or comparing them against each other. In a sense, we take the probabilistic approach as given. Every work of research needs a starting point, and this is ours, in essentially every chapter.

As seen in Chapter 1, one of the most basic implementation of the probabilistic approach is obtained by supplementing the usual possible-world semantics with probability measures on propositions. More in detail, assuming a set of possible worlds or states S we can associate each state $s \in S$ with a function from propositions (subsets of S) to real numbers $P_s : \mathcal{P}(S) \to \mathbb{R}$ such that P_s assigns 1 to tautologies and is additive.[1] This simple enough machinery allows us to define the semantics of, say, *probably* as follows:

probably(p) is true in s *iff* $P_s(p) > 0.5$

This chapter moves from here. In Section 2.2 we look at how this basic semantic definition relates to both empirical studies and linguistic intuitions highlighting phenomena of context-sensitivity of uncertainty expressions. The phenomena are the so-called *Alternative Outcome Effect* (Teigen, 1988; Windschitl & Wells, 1998) and the QUD-sensitivity hypothesized by Lassiter (2011a), which we explore in more detail

[1] Additivity means that $P_s(p \cup q) = P_s(p) + P_s(q)$ for any $p, q \in \mathcal{P}(S)$with$p \cap q \neq \emptyset$.

here. This sets the stage for Section 2.3, where we report the experimental design and data analysis of an empirical study (Experiment 1) run to investigate these phenomena. Section 2.4 concludes the chapter.

A note on pre-registration. This chapter is based on a research project pre-registered on the Open Science Foundation portal. This means that, in the spirit of open, transparent and reproducible research (see Munafò et al., 2017), all the details of the experimental design and planned statistical analysis have been recorded in an immutable and publicly available online form prior to the actual data collection.[2] No strategic data cleaning or *p*-hacking were involved in the data analysis. Further analyses going beyond the pre-registered project are explicitly marked as exploratory, and as such should not be considered as conclusive results but as starting points for future research.

2.2 Background

In Chapter 1 we briefly mentioned a series of psychological studies highlighting several sources of contextual variability involved in people's use and interpretation of probability expressions. In this section we look in slightly more detail at two of these studies, i.e. the works by Teigen (1988) and Windschitl and Wells (1998). The reason behind this choice resides mainly in the fact that the results reported in these works play a role in a more recent trend of linguistic research about probability expressions (Yalcin, 2010; Lassiter, 2011a), exemplifying the interdisciplinarity between psychology and linguistics that we deem so necessary for the understanding of the vocabulary of uncertainty and that we strive for in this dissertation too.

Alternative Outcome Effect. Both Teigen (1988) and Windschitl and Wells (1998) find empirical evidence for essentially the same phenomenon, which the latter call the *Alternative Outcome Effect*. In a nutshell: given an event p, the way in which the space of possible events (or outcomes) alternative to p is structured has an effect on the way in which people talk about p using probability expressions. To get a more concrete intuition, consider the following example, adapted from Yalcin (2010). Imagine a fair lottery, with exactly one winning ticket in a total of 1000 sold tickets. Suppose that John has bought 450 tickets (as unusual as it might seem). Therefore, the probability of him winning is equal to $\frac{450}{1000} = 0.45$. Now, can we appropriately describe this situation by saying, for example, that *John will probably win the lottery*? According to the Alternative Outcome Effect the answer to this question can be influenced by the way in which the remaining 550 tickets, those which John didn't buy, are distributed. For simplicity we can consider the following two extreme scenarios:

a. John bought 450 tickets. Kate bought all the remaining 550 tickets.

b. John bought 450 tickets. 550 other people bought one of the remaining tickets each.

[2] The project is publicly accessible at `https://osf.io/5u3gw`.

The intuition (backed up by experimental evidence) is that a sentence such as *John will probably win the lottery* is harder to assert (if not false) in the first scenario (which we will call *dual*) where the remaining 550 tickets were all bought by a single agent than in the second scenario (*plural*) where the remaining 550 tickets are uniformly distributed over an equal number of agents. What makes this result interesting is that the appropriateness of a sentence of the form *probably(p)* in a situation seems to depend, at least intuitively, on the likelihood of the event *p* to obtain, which is also the assumption behind the semantic definition given in the introduction. However, the likelihood that John will indeed win the lottery is exactly the same in both scenarios.

There are several possible ways to approach the Alternative Outcome Effect and to "solve the puzzle". For example, we can question the assumption that the likelihood of *p* is what matters when evaluating *probably(p)*, or that it is the only thing that matters. Alternatively, and perhaps more conservatively, we can keep that assumption and explain people's linguistic behavior as the result of the fact that people's intuitions about ratios and probabilities seem to be (and have been shown to be) often vague and imprecise and rely more on heuristics than on quantitative reasoning. Simplifying, this is the strategy adopted by Windschitl and Wells (1998) too: even if a large body of translation studies lead us to agree that *probably(p)* essentially means that *p* is more likely than not, people's linguistic behavior when probability expressions are involved follows what the authors call an *associative* reasoning system, "relatively quick and spontaneous [...] accompanied by an intuitive or gut-level sense [..., it] operates according to principles of similarity and contiguity", as opposed to a rule-based reasoning system, which "operates according to formal rules of logic and evidence" (Windschitl & Wells, 1998).

It does not fall within the scope of our work to discuss this specific proposal. What matters for us is that recent works in the logico-philosophical tradition of linguistics have accepted the challenges posed by the empirical findings of Teigen, Windschitl & Wells and others, in the context of the development of a formal semantics for probability expressions. We refer in particular to the works by Yalcin (2010) and Lassiter (2011a), who independently from each other take on the issue raised by the Alternative Outcome Effect. Both authors essentially reverse the traditional point of view adopted in psychology. Rather than assuming that *probably* has a fixed meaning such as "more likely than not", and that contextual factors affect people's linguistic behavior because people are "bad" at probabilistic reasoning, Yalcin and Lassiter put forward the hypothesis that perhaps contextual factors can directly affect the meaning (which therefore is not fixed, but context dependent) of *probably* and similar expressions.

Threshold semantics for *probably*. A concrete proposal for how to incorporate some form of context sensitivity in the semantics of *probable*, *probably*, *likely* and similar expressions is to treat them as relative expressions, much like vague relative adjectives (Kennedy, 2007). Simplifying, the adjective *tall* can be analyzed in terms of a context-dependent threshold semantics: *tall(j)* is true *iff* $\text{height}(j) > \theta_{\text{tall}}$, where θ_{tall} crucially depends on the context, in such a way that, for example, if John is $1.85m$ then *John is tall* can be evaluated as true if the (explicit or implicit) comparison is the average height of European males, but can also be evaluate as false if the comparison is the height of NBA players. The adjective/adverb *probable/probably/likely* can be analyzed along the same lines:

$$\text{probably}(p) \text{ is true } \textit{iff } P(p) > \theta_{\text{probably}}$$

This definition is obviously reminiscent of the one given in the introduction of this chapter, except for the fact that the probability of the proposition p is not compared to a fixed value such as 0.5, but to a context-dependent threshold θ_{probably}. But how is θ_{probably} computed? And, in particular, how does this semantics help in explaining the Alternative Outcome Effect? A possibility, briefly mentioned by Yalcin and more extensively discussed by Lassiter is that the threshold depends on the number of alternative outcomes in the state space. This dependency can in turn be made precise in different ways. The simplest way is perhaps to compute the threshold as the average probability of the events in the state space or, in other words, as the probability that every event in the state space would have if they were all equally likely: $\theta_{\text{probably}} = \frac{1}{|A|}$ where $|A|$ denotes the cardinality of the set of the alternative events. This approach has no trouble explaining the example of the Alternative Outcome Effect discussed above. In the dual scenario (a) there are two possible outcomes (either John wins or Kate wins), hence the threshold is equal to $\frac{1}{2} = 0.5$; now, $P(\text{"John wins"}) = 0.45$, which is smaller than the threshold, hence the sentence is evaluated as false. On the other hand, in the plural scenario (b) there are 551 possible outcomes (either John wins the lottery or one of the other 550 people does), hence the threshold is equal to $\frac{1}{551} \simeq 0.002$, which is much smaller than $P(\text{"John wins"}) = 0.45$; hence the sentence is evaluated as true.

Lassiter goes a step further and suggests that the set A of the alternative outcomes which matter for the computation of the threshold can be affected by the QUD in the conversation (Roberts, 1996, 2012). Different QUDs can structure the event space in different ways. For example, if the interlocutors are talking specifically about John, one may ask a polar question such *Will John win the lottery?* and this will likely lead the interlocutors to conceptualize the event space as containing two alternative outcomes, i.e. either John wins or he does not win. Conversely, a more general conversation about the lottery can be introduced by a *wh*-question such as *Who will win the lottery?*, which will likely lead the interlocutors to conceptualize the event space as containing several different alternatives. Empirical evidence for Lassiter's idea comes from an informal survey conducted by Lassiter, in which native English speaking consultants provided intuitions about the acceptability of sentences such as *John will probably win the lottery* and *It's likely that John will win the lottery* in situations analogous to the plural scenario discussed above but with two different QUDs. Specifically, different QUDs were introduced by manipulating the focal structure in the sentences. For example:

(1) a. **John** will probably win the lottery.
 b. John will probably **win** the lottery.

Where boldface font indicates the focused material. Assuming the same background scenario in which John has 450 tickets and 550 other people have one ticket each, Lassiter's finding is that people are more inclined to judge as true or acceptable a sentence with the focus structure exemplified in (1-a), where the stress on *John* can lead the listener to infer that the QUD is likely *Who will win the lottery?*, rather than with the focus structure exemplified in (1-b), where the stress on *win* makes the sentence congruent with a polar question such as *Will John win or not?*. The explanation is that under a *wh* QUD every alternative matters, hence the likelihood of John winning $P(\text{"John wins"}) = \frac{450}{1000}$ is compared with the average probability $\frac{1}{551}$ exactly as before, and the sentence is evaluated as true. Conversely, under a polar QUD the number of alternatives in the scenario is "overridden" by the only two salient alternatives in the QUD: either John wins or he does not, and the sentence is evaluated as false because $P(\text{"John wins"}) = \frac{450}{1000}$ which is smaller than $\frac{1}{2}$.

Before concluding, it should be noticed that the proposal of setting the threshold to the average probability of the salient alternative outcomes is easy to dismiss, according to Yalcin, if we consider a slightly more contrived variation of the scenarios involved. For example:

c. John bought 450 tickets. Kate bought 500 tickets. 50 other people bought one ticket each.

Yalcin's intuition is that *John will probably win the lottery* is evaluated as false in this scenario, despite there being 52 alternative outcomes and $P(\text{"John wins"}) = 0.45$ being bigger than $\frac{1}{52} \simeq 0.02$. Appealing to different QUDs does not seem to help either. We are inclined to agree with Yalcin that the simple proposal of equating θ_{probably} to $\frac{1}{|A|}$ might be in fact too simple. However, we think of it as a first step in the process of formalizing the (likely) more complicated effects of the ways in which the state space is structured in the context or conceptualized in the conversation on the meaning of probability expressions. But the fact that *some* effects appear to be present should be the starting point, and it is certainly worthy of deeper experimental investigations which go beyond linguists' intuitions and informal surveys.

2.3 Experiment 1

Introduction. Which factors play a role in determining the truth values of sentences such as *John will probably (not) win the lottery* and *John has certainly wasted his money*?

Let us assume a probabilistic semantics for *probably* and *certainly* of the kind introduced above:

probably(p) is true *iff* $P(p) > \theta_{\text{probably}}$
certainly(p) is true *iff* $P(p) > \theta_{\text{certainly}}$

An obvious first answer to our question is: the likelihood that the event p happens should play a role in determining the truth value of *probably(p)* and *certainly(p)*. This is our first hypothesis. It is not theoretically very interesting, but it provides a baseline benchmark for the experimental setting reported in this section. If we do not manage to substantiate this hypothesis with our data, this might be a sign that our procedure is not as sound as we would like it to be.

In order to test this hypothesis, we need to manipulate the likelihood of particular events to happen. In this setting (as well as in most of this dissertation) we do so

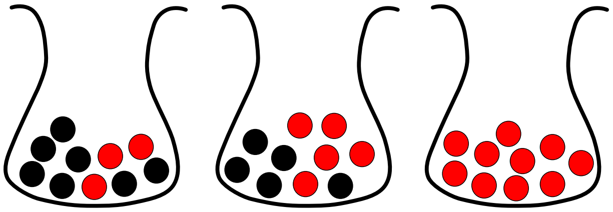


Figure 2.1: Examples of urns with different likelihood of drawing a red ball: from left to right $3/10$, $5/10$, $10/10$.

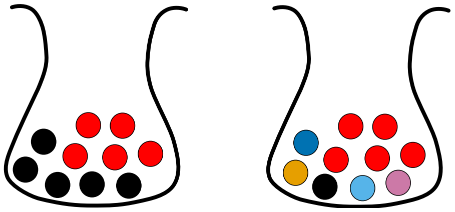


Figure 2.2: Examples of urns with same likelihood of drawing red but different scenarios: dual on the left, plural on the right.

by manipulating quantities and colors of balls contained in an urn. More precisely, imagine an urn containing 10 balls of two different colors, e.g. 8 balls are red and 2 are black. If we consider the event that a randomly drawn ball is red, then its likelihood is $8/10 = 0.8$. Conversely, the event that a randomly drawn ball is black has likelihood $2/10 = 0.2$. In this chapter we will often talk about likelihood *values*, referring to the quantities of balls of a certain color contained in the urn (see Figure 2.1 for some examples of pictures used in the experimental stimuli).

More interesting answers to our opening questions come from the results and intuitions found in the psychological and linguistic literature summarized in the previous section. Even if we do not know exactly if/how different ways of conceptualizing the set of alternative events contributes in determining the semantic thresholds for probability expressions, it seems reasonable to at least hypothesize that different scenarios and/or QUDs should have an effect (be it semantic or pragmatic) on the production of these expressions.

When it comes to scenario, the hypothesis is that *probably* and *likely* are easier to assert in plural scenarios than in dual scenarios, even if the likelihood of the event is the same. This is essentially the result found by Teigen (1988) and Windschitl and Wells (1998) and replicated in Herbstritt (2015). Our expectation is then to be able to replicate this result here as well. The urn setting allows us to easily implement the distinction between dual and plural scenarios. Given a certain quantity of red balls N, a dual scenario is implemented with an urn containing N red balls and $10-N$ balls of a single different color, whereas a plural scenario is implemented with an urn containing N red balls and $10-N$ balls of several different colors (see Figure 2.2 for some examples).

Moving to QUDs, we follow the intuition and the result of the informal survey reported by Lassiter (2011a). Our hypothesis is then that *probably* and *likely* are easier to assert when responding to a *wh*-question than to a polar question, even if the likelihood of the event is the same. Now, manipulating QUDs of a conversation in an artificial experimental setting might not be a trivial matter. In this work we attempt to do so by carefully designing a setting where the interactive and cooperative nature of the conversation is essential to the task: participants are not asked to passively evaluate sentences, but they actively take on the roles of interlocutors who share a common goal and can only achieve it through linguistic cooperation. From this point of view, asking questions can be seen as a tool to achieve a goal, and different kinds of questions (*wh*, polar, ...) can become tools to achieve different goals. In this way we hope to create an environment where, much like in real-life conversation, QUDs actually matter for the participants. The full details of our experimental setting are discussed in the next section.

2.3.1 Setup and procedure

Participants. We recruited 50 workers with IP addresses located in the USA on Amazon's Mechanical Turk (AMT) crowd-sourcing service. The workers were paid 2.00 USD for their participation, amounting to an average hourly wage of approximately 10 USD.[3]

Materials and procedure. The experiment was presented to the participants as a game with two players. The players cooperate in order to place bets on the color of a ball randomly drawn from an urn. For example, imagine looking inside an urn and observing that it contains 10 balls, 8 of which are red and the remaining 2 are blue. Imagine drawing a ball at random from the urn; before looking at the ball, place your bet: will the ball be red? or blue? In our game, only one player can look inside the urn to directly know its contents, but she cannot place any bet. Her job is to send a message to the other player, giving him a hint about the contents of the urn. For this reason, we call her the *sender*. The other player, whose job is to place the bet, has no direct access to the contents of the urn and can only rely on the information received from the sender. We call him the *receiver*.

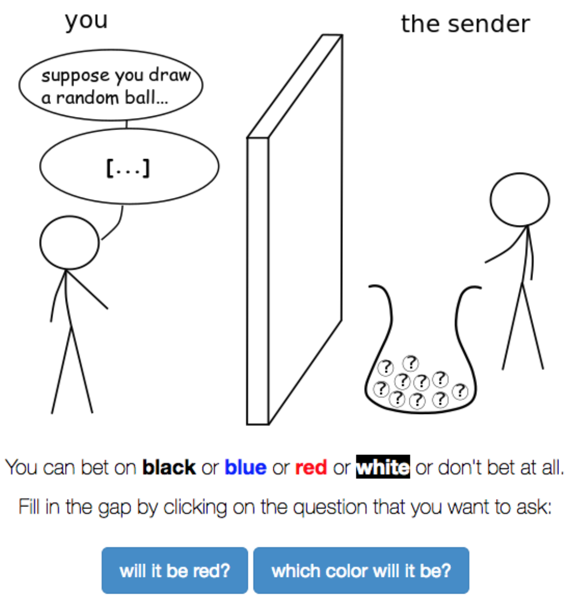
More in detail, this is how a match of this game would play out. First, the receiver is given a betting option, for example

> *You can bet on red, or bet on blue, or bet on green, or not bet at all.*

Second, the receiver asks exactly one question to the sender in order to get some information about the contents of the urn, for example

> *Which color will a randomly drawn ball be?*

[3] All the experiments reported in this dissertation were implemented and run within the psiTurk framework (Gureckis et al., 2016). Code and data for the study reported in this chapter are publicly available at `https://github.com/mic-he/ProbExp-PhD/tree/master/chapter2`.



You can bet on **black** or **blue** or **red** or **white** or don't bet at all.

Fill in the gap by clicking on the question that you want to ask:

Figure 2.3: Example of a training trial.

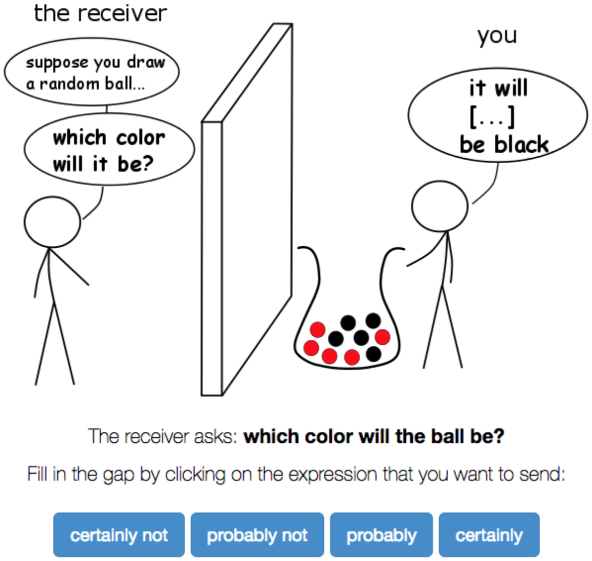
Third, the sender, who can look inside the urn, reads the question and sends one message to the receiver, choosing from a list of options (more about this below). For example, the sender can send a message such as *A randomly drawn ball will probably be red*. Finally, the receiver reads the message and decides if/how she wants to bet.

The participants to the experiment first read a cover story introducing them to this setting, then completed two phases. In the first phase participants played in the role of receiver. In the second phase they played in the role of sender. We refer to the first phase as *training*, because its main goal was mostly to get the participants acquainted with the the game. After reading a set of instructions describing the first phase in more detail, participants completed three trials in the role of receiver. In each trial, participants were assigned to a betting option, selected at random among two possible variants. The *single* variant mentioned only one color, e.g. *You can bet on red or not bet at all*; the *multiple* variant mentioned four different colors, e.g. *You can bet on black or blue or red or white or not bet at all*. Each participant always saw at least one betting option (and at most two) for each variant.

In each trial participants were then required to ask a question to the sender. This was a forced choice task. Participants were asked to complete a sentence of the form

> *Suppose you draw a random ball, ...*

by clicking on either of two possible continuations. One continuation contained a polar question (e.g. *Will it be red?*); the other continuation contained a *wh*-question (e.g. *Which color will it be?*). We recorded participants' choices of question. Figure 2.3 contains a screenshot taken from one of the training trials. After clicking on either



The receiver asks: **which color will the ball be?**

Fill in the gap by clicking on the expression that you want to send:

Figure 2.4: Example of a critical trial.

question, participants were told that the other player had received their question and answered with a sentence such as *It will probably be red* or similar. Finally, participants were asked to decide if/how they wanted to bet, based on the sender's answer.

The training phase was followed by a screen providing brief feedback to the participants and introducing them to the second phase, which we call *production*, where they would play in the role of sender.

The production phase was the critical experimental phase, aimed at collecting production data in each combination of likelihood values (3/10, 4/10, 5/10), scenarios (*dual*, *plural*) and QUDs (*polar*, *wh*). Participants completed 12 critical trials, one for each possible combination, in random order. Additionally, the critical trials were interspersed with four control trials, in which only trivial likelihoods were displayed (0/10 and 10/10) together with random combinations of scenarios and QUDs.

In each trial, participants observed a picture which displayed the receiver asking the question corresponding to the current QUD condition and the urn containing the number of balls corresponding to the current value condition and structured according to the scenario condition. Participants were asked to complete the message to be sent to the receiver, of the form

It will … be red

by clicking on one of four possible expressions: *certainly not*, *probably not*, *probably*, *certainly*. We recorded participants' choices of expressions. Figure 2.4 contains a screenshot taken from one of the critical trials.

The production phase was followed by a screen giving feedback to the participants

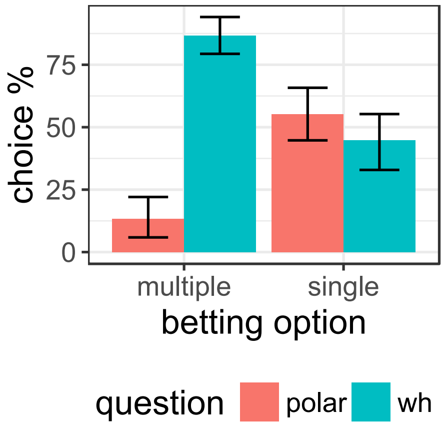


Figure 2.5: Question choice percentage for each betting option. Black bars represent bootstrapped 95% confidence intervals.

and a final form which participants could use to leave their feedback about the experiment.

2.3.2 Results

Exclusion criteria. First of all, we excluded data points obtained from two participants whose (self-reported) native language was different from English.

Subsequently, we turned to the data collected in the control trials of the production phase. As mentioned above, each participant completed four control trials, each displaying only trivial likelihood values. There is a rather obvious association between expressions and trivial values: observing 0 red balls out of 10 should correspond to the message *The ball will* certainly not *be red*; conversely, observing 10 red balls out of 10 should correspond to the message *The ball will* certainly *be red*. Because of this, we expected participants to behave in the obvious way at the very least in half of the trials. That is to say, we decided to exclude data points obtained from participants who had not responded as expected in 3 or more control trials.[4] As it turned out, no participant failed more than one control trial, with five participants failing exactly one. Therefore, no data points were removed based on participants' performance in control conditions.

Finally, we checked the feedback given at the end of the experiment. Nothing worthy of notice emerged from this exploration, so no data points were removed based on participants' feedback.

Training data. We begin with the data collected in the training trials. Our goal here is mainly to explore whether our intuition is on the right track, that a *wh*-question is more

[4] This decision was made before running the experiment. See pre-registered project at `https://osf.io/5u3gw/`.

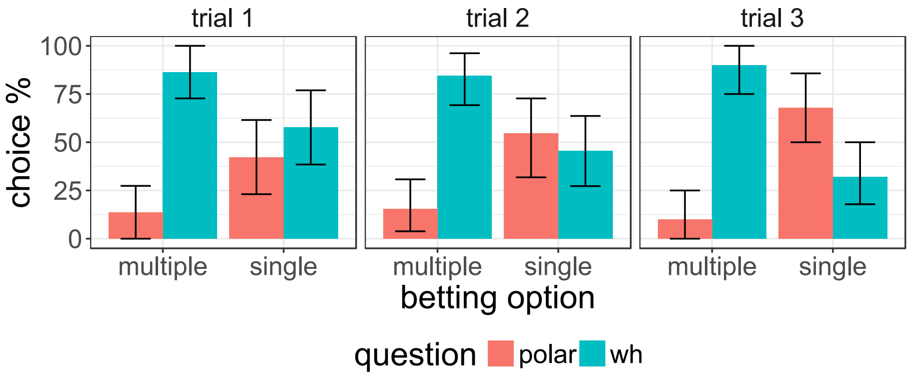


Figure 2.6: Question choice percentage for each betting option, split by trial. Black bars represent bootstrapped 95% confidence intervals.

appropriate in situations with multiple betting options and, conversely, a polar question is more appropriate in situations with a single betting option. Since the experiment was *not* specifically designed to address this question, we will not report any statistical analysis of this data and limit ourselves to a qualitative visual exploration of the results.

Figure 2.5 displays overall percentages of participants' choices of questions in each betting option condition. We can see that the data displays the expected pattern. However, while there seems to be a clear difference in question choice in the multiple condition (where *wh*-questions were chosen approximately 85% of the times), the difference in the single condition appears to be much less clear-cut (polar questions were chosen approximately 55% of the times). This observation is not particularly puzzling if we keep in mind that a *wh*-question about a set of outcomes, e.g. *Which x is P?*, semantically entails the polar questions about each specific outcome, e.g. *Is a P?*, *Is b P?*, ..., in the sense that a complete answer to the former contains enough information to completely resolve the latter too (Groenendijk & Stokhof, 1997). This means that, from a semantic perspective, asking a *wh*-question in the single condition can be just as appropriate as asking a specific polar question.

However, we do believe that polar questions are *pragmatically* more efficient in the single option: they are shorter and more to the point. Therefore, some tension might be expected. We can investigate this issue further by taking a look at the same data split by trial (Figure 2.6). We can see that participants' choice percentages in the multiple condition were approximately the same in the three trials, with a slight increase of *wh*-questions in the third trial, where they were chosen 90% of the times (compared to approximately 86% and 85% of the first and second trial respectively). More interestingly for our exploration, however, is the fact that the same is not true for the single condition. In trial 1 the pattern is the opposite of what we would expect: polar questions where chosen approximately 42% of the times and *wh*-questions were chosen approximately 58% of the times. In trial 2 the percentages get closer to each other, but with the reverse (and expected) pattern: polar questions 55% and *wh*-question 45%.

Finally, in trial 3, there seems to be a clear-cut difference in the expected direction between polar questions (68%) and *wh*-questions (32%). This pattern suggests that some learning effect might be happening here, as it becomes clearer that there are two betting options and two kinds of questions, so a stable association is eventually established. At this point this is nothing but a suggestive speculation, and we leave an experimental investigation of this issue for another occasion. In any case, we believe that by the end of the training trials participants were acquainted with the general setting of the game and that perhaps they might have developed an intuition about efficient use of questions in the role of receiver.

Production data. We can turn now to the data collected in the critical trials. The complexity of the data is higher, as each participant completed twelve trials, one for each combination of critical values $(3,4,5)$, scenarios (*dual*, *plural*) and QUDs (*polar*, *wh*), plus four additional control trials (values equal to 0 and 10). We begin with a visual exploration of the data. On the basis of the hypotheses that drove the design of the experiment, we have the following expectations concerning our three manipulated variables:

1. all other things being equal, we expect the value 0 to be associated with comparatively higher choice percentages of *certainly not*; the value 3 to be associated with higher percentages of *probably not* and 5 with higher percentages of *probably*, with 4 in between them; finally, we expect 10 to be associated with higher percentages of *certainly*;

2. all other things being equal, we expect plural scenarios to be associated with comparatively higher choice percentages of *probably* and dual scenarios to be associated with comparatively higher choice rates of *probably not*, at least for likelihood values sufficiently bigger than 0;

3. all other things being equal, we expect *wh*-questions to be associated with comparatively higher choice percentages of *probably* and polar questions to be associated with comparatively higher choice percentages of *probably not*, at least for likelihood values sufficiently bigger than 0.

First, we look at expression choices in each value condition, aggregated by scenario and QUD, as displayed in Figure 2.7. Aside from some noise in the control conditions, we can clearly observe the expected pattern. This feature is not especially surprising, and we can think of it as a basic sanity check for our experimental setting.

Second, we can visualize the potential effect of scenario by splitting the same data along the scenario condition in addition to the value condition (Figure 2.8). The displayed pattern is less clear here. On the one hand, if we compare only the quadrants for the critical values $(3,4,5)$ in the top panel (dual scenario) with the ones in the bottom panel (plural scenario) we can observe a tendency of the data to follow the expected pattern: given the same value, the plural scenario is associated with comparatively higher choice percentages of *probably* compared to the dual scenario. On the other

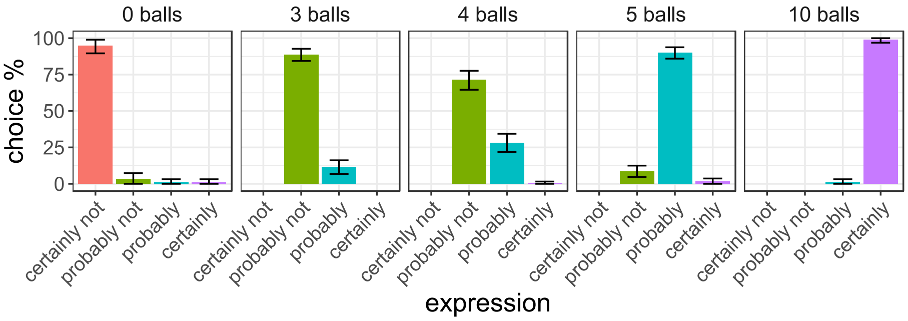


Figure 2.7: Expression choice percentage for each value. Black bars represent bootstrapped 95% confidence intervals.

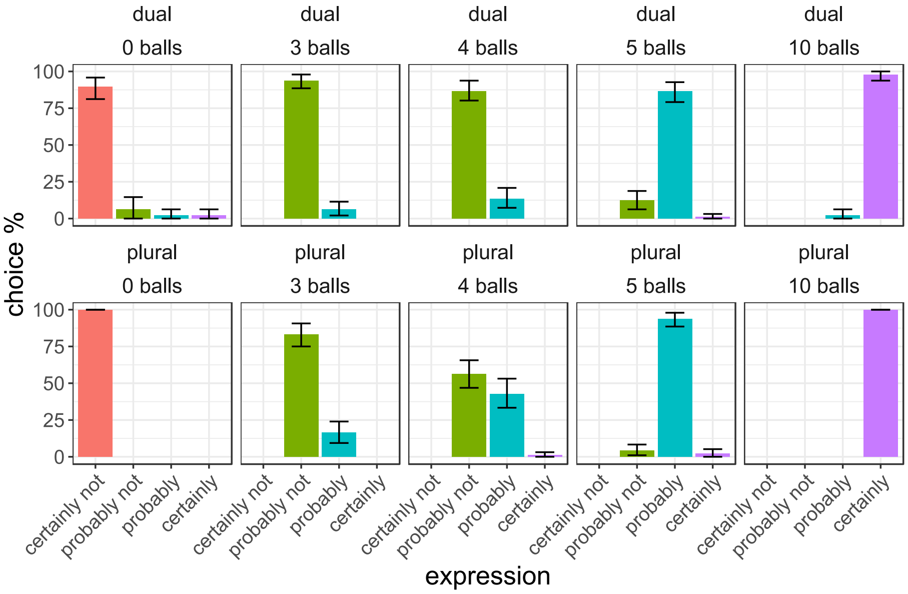


Figure 2.8: Expression choice percentage for each value, scenario. Black bars represent bootstrapped 95% confidence intervals.

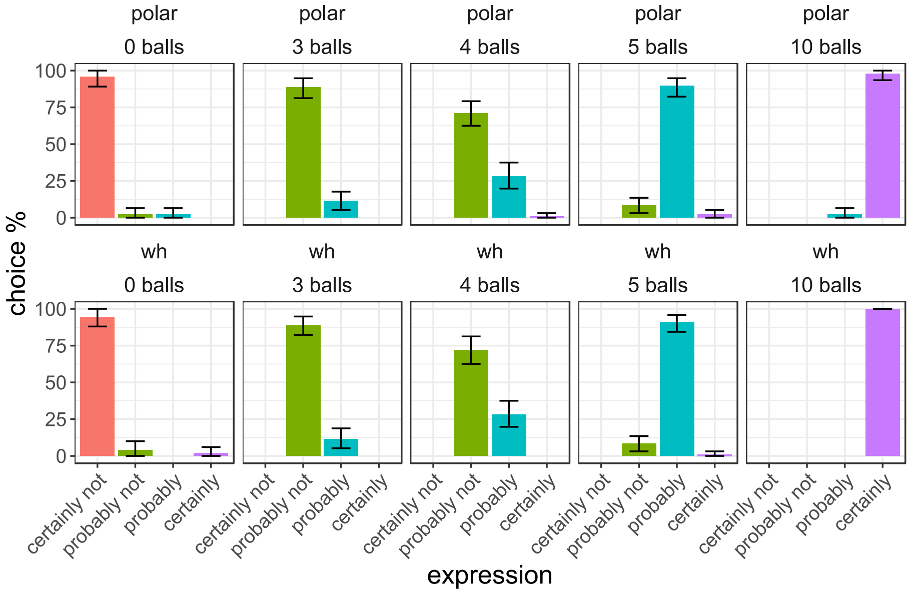


Figure 2.9: Expression choice percentage for each value, QUD. Black bars represent bootstrapped 95% confidence intervals.

hand, the data relative to the control condition with value equal to zero (leftmost quadrants) seems to exhibit a behavior opposite to what we would expect, with participants' choices of *certainly not* increasing from dual to plural scenario.

Third, we can visualize the potential effect of QUD by splitting the same data along the QUD condition in addition to the value condition (Figure 2.9). In this case visual inspection does not seem to reveal any noteworthy difference between the top panel (polar QUD) and the bottom panel (*wh* QUD).

Analysis. The plots in Figures 2.7, 2.8, 2.9 display percentages of expression choices. The raw data, however, were counts. Moreover, notice that the four possible expressions can be naturally ordered in terms of "strength":

certainly not < *probably not* < *probably* < *certainly*

Consequently, our dependent variable is the ordinal factor `expression` of count data for each of the four levels `certainly not`, `probably not`, `probably` and `certainly`, in this order. The explanatory variables, corresponding to the manipulated conditions, are the metric factor `value` $(0, 3, 4, 5, 10)$ and the categorical factors `scenario` (`dual`, `plural`) and `qud` (`polar`, `wh`).

We analyzed our data with Bayesian mixed effect ordinal regression, as implemented in the R package `brms` (Bürkner, 2017). First, we fit the full model which regresses `expression` against `value`, `scenario` and `qud`, together with all their possible interactions and the maximal random effect structure, i.e. by-participant random

	lower	mean	upper	
`value`	2.501	3.136	3.772	*
`scenario-plural`	-5.562	-2.565	0.283	
`qud-wh`	-2.408	-0.046	2.231	
`value:scenario`	0.412	1.113	1.852	*
`value:qud-wh`	-0.445	0.063	0.648	
`qud:scenario`	-3.93	-0.349	3.675	
`value:scenario:qud`	-0.980	-0.047	0.860	

Table 2.1: Mean values and HDIs for model coefficients for the main effects of `value`, `scenario` and `qud` and their interactions. Coefficients credibily different from zero are marked with *.

intercepts and slopes for all explanatory variables.[5] In the formula notation adopted in `brms`:

$$\texttt{expression} \sim \texttt{value*scenario*qud+(1+value*scenario*qud|id)} \qquad (2.1)$$

where `id` is a categorical factor containing participants' AMT user ids. In order to test our hypotheses about the influence of manipulated conditions on participants' choices, we checked whether the estimated posterior probability mass of the model coefficients for main effects of `value`, `scenario` and `qud` is credibly different from 0 (in terms of 95% highest density intervals) in the expected direction.

Table 2.1 contains mean values and Bayesian 95% highest density intervals (HDIs) for the coefficients of the full model defined in Equation 2.1.[6] Coefficients credibly different from zero have been marked. First of all, we can observe that `value` has a coefficient credibly bigger than zero. This means that, given our data, we can reasonably believe that the manipulation of the likelihood values has an effect on participants' choices of uncertainty expressions: the higher the value, the more likely it is to choose a stronger expression. This is in line with our expectations and with what we observed about Figure 2.7.

Moving to `scenario`, the picture emerging from the regression is harder to interpret. The coefficient for `scenario` is not credibly different from zero, but has a tendency to be smaller than zero. This contrasts with our expectations: if anything, we would expect `scenario-plural` to have a positive effect, i.e. to make stronger expressions more likely. Additionally, we can see that the coefficient for the interaction between `value` and `scenario` is credibly bigger than zero. A possible explanation for these results can be the apparent contrast, highlighted in our informal observations about Figure 2.8, between the data collected in the control trials and the critical trials: the fact that participants chose *certainly not* more often in the condition with

[5] More in detail, ordinal regression models were obtained calling the `brm` function and specifying `cumulative` as family parameter. For simplicity, we stuck with the default prior assumptions of the `brms` package (at the time of the analysis), i.e. improper flat priors over the reals.

[6] Bayesian 95% highest density intervals specify intervals spanning 95% of the distribution, such that every point inside the interval has higher credibility than any point outside the interval (Kruschke, 2014).

	`full`	`no value`	`no scenario`	`no qud`
LOO-IC	533.04	1930.95	573.79	514.34
SE	65.98	28.97	57.15	62.03
	`only value`	`only scenario`	`only qud`	`intercept`
LOO-IC	564.47	1923.41	1928.09	1922.10
SE	55.52	28.87	27.93	27.60

Table 2.2: LOO-IC scores with standard errors (SE). Lower is better.

value=0 and plural scenario (compared to dual) might have driven down the coefficient for `scenario`, whereas the fact that participants' choices of *probably* (vs *probably not*) were higher in the critical conditions with plural scenario (compared to dual) might be responsible for the credible effect of the interaction between `value` and `scenario`. We will return to this explanation in the next section, where we will attempt further exploratory analysis of the data.

Finally, none of the other coefficients reported in Table 2.1 are credibly different from zero. In particular, given our data, we cannot conclude that the manipulation of QUD has any effect on choices of uncertainty expressions.

To further examine these results we compare the full model (Equation 2.1) with simpler models obtained by dropping one, two and finally three of our explanatory variables. We compare the models in terms of Leave-One-Out Information Criterion (LOO-IC), as implemented in the R package `loo` (Vehtari, Gelman, & Gabry, 2016). In particular, we check whether dropping any of the explanatory variables results in higher (i.e. worse) or lower (i.e. better) LOO-IC score. In the former case, this would be evidence that the dropped variable is relevant for the posterior predictive success of the model and should not be dropped. In the latter case, this would be evidence that the variable can be dropped without hurting the predictive success of the model.

The LOO-IC scores resulting from the model comparison are reported in Table 2.2. We can observe that the model dropping `qud` has the lowest LOO-IC scores. However, taking into account the standard errors of the scores, we can observe that the interval corresponding to the `no qud` model (514.34 ± 62.03) overlaps with the interval corresponding to the `full` model (533.04 ± 65.98), which means that we cannot conclude that the former is credibly a better model than the latter.

Inferred values of the coefficients of the `no qud` model are summarized in Table 2.3. We can observe that they are in line with the results of the full regression model, which confirms our previous analysis: both `value` and `scenario` are relevant factors in explaining the rates of participants' choices of expressions, with `value` having a credible effect in accordance with our expectations and `scenario` behaving in a somewhat more complex way. Contrary to our expectations, the same is not true for `qud`, which appears to be irrelevant.

Further exploratory analysis. As we mentioned above when we inspected Figure 2.8 and commented on the coefficients for the full model, there seems to be a tension between the patterns found in the data in the control conditions (in particular with value=10) and the critical conditions, especially when comparing dual and plural

	lower	mean	upper	
value	2.600	3.162	3.751	*
scenario-plural	-4.600	-2.396	-0.002	*
value:scenario	0.431	1.020	1.590	*

Table 2.3: Mean values and HDIs for model coefficients for the main effects of value and scenario and their interactions. Coefficients credibily different from zero are marked with *.

	lower	mean	upper	
value	4.643	6.203	7.987	*
scenario-plural	-2.223	3.341	8.911	
qud-wh	-2.529	2.801	8.691	
value:scenario	-1.311	-0.026	1.280	
value:qud-wh	-1.833	-0.606	0.747	
qud:scenario	-10.312	-3.598	2.987	
value:scenario:qud	-0.820	0.679	2.316	

Table 2.4: Mean values and HDIs for model coefficients for the main effects of value, scenario and qud and their interactions, excluding wrong responses in control conditions. Coefficients credibily different from zero are marked with *.

scenarios. The data seems to show the expected pattern in the critical conditions (i.e. plural scenario has positive effect on expression choice) but the opposite pattern in the control condition. Here we attempt to further investigate this issue by fitting the same regression models on smaller subsets of the data obtained applying increasingly stricter exclusion criteria, specifically aimed at minimizing (when not annulling altogether) the importance of the noise in the control conditions.[7]

We consider three possible approaches. 1) The less radical approach is to simply exclude from the analysis the “wrong” responses in the control conditions (i.e. expression choices different from *certainly not* with value=0 and choices different from *certainly* with value =10). 2) The intermediate approach is to adopt a stricter exclusion criterion and discard every data point obtained from participants who gave at least one wrong response in a control trial. 3) The more radical approach is to discard all the data points obtained in the control trials.

Fitting the full regression model to the data sets obtained following approaches 1 and 2 gives similar results. For this reason, we report here only the results obtained with the former. Model coefficients are summarized in Table 2.4. We can see that value has the only coefficient credibly different from zero, which is not surprising. Perhaps more interestingly, the coefficient corresponding to the interaction between value and scenario is not credibly different from zero anymore, whereas it was with the full data set (see Table 2.1). This fact supports our hypothesis that control data

[7] The analyses reported in this paragraph go beyond the pre-registered project.

	lower	mean	upper	
`value`	6.562	9.735	13.294	*
`scenario-plural`	7.508	18.341	30.709	*
`qud-wh`	0.687	11.327	22.910	*
`value:scenario`	5.922	-3.306	-0.872	
`value:qud-wh`	-5.109	-2.475	-0.064	
`qud:scenario`	-26.137	-12.766	-0.777	*
`value:scenario:qud`	-0.197	2.712	5.583	

Table 2.5: Mean values and HDIs for model coefficients for the main effects of `value`, `scenario` and `qud` and their interactions, excluding all data from control trials. Coefficients credibily different from zero are marked with *.

was "driving up" the coefficient for the interaction. Moreover, both the coefficients for `scenario` and `qud` display a tendency towards values bigger than zero (although smaller values cannot be rationally excluded given our data). All of this seems to point to a data set which exhibits a behavior slightly closer to our starting expectations.

Let us conclude our analysis by briefly discussing approach 3. Fitting the full regression model on the data restricted to the critical trials only yields the coefficients summarized in Table 2.5. We can see that `value`, `scenario` and `qud` all have coefficients credibly bigger than zero. The latter, although in line with our expectations, is particularly puzzling: recall our visual inspection of Figure 2.9, where we could not observe any difference between QUD conditions. A possible explanation could be put forward observing that the model infers mass credibly smaller than zero for the interaction between `scenario` and `qud`, which could indicate that the effects of different QUDs, if present, cancel out depending on the kind of scenario. Looking at the full plot of our production data split by value, scenario and QUD (Figure 2.10) we can see that this might be the case. Moreover, fitting a model on the same data which keeps the three explanatory variables but drops their interactions results in a coefficient for `qud` which is not credibly different from zero anymore.

2.4 Conclusion

The main goal of this chapter was to investigate the question of how the availability of alternative events in the state space and/or the way in which the alternatives are conceptualized in the conversational context might affect the production of uncertainty expressions. As we have seen, the Alternative Outcome Effect already observed by Teigen (1988) and Windschitl and Wells (1998) was relatively easy to replicate, whereas the QUD Effect hypothesized on the basis of (Lassiter, 2011a) proved to be more elusive.

About this last point, it has to be noted that the results of the exploratory analysis above leave the door open to the possibility that the QUD effect could be better investigated in a slightly different experimental setting, in which the interaction (and perhaps confusion for the participants?) between scenario and QUD is not present. As already mentioned, this is nothing but a suggestive starting point for future research, in that it goes beyond the scope of the pre-registered analysis. Let us remark on the fact that this is part of the importance of pre-registrations: since the very beginning of this research it has been explicit and clear (to us and, potentially, the scientific community) which its two research questions were and what would constitute legitimate ways to answer them.

In any case, we believe that the value of this chapter goes beyond the report of the (half and half) success or failure of the empirical investigation regarding the specific linguistic issue at hand. First of all, because the experimental setting with urns and colored balls introduced in this chapter will come back several times in this dissertation, and the fact that we could replicate (at least) the Alternative Outcome Effect speaks in favor of our choice of manipulating uncertainty and context in this way.

Finally, and more in general, the issue discussed in this chapter is but one of the several issues at the interface between semantics and pragmatics which have been raised by philosophers and logicians, perhaps noticed and sometimes investigated by experimental psychologists and linguists and which constitute the natural field of research for a study of uncertainty expressions integrating formal, experimental and modeling approaches. With this chapter, and with this dissertation, we would like to advocate such integration of approaches originating from different research traditions, and we hope to contribute to its development.

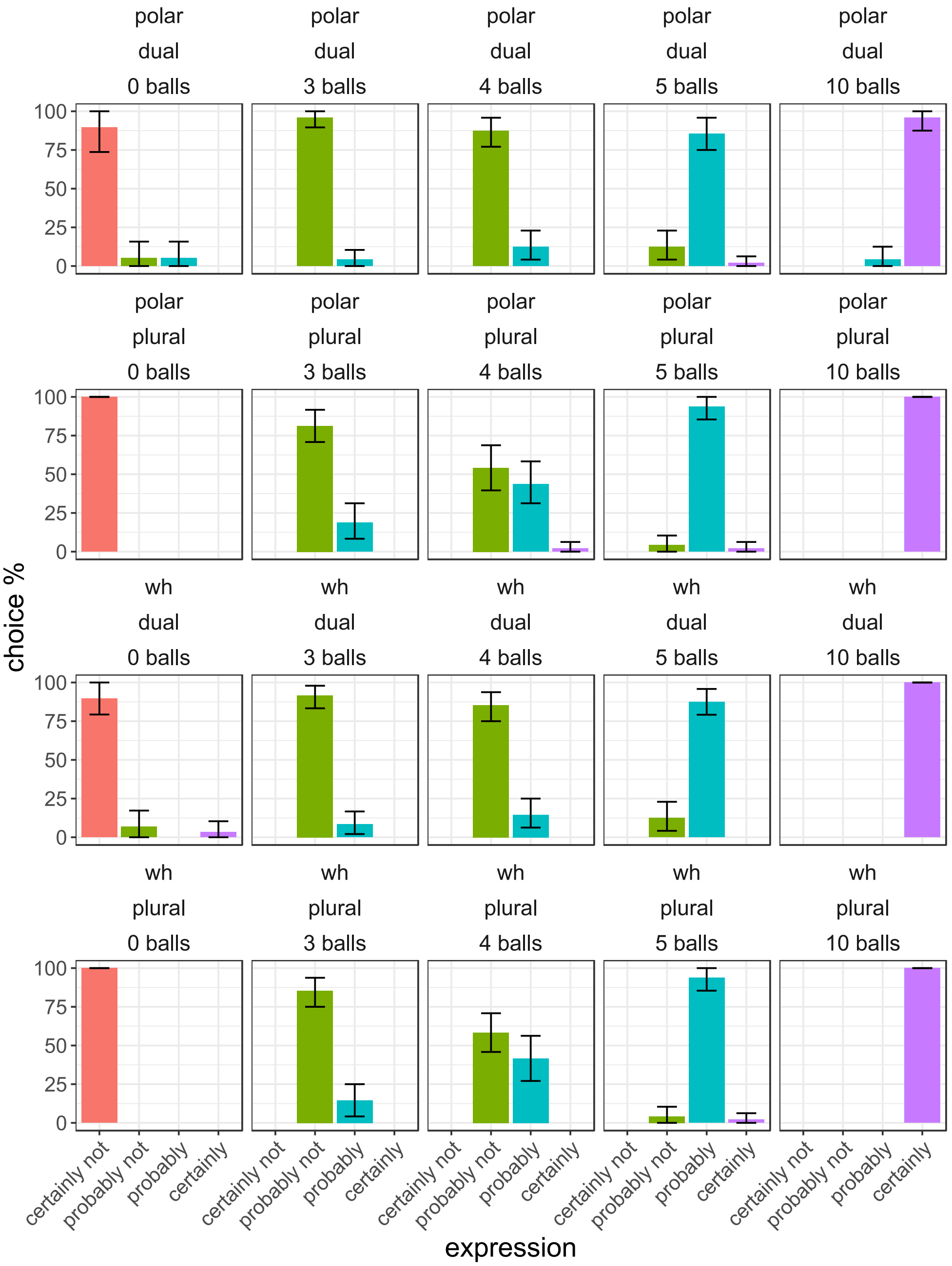


Figure 2.10: Expression choice percentage for each value, scenario and qud. Black bars represent bootstrapped 95% confidence intervals.

Chapter 3

Uncertainty adverbs & adjectives

Probably the best beer in the World.
(Carlsberg, about themselves)

3.1 Introduction

In Chapter 2 we investigated possible sources of context-sensitivity of uncertainty expressions, and in particular how the different ways in which the state space of alternative events is presented in the context or conceptualized in the conversation might affect pragmatic use and interpretation of uncertainty expressions.

In this chapter we look more specifically at the QUD sensitivity of uncertainty expressions and in particular at how this feature can shed light on a specific issue concerning uncertainty expressions in a "post-Kratzerian" world, namely the account of the differences between adverbial forms of uncertainty expressions (*possibly*, *probably*, *certainly*,. . .) and adjectival forms (*possible*, *probable*, *certain*,. . .).[1]

The sheer fact that *there are* such differences can be noteworthy by itself. In fact, as anticipated in Chapter 1, the tendency that can be observed in the literature, especially in the logico-philosophical tradition originating from Kratzer's account, is to focus on the logical properties of epistemic modals such as *possibly* and *possible* and probability expressions such as *probably* and *probable* abstracting away from their different grammatical categories. And, at first glance, it seems intuitive that the two expressions in each of the pairs just mentioned have the same logical properties. However, in Chapter 1 we have also seen that a number of authors have focused on investigating uncertainty expressions especially from a point of view which highlights the differences

[1] A preliminary version of the work reported in this chapter was presented in a talk with the title *On the difference between modal adverbs and adjectives*, given by the author at the workshop New Ideas in Semantics and Modeling (Paris, September 8, 2016).

between uncertainty adverbs and adjectives. In this chapter we adopt a similar point of view.

The main source of inspiration for this chapter is a body of relatively recent works by Lavi Wolf and colleagues. In particular, we refer mainly to Wolf's dissertation (Wolf, 2014) and a number of surrounding publications (Wolf & Cohen, 2009; Wolf, 2015; Wolf et al., 2015). We believe that the work by Wolf and colleagues has two important merits. First of all, the authors collect and systematize a large number of empirical observations about the syntactic / distributional and semantic / pragmatic / conversational differences between uncertainty adverbs and uncertainty adjectives found in the literature (e.g. Bellert, 1977; Nuyts, 2001a, 2001b; Nilsen, 2004; Piñón, 2006; Ernst, 2009). Second, the authors put forward an original and thought-provoking proposal to account for said differences, which we will refer to as the Speech Act Modifier (SAM) theory of uncertainty adverbs.

3.2 Background: the Speech Act Modifier theory

In this section we briefly summarize the main idea behind SAM theory, and review the body of data collected in support of the theory.

The theory. In a nutshell, the main claim of the SAM theory is that virtually all the differences between uncertainty adverbs and adjectives can be analyzed (hence, explained) in terms of one fundamental difference: while uncertainty adjectives (*It's possible that...*) can be analyzed as compositional sentential operators in a somewhat traditional way, uncertainty adverbs (*Possibly,...*) are better analyzed as expressive operators (Potts, 2007) modifying assertive force at speech-act level. Wolf develops a semi-formalized speech-act framework in which assertions are associated with both a propositional content *C* and an assertive strength *S*. Roughly speaking, the propositional content is the compositionally-derived semantic meaning of the assertion, and the assertive strenght can be thought of as the subjective credence associated by the speaker to the content and modeled as a value assigned to propositions by a probability measure *P*. Normally, a sincere non-modal (*bare*) assertion can be seen as assigning a default "high enough" strength to a propositional content, e.g.:

(1) a. The dog is on the lawn.
b. P("dog on the lawn") $\geq$ high

where *high* is a conventionally or contextually determined threshold. According to this analysis, a sincere assertion of (1-a) assigns a high strength to the propositional content that the dog is on the lawn. This content is communicated with relatively high subjective credence. Now, according to the first half of Wolf's theory, uncertainty adjectives give rise to assertions which behave in the default way, i.e. they modify the propositional content but not the assertive force, e.g.:

(2) a. It's possible that the dog is on the lawn.
b. $P[P$("dog on the lawn") $> 0] \geq$ high

In this example the threshold of the assertive strength is the default *high* value of sincere assertions, whereas the propositional content has been compositionally modified by the modal in a standard (probabilistic) way. According to this analysis a sincere assertion of (2-a) assigns a high strength to the content that there is a chance that the dog is on the lawn. A similar analysis is put forward for the adjective *probable*:

(3) a. It's probable that the dog is on the lawn.
b. $P[P(\text{“dog on the lawn”}) > 0.5] \geq \text{high}$

Now, the following example should immediately clarify what the difference is between adjectives and adverbs according to SAM theory:

(4) a. The dog is possibly on the lawn.
b. $P(\text{“dog on the lawn”}) > 0$

According to the second half of the theory, the adverb *possibly* has no compositional effect on the propositional content, which is the same as in the non-modal claim in (1-a) (i.e. that the dog is on the lawn). Instead, the effect of *possibly* is on the assertive strength assigned to the propositional content. A sincere assertion of (2-a) has a much lower threshold of assertive strength than the default “high enough” value, namely 0. In other words, the asserted propositional content is the same as in the non-modal claim, but the subjective credence associated with it is much lower than the default. A similar analysis is given for the adverb *probably*:

(5) a. The dog is probably on the lawn.
b. $P(\text{“dog on the lawn”}) > 0.5$

Having outlined the core concepts of SAM theory, we can move to the body of linguistic data collected by Wolf as evidence in support of the theory. We begin with a set of observations grouped under the category of *embedding patterns*, highlighting differences between the syntactic distributions of uncertainty adjectives and uncertainty adverbs.

Embedding patterns. There seems to be a reasonably strong consensus in the literature that uncertainty adjectives are much easier to embed than uncertainty adverbs in a number of syntactic contexts. Bellert (1977) observes that uncertainty adverbs are harder to embed under negation and questions:

(6) a. It's not probable that John will come to the party.
b. *Not probably, John will come to the party.

(7) a. Is it probable the John will come to the party?
b. *Will John probably come to the party?

Notice that variants of (6-b) were the adverb scopes above negation is perfectly acceptable:

(8) a. John will probably not come to the party.
b. Probably, John will not come to the party.

Similarly, Piñón (2006) observes that uncertainty adverbs are harder to embed in the antecedent of a conditional:

(9) a. If it's probable that the socialists will win the elections, the rich will worry about a luxury tax.
b. *If the socialists will probably win the elections, the rich will worry about a luxury tax.

Finally, Papafragou (2006) observes that modal auxiliaries are hard to embed under factive verbs such as *be surprising*. Wolf claims that the same applies to uncertainty adverbs, but not adjectives:

(10) a. It's surprising that it's probable that the socialists will win.
b. ?It's surprising that the socialist will probably win.

Examples like these seem to make a relatively strong case in favor of SAM. If uncertainty adverbs are speech-act modifiers, then by definition they are applied at the level of the full speech-act, taking the widest possible scope, and hence do not partake in the compositional computation of the content of the sentence: they cannot be easily negated or otherwise embedded in more complex constructions.

However, informal online searches suggest that the observed patterns might not be as stable as the literature has taken them to be. Here are a few examples, where uncertainty adverbs appear in the scope of questions:

(11) Is it possibly the case that Sheldrake is even more skeptical than the skeptics? (Sheldrake, 2004)

(12) Gramps has a function that can compute whether someone is probably alive based on their event dates. (`gramps-project.org`, 2013)

Or in the scope of factive verbs:

(13) It is pretty incredible that your paternal line probably goes back to that Scottish hero who lived almost 1,000 years ago. (Forum post on *The Apricity* available at `goo.gl/LFzUc8`, 2014)

Or in the antecedent of a conditional:

(14) If something probably will happen, it's likely. If it probably won't happen, don't get your hopes up. (`vocabulary.com`, 2018)

While we are certainly aware that these examples are to be taken with a grain of salt, we believe that they justify the adoption of a skeptic point of view on the stability of the embedding patterns discussed above. By this we mean that even if these examples are certainly not enough to conclusively dismiss the SAM theory of uncertainty adverbs, they partially weaken its case, as the strict syntactic distinction between uncertainty adverbs and adjectives posited by the SAM theory does not appear to be mirrored in an equally strict and stable difference in their distribution.

The goal of this chapter is to show that if a syntactic distinction between uncertainty adverbs and adjectives is not warranted by observations about the syntactic distribution or the embedding patterns, then there is no reason to assume a syntactic distinction at all. As mentioned in the introduction, our focus is on the semantic/pragmatic/conversational differences between uncertainty adverbs and adjectives rather than on the embedding patterns discussed in this section. Specifically, we set out to show that the observable semantic/pragmatic/conversational differences between uncertainty adverbs and adjectives can be explained on the basis of a more economic (and independently motivated) hypothesis.

Having clarified all this, let us turn to the other observable differences between uncertainty adverbs and adjectives, as collected by Wolf and colleagues.

Interaction with tense. Wolf observes that uncertainty adverbs tend to take wide scope over tense operators, whereas uncertainty adjectives can more easily scope below tense:

(15) a. In 1979, it was probable that the socialists won the elections.
b. Given people's uncertain beliefs *in 1979*, they judged the likelihood of a socialist victory (in 1979) to be high.

(16) a. In 1979, the socialists probably won the elections.
b. Given our uncertain beliefs *now*, we judge the likelihood of a socialist victory (in 1979) to be high.

The intuition is that in (15-a) the body of beliefs relevant to assess the probability of the claim is shifted in the past, in such a way that the sentence can be paraphrased as (15-b). On the other hand, in (16-a) the relevant body of beliefs seems to be fixed to the present and the sentence is better paraphrased as (16-b). The SAM account of this pattern simply states that uncertainty adverbs are as hard to embed under tense as they are hard to embed under negation etc., and that they tend to take wide scope over tense.

Before moving to the next group of observations, let us notice that if we embed both (15-a) and (16-b) under a tensed attitude verb (e.g. *The socialists believed that...*) the difference between the two readings disappears:

(17) a. The socialists believed that it was probable that they won the elections.
b. The socialists believed that they probably won the elections.
c. Given what the socialists believed *back then*, the likelihood of victory was high.

Here, both (17-a) and (17-b) can be appropriately paraphrased as (17-c), regardless of which category of modals they contain (adjective or adverb). We believe that this phenomenon needs explaining, and it is not obvious to us how such an explanation would look like within the framework of the SAM theory.

Speaker inference. First, Nuyts (2001a, 2001b) observes that sentences containing uncertainty adverbs tend to be perceived as more subjective than sentences containing corresponding uncertainty adjectives. In other words, uncertainty adverbs are *speaker oriented* (Jackendoff, 1972) in that they convey (semantically, pragmatically or otherwise) that the speaker is the agent whose opinion is reported by sentences containing said adverbs. We refer to this phenomenon as *speaker inference*. As an illustration, consider how Bob's possible follow-up questions in (19) sound less felicitous (or at least more redundant) than the ones in in (18):

(18) a. *Alice:* It's probable that the dog is on the lawn.
b. *Bob:* Who says so? / Whose opinion is this?

(19) a. *Alice:* The dog is probably on the lawn.
b. *Bob:* #Who says so? / #Whose opinion is this?

The SAM theory comes with a built-in explanation of this observation. The assertive force modified by the uncertainty adverbs by itself refers to the speaker (we can think of it as the subjective speaker's confidence in the communicated content), hence uttering a sentence such as (19-a), where the assertive force is explicitly modified by *probably*, will most typically convey that the speaker's subjective belief about the subject matter is what is being communicated. Hence the redundancy of the follow-up questions in (19-b). On the other hand, if a sentence such as (18-a) is to be analyzed according to a more classic quantificational approach, then it could be paraphrased approximately as *The likelihood of the dog being on the lawn is bigger than a certain threshold, given a certain relevant body of beliefs*. We can see how easily the questions arise: *which/whose* relevant body of beliefs? Hence the availability of the follow-up questions in (18-b).

We believe that there are at least two interesting observations related to this phenomenon. First, we notice that bare, non-modal claims seem to give rise to a speaker inference as well, at least in certain contexts:

(20) a. *Alice:* The dog is on the lawn.
b. *Bob:* #Who says so? / #Whose opinion is this?

It seems to us that Bob's follow-up questions in (20-b) are just as redundant after (20-a) as they are after (19-a): in a such a minimal context, Bob will typically infer that Alice is reporting her own opinion. Now, this observation is not a direct objection to the SAM theory. But it can be seen as a hint that something different from SAM's explanation might be happening with the speaker inference: if the inference is not unique of modal claims, as shown here, then perhaps an independent explanation is viable, which accounts for the inference arising from both modal and bare claims.

Furthermore, it is easy to see that if we embed sentences such as (18-a) and (19-a) once again under *believe*, then the observed pattern disappears as the relevant body of beliefs is explicitly set to the attitude holder's:

(21) a. Alice believes that it's probable that the dog is on the lawn.
b. Alice believes that the dog is probably on the lawn.
c. According to Alice's beliefs, there are good chances that the dog is on the lawn.

Both (21-a) and (21-b) can be adequately paraphrased along the lines of (21-c).

Nielsen's contrast. Second, Nilsen (2004) notices that uncertainty adverbs and adjectives behave differently in conjunctive constructions such as the following:

(22) It's possible that Le Pen will win even though she certainly won't.

(23) #Le Pen will possibly win even though she certainly won't.

The intuition here is that while the sentence in (22), containing the adjective *possible* in its first conjunct, is perfectly acceptable, the sentence in (23), with the adverb *possibly* in the first conjunction, sounds less felicitous, if not contradictory. SAM theory's explanation of this fact goes as follows. Both (22) and (23) can be analyzed as containing two assertions each. In (22), the first conjunct asserts (with default assertive force) the content that there is a chance for Le Pen to win; the second conjunct asserts (with very strong assertive force) the content that Le Pen will not win. This conjunction is not contradictory, as two different contents are asserted: that Le Pen will not win, and that her victory is not impossible. Things are different in (23) because the two conjuncts here assert contradictory contents. Remember that according to SAM theory uncertainty adverbs are not part of the asserted content: the first conjunct asserts (with weak assertive force) that Le Pen will win, but the second conjunct asserts (with very strong assertive force) that Le Pen will not win. Hence, the whole conjunction sounds contradictory.

Agreement patterns. Finally, Wolf observes that uncertainty adverbs and adjectives have different conversational effects, in terms of the body of information that is made available in the conversation by utterances of sentences containing either kind of expression. Following Papafragou (2006), one way to investigate these conversational effects is by using agreement and disagreement devices, as in the following examples:

(24) a. *Alice:* It's probable that John is at home.
b. *Bob:* I agree.
c. Bob agrees that it's probable that John is at home.

(25) a. *Alice:* John is probably at home.
b. *Bob:* I agree.
c. Bob agrees that John is at home.

It has to be noted that a substantial amount of research about modality in the logico-philosophical tradition of linguistics has been devoted to investigating all sorts of issues concerning the behavior and interactions of modals and agreement/disagreement devices (e.g. Stephenson, 2007; von Fintel & Gillies, 2008, 2009; MacFarlane, 2009, 2010; Yanovich, 2014). In this respect, Papafragou's and Wolf's assumption seems to be that agreement/disagreement devices such as *I agree* usually target the full propositional content expressed by the sentence which they are used to react to and respectively affirm or deny such content. Hence, we can "reason backwards" and use the conversational effects of these reactions to inspect the content expressed by the sentences they are used to react to. For example: (24-c) is intuitively an appropriate way to paraphrase Bob's agreeing reaction to (24-a), indicating that the content expressed by Alice's utterance is the full modal proposition that it is probable that John is at home.

Crucially, things are different in (25). Here the best paraphrase of Bob's utterance *I agree* seems to be the one in (25-c): intuitively, Bob is agreeing with the statement that John is at home and the uncertainty expressed by *probably* is left out. Given Wolf's assumption, all this can only mean that *probably* does not contribute to the propositional content of (25-a), whereas *probable* does contribute to the propositional content of (24-a). And this is perfectly in line with SAM's main claim that uncertainty adverbs, but not uncertainty adjectives, are assertive force modifiers and not compositional operators.

3.3 Proposal: context-dependent semantics

In this section we build towards our proposal for a unified context-dependent semantics for uncertainty adverbs and adjectives and we evaluate it against the data presented in Section 3.2. More in detail, Section 3.3.1 expands the idea that uncertainty adverbs and adjectives generally have different at-issue contents (see below), in that they are typically used to answer different QUDs. We claim that this is the core difference between the two categories of uncertainty expressions, on the basis of which we can explain the semantic, pragmatic and conversational phenomena presented above. A detailed explanation of how our approach can deal with these phenomena is given in Section 3.3.2.

3.3.1 Agreement, at-issueness and QUD-sensitivity

Agreement patterns and at-issue content. Let us go back to the agreement and disagreement response patterns exemplified in (24) and (25). Roberts (2015, 2017) argues that agreement/disagreement responses to sentences containing epistemic auxiliaries (e.g., *might*) should be analyzed under the assumption that devices such as *I agree* do not typically target the whole content of the utterance but rather they typically target only the part of the content which is at-issue in the context of the utterance (Roberts, Simons, Beaver, & Tonhauser, 2009; Simons, Tonhauser, Beaver, & Roberts, 2010), or in other words the part of the content that is intended to address the QUD in the context of the utterance (Roberts, 1996, 2012). If this is true, then our intuitions about the meaning of the agreement/disagreement sentences exemplified in (24) and (25), repeated below as (26) and (27), can be leveraged to infer not (or not only) the content of the full sentences that these devices are used to respond to, but the part of the content at-issue in the context of utterance:

(26) a. *Alice:* It's probable that John is at home.
b. *Bob:* I agree.
c. Bob agrees that it's probable that John is at home.

The sentence in (26-c) seems to be a good paraphrase of Bob's reaction *I agree* in this context. What exactly is Bob agreeing to in (26-b)? Intuitively, he is agreeing that it is probable the John is at home. Then, given our assumption about agreement devices, we can conclude that the content at-issue in the context of this exchange is the likelihood, or probability, that John is at home, or at least that the likelihood is part of the at-issue content. By the same token, the sentence in (27-c) below seems to be a good paraphrase of Bob's reaction *I agree* to Alice's initiative in (27-a):

(27) a. *Alice:* John is probably at home.
b. *Bob:* I agree.
c. Bob agrees that John is at home.

What is Bob agreeing to in (27-b)? Intuitively, he is agreeing that John is at home. Hence, what is at-issue in this context are John's whereabouts.

From at-issue content to most likely QUD. We believe that these observations about the different typical at-issue contents of uncertainty adverbs and adjectives allow us to work our way to the QUDs that the two kinds of modal constructions are most typically used to address. It is a form of backward reasoning: given the fact that a certain construction is typically associated with a certain at-issue content (e.g., the likelihood of an event) we infer that that construction is typically used to address a QUD which is best answered by that content (e.g., *How likely is it that...?*).

In general, our hypothesis is that uncertainty adverbs are typically used to address QUDs which are directly about the prejacent (e.g. polar or *wh* question about the outcome of an uncertain event) whereas uncertainty adjectives are typically used to address QUDs about the likelihood of the prejacent (e.g. question about the chances of an uncertain event to obtain). For example, it seems to us that Alice's question in (28-a) (about the likelihood of John winning) exemplifies the most typical kind of question which can be assumed to be under discussion in a context where Bob has asserted a claim about the likelihood of John winning such as the one in (28-b):

(28) a. *Alice:* How likely is it that John will win the lottery?
b. *Bob:* It's probable (that John will win).

On the other hand, Bob's claim in (29-b) (which is directly about John) is most typically taken to answer a QUD such as the one in (29-a), which is a *wh* question directly about the outcome of the lottery:

(29) a. *Alice:* Who will win the lottery?
b. *Bob:* Probably John. / John will probably win.

These observations are generalized in Table 3.1, which displays the intuitively most typical associations between QUD types, at-issue contents and produced answers.

QUD	*at-issue content*	*example answers*
how likely is *p*?	likelihood of *p*	*it's possible that p*, *p is likely*, ...
which *x* is *P*?	denotation of *P*	*a is P*, *b is probably P*, ...

Table 3.1: Association between QUD types and typical at-issue contents and produced answers.

Context-dependent semantics. The remainder of this chapter explores the hypothesis that the discussed semantic, pragmatic and conversational differences between uncertainty adjectives and uncertainty adverbs can be accounted for on the basis of the typical associations between either category of modals with different conversational contexts (QUD/at-issue content) as summarized in Table 3.1.

As we mentioned in the introductory section, our starting point is a probabilistic threshold semantics for possible/possibly and probable/probably. In its most simple implementation, such a semantics for these expressions could be expressed as follows:

possible/possibly(p) is true in s *iff* $P_s(p) > 0$
probable/probably(p) is true in s *iff* $P_s(p) > 0.5$

Crucially, adverbs and adjectives receive a unified semantic treatment. The entry for *possible/possibly* states that sentences of the form *possible(p)* or *possibly(p)* are evaluated as true in a state s if and only if the probability of the prejacent $P(p)$ in that state is bigger than 0. Similarly, the entry for probable/probably states that sentences of the form *probable(p)* or *probably(p)* are evaluated as true in a state s if and only if the probability of the prejacent $P(p)$ in that state is bigger than 0.5.

There does not seem to be much room for context dependency in these semantic entries: 1) the definitions refer to the state of evaluation s but do not contain any reference to the context of utterance; and 2) the definitions refer to the probability measure P, implicitly assuming it as given. Instead, if we want to capitalize on our observations about typical associations between QUDs, at-issue contents and types of expressions we need to make sure that the definitions make some reference to the context of utterance; in particular, and this anticipates the core idea at the center of our proposal, the reference to the context should be the feature of the semantics which determines the probability distribution P.

This (crucial) component of our proposal is heavily inspired by the contextualist theory of epistemic modals proposed by Roberts (2015, 2017).[2] In a nutshell, the main idea is that an epistemic modal auxiliary such as *might* is ambiguous and that the crucial role in the disambiguation is played by the context of utterance and in particular by the body of beliefs or evidence which is relevant in the context. More in detail, the modal is always interpreted relative to a so-called *center*, encompassing an agent (or group of agents) at a certain time with a certain epistemic state, which functions as an indexical anchor for the modal. Different conversational contexts (and in particular different QUDs, hence different at-issue contents) make it so that different centers are salient and that consequently different readings of *might* can arise. Simplifying Roberts' notation, we can capture the proposal with the following semi-formal semantic entry:

Utterance
might$_{i,j}(p)$ in context D, world and time of evaluation w, t_j
Presupposed content
1. $\exists\ c_{i,j} \in D$, i.e. an agent or group of agents d_i at time t_j;
2. $S_{i,j} \neq \emptyset$ is the smallest set of propositions supposed by d_i at t_j in w.

[2] For references on "standard" contextualist theories on epistemic modality see for example (Weatherson & Egan, 2009).

Proffered content
$p \cap (\bigcap S_{i,j}) \neq \emptyset$

Some words of clarification are in order here. First, this semantic entry is clearly contextualist, as it defines the meaning of *utterances* of a modal claim relative to a context D. For our purpose, it is enough to assume that D contains salient centers, i.e. agents or groups of agents at certain times. Second, the existence of the center $c_{i,j}$ determining the modal base $S_{i,j}$ is *presupposed* much like the existence of reference for indexical terms such as *I* and *now* is presupposed: the presupposed content constrains the contexts in which an utterance of the given expression is felicitous. If it is felicitous, then its propositional content, which enters further compositional processes, is computed according to the rule specified by the expression's *proffered* content. In this case, that the prejacent p is consistent with the modal base.

To see how this entry works, and in particular how it interacts with the QUD in the conversation, we can use the classic example of Pascal's and Mordecai's game of Mastermind discussed by von Fintel and Gillies (2008). The game involves two players: a codemaker, in this case Mordecai, has perfect knowledge of an array of colored pegs and a codebreaker, Pascal, tries to guess the sequence of pegs in a limited period of time by asking a constrained set of questions. After a few rounds of the game, the codemaker Mordecai gives the following hint to Pascal:

(30) *Mordecai:* There might be two reds.

What does Mordecai's utterance in (30) mean exactly? It could mean that given what Mordecai himself knows (i.e. the complete sequence of pegs), two reds cannot be ruled out. Or it could mean that given the body of common knowledge shared by Mordecai and Pascal, two reds cannot be ruled out. Or it could also mean that given what *Pascal* believes (i.e. he has inferred with his questioning), Pascal cannot rule out two reds. Or, perhaps too extremely?, it could mean that given the rules of the game two reds cannot be excluded *a priori*.

As we said, a crucial role in solving this ambiguity is to look at the QUD in the conversation. Different QUDs makes different centers salient in the context of utterance. For example, if Mordecai wants to help Pascal without giving too much away then Mordecai might want to summarize the exchange of information happened so far. In this case, Mordecai takes his utterance to be answering a question about the common knowledge shared between him and Pascal, hence Mordecai takes his utterance of *might* to be anchored to the center consisting of him and Pascal at the time of utterance, and the resulting modal base is the intersection of the information exchanged so far. More subtly, Mordecai might want to bring to Pascal's attention that, given his questioning so far, Pascal cannot exclude two reds: in this case Mordecai takes his utterance of *might* to answer a QUD about Pascal's epistemic state, hence anchoring the modal to the center consisting of Pascal alone. Finally, an especially reticent Mordecai might not want to give absolutely anything away, and takes his utterance to simply be about the rules of the game: the modal is anchored to a center whose specific constitution is not important, as long as the intersection between the agents' epistemic states contains exactly the rules of the game.

At its core, our proposal is to adapt Roberts' semantics for *might* to the pairs of adverbs/adjectives *possible/y* and *probable/y* and to derive their differences on the basis of the interactions between the semantics and the typical association between expression types and QUD/at-issue contents. The semantic entry can be stated as follows, where *EXP* stands for any uncertainty expressions among *possible*, *possibly*, *probable*, *probably*:

> **Utterance**
> $EXP_{i,j}(p)$ in context D, world and time of evaluation w, t_j
> **Presupposed content**
> 1. $\exists\, c_{i,j} \in D$, i.e. an agent or group of agents d_i at time t_j;
> 2. $P_{i,j}$ is a probability measure assigning probability mass to propositions given d_i's credence at t_j in w .
> **Proffered content**
> $P_{i,j}(p) > \theta_{EXP}$

The main difference with the entry for *might* above is that here we adopt a quantitative threshold-based probabilistic semantics for uncertainty expressions rather than a qualitative possible-world semantics.

3.3.2 Back to the data

We can now look back at the data discussed in Section 3.2. Can we explain the data on the basis of the theory developed in the previous sub-section and without assuming a syntactic distinction between adverbs and adjectives?

Speaker inference. The first phenomenon that needs explaining is the speaker inference associated with uncertainty adverbs. The relevant examples are repeated here as (31) and (32), where the redundancy (or lack thereof) of a follow-up question such as *Who says so?* shows that uncertainty adverbs (but not uncertainty adjectives) are speaker oriented: they imply that *the speaker says so*, i.e. she is the agent whose opinion is reported in the modal claim.

(31) a. *Alice:* It's probable that the dog is on the lawn.
b. *Bob:* Who says so? / Whose opinion is this?

(32) a. *Alice:* The dog is probably on the lawn.
b. *Bob:* #Who says so? / #Whose opinion is this?

Let us show how this difference can be explained on the basis of our proposal. Our strategy is to look at the intuitively most likely conversational context (QUD/at-issue content) of the exchanges in the two examples. Let us begin with the utterances containing uncertainty adjectives, such as (31-a). We have observed above that they are typically used to answer a question about the likelihood of the prejacent, *How likely is it that the dog is on the lawn?*. This question appears to be about the objective chance of the dog being on the lawn. In other words, the question requires Alice to answer by reporting the likelihood of the event according to that which is (or which Alice takes to be) a body of objective evidence and belief shared by the relevant group of agents. In such a conversational context, Alice cannot be taken to be anchoring her modal claim anywhere else than this body of evidence and belief. However, and this is crucial, the listener typically *does not* share the same body of evidence and belief: otherwise, the question about the likelihood would not have been under discussion in the first place. In other words, the listener typically interprets Alice's answer by inferring that it must be the report of a group of agents' shared beliefs, but crucially a group he does not belong to. Hence, the follow-up question *Who says so?* is perfectly acceptable in the context.

Next, let us move to the example in (32), containing an uncertainty adverb (*probably*). Our goal is to answer the question why Alice's utterance in (32-a) gives rise to the speaker inference. First of all, let us think about that contrast between the two alternative utterances in (31-a) and (32-a). We have observed that the former is typically used to answer a QUD about the likelihood of the prejacent *The dog is on the lawn*. In other words, if a speaker chooses to say something along the lines of (31-a), the listener reasonably take her to be addressing such a QUD about the objective body of objective evidence and belief shared by the relevant group of agents. By contrast, if the speaker chooses to utter an alternative form containing an uncertainty adverb (such as (32-a)), the listener might reason about what the speaker could have said but did not, and the listener might infer that the speaker is addressing a different QUD. Which one?

We believe that utterances such as (32-a) are typically used to answer a direct question about the dog's whereabouts, *Where is the dog?*. We notice that in this respect, such utterances are similar to bare, non-modal utterances:

(33) a. *Bob:* Where is the dog?
b. *Alice:* The dog is on the lawn.
c. *Alice:* The dog is probably on the lawn.
d. *Bob:* #Who says so? / #Whose opinion is this?

Now, bare, non-modal, utterances typically go hand-in-hand with direct factual QUDs about "what the world is like": the best way to answer Bob's question in (33-a) is clearly by specifying the dog's whereabouts, as in Alice's answer in (33-b), no uncertainty involved. On the other hand, if Alice *is* uncertain, the best she can do is to still mention the most likely location (*the lawn*) but also express her uncertainty with the modal claim, such as the one in (33-c). So utterances containing uncertainty adverbs are typically used to answer the same kind of default questions (direct and factual *wh*-questions) as bare, non-modal claims. Moreover, as already observed above and shown again here in (33-d), bare, non-modal claims give rise to same speaker inference as utterances containing uncertainty adverbs: Bob's follow-up questions in (33-d) sound just as redundant after (33-b) as after (33-c).

We conclude that in the given conversational context, one where the QUD is (or is inferred to be, by contrastive reasoning or otherwise) a factual question about the world, listeners will typically assume that the speaker's answer reflects the speaker's own beliefs. In other words, the default anchor of utterances about the world is the minimal group of agents containing exactly the speaker, at the time of the utterance, unless otherwise explicitly or implicitly specified. Summing up, there's nothing special to the speaker inference of utterances containing uncertainty adverbs: being typically used to answer direct factual questions about some event, they behave as non-modal claims in reporting the speaker's beliefs (at the time of the utterance).

Finally, let us show how we can explain the observation that the difference discussed above disappears when the modal claims are embedded under epistemic attitude expressions such as *believe*. The relevant example is repeated here as (34):

(34) a. Alice believes it's probable that the dog is on the lawn.
b. Alice believes that the dog is probably on the lawn.
c. According to Alice's beliefs, there are good chances that the dog is on the lawn.

The observation is that the speaker inference is missing: no matter who utters which sentence, be it (34-a) or (34-b), the resulting interpretation will be along the lines of the paraphrase in (34-c), i.e. that *Alice's* beliefs on the subject matter (and not the speaker's) are being reported in both the modal claims. Why does this happen? Our answer is that both (34-a) and (34-b) are typically used to answer the same QUD: *What does Alice believe?*. In a context where this is the QUD, the at-issue content is clearly about Alice's body of evidence and beliefs, hence both *probable* and *probably* receive the same anchor, hence the same interpretation.

Nielsen's contrast. One fact that needs explaining is why uncertainty adjectives and adverbs seem to behave differently in complex constructions such as Nielsen's conjunctions, repeated here as (35) and (36):

(35) It's possible that Le Pen will win even though she certainly won't.

(36) #Le Pen will possibly win even though she certainly won't.

The observation is that (35) is an acceptable sentence, whereas (36) sounds less felicitous (if not altogether contradictory). In order to put forward an explanation of this fact, we ask ourselves: what is the at-issue content in each of these sentences? Le us begin with (36), and explain why it sounds contradictory. Given what we observed above, the typical at-issue content in sentences containing uncertainty adverbs is the status of the prejacent. In this case, whether Le Pen will win the elections or not. Then, both modals in (36) typically receive default anchoring to the speaker at the time of the utterance, such that the sentence could be paraphrased along the lines of (37):

(37) As far as my current evidence and beliefs go, Le Pen has a chance of winning and she doesn't have one.

which is quite the contradiction. Similarly, we can explain why (35) sounds better. The first conjunct is typically used to address a QUD about the objective chance of the prejacent, e.g. *How likely is it that Le Pen will win?* and as such the modal *possible* is interpreted to an objective or shared body of evidence and belief. On the other hand, the second conjunct contains an uncertainty adverb, which is typically anchored to the speaker. Hence, (35) can be paraphrased along the lines of (38), which is not contradictory:

(38) It is generally believed that Le Pen has a chance, but I personally believe she doesn't.

Interaction with tense. The examples discussed above are repeated here as (39-a) and (40-a):

(39) a. In 1979, it was probable that the socialists won the elections.
b. Given people's uncertain beliefs *in 1979*, they judged the likelihood of a socialist victory (in 1979) to be high.

(40) a. In 1979, the socialists probably won the elections.
b. Given our uncertain beliefs *now*, we judge the likelihood of a socialist victory (in 1979) to be high.

The observation is that the uncertainty adjective *probable* can easily scope below the past tense, as exemplified in (39-a), whereas the adverb *probably* tends to scope above the tense, as in (40-a). Once again, we ask ourselves which QUD these sentences are typically used to address. Starting with (40-a), this strikes us as the sort of claim that a(n uncertain) historian would make: the content at-issue here is about the outcome of the elections in 1979; the historian is trying to answer, to the best of her knowledge, a QUD along the lines of *Who won the elections in 1979?*. As noticed above, sentences containing uncertainty adverbs such as *probably* are typically used to answer this type of factual questions, and as a consequence the uncertainty adverbs receive a global anchoring to the speaker at the time of the utterance. Hence, the interpretation suggested by (40-b). On the other hand, we observed how sentences containing uncertainty adjectives (*probable*) are typically used in contexts where the likelihood of the prejacent is at-issue as well. This is the case in examples such as (39-a), where a likely QUD is something along the lines of *How likely was it that the socialists won the elections?*. If this is the case, then the speaker is taken to be anchoring her modal claim to a group of agents at the specified time, whose beliefs are (or she takes to be) relevant to answer the question how likely was a socialist victory believed to be back then. Hence, the interpretation suggested by (39-b).

Finally, suppose that the QUD is about the socialists' beliefs about the outcome of the elections in 1979:

(41) a. What did the socialists believe about the elections in 1979?
b. The socialists believed that it was probable that they won the elections.
c. The socialists believed that they probably won the elections.
d. Given what the socialists believed *back then*, the likelihood of victory was high.

These examples allow us to clarify why the difference between uncertainty adverbs and adjectives in interaction with tense seems to disappear when both are embedded under verbs of belief. The explanation is simple and follows the same strategy adopted with present-tensed modal claims. If a modal claim is embedded under a construction such as *The socialists believed that...* then presumably the sentence is typically being uttered to answer a QUD about the socialists' beliefs at the (specified or unspecified) time, as exemplified by (41-a). In other words, the socialists' beliefs are part of the at-issue content of the sentence. It is then natural to assume that typically the speaker is anchoring her modal claim (regardless of it containing an uncertainty adjective or adverb) to the group of socialists at the (specified or unspecified) time. It is *their* beliefs *back then* that the speaker is reporting. Hence, both (41-b) and (41-c) can be appropriately paraphrased as (41-d).

3.4 Conclusion

In this chapter we sketched a semi-formal theory of the context-dependency (in particular, the QUD-sensitivity) of uncertainty expressions with the goal of investigating the differences between uncertainty adverbs (*possibly*, *probably*) and adjectives (*possible*, *probable*). The theory, heavily inspired by (Roberts, 2015, 2017), posits that different grammatical forms of uncertainty expressions are typically used to address different QUDs and as a consequence they are typically assumed to be indexically anchored to different (groups of) agents and hence different bodies of evidence or beliefs. We argued that a number of observable semantic/pragmatic/conversational differences between adverbs and adjectives can be explained on the basis of this QUD-sensitivity.

As we mentioned at the beginning of the chapter, our investigation was inspired by the research of Wolf and colleagues, whose SAM theory and the data collected in support of the theory have been in the background of the whole chapter. More specifically, the linguistic phenomena involving uncertainty adverbs and adjectives collected by Wolf and colleagues were the baseline benchmark of success of our own proposal. Can we explain all the observed phenomena in terms of QUD-sensitivity? As far as the semantic/pragmatic/conversational differences go, we hope to have shown that our approach fares just as well as SAM theory (and sometimes even better).

Therefore, is a strong distinction such as the one assumed in SAM theory still necessary? How far can pragmatics go in explaining *all* the differences between uncertainty adverbs and adjectives? These questions are harder to answer. We have seen that our theory does not say anything about the syntactic differences such as the different distributions of adverbs and adjectives within the scope of interrogative contexts or conditional constructions or factive verbs. On the other hand, we have argued that Wolf's and colleagues' data about these phenomena might not be as stable as they claim, which seems to weaken the case for the strong distinction assumed in SAM theory. However, the patterns involving negation do seem to be strong, and it is hard to find genuine examples of negated uncertainty adverbs: *impossibly*, *improbably* and similar certainly exist, but they do not appear be typically used with the meaning of NOT(*possibly*) and NOT(*probably*).

The distributional properties of uncertainty adverbs and adjectives are an intricate matter and, as things stand at the moment, we are unable to completely rule out that some reference to the grammatical category or syntactic role of the expressions involved might be necessary to completely understand their behavior. One conclusion which we want to draw in this chapter has a conditional form. *If* the distributional properties of uncertainty adverbs and adjectives are not enough to justify a strong syntactic distinction such as the one assumed by SAM theory, *then* we can safely do without such a strong distinction, because we have shown that the remaining semantic, pragmatic and conversational differences can be successfully explained in terms of QUD-sensitivity.

Let us conclude by pointing out that even if our proposal will eventually need to resort to some hard grammatical assumption about the embeddability of uncertainty adverbs, we believe that our proposal should still be preferred to SAM theory, in that it would still be less stipulative and more explanatory. Shall we need to stipulate that *probably* is not licensed inside the scope of negation? Notice that SAM theory is based on a similar (if more general) stipulative assumption. We, on the other hand, have also given a functional explanation of the other observable differences between uncertainty adverbs and adjectives which is based on general and independently motivated principles at play in cooperative conversations.

Chapter 4

Higher-order uncertainty and rational beliefs

The only thing that makes life possible
is permanent, intolerable uncertainty:
not knowing what comes next.
(Ursula K. Le Guin, The Left Hand of Darkness*)*

4.1 Introduction

Objective and subjective uncertainty. In the previous chapters of this dissertation we referred several times to concepts such as likelihood and credence, or objective and subjective uncertainty, but we kept the discussion of these terms, and in particular of the distinction between objective and subjective, at a rather intuitive level. For example, in Chapter 3 we talked about how different conversational contexts (QUDs/at-issue contents) might have an influence on whether utterances of probability expressions are taken to report about an intersubjective or objective body of credence rather than the speaker's subjective uncertain beliefs, but the distinction between the two kinds of uncertainty was not spelled out formally. Moreover, the experimental study reported in Chapter 2 was concerned exclusively with manipulations of the objective likelihood of drawing balls of a certain color.

However, we have seen in Chapter 1 that the distinction between objective and subjective uncertainty seems to play a crucial role in people's use and interpretation of simple and complex probability expressions (Moss, 2015; Ülkümen, Fox, & Malle, 2016; Lassiter, 2018). In this short chapter we introduce a formal distinction which helps us capture at least some of the intuitions surrounding the distinction between objective and subjective uncertainty in a precise way, which is also quite easy to implement and manipulate in experimental settings. Crucially, the setting introduced here is also adopted in the experimental studies reported in Chapter 5 and Chapter 7.

Instead of talking about objective and subjective uncertainty, we distinguish between what we call different *layers* or *orders* of uncertainty. Before defining explicitly what we mean by this, notice that the proposed distinction between orders of uncertainty is not supposed to be 100% equivalent (and hence replace) the intuitive distinction between objective and subjective uncertainty. Rather, we think of it is a precise and formal way to capture some of the intuitions related to the intuitive distinction. Having clarified this, we can dive into the details.

The first, most basic layer of uncertainty is perfectly exemplified by a situation where an agent knows the bias of a coin, e.g. 0.7, and the agent tries to predict the outcome of a single coin toss. The agent has perfect information about the chance of getting heads (or tails). And yet, prior to a coin toss, the agent will be uncertain. Analogously, recall the manipulation of the quantity of balls of a certain color in the urn setting of the experiment described in Chapter 2. If an agent knows that an urn contains ten balls and that exactly seven of them are red and three of a different color, the agent has perfect information regarding the chance of randomly drawing a red ball from the urn. And yet, prior to the actual draw, the agent will be uncertain: will the ball be red or not? Much like the bias of a coin, the quantity of red balls in an urn can be seen as an objective feature of the world, the result of a genuine and in-principle inscrutable stochastic event.[1] For this reason, we think of this first layer of uncertainty as akin to what we usual have in mind when we refer to objective uncertainty: it is not due to lack of factual knowledge, as the agent knows everything there is to know about the urn, and every rational agent in the same situation will have the same sort of uncertainty.

Higher-order uncertainty. Higher-order layers of uncertainty are possible. In this work we limit our investigations to a second layer of uncertainty, corresponding to situations where an agent does not have perfect information about the objective chance of some event, and instead the agent is uncertain between a set of possible values for the objective chance. For example, imagine an agent playing with the same urn containing seven red balls and three blue balls, except this time the agent cannot directly look inside the urn and know its content. Instead, the agent is allowed to draw a certain limited number of balls from the urn, say eight, and look only at those. Based on the number of red balls among the eight sampled balls, the agent can form an uncertain belief about how many red balls there might be in total in the urn, but this one-shot process is not enough to acquire perfect information of the urn. For example, if there are six red balls among the sampled eight, then the agent might believe that the urn contains six red balls in total, or seven, or eight.

In this case, our intuition is that the kind of uncertainty involved can be called "subjective" as it is due to the agent's lack of factual knowledge: the objective chance of drawing red is still a feature of the world, potentially accessible, but the agent has only limited access to it, hence he or she is uncertain about the objective chance. This layer of uncertainty is akin to what we usually have in mind when we refer to subjective uncertainty, as it can be influenced by the agent's body of evidence and can be reduced by acquiring true information. In fact, with better access or more repetitions of the

[1] We gloss over the philosophical intricacies revolving around the question of which events are genuinely stochastic and which are not. If the reader finds coin flips and ball draws to be not enough genuinely stochastic, he or she is free to concoct a creative scenario in which the results of coin flips or ball draws are situated at the end of a causal chain initiated by, say, the decay of a certain radioactive particle.

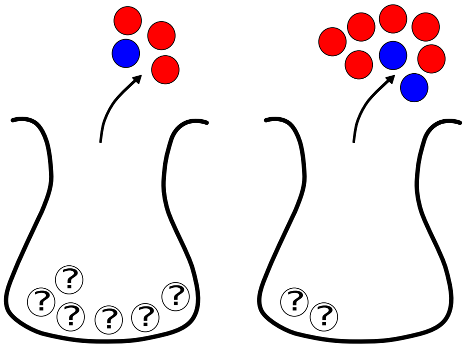


Figure 4.1: Examples of partial observations of the contents of the urn. Left hand side: $a = 4$, $o = 3$; right hand side: $a = 8$, $o = 6$.

observation process the agent could acquire better knowledge of the objective chance, decreasing his or her amount of higher-order uncertainty.

Let us expand on the urn example, and introduce the formal setting which underlines the experimental and modeling work presented in this chapter (and the following ones). The urn always contains ten balls of two different colors, e.g. red and blue. Any number $s \in S = \{0, \ldots, 10\}$ of balls in the urn can be red. S is the state space, or universe of the discourse. The ratio $s/10$ is the objective chance of a randomly drawn ball to be red. As we said, agents who know the objective chance have perfect information about the contents of the urn, and yet they are uncertain when it comes to making predictions about, e.g. the color of a ball drawn at random from the urn. Because we want to model higher-order uncertainty, agents typically do not know the objective chance: they cannot directly observe s. Instead, they draw a certain number of balls from the urn –we call this "access", denoted with a, and count how many of the drawn balls are red –we call this "observation", denoted with o (see Figure 4.1). The ratio o/a of observed over accessed red balls provides an approximation of the objective chance, and higher access values give rise to better approximations. For example, drawing four balls from the urn and observing that three of them are red or drawing eight and observing that six are red correspond to the same proportion of observed red balls (75%) but the latter case intuitively provides more information. If a (rational) agent observes $^6/_8$ red balls, he or she will have more precise beliefs about the contents of the urn than another agent who has only observed $^3/_4$ red balls, despite the proportion of observed red balls being the same.

The move from different observations of the contents of the urn to more or less precise beliefs of a rational agent is intuitive enough, but how can it be made precise enough to enter a formal model of language use and interpretation such as the one presented in the following chapters?

Uncertain rational beliefs. The inspiration to answer this question comes from a paper by Goodman and Stuhlmüller (2013), in which the authors adopt a similar setting with partial observations of the contents of an urn in their pragmatic model of scalar im-

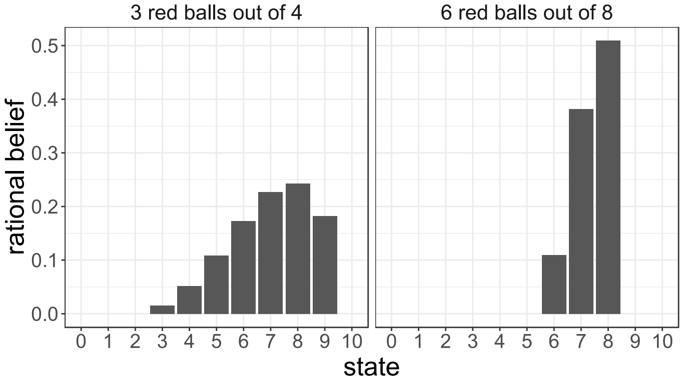


Figure 4.2: Examples of rational belief distributions given two partial observations of the urn and assuming a uniform prior over states.

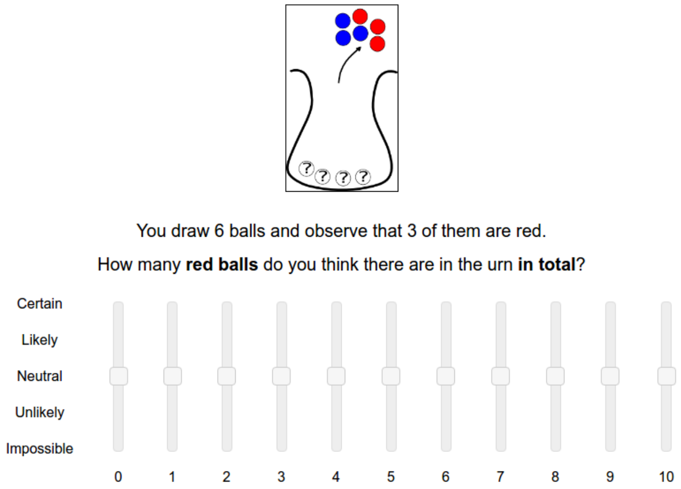
plicatures. The belief formation of an ideally rational agent who draws *a* balls from an urn containing ten balls and observes that *o* are red can be modeled as Bayesian update of the agent's prior belief distribution over the state space given the hypergeometric model of the urn (Goodman & Stuhlmüller, 2013):[2]

$$\text{rat.bel}(s \mid v = \langle o, a \rangle) \propto \text{Hypergeometric}(o; a, s, 10) \cdot \text{prior}(s) \tag{4.1}$$

Figure 4.2 displays the belief distributions computed for the two observations of our running example (3/4 and 6/8 red balls), assuming for the time being a flat prior distribution over states. We can see that both distributions have the same mode equal to 8 but the right hand side distribution has lower entropy, being more closely concentrated around the mode. This reflects the intuition that the agent's beliefs are more precise.

Equation 4.1 defines a normative model: it tells us what rational agents *should* believe about the contents of the urn given their partial observations. This brings us to the goal of this chapter. Goodman and Stuhlmüller (2013) assume the normative model of belief formation defined in Equation 4.1 in their model. But do people really form uncertain beliefs according to Equation 4.1? How lightheartedly can we assume the normative model as a crucial component of *our* model? It is a controversial question to what extent people deviate from the norms of probabilistic reasoning (e.g. Tversky & Kahnemann, 1974; Gigerenzer & Goldstein, 1996; Jones & Love, 2011; Sanborn & Chater, 2016). For this reason, we ran an experimental study to estimate participants' posterior beliefs about the contents of the urn with a slider bin rating task (e.g. Kao, Wu, Bergen, & Goodman, 2014; Degen, Tessler, & Goodman, 2015; Franke et al., 2016).

[2] The proportionality sign between the two sides of the equation indicates that they are to be equal up to a normalizing constant, which in this case is: $\sum_{s'} \text{Hypergeometric}(o \mid a, s') \cdot \text{prior}(s')$.



You draw 6 balls and observe that 3 of them are red.

How many **red balls** do you think there are in the urn **in total**?

Figure 4.3: Sample stimulus and input slider bins.

4.2 Experiment 2: measuring uncertain beliefs

Participants. 104 self-reported English native speakers with IP addresses located in the USA were recruited on Amazon's Mechanical Turk. The workers were paid 1 USD for their participation, amounting to an average hourly wage of approximately 10 USD.[3]

Materials and procedure. A short tutorial introduced the participants to the overall urn setting, to the graphical and textual representation of partial observations of the urn and to the experimental task. Participants learned that they would exclusively be presented with urns containing ten balls of at most two colors (e.g. red and blue) and that they would not be able to look inside the urns. Instead, they would draw a certain number of balls (*access*) and observe the distribution of colors in the sample (*observation*). On the basis of their observation, they would try to guess the exact contents of the urn.

The experimental conditions were all the 65 logically possible combinations of access values from 1 to 10 and observation values from 0 to 10. After the tutorial, each participant completed 13 experimental trials. Each trial was randomly associated with one of the 65 conditions which was not previously selected for the participant. In each trial we displayed a picture representing the associated observation of the urn, together with a short description. We asked the participants to answer the question "How many red balls do you think there are in the urn in total?". In order to provide their answers, participants adjusted 11 sliders, one for each possible quantity of red balls in the urn (i.e. from 0 to 10) and ranging from *Impossible* to *Certain*, expressing the in-

[3] Code and data relative to this study are publicly available at `https://github.com/mic-he/ProbExp-PhD/tree/master/chapter4`.

tuitive likelihood associated with each quantity of red balls having made the current observation of the urn. We recorded slider ratings as discrete values ranging from 0 (*Impossible*) to 1 (*Certain*) with a step of 0.01. Figure 4.3 displays a screenshot taken from an experimental trial.

Results. We discarded the data points obtained from 3 participants because they had selected *Impossible* for all the sliders in at least one observation condition. For each of the remaining 101 participants, our data consist of 13 vectors (one for each trial) of 11 ratings each (i.e. real numbers, one for each slider). For each participant and each vector we normalize the ratings so that the values in each vector sum to one. Next, for each of the 65 conditions, we calculate the average vector of ratings across participants. We obtain 65 discrete distributions, one for each possible observation of the urn, which we take be an approximation of the central tendency of beliefs held by all participants given the corresponding observation of the urn.

These distributions can be easily visualized. For example, Figure 4.4 displays histograms representing the measured distributions for 15 observation conditions. These are especially interesting for us as they will play a role in the linguistic experiments reported in the following chapters of this dissertation. The plotted distributions seem to display a reasonable pattern: they get more and more peaked, towards reasonable values, as the number of accessed balls increases, i.e. as the agent's level of higher-order uncertainty decreases. We think of this feature as a sanity check for our present and later experimental setting: the behavior of the participants does not appear to be especially surprising with respect to our expectations. But are these results compatible with the assumption that agents form their beliefs about the urn according to the normative model sketched in Equation 4.2? In order to answer this question we analyze our data with a Bayesian hierarchical model.

4.3 Model: rational beliefs about the urn

The data-generating model described in this section is based on the model developed by Franke et al. (2016) to "estimate subjective beliefs from empirical data". In particular, we adopt the same likelihood function for observed slider ratings proposed by Franke et al. in a setting similar to ours. More in detail, Franke et al. build a model in which population-level average beliefs are the central tendency of the experimental participants' individual-level beliefs, which in turn determine the likelihood of observing each particular slider rating in the data. We follow the same approach. The difference between our work and Franke et al.'s is that their goal is to infer, via Bayesian posterior inference, likely population-level beliefs in domains where no obvious normative belief distribution exists, whereas our goal is to test the assumption that the observed slider ratings were produced by agents who held the normatively correct Bayesian beliefs in each experimental condition. Figure 4.5 depicts the model as a probabilistic graphical model, following the conventions outlined by Lee and Wagenmakers (2014), together with the full formal definition of the model.

The normatively correct Bayesian beliefs are defined by Equation 4.1 above, re-

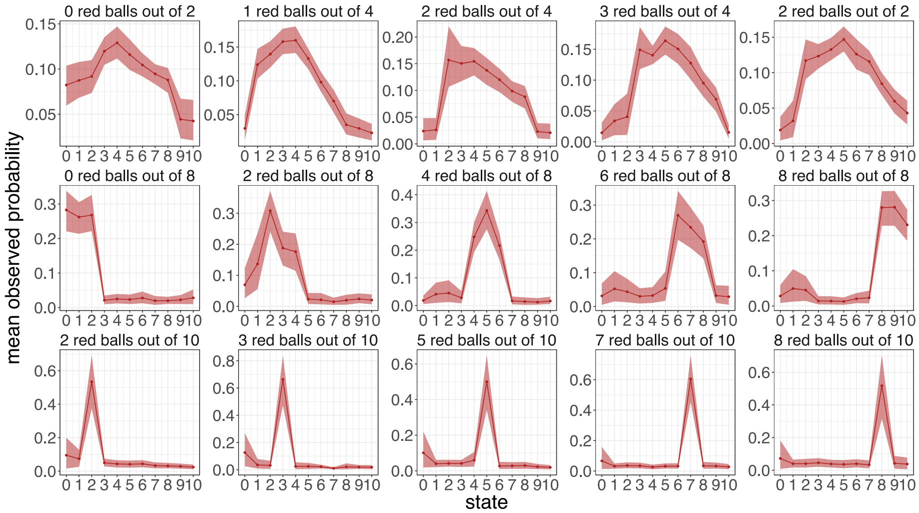


Figure 4.4: Measured belief distributions (red lines) in 15 observation conditions together with 95% bootstrapped confidence intervals (ribbons).

peated here as 4.2:

$$\text{rat.bel}(s \mid v = \langle o, a \rangle) \propto \text{Hypergeometric}(o; a, s, 10) \cdot \text{prior}(s) \tag{4.2}$$

As mentioned above, the distribution *rat.bel* expresses how likely it is, according to the agent, that there are s red balls in an urn containing ten balls, given that the agent has drawn a balls from the urn and observed that o of them were red. The agent's prior beliefs about the contents of the urn play a role here, because Equation 4.2 is nothing but an application of Bayes' rule to the hypergeometric model of the urn. The prior on S is defined for convenience as a beta-binomial distribution:[4]

$$\text{prior}(s) \sim \text{Betabinomial}(s; \alpha, \beta, 10) \tag{4.3}$$

The parametrization of the beta-binomial in terms of $\alpha = \omega \cdot (\kappa - 2) + 1$ and $\beta = (1 - \omega) \cdot (\kappa - 2) + 1$ is taken from Kruschke (2014), whereas the prior structure on the hyperparameters with $\kappa - 2 \sim \text{Gamma}(0.01, 0.01)$ and $\omega \sim \mathcal{U}(0, 1)$ reflects our noncommittal stance on the prior and enforces a roughly flat distribution.

We make the strong assumption that the distribution *rat.bel* defined here is held by every participant in every condition of our experiment. For a given condition (hence, belief distribution), the model predicts a likelihood of observing a particular slider value. Assuming for simplicity that each slider rating is independent of each other, we model each slider rating as a noise-perturbed realization of the corresponding probability mass prescribed by *rat.bel*. This is captured by Equation 4.4, where the likelihood

[4] A variable with a beta-binomial distribution is distributed as a binomial distribution with parameter p, where p is distributed as a beta distribution with shape parameters $alpha$ and $beta$. The beta-binomial distribution is the conjugate prior of the hypergeometric distribution (Peskun, 2016).

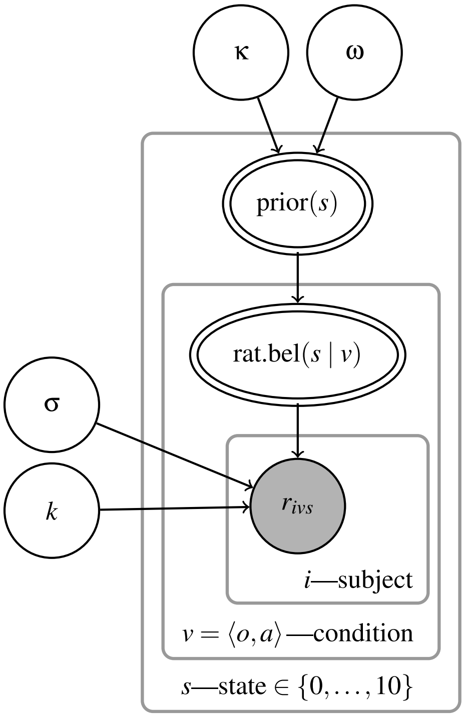

κ
ω
prior(s)
rat.bel($s \mid v$)
σ
k
r_{ivs}
i—subject
$v = \langle o, a \rangle$—condition
s—state $\in \{0, \ldots, 10\}$

Figure 4.5: The data-generating model as a probabilistic graphical model. White nodes represent latent variables, shaded nodes represent observed variables. Single-bordered nodes represent stochastic dependence, double-bordered nodes represent deterministic dependence. Boxes indicate scope of indices.

of observing the rating r_{ivs} given by participant i in condition $v = \langle o, a \rangle$ for slider s is defined as follows:

$$\text{logit}(r_{ivs}) \sim \text{Norm}(\text{logit}(\text{rat.bel}(s \mid v = \langle o, a \rangle), k), \sigma) \tag{4.4}$$

Both $\text{logit}(r_{ivs})$ and $\text{rat.bel}(s \mid v)$ are mapped from the unit interval to the reals and the observed value is modeled as a realization of the predicted value with normally distributed noise with standard deviation σ. The parameter k modulates the steepness of the logit transform of $\text{rat.bel}(s \mid v)$, allowing for the possibility that participants may be affine ($\kappa > 1$) or averse ($\kappa < 1$) to realizing extreme slider ratings close to 0 or 1. Finally, the prior structure on the parameters σ and k is taken from (Franke et al., 2016).

In order to address our starting question about the plausibility of the normative model of belief formation, we compare our experimental data with model predictions. These are generated as posterior predictive distributions by implementing the model in the probabilistic programming language JAGS (Plummer, 2003) and estimating the joint posterior distribution over the model's parameter values given our data. More in detail, we collect two Markov Chain Monte Carlo (MCMC) chains of 2500 samples from the posterior distributions after dicarding the first 2500 samples —the so-called *burn-in* period (Kruschke, 2014).[5] We checked convergence via $\hat{R}$ (Gelman & Rubin, 1992).[6] Model predictions are produced by generating hypothetical data from the model with parameter values randomly drawn from their inferred posterior distribution (i.e., conditional on the data). For each of the 2500 sample vectors of parameter values the model generates a set of posterior predictive distributions for the population-level beliefs. As an illustration, mean values of the generated posterior distributions corresponding to 15 experimental conditions are visualized in Figure 4.6, together with 95% highest density intervals, and superimposed on the experimental data already displayed in Figure 4.4.

The predictions can be visually compared with experimental data via Bayesian posterior predictive checks (PPCs). We look for shortcomings of the model in the form of discrepancies between predicted and actual data, i.e. points in the plots displayed in Figure 4.6 where the red and grey ribbons do not overlap for a given state value: in these cases the data is still unexpected for the model trained on the data. The only obvious discrepancies can be observed in the 8/8 condition (rightmost panel of the middle row of Figure 4.6), where the model clearly underpredicts the probability of state 8 and 9. This seems to follow a tendency of the model to be more cautious than the observed data, so to speak, in the access=8 condition (middle row of Figure 4.6). However, we can observe that in the vast majority of data points the model appears to be able to adequately predict the data, suggesting that the normative model of rational belief formation adopted in the model can be a rough but good enough approximation to the population-level belief distributions underlying participants' choices given partial

[5] A so-called *Markov* process can be defined as a process in which each step has no dependence on (or memory of) any state or step before the current one; a sequence of such steps is a Markov chain. The *Monte Carlo* part of the name is evocatively used to refer to any simulation sampling many random values from a distribution. MCMC chains are sequences of samples generated through a Markovian random walk in order to estimate the unknown posterior distribution of interest.

[6] The $\hat{R}$ convergence statistics is based on the stability of outcomes between and within chains of the same length.

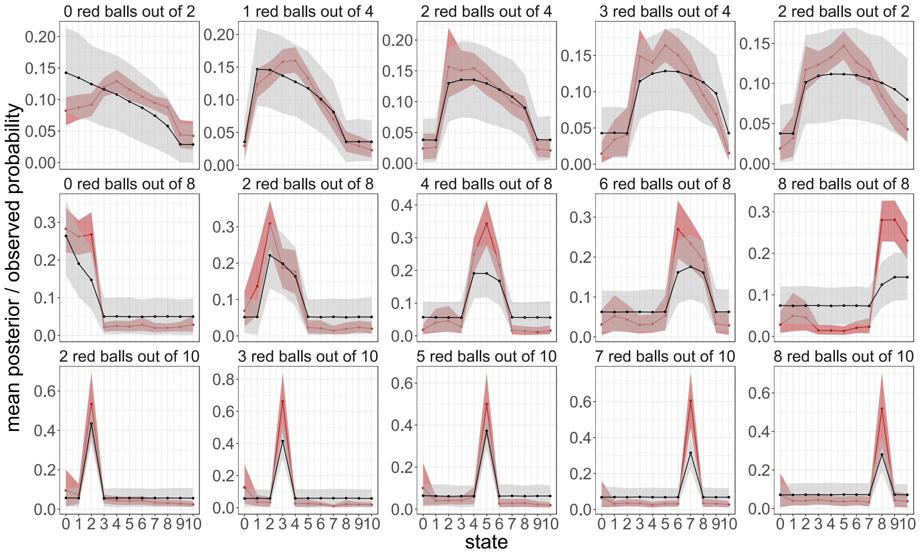


Figure 4.6: Posterior predictive distributions (black lines) in each of 15 observation conditions together with Bayesian 95% highest density intervals (grey ribbons). Red lines and ribbons display mean values and 90% CIs of observed data.

observations of the urn.

4.4 Conclusion

The goal of this chapter was to introduce an urn and balls scenario in which (some aspects of) the distinction between subjective and objective uncertainty can be captured and experimentally manipulated. At the same time, we were interested in testing whether experimental participants in an urn-based scenario would form beliefs about the contents of the urns according to a particular normative model of rational belief formation.

To do so we ran an experimental study in which we measured participants' guesses about the contents of an urn after being exposed to different partial observations of said contents. The data showed encouraging patterns, and it was integrated into a cognitive Bayesian model based on the assumption that observed experimental data can be generated as a noisy realization of population-level beliefs which, in turn, conforms to the normative model. The model trained on the observed data yields Bayesian posterior predictive distributions for each of the experimental conditions and these distribution were visually compared to the observed data. Despite some discrepancies, observed data and model predictions align quite well, which we take as an indication that participants' behavior is compatible with the starting rationalistic assumption.

In view of this result, we can conclude that the normative model of rational belief formation defined in 4.2 can be a seen as a good-enough approximation of actual agents' behavior and for this reason we will adopt it, together with the urn-based scenario, in our data-driven pragmatic models of language use and interpretation.

Chapter 5

Simple uncertainty expressions: data

> Phoebe*: You're sure? You're absolutely sure?*
> Monica*: Well, no, but she probably does.*
> Phoebe*: Oh, probably? Yeah, I don't like that word.*
> *Look, I know what probably* ***really*** *means.*
> *(Friends, season 7, episode 10)*

5.1 Introduction

Let us take stock. Chapter 2 was dedicated to Experiment 1. We gathered and analyzed data about the context sensitivity of uncertainty expressions: the importance of the context should now be apparent but the question remains of how uncertainty expressions are pragmatically used to communicate in context. Chapter 3 contained theoretical/philosophical speculations in this direction, in that we investigated the way in which conversational contexts (QUDs/at-issue content) can affect the use of uncertainty expressions. We spelled out a semi-formal model, but the discussed empirical data was nothing more than (our and other authors') armchair intuitions about the relevant phenomena. Chapter 4 introduced and investigated (Experiment 2) the formal distinction between different layers or orders of uncertainty, following up on our early (Chapter 1) intuition that they play a role in the pragmatic use of uncertainty expressions.

This chapter moves from there, and can be seen, together with the next one, as the focal point of the dissertation, as different lines converge here. We report first and foremost on empirical investigations (Experiment 3) into how people use and interpret simple uncertainty expressions under higher-order uncertainty. We show that our data vindicates our intuitions. Interesting result on its own, we nonetheless aim at explaining the how and why, in a rationalistic fashion: how and why do we use simple uncertainty expressions in this way? Our answer takes the shape of a probabilistic pragmatic model, based on RSA, whose detail will be spelled out in Chapter 6.

Complex expressions will be investigated (both data —Experiment 4, and modeling) in Chapter 7.

5.2 Experiment 3

Goal. We ran two experiments on AMT collecting human data on the production (Experiment 3a) and interpretation (Experiment 3b) of simple uncertainty expressions under higher-order uncertainty. The main goal of Experiment 3a was to test the hypothesis that different levels of higher-order uncertainty, of the kind introduced in Chapter 4, play a role in speakers' production of uncertainty expressions. Our intuition is that they do. Moreover, we believe that part of the communicative effect of our utterances containing uncertainty expressions is indeed the transfer of higher-order uncertain information: if I say that a ball drawn at random from the urn will probably be red, I'm not only communicating a vague estimate of the likelihood of drawing red, but also something about my degree of credence in that estimate. If this is so, then we can reasonably expect that listeners will be able to interpret these utterances accordingly. This is an empirically open question: can listeners infer the communicated information about the speaker's level of higher-order uncertainty alongside the communicated information about the world? Answering this question was the main goal of Experiment 3b.

Design. Participants in both experiments read a short cover story introducing them to the general setting of the experiment, which was fictitiously described as a game in which the players would cooperate with each other. The exact description read as follows:

> "This experiment is an interactive two player game of chance. The players cooperate to guess the contents of an urn. Both players know that the urn always contains 10 balls of different colors (for example, red and blue). But only one player (the sender) is allowed to draw a certain number of balls from the urn and look at them. The sender puts the balls back into the urn and gives it a nice shake, then the sender draws a new ball from it. Before looking at it, the sender sends a message to the other player (the receiver). The receiver reads the message and tries to guess the exact contents of the urn."

This description summarizes three important elements of the experimental design, which are intended to meet three desiderata for a scenario in which reasoning about higher-order uncertainty could affect language use. First of all, the urn setting together with the partial observation procedure allow us to easily manipulate both the likelihood of the event (the random ball is red) and the level of higher-order uncertainty in a flexible and precise way, which is also easy to visualize and intuitive for the participants. Second, the game-like cooperative scenario has the goal of prompting participants to reason about the communicative effect of their utterances on other agents (in Experiment 3a) and to interpret other agents' utterances by reasoning about what they could have intended to communicate (in Experiment 3b). Finally, by explicitly stating that the participants should coordinate in order to guess the exact contents of the urn, we make it clear that the purpose of the conversation is the transfer of information about the world (e.g., how many red balls there are in the urn), while at the same time communicating (or inferring) the level of uncertainty of the transferred factual information.

5.2.1 Experiment 3a: production

Participants. 84 self-reported English native speakers with IP addresses located in the USA were recruited on Amazon's Mechanical. The workers were paid 0.75 USD for their participation, amounting to an average hourly wage of approximately 10 USD.[1]

Materials and procedure. After the introductory phase described above, participants completed a few familiarization trials in which they played in the role of receivers, reading a messages sent by another (fictitious) player and trying to estimate the number of red balls contained in the urn. Subsequently, they moved to the main experimental phase, in which they took on the role of senders. In each trial of the main phase, participants observed a picture representing an observation of the urn and they were asked to make a prediction about the color of a ball drawn at random from that same urn.

high	0/2	1/4	2/4	3/4	2/2
low	0/8	2/8	4/8	6/8	8/8
none	2/10	3/10	5/10	7/10	8/10

Table 5.1: Experimental conditions in Experiment 3a. The fractions represent observations of the urn. The labels on the left refer to levels of higher-order uncertainty.

The experimental conditions were 15 observations of the urn, as summarized in Table 5.1. Each fraction in the table represents a possible observation: the denominator is the number of drawn balls, the numerator is the number of red balls observed among them. Our choice of experimental conditions allows us to cover a reasonably wide range of proportions of observed red balls under different levels of higher-order uncertainty. For example, we can realize the same proportion of 1/2 under three levels of higher-order uncertainty, namely high (corresponding to access=4), low (access=8) and none (access=10). For other proportion values the symmetry is not perfect, but we can still obtain close enough values in the three different levels of uncertainty (for example, 1/4 and 2/8 correspond to a proportion of 25%, which fits in between 2/10 (20%) and 3/10 (approximately 33%).

The main phase consisted of 9 trials, 3 for each level of higher-order uncertainty, in random order; each trial in each level was randomly associated with one experimental condition which was not previously selected in that level. In each trial participants were shown a picture representing the selected condition and they were told to imagine making the depicted observation of the urn, then putting all the balls back in the

[1] Code and data relative to this study are publicly available at `https://github.com/mic-he/ProbExp-PhD/tree/master/chapter5-6`.

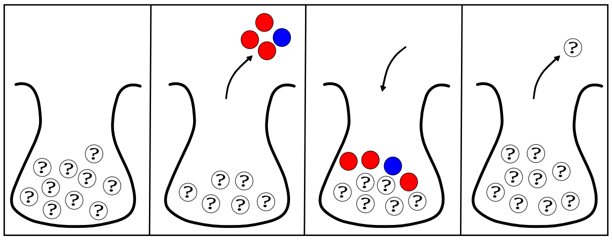


Figure 5.1: Example of picture displayed to participants in the production trails. From left to right, it represents the unknown urn, a partial observation of 3 red balls out of 4, and the random draw of a new ball whose color needs to be predicted by the participant.

urn (see Figure 5.1). Participants were then asked to send a message to another (fictitious) player, containing a prediction: will a ball drawn at random from the urn be red? Crucially, this message must be of the form

> *The next ball will [...] be red*

where the gap must be completed by the participant selecting an item from a drop-down menu containing the followgin messages:

certainly, *probably*, *possibly*, *probably not*, *certainly not*

We recorded counts of participants' choices of expressions in each observation condition.

Results. Figure 5.2 displays the collected data as percentages of choices of expressions in each condition. We look at the plot with a precise research question in mind: which factors play a role in participants' choices? Visual inspection reveals a number of interesting features. The clearest pattern observable in the data is that the proportion of observed red balls over the number of drawn balls seems to have an effect on expression choice. For example, proportion values between 0 and 1/3 (two leftmost columns in the plot) seem to associate mostly with the expression *probably not*, which is the most frequently chosen expression in all conditions except two (where is the second most frequently chosen); proportion values of exactly 1/2 seem to invariably correspond to modal choice of *possibly*; and proportion values between 3/4 and 1 seem to correspond to modal choice of *probably* (although the pattern is less clear in this case).

These informal observations lead us to formulate the rather intuitive hypothesis that higher proportions of observed red balls correspond to higher choice rates of stronger expressions. In order to substantiate this hypothesis, we analyzed our data with Bayesian ordinal regression models using the R package `brms` (Bürkner, 2017), following a similar approach to the one adopted in Chapter 2.[2] In fact, our raw data

[2] Ordinal regression models are obtained calling the `brm` function and specifying `cumulative` as family parameter. For simplicity, we kept the default prior assumptions of the `brms` package, i.e. improper flat priors over the reals. Unless stated otherwise, this holds for subsequent analyses in the chapter as well.

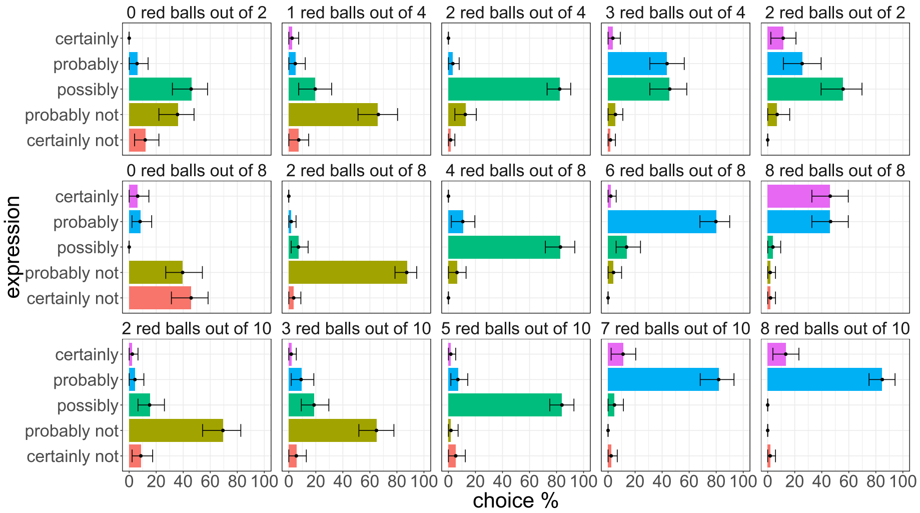


Figure 5.2: Percentages of expression choices in each observation condition, together with bootstrapped 95% confidence intervals.

were counts of expression choices in each condition, where the expressions can be naturally ordered as follows:

certainly not < *probably not* < *possibly* < *probably* < *certainly*

We fit a model regressing the dependent variable `expression` (an ordinal factor containing counts of each expression *certainly not*, *probably not*, *possibly*, *probably*, *certainly*, in this order) against the metric factor `proportion` (containing the proportions of observed red balls over accessed balls in each condition, i.e. the ratio $^{o}/_{a}$). In the formula notation adopted in `brms`:

$$\texttt{expression} \sim \texttt{proportion} \tag{5.1}$$

In order to test the hypothesis about the influence of the proportion of observed red balls on participants' choices, we checked whether the estimated posterior probability mass of the model coefficient for the main effect of `proportion` is credibly bigger than 0 (in terms of 95% highest density intervals). This turned out to be the case, specifically the mean inferred value for the coefficient resulted to be equal to 6.24 , with HDI (see footnote 6) between 5.59 and 6.88. This is a rather intuitive result, which we can interpret as a sanity check for our experimental setting.

Following our intuition that proportion is not *all* that matters, the main goal of the production task was to collect data in situations of different levels of higher-order uncertainty –obtained here by manipulating different observations of the urn. Looking at the plot from this perspective, we can observe that the same (or very close) proportion values in combination with different access values seem to give rise to different expression choices. For example, compare the choices of *probably* and *possibly* in the conditions corresponding to 3 observed red balls out of 4 and 6 out of 8 (fourth column from the left, top and middle quadrants). Despite the proportion being the same (a reasonably high 0.75 chance) it seems that only the group who observed this proportion in the lower uncertainty situation (access=8) reliably chose *probably*, whereas the group who observed the same proportion but in the higher uncertainty situation (access=4) was undecided between *probably* and *possibly*. Similar differences can be observed comparing the distributions of expression choices recorded with a proportion of 0 and access=2 or access=8, and similarly with a proportion of 1 and access=2 or access=8.

We tried to substantiate these observations with another regression model. Before diving into the details of this second model, however, a cautionary note is in order about the goal of the analysis itself. In the previous paragraphs about the proportion model we formulated a precise hypothesis about the expected effect of the explanatory variable `proportion` on the dependent variable `expression`: higher proportions of observed red balls should correspond to higher rates of stronger expressions. Consequently, the goal of the analysis has been to test this hypothesis by computing the posterior distributions over the coefficient representing the explanatory variable in a regression model. Things are different here. It is true that we expect that both access and observation might influence participants' choices of expressions (to some degree this is also apparent from visual inspection). And for what concerns observation, it seems intuitive to suppose a "linear" —so to speak— behavior: higher values should correspond to higher choice rates of stronger expressions. When it comes to access, however, such a linear behavior is not what seems to emerge from visual inspection. In general, it is more difficult to formulate *a priori* a precise hypothesis about the role of access and its interactions with observation. As a consequence, the goal of the regression model spelled out below was uniquely to test, in general, whether our data allows us to reasonably believe that both access and observation have an effect on participants' choices. A precise investigation of the role played by access, observation and their interaction is the goal of the next chapter, where we develop a theory-driven computational model of use and interpretation of uncertainty expressions.

Having clarified all this, we can finally move to our second model, regressing the dependent variable `expression` against `access` and `observation` (both metric). In formula notation:

$$\texttt{expression} \sim \texttt{access+observation} \tag{5.2}$$

Table 5.2 contains mean values and HDIs for the coefficients of the model defined in Equation 5.2. Coefficients credibly different from zero have been marked. First, we can observe that `observation` has a coefficient credibly bigger than zero. This means that, given our data, we can reasonably believe that the manipulation of the observation value has an effect on participants' choices of expressions: the higher the value, the more likely it is to choose a stronger expression. This is intuitive enough, and in line with our expectations. As for `access`, we can observe that the coefficient is credibly smaller than zero. As clarified above, this does not tell us much about the role of access in participants' choices. However, it allows us to be confident enough that our manipulation of access too had an effect.

In order to deepen our understanding of this result, we performed a number of model comparisons on the basis of the Leave-One-Out Information Criterion (LOO-

	lower	mean	upper	
`observation`	1.02	1.12	1.23	*
`access`	-0.54	-0.47	-0.40	*

Table 5.2: Mean values and HDIs for model coefficients for the main effects of `observation` and `access`. Coefficients credibly different from zero are marked with *.

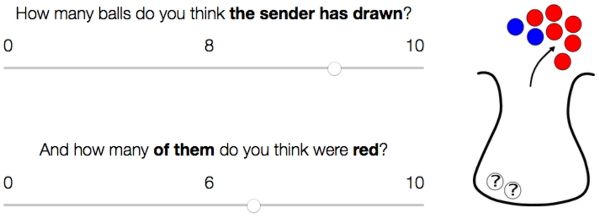
IC), which allowed us to compare different Bayesian regression models sharing the same dependent variable —in this case `expression` (Vehtari et al., 2016). A comparison between our first model using only `proportion` as predictor and the model using `access` and `observation` resulted in a preference for the latter, despite the added complexity: the LOO-IC scores are equal to 1684.05 (SE=52.01) and 1524.97 (SE=64.24), respectively —lower is better. Moreover, the factor `access` seems to be crucial in explaining our data: a simpler model regressing `expression` against `observation` alone resulted in a substantially higher score (LOO-IC=1730.52, SE=49.00). Finally, a more complex model including the interaction between `access` and `observation` as predictor resulted in essentially the same score (LOO-IC=1526.37, SE=64.12), which we take as indication that the interaction (in the strict sense assumed in the regression models) does not play a substantial role in explaining the data. From this set of comparisons we can conclude that, as far as regression models go, our best option to model our data is to include main effects of both access and observation. This is in line with our expectations.

However, it is still not clear whether the different levels of higher-order uncertainty induced by the partial observations of the urn directly played a role in participants' choices or not. It could be argued that even if higher-order uncertainty played a role in the belief formation, perhaps participants made their choices without taking the full distributions into account but only a flattened-out summary value approximating the objective chance that a randomly drawn ball ball will be red, i.e. expressing their first-order uncertainty about the event.

	`obs.+acc.`	`mode`	`ev`
LOO-IC	1524.97	1646.57	1557.16
SE	64.24	54.31	62.90

Table 5.3: LOO-IC scores and standard errors (SE) of the models with `obs.+acc.`, `mode` and `ev` as predictors —lower is better.

In an attempt to dismiss this interpretation we compared the regression model defined above (`expression`∼`obs.+acc.`) with two models trying to explain participants' expression choices on the basis of a single summary value of the empirically measured participants' beliefs. For each of the experimental condition of the production task we computed the mode and the expected value (`ev`) of the corresponding probability distribution, empirically measured as reported in Chapter 4. We fitted two ordinal Bayesian regression models explaining `expression` respectively with the metric factors `mode` and `ev`. As apparent from Table 5.3, the comparison between these



How many balls do you think **the sender has drawn**?

And how many **of them** do you think were **red**?

Figure 5.3: Input sliders in the interpretation tasks, observation trials. The picture on the right provided immediate and interactive visual feedback, dynamically displaying the current slider selection (see the midpoint of the slider).

models and the original model with access and observation in terms of LOO-IC resulted in a slight preference for the original model. More in detail, the `mode` model can be dismissed as strictly worse than the original model, because the lower boundary of the credibility interval around its LOO-IC score ($1646.57 - 54.31 = 1592,26$) is not low enough to overlap with the upper boundary of the credibility interval around the score of the original model ($1524.97 + 64.24 = 1589,21$). However, the same does not hold for the `ev` model: its LOO-IC score is higher (read: worse) than the one for the original model, but the credibility intervals of the two scores noticeably overlap.

Let us take stock. We have provided statistical evidence that our manipulation of both observation and access had an effect on participants' choices of simple uncertainty expressions. Moreover, a series of model comparisons allowed us to reasonably believe that the decision process involved in participants' choice was likely not limited to summarizing the beliefs induced by the observation of the urn and choosing based on this. We can conclude that different levels of higher-order uncertainty matter for the production of simple uncertainty expressions, as originally hypothesized. But what was the exact role of observation and access in participants' decision processes? An attempt to answer this question is provided in the form of a theory-driven computational model of pragmatic language use (and interpretation), whose details are spelled out in the next chapter. Before turning to the model, we report on design and results of the interpretation experiment.

5.2.2 Experiment 3b: interpretation

Participants. 145 self-reported English native speakers with IP addresses located in the USA were recruited on Amazon's Mechanical Turk. Workers were paid 1 USD for their participation, amounting to an average hourly wage of approximately 10 USD.

Materials and procedure. After the introductory phase, participants completed a few familiarization trials in which they played in the role of sender, making partial observations of the urn and sending a message to the receiver. In the main experimental phase participants took on the role of receiver.

The experimental conditions in Experiment 3b were the same five input expressions that participants could select in Experiment 3a, i.e. *certainly*, *probably*, *possibly*,

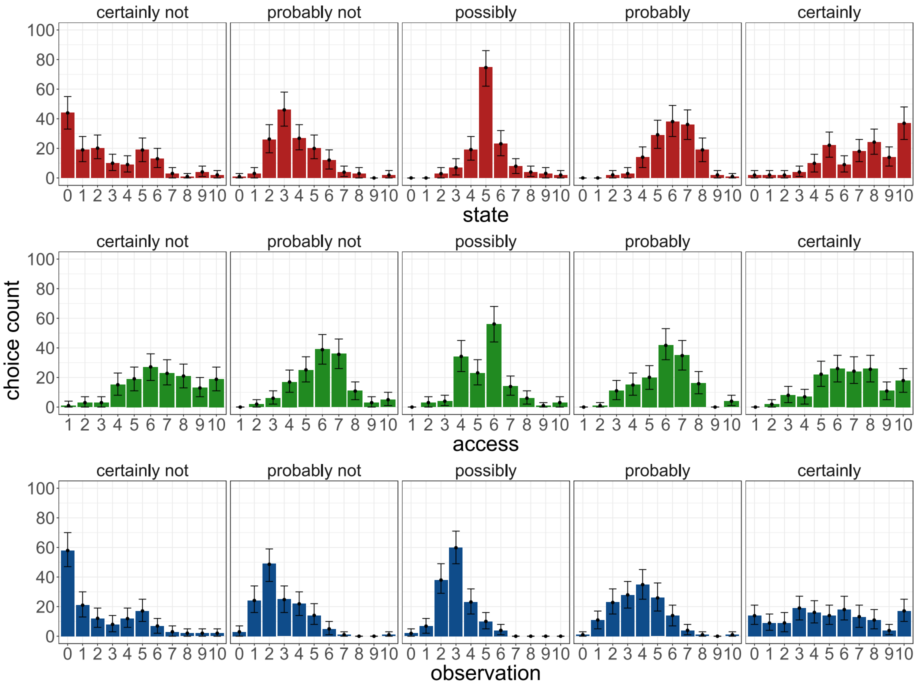


Figure 5.4: Counts of state, access and observation value choices in each expression condition, together with bootstrapped 95% confidence intervals.

probably not and *certainly not*. The main experimental phase consisted of 10 trials: for each expression, participants completed 2 different trials, in a perfectly balanced design. In each trial, we displayed the expression to the participant in the form of a message sent by the sender and we asked the participant to interpret the message. We alternately recorded participants' interpretations alongside two axes of communicative effect: half of the trials ("state" trials) recorded participants' answer to the question

> *How many red balls do you think there are in the urn?*

expressed by adjusting a discrete slider ranging from 0 to 10; half of the trials ("observation" trials) recorded participants' answer to the questions

> *How many balls do you think the sender has drawn? And how many of them do you think were red?*

expressed by adjusting two discrete sliders ranging from 0 to 10 (see Figure 5.3).

Results. We dropped the data points provided by one participant who had selected both access=0 and observation=0 for at least one expression, which is incompatible with the instructions provided at the beginning of the experiment.

	lower	mean	upper	
probably not	0.67	1.12	1.58	*
possibly	2.09	2.56	3.03	*
probably	2.96	3.40	3.86	*
certainly	4.08	4.53	5.00	*

Table 5.4: Mean values and HDIs for model coefficients for the main effects of expression on state. Coefficients credibly different from zero are marked with *.

The bar plots displayed in Figure 5.4 show counts of 144 participants' choices of state, access and observation values in each expression condition. Visual inspection reveals a number of interesting features of the data. First, we look at the choice of state values (i.e., the quantity of red balls in the urn) displayed in red (top row of the picture). The interpretations of the participants appear to be consistent with what we might expect: we observe a symmetric behavior of the pairs of basic and negated messages, with *certainly not* and *certainly* associated with the extreme values (respectively 0 and 10 red balls in the urn), *probably not* and *probably* most frequently associated with 3 and 6-7 red balls, and *possibly*, exactly in the middle, associated with 5. Similar considerations also appear to hold for participants' choices of observation values, displayed in blue (bottom row of the picture).

In order to substantiate these observations from a statistical perspective, we fit a Bayesian linear regression model (once again using brms) with the metric factor state as dependent variable and the categorical factor expression as predictor:[3]

$$\texttt{state} \sim \texttt{expression} \tag{5.3}$$

Looking at the posterior distribution on the coefficients for the main effect of expression we can confirm the results of our visual inspection, as apparent from Table 5.4: all the levels in the expression factor have coefficient credibly different from zero and, taking for example *certainly not* as reference level, the direction of the difference for each coefficient is coherent with our observations: stronger expressions are associated with higher state values. This is a rather intuitive and unsurprising result, which we can take as a sanity check for our experimental procedure.[4]

Next, we observe that the counts of access values, displayed in green (middle row), display an interestingly different pattern: the distributions associated with *certainly* and *certainly not*, instead of being symmetric, appear to be quite similar to each other and the same holds for *probably* and *probably not*. In other words, the same expressions are associated with comparable access values (regardless the presence of negation), hence ultimately with the same levels of higher-order uncertainty. Bayesian regression analysis confirms these results. We fit a linear model with the metric factor access as dependent variable and the categorical factor expression as predictor:

$$\texttt{access} \sim \texttt{expression} \tag{5.4}$$

[3] Linear regression models are obtained calling the brm function and keeping the default gaussian family parameter. We also kept the default prior assumptions of brms.

[4] Similar results hold for participants' choices of observation values as well. The details of the regression analysis are omitted for readability.

	lower	mean	upper	
probably not	-1.12	-0.72	-0.30	*
possibly	-1.65	-1.24	-0.84	*
probably	-1.17	-0.75	-0.35	*
certainly	-0.36	0.06	0.47	

Table 5.5: Mean values and HDIs for model coefficients for the main effects of `expression` on `access`. Coefficients credibly different from zero are marked with *.

The mean values and HDIs for the coefficients of the levels in `expression`, taking *certainly not* as reference levels are summarized in Table 5.5. First, we can observe that *certainly* has the only coefficient which is not credibly different from zero. This in line with our expectations and the results of visual inspection: *certainly not* and *certainly* roughly correspond to the same level of higher-order uncertainty. Similarly, *probably* and *probably not* have very similar coefficients, both credibly smaller than zero: in other words, they are equally associated with lower access values, i.e. higher uncertainty of the speaker. Finally, *possibly* has an even smaller coefficient, which is in accordance with the intuition that *possibly*, being the less informative expression of the set, is interpreted as communicating that the speaker is more uncertain.

5.3 Conclusion

In the Introduction of this dissertation we put forward the hypothesis that (at least two) distinct layers or orders of uncertainty play a role in our use and interpretation of simple uncertainty expressions. Our intuition, already found in (Moss, 2015), was that the objective chance of the event under discussion to obtain certainly matters, but it does not appear to be all that matters. The speaker's higher-order uncertainty about the objective chance seems to be crucial too, in such a way that listeners take this fact into consideration when they interpret utterances containing simple uncertainty expressions.

The results discussed in this chapter provide empirical evidence supporting our hypothesis. Take the most glaring case in the production task: participants who observed 6 red balls out of 8 (objective chance=75%) almost unanimously chose *probably* to describe the situation to the listener, whereas participants who observed 3 red balls out of 4 (objective chance=75% as well) were visibly undecided between *possibly* and *probably*. And from the point of view of the listeners, there seems to be little doubt that speakers who had chosen *certainly* or *certainly not* were taken to be the most knowledgeable (in terms of inferred access values), followed by those who had chosen *probably* or *probably not* and finally those who had chosen *possibly*.

But why does this happen? At the end of Chapter 4 we reached the conclusion that participants in our experimental settings form beliefs in an approximately rational way. Now we know that the higher-order belief distributions held by the participants likely play a role in their communicative choices. But how? What exactly is the role played by the objective chance of the event and what the role played by higher-order levels of uncertainty? In the vocabulary adopted in our setting, how do state, observation and access values contribute to the speakers' choice of expressions?

Chapter 6

Simple uncertainty expressions: model

All models are wrong, but some are useful.
(Box, 1979)

6.1 Introduction

The experimental results of the previous chapter are interesting on their own, but they leave the question open how and why we use uncertainty expressions in the way observed in the data. Our answer takes the shape of a probabilistic pragmatic model, based on the Rational Speech Act (RSA) theory (e.g. Frank & Goodman, 2012; Franke & Jäger, 2016; Goodman & Frank, 2016), whose details are spelled out in this chapter.

RSA can be described as a probabilistic formalization of Gricean pragmatics (Grice, 1975; Levinson, 2000), which incorporates insights from decision theory and game-theoretic pragmatics (Benz, Jäger, & Van Rooij, 2005; Franke, 2017). Why did we choose to model the use and interpretation of uncertainty expressions within the RSA framework? First, because the probabilistic nature of RSA makes it especially useful for investigating more nuanced and fuzzy pragmatic phenomena. This is especially true in situations where agents might not have perfect information, such as the situations in which agents use uncertainty expressions, investigated in this dissertation. Moreover, the computational nature of RSA models means that they can be implemented and used to derive precise quantitative predictions about the modeled phenomena. Therefore, RSA models are explicitly testable against empirical data, making them an ideal tool for our purposes.

More concretely, our model is a variation of on the RSA model developed by Goodman and Stuhlmüller (2013) to account for scalar implicature inferences under uncertainty. The main motivation for our choice stems from the observation that the interpretation of uncertainty expressions appears to be affected by considering alternative utterances that the speaker could have made but did not, similarly to what happens with scalar implicature inferences (e.g., Levinson, 1983; Geurts, 2010). As an illustration, consider how an utterance of (1-a) will often suggest that (1-b) is true, because otherwise the speaker would rather have uttered (1-c).

(1) a. The next ball drawn from this urn is probably red.
b. ⇝ It is not certain that the next ball drawn from this urn is red.
c. The next ball drawn from this urn is certainly red.

As we will see, RSA modeling is a convenient framework for modeling the recursive listener's reasoning about the speaker's likely choice of alternative utterances and for modeling the communicative effects of such reasoning, especially when it happens under uncertainty. This brings us to another concrete motivation for our modeling choice, namely the fact that Goodman and Stuhlmüller adopt the partial observation of the urn procedure to model different uncertainty situations and, more importantly, they propose the normative model of rational belief formation to describe agents' beliefs in such uncertainty situations (see Chapter 4).

Before going into the details of our model of simple uncertainty expressions, let us introduce the main ingredients of RSA modeling in a much simpler scenario, for the benefit of the readers who might not be familiar with them.

6.2 Rational Speech Act model

As we said, RSA probabilistic models incorporate insights from Gricean pragmatics (Grice, 1975; Levinson, 2000) and game-theoretic pragmatics (Benz et al., 2005; Franke, 2017). In a nutshell, the speaker's behavior (language use) and the listener's behavior (interpretation) are formalized as recursive Bayesian inferences. The speaker is modeled as an approximately rational pragmatic agent who chooses the best message to send to the listener given the situation. The listener is modeled as a pragmatic reasoner who infers the intended meaning by reasoning about the speaker's behavior.

What does it mean that the speaker chooses the best message given the situation? In this simplified model of communication, there are three main factors contributing to the speaker's choice: 1) the speaker's belief about the world, 2) the literal meaning of the messages, and 3) the goal of communication. Let us intuitively illustrate these concepts with a basic version of RSA, modeling pragmatic language use in simple reference games (Wittgenstein, 2010; Frank & Goodman, 2012). Starting from 1), the "world" of the game consists of exactly three objects: a blue square, a blue circle and a green square (Figure 6.1). At each round of the game, only one of these objects is chosen at random to be the target. Or, so to speak, to be the "true" state of the world. Crucially, only the speaker knows which object is the target at any given point. In this simple version of the model, the speaker has no uncertainty and her beliefs are always precise and true: the speaker knows, for example, that the object in the middle ● is the true state.

Clearly, the listener does not know this. The goal of the speaker is to communicate her beliefs to the listener, i.e. to make it so the listener is able to identify the true state with reasonably high confidence. What makes these games interesting is that in each situation the speaker can communicate using only a very limited array of expressions.

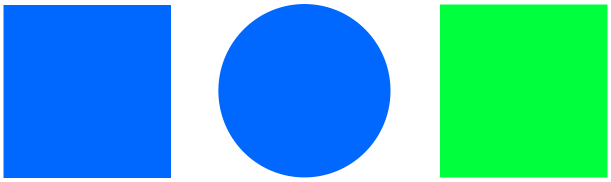


Figure 6.1: The universe of discourse in a simple reference game.

Crucially, the speaker cannot say “blue circle" in order to refer to 🔵. In this particular case, the speaker can only send one atomic message to the listener, choosing among the following possibilities:

blue, *circle*, *square*, *green*

Which message should the speaker send? This depends first and foremost on the literal meaning of the messages, which brings us to 2). Assuming the most simple extensional semantics available on the market, we can say that the literal meaning of a message corresponds to the set of objects of which the message can be truthfully predicated. For example, the meaning of *blue* is the set of blue objects (blue square and blue circle), the meaning of *circle* is the set of circular objects (blue circle), etc.:

⟦blue⟧ = {🟦, 🔵}   ⟦circle⟧ = {🔵}

⟦square⟧ = {🟦, 🟩}   ⟦green⟧ = {🟩}

Based on this, how can the speaker refer to the target object 🔵? The messages *square* and *green* are easy to disregard, as they are both semantically false in this situation. But how can the speaker choose between the remaining *blue* and *circle*? After all, they can be both truthfully predicated of 🔵. From this point of view, the two messages are equivalent. Or are they not?

Intuitively, the message *blue* is ambiguous, as it can be truthfully predicated of two different objects (🟦 and 🔵) whereas the message *circle* is not. Let us assume a naive semantic listener, who has no prior preference for any object and simply interprets each message exclusively on the basis of its literal meaning. If such a literal listener receives the message *circle* he will have no doubt about its reference: the message is true exactly of 🔵. On the other hand, if the literal listener receives the message *blue*, he will be uncertain whether the reference is 🟦 or 🔵. In a sense, he will have to flip a fair coin and decide based on the outcome.

These intuitions are formalized in RSA with the concept of informativeness of a message, defined for example in terms of entropy. We do not need to spell this out in details here. It should be clear from the previous observations that, in the given situation, both *circle* and *blue* are true of 🔵, but *circle* is more informative, as it reduces the literal listener’s uncertainty more than *blue*.

These differences in informativeness can be exploited by the speaker to achieve her goal, which brings us to 3). The goal of communication is the transfer of information from the speaker to the listener, as successfully and efficiently as possible. With respect to this goal, messages can be more or less useful: in this particular game, more informative messages are more useful than less informative messages. Let EU(message; object) refer to the expected utility of a message given the target object:

$$\text{EU}(\text{message; object}) = \text{info}(\text{message; object}) \tag{6.1}$$

Assuming that our speaker is (approximately) rational she could reason as follows:

> "If I want to refer to 🔵 and I say *blue*, there is approximately a 50% chance that a naive listener gets it right and approximately 50% chance that he gets it wrong; but if I say *circle*, the chance that the listener gets it right approaches 100%; therefore, I should expect *circle* to be much more useful, and I should likely say *circle*."

This reasoning is formalized in RSA by defining a probabilistic rule of speaker behavior, which assigns a probability to each message given each situation. Ideally, a perfectly rational speaker would always choose the best (or a best) message in terms of EU in each situation. Instead, in order to make the model more flexible, the usual practice is to define the speaker's probability distribution as a softmax function of EU, where a parameter λ regulates how "ideally rational" the speaker's behavior is (more on this below):

$$\text{speak.prob}(\text{message} \mid \text{object}) \propto e^{\lambda \cdot \text{EU}(\text{message; object})} \tag{6.2}$$

Equation 6.2 defines the behavior of a pragmatic speaker who assumes a basic level of literal interpretation. However, the goal of RSA modeling is not only to capture pragmatic language use, but also pragmatic language interpretation. The naive listener introduced above is to be thought of as a theoretical construct, necessary to ground the recursive process of reasoning, but with no ambition to model any real agent. Instead, a more realistic model of interpretation is approximated by the definition of a *pragmatic* listener. That is, an interpreter who receives a message and reasons on the basis of what he knows about the world and about the speaker's behavior, in order to infer what could have been the intended meaning of the message, i.e. what object the speaker was trying to refer to. This inference is formalized in RSA applying the Bayes' Rule to the speaker probabilities, updating the pragmatic listener's prior about the world:

$$\text{listen.prob}(\text{object} \mid \text{message}) \propto \text{speak.prob}(\text{message} \mid \text{object}) \cdot \text{prior}(\text{object}) \tag{6.3}$$

For example, suppose the listener receives the message *square*. He could reason as follows:

> "The speaker said *square*; the message could refer either to 🟦 or to 🟩; however, had the speaker wanted to refer to 🟩, she could have also said *green*; assuming the speaker is rational, she would have likely known that *green* would be a more useful message in order to refer to 🟩, because *green* is not ambiguous in this world; hence, the speaker would have likely said *green*; but she didn't; hence, I can conclude that the speaker likely didn't intend to refer to 🟩; but then she is likely referring to 🟦."

This kind of pragmatic reasoning in reference games has been empirically attested and successfully modeled within the RSA framework (Frank & Goodman, 2012; Qing & Franke, 2015). Further refinements and extensions of the RSA model summarized in this section, together with other probabilistic computational models inspired by RSA, have been developed to investigate and explain a large variety of semantic/pragmatic phenomena. For example, scalar implicatures (Goodman & Stuhlmüller, 2013); adjectival vagueness (Qing & Franke, 2014; Lassiter & Goodman, 2017; Tessler, Lopez-Brau, & Goodman, 2017) and vague quantifiers (Franke, 2014; Schöller & Franke, 2015); irony (Kao & Goodman, 2015), metaphor (Kao, Bergen, & Goodman, 2014) and hyperbole (Kao, Wu, et al., 2014); generics (Tessler & Goodman, 2019). In the next section we follow this tradition and introduce an RSA model of the pragmatic use and interpretation of simple uncertainty expressions.

6.3 A model of uncertainty expressions

As already mentioned, the model introduced in this chapter extends in a conservative way the RSA model developed by Goodman and Stuhlmüller (2013). The main difference between our model and Goodman and Stuhlmüller's resides obviously in the object of investigation: derivation of scalar implicatures for Goodman and Stuhlmüller, pragmatic use and interpretation of uncertainty expressions for us. Moreover, unlike Goodman and Stuhlmüller, we not only investigate model predictions and empirical data concerning listeners' inferences of factual information about the state of world, but also model predictions and empirical data concerning listeners' inferences about the speaker's higher-order uncertain knowledge state.

Rational speaker. Mirroring the experimental setting introduced in the previous chapter, the universe of the discourse is the set of natural numbers $S = \{0, \dots, 10\}$, where for any $s \in S$ the ratio $s/10$ is the objective chance that a ball drawn at random from the urn will be red. We want to model pragmatic communication about the objective chance, i.e. the contents of the urn, under higher-order uncertainty.

Let us begin with the behavior of the rational speaker. Recall three main ingredients of idealized communication introduced above. First of all, we need to model the speaker's belief about the world. This was already done in Chapter 4. Following the partial observation procedure, the speaker draws a number of balls from the urn, denoted with a (*access*) and counts of how many of them are red —this quantity is denoted with o (*observation*). On the basis of this observation, the speaker forms a rational belief about the contents of the urn, which is expressed as a discrete probability distribution over S. Assuming that the speaker has a prior belief distribution over S, the posterior beliefs are defined as in Chapter 4:

$$\text{rat.bel}(s|o,a) \propto \text{Hypergeometric}(o|a,s,10) \cdot \text{prior}(s) \tag{6.4}$$

An illustration of this definition was already given in Chapter 4 in Figure 4.2, repeated here as Figure 6.2. The two bar plots display the belief distributions computed with Equation 6.4 that a rational agent should have after having made the corresponding

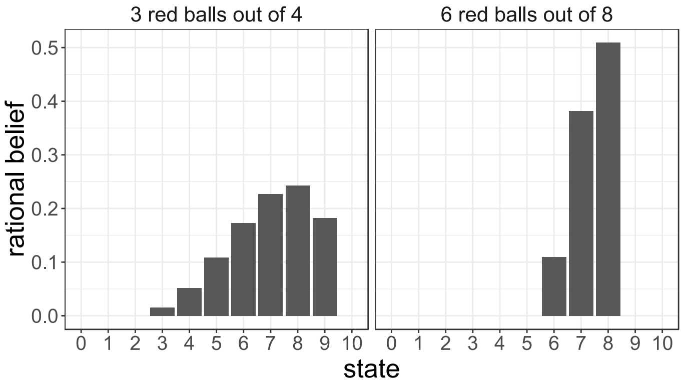


Figure 6.2: Examples of rational belief distributions given two partial observations of the urn.

partial observations of the urn (respectively 3/4 and 6/8 red balls), assuming flat priors. As already noted in Chapter 4, the distributions have the same mode equal to 8, but the right hand side distribution has lower entropy, which reflects the intuition that the agent's beliefs are more precise.

In this illustration of Equation 6.4 we have lightheartedly assumed flat prior beliefs, which is perfectly acceptable in the context of abstract modeling. However, it is important to keep in mind that our goal is to model empirical data. From this perspective, we as modelers are uncertain about participants' actual prior beliefs over the state space. What intuitions about the urn might the participants have had before starting the experimental task? Assuming flat priors could be a too simple answer to this question. A flexible yet manageable representation of our uncertainty needs to be included in the model. For convenience, the prior distribution over states is assumed to be a discrete beta-binomial distribution between 0 and 10 with shape parameters α_s and β_s, free in the model.[1] Credible values for the shape parameters will be inferred by conditioning on experimental data in Section 6.4.

The second ingredient is the literal meaning of messages. The uncertainty expressions available to the speaker are the same as in the production experiment reported in Chapter 5:

certainly not, *probably not*, *possibly*, *probably*, *certainly*

We assume that the messages can be formalized as the composition of an uncertainty expression with an optional negation (*not*) and a simple sentence expressing the focal event, in this case *The next ball will be red*. As for the meaning, we assume a simple implementation of a probabilistic threshold semantics for uncertainty expression (Swanson, 2006; Yalcin, 2007, 2010; Lassiter, 2010, 2011a; Moss, 2015). We frame our analysis within the logic for reasoning about knowledge and probabilistic beliefs by (Fagin & Halpern, 1994). In a nutshell, each possible world, or state, s in

[1] See Chapter 4, footnote 4 for the definition of beta-binomial distribution.

a model for this logic is associated with a classic valuation function V_s which assigns truth values to atomic propositional letters and with a probability measure $\mu_{i,s}$. For each agent i, the probability measure $\mu_{i,s}$ assigns a probability value to each subset of states in the model.[2] Now, given a subset of states representing the proposition A, the value $\mu_{i,s}(A)$ is the level of credence assigned to the proposition A by agent i in state s.

This machinery allows us to define the semantics of the messages in our model in a straightforward way. Let us look at *probably* in the following examples, where the a-variant is the sentence to be analyzed, the b-variant is a formal characterization of its meaning (ignoring tense for simplicity) and the c-variant a gloss of the b-variant in natural language:

(2) a. The next ball drawn from this urn will be red. (= RED)
b. $[\![\text{RED}]\!] = \{s \mid V_s(\text{RED}) = 1\}$
c. The set of all worlds in which the next draw is red.

(3) a. It is probable that the next ball drawn from this urn will be red.
b. $[\![\text{probably}_i\ (\text{RED})]\!] = \left\{s \mid \mu_{i,s}([\![\text{RED}]\!]) > \theta_{\text{probably}}\right\}$
c. The set of worlds in which agent i assigns a level of credence higher than the semantic threshold θ_{probably} to the proposition that the next draw will be red.

Generalizing a bit, we can say that if X is an uncertainty expression and p a simple sentence (such as RED or not RED), the meaning of $X(p)$ is the set of states where the probability of p being the case is higher than a threshold θ_X associated with expression X, i.e. $\mu_{i,s}(p) > \theta_X$. In our setting p is always instantiated with *The next ball will (not) be red*, i.e. (not)RED. Moreover, given our urn scenario, the only possible states are those such that $\mu_{i,w}(\text{RED}) \in \{{}^{s}\!/_{10} \mid s \in S = \{0, \dots, 10\}\}$, i.e. the states corresponding to each possible value of objective chance that a randomly drawn ball will be red. This allows us to simplify a bit and formulate the the following semantic definitions directly in terms of states:

$$[\![\text{certainly}(p)]\!] = \{s \in S \mid {}^{s}\!/_{10} > \theta_{certainly}\}$$
$$[\![\text{probably}(p)]\!] = \{s \in S \mid {}^{s}\!/_{10} > \theta_{probably}\}$$
$$[\![\text{possibly}(p)]\!] = \{s \in S \mid {}^{s}\!/_{10} > \theta_{possibly}\}$$

For negated sentences we can easily derive the following definitions:

$$[\![\text{certainly not}(p)]\!] = \{s \in S \mid {}^{s}\!/_{10} < 1 - \theta_{certainly}\}$$
$$[\![\text{probably not}(p)]\!] = \{s \in S \mid {}^{s}\!/_{10} < 1 - \theta_{probably}\}$$

Notice that we are not fixing any value for the semantic thresholds *a priori*. The thresholds are free parameters in the model, whose credible values will be inferred by conditioning on experimental data in Section 6.4 (cf. Schöller & Franke, 2017).

[2] Strictly speaking, each world s and agent i are associated with a probability space $\langle \Omega_{i,s}, \chi_{i,s}, \mu_{i,s} \rangle$ such that $\Omega_{i,s} \subseteq W$ is a subset of possible worlds, $\chi_{i,s}$ is the usual σ-algebra of measurable subsets of $\Omega_{i,s}$ and $\mu_{i,s}$ is a probability measure defined on the elements of $\chi_{i,s}$. If a set of possible worlds $A \subseteq W$ is not in $\chi_{i,s}$, we resort to $\mu^*_{i,s}(A)$, where μ^* is the inner measure of $\mu_{i,s}$ as the probability measure that describes agent i's belief in A at world s.

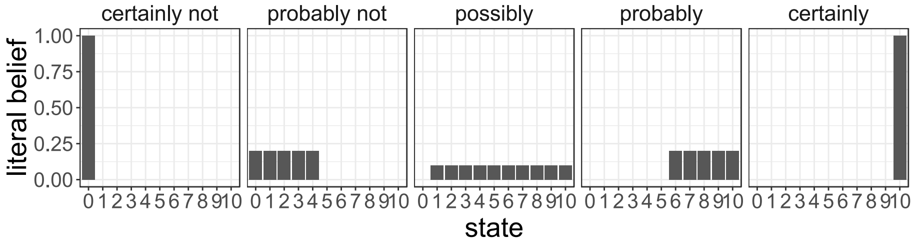


Figure 6.3: Examples of literal belief distributions over states as a function of the received message, assuming flat priors and setting $\theta_{certainly} = 0.99$, $\theta_{probably} = 0.5$, $\theta_{possibly} = 0.01$.

The literal meaning of the messages is assumed to be the most basic, purely semantic, level of interpretation, which grounds the recursive process of pragmatic language use and interpretation, as usual in RSA and related models. This level is modeled as an idealized naive listener who receives a message *m* and interprets it uniquely on the basis of its literal meaning, prior to any pragmatic enrichment or inference. Formally, the literal listener simply updates her prior belief over *S* on the assumption that the received message *m* is literally true:

$$\text{lit.bel}(s|m) \propto \delta_{s \in [\![m]\!]} \cdot \text{prior}(s) \tag{6.5}$$

The δ function returns 1 if the condition $s \in [\![m]\!]$ is met (i.e., the message *m* is true in *s*), and 0 otherwise. As an illustration, Figure 6.3 displays the belief distributions of the naive listener as a function of the received message, having fixed reasonable values for the threshold parameters and assuming flat priors. This captures basic intuitions about the meaning of uncertainty expressions: the most informative expressions are *certainly* and symmetrically *certainly not*, the least informative is *possibly*, with *probably* and *probably not* exhibiting an intermediate behavior.

The third ingredient is the goal of communication. Our assumption is that the speaker chooses her messages with the goal of maximizing the information transferred to the listener. Clearly, not all the information matters all the time: the goal of the speaker should be to maximize the *relevant* information. One way in which the concept of relevance can be incorporated in RSA modeling is by fixing the QUD in the conversation (Roberts, 1996, 2012; Kao, Wu, et al., 2014; Lassiter & Goodman, 2017). In our particular setting, we assume that the QUD for evaluation of a message $X(p)$ is "What is the probability of p?". Therefore, from the speaker's perspective, transferring relevant information to the listener amounts to sending messages which bring the listener's belief about the contents of the urn as close as possible to the speaker's own belief about the contents of the urn. In other words, the speaker's goal is to minimize the distance between the two distributions expressing her own beliefs and the listener's beliefs. Relative to this goal messages can be more or less useful, depending on the situation. To formalize all this, the expected utility (EU) of a message *m* given an observation of the urn $\langle o, a \rangle$ is computed as the negative Hellinger distance (HD) between the speaker's belief distribution given $\langle o, a \rangle$ and the literal listener's belief distribution given m:[3]

$$\mathrm{EU}(m;o,a) = -\mathrm{HD}[\mathrm{rat.bel}(s|o,a),\ \mathrm{lit.bel}(s|m)] \tag{6.6}$$

	3 red balls out of 4	*6 red balls out of 8*
certainly not	-1.00	-1.00
probably not	-0.91	-1.00
possibly	-0.46	-0.69
probably	-0.44	-0.51
certainly	-1.00	-1.00

Table 6.1: Examples of EU of each message given two partial observations of the urn, rounded to two decimal places.

As an illustration, Table 6.1 contains EU of messages computed for two partial observations of the urn, namely $^3\!/_4$ and $^6\!/_8$. The values in the table are negative HD between each rational belief distribution displayed in Figure 6.2 and each literal belief distribution displayed in Figure 6.3. The most interesting rows in the table are the ones corresponding to *possibly* and *probably*. We can observe that in both uncertainty situations these messages have comparatively high EU. In more intuitive terms, *possibly* and *probably* are more useful in these situations, compared to the other three messages. However, while the EUs of *possibly* and *probably* are very close to each other in the $^3\!/_4$ situation, the EU of *possibly* drops in favor of *probably* in the $^6\!/_8$ situation. In order to see why this happens, we can look again at the distributions representing an agent's rational beliefs in the $^3\!/_4$ and $^6\!/_8$ conditions displayed in Figure 6.2 above. If we compare the distribution in the $^6\!/_8$ condition with the literal beliefs induced by *possibly* and *probably* in a naive listener (Figure 6.3) we can see that both are compatible with the speaker's rational belief in the $^6\!/_8$ situation, but the belief distribution induced by *probably* is visibly more similar to the speaker's rational beliefs, hence the message is predicted to be more useful. In other words, it is hard to decide whether *possibly* or *probably* is the more useful message in the $^3\!/_4$ situation, whereas a clearer "winner" seems to emerge in the $^6\!/_8$ situation.

The speaker's behavior depends on the EU of messages. Ideally, a perfectly rational speaker would always choose the best (or *a* best) message in each situation. As usual in RSA and related models, this ideal behavior can be made more flexible by defining the speaker's choice probabilities as a softmax function of EU, as shown in Equation 6.7. The parameter λ, sometimes referred to as the "rationality parameter", regulates how close the speaker's behavior is to the ideal behavior: as λ grows, choice probabilities approach EU-maximization behavior. λ is a free parameter in the model and its credible value will be inferred by conditioning on experimental data in Section 6.4.

$$\mathrm{speak.prob}(m|o,a) \propto e^{\lambda \cdot \mathrm{EU}(m;o,a)} \tag{6.7}$$

[3] Goodman and Stuhlmüller (2013) use Kullback-Leibler (KL) divergence as a measure of distance between speaker and listener beliefs. We prefer Hellinger distance in our setting because utilities computed in terms of KL lead to speakers who never send literally false messages, whereas HD allows for pragmatically "true enough" messages to be sent. The Hellinger distance between two discrete distributions P and Q is defined as $\mathrm{HD}(P,Q) = {}^1\!/\!\sqrt{2} \cdot \sqrt{\sum_i \left(\sqrt{P_i} - \sqrt{Q_i}\right)^2}$.

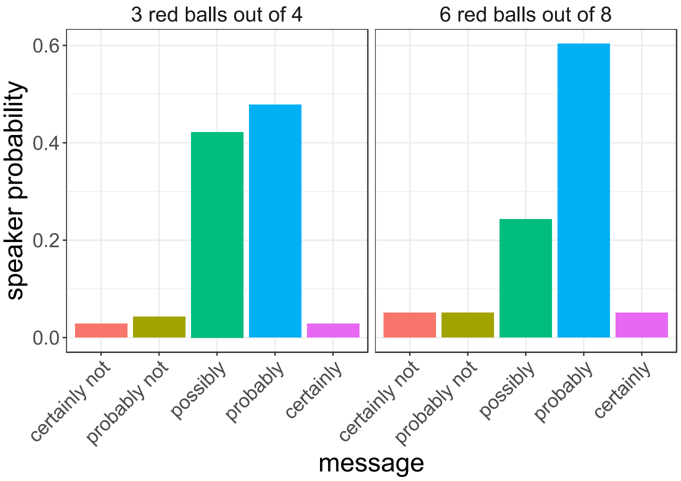


Figure 6.4: Examples of speaker's distributions over messages given two partial observations of the urn.

As an illustration, Figure 6.4 displays speaker's probabilities of sending each message given the two situations of our running example (3⁄4 and 6⁄8), setting $\lambda = 5$. Consistently with our observations about the EU of messages, we can observe that there are essentially two competing messages in both situations, i.e. *possibly* and *probably* and that the choice probabilities of these two messages are close to each other in situation 3/4, whereas they noticeably diverge in the 6/8 situation. This observation is very promising, as it showcases that this version of the model, despite not being optimized on the data yet, can already capture, at least qualitatively, one of the most interesting features of our experimental data. That is, the fact that a crucial role in participants' use of uncertainty expressions is played not only by the observed ratio of red balls but also by the different levels of higher-order uncertainty associated with that observation. Figure 6.4 displays precisely an example of this behavior: the 6/8 observation makes the speaker noticeably more inclined to say that a randomly drawn ball will *probably* be red rather than just *possibly*, whereas the 3/4 observation does not, despite corresponding to the same observed proportion (cf. Chapter 5, Figure 5.2, fourth column).

Rational listener. We round off the exposition of our model by introducing the definition of the pragmatic *listener*, who receives the message sent by the speaker and interprets it by reasoning about how a pragmatic speaker could have used the message, on the assumption that she has formed a rational belief abut the contents of the urn for a particular observation. More formally, we compute the interpretation probabilities of the pragmatic listener as joint Bayesian inference over state (s), access (a) and observation (o) values, given the speaker probability distribution, the hypergeometric model of the urn and the priors:

$$\begin{aligned}\text{listen.prob}(s,o,a|m) \propto\ & \text{speak.prob}(m|o,a)\cdot\text{Hypergeometric}(o|a,s,10)\cdot \\ & \text{prior}(a)\cdot\text{prior}(s)\end{aligned} \tag{6.8}$$

$$\text{listen.prob}(s|m) = \sum_{\langle o,a\rangle} \text{listen.prob}(s,o,a|m) \tag{6.9}$$

$$\text{listen.prob}(o,a|m) = \sum_{s} \text{listen.prob}(s,o,a|m) \tag{6.10}$$

Equations 6.9 and 6.10 are obtained by marginalizing the distribution defined in 6.8 respectively over pairs $\langle o,a\rangle$ and over states s. Similarly to what we said about the speaker's prior over state values in the previous section, the listener's prior over access values prior(a) from Equation (6.8) is subject to modeler's uncertainty. We assume the same structure as for prior(s), namely a beta-binomial prior with free parameters α_a and β_a (once again, these are free in the model). The speaker's distribution defined in Equations 6.7 and the listener's distributions defined in 6.9 and 6.10 allow us to generate the model predictions which we compare to our experimental data.

6.4 Model evaluation and criticism

In the previous section we referred multiple times to the free parameters in our model. Formally, these are nothing but variables in the mathematical definitions of the model, whose values are not fixed *a priori*. We can think of these parameters as "knobs" that can be adjusted to modify the predictions of the model to various degrees, in particular in order to make the model better fit the empirical data (more on this below). Clearly, more knobs lead to more flexible models, but a balance is needed between flexibility and informativeness: the extreme case of a model with a free parameter for each data point would eventually perfectly fit any empirical data, but would be of little (or none at all) theoretical interest. For this reason, we have tried to make it clear in the previous section that the adoption of free parameters in our model corresponds for the most part to modeling choices about which we, as modelers, are uncertain: what are the intuitions of the agents prior to an observation of the urn?, what are the semantic thresholds regulating the meaning of simple expressions?, etc. Instead of trying to answer these questions *a priori*, we let the values vary within a reasonable range.

Different combinations of answers to these questions lead to slightly different versions of the model, which make slightly different predictions. But then how can we evaluate our model? Which one is *the* model whose predictions we should compare to empirical data? The Bayesian answer to these questions is that *all* of them are. The model comes with built-in uncertainty, reflecting the modelers' uncertainty that we have talked about. Each of the slightly different versions of the model is legitimate. Crucially, though, each version yields slightly different predictions which will be more or less close to the empirical data. In other words, each version will be more or less *credible* given the empirical data. This is where Bayesian inference of parameter values comes into play.

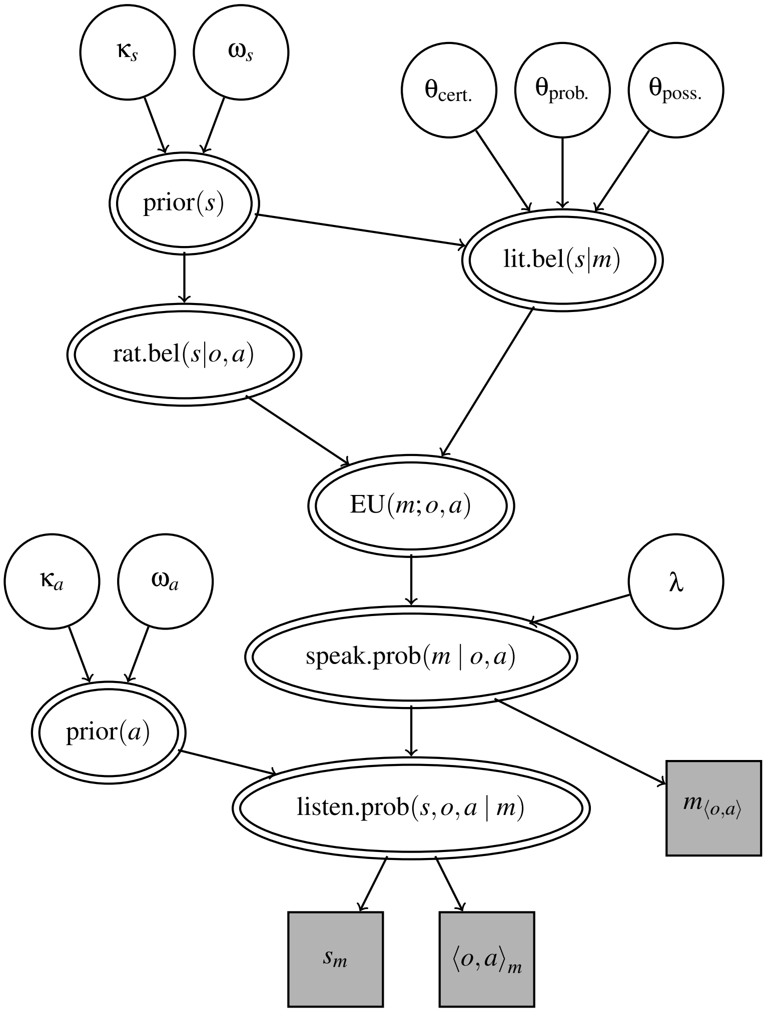


Figure 6.5: The data-generating model. White nodes represent latent variables, shaded nodes represent observed variables. Single-bordered nodes represent stochastic dependence, double-bordered nodes represent deterministic dependence.

Bayesian inference. Our model can be schematically seen as consisting of two parts, i.e. a likelihood function $P(D \mid \vec{\theta})$ defining the conditional probability of observing data D given the vector of parameter values $\vec{\theta}$; and a prior distribution over these parameters $P(\vec{\theta})$.[4] In order to determine how likely (or credible) each combination of parameter values is given a particular set of observed data, we compute the posterior probability $P(\vec{\theta} \mid D)$ with Bayes' rule (Kruschke, 2014; Lee & Wagenmakers, 2014).[5]

$$\underbrace{P(\vec{\theta} \mid D)}_{\text{posterior}} \propto \underbrace{P(\vec{\theta})}_{\text{prior}} \cdot \underbrace{P(D \mid \vec{\theta})}_{\text{likelihood}} \tag{6.11}$$

We used the data collected in Experiment 3 (Chapter 5) to infer credible values for the free parameters of the model. These are the semantic thresholds $\theta_{\text{certainly}}$, θ_{probably} and θ_{possibly}, the shape parameters α_s, β_s and α_a, β_a of the beta-binomial model of participants' prior over state and access values respectively, and the rationality parameter λ. This complex joint posterior distribution over parameter values was estimated via MCMC sampling (see footnote 5), performed by implementing the model in the probabilistic programming language JAGS and conditioning on empirical data (Kruschke, 2014; Plummer, 2003).[6]

More in detail, the JAGS model assumes that the empirically observed counts of expression choices, of state, access and observation values in each experimental condition are samples from multinomial distributions with weights equal to the probabilities predicted by the RSA model (i.e., the functions *speak.prob* and *listen.prob*) in the corresponding condition. This corresponds to the likelihood part in Equation 6.11. As for the prior distributions over parameter values, we remained relatively uncommitted, and we assumed flat distributions with support $[0;1]$ for the semantic thresholds and $[0;20]$ for λ; the prior distributions over state values and access values were both defined for convenience as beta-binomial distributions, parametrized in terms of mode ω concentration κ (Kruschke, 2014):

$$\begin{gathered}
\theta_{certainly} \sim \mathcal{U}(0,1) \quad \theta_{probably} \sim \mathcal{U}(0,1) \quad \theta_{possibly} \sim \mathcal{U}(0,1) \\
\lambda \sim \mathcal{U}(0,20) \\
\kappa_{s,a} \sim \text{Gamma}(0.01, 0.01) \quad \omega_{s,a} \sim \mathcal{U}(0,1) \\
\alpha_{s,a} = \omega_{s,a} * (\kappa_{s,a} - 2) + 1 \qquad \beta_{s,a} = (1 - \omega_{s,a}) * (\kappa_{s,a} - 2) + 1 \\
\text{prior}(s) = \text{Betabinom}(s|\alpha_s, \beta_s, 10) \qquad \text{prior}(a) = \text{Betabinom}(a|\alpha_a, \beta_a, 10)
\end{gathered}$$

Figure 6.5 displays the full data-generating model as a probabilistic graphical model (Lee & Wagenmakers, 2014). The square shaded nodes at the bottom represent observed variables, i.e. the empirically observed counts of message choices, and of state, access, observation values. These are connected to the nodes representing speaker's

[4] The use of $\vec{\theta}$ here should not be confused with the semantic thresholds $\theta_{\text{certainly}}$, θ_{probably} and θ_{possibly} regulating the literal meaning of messages. These are particular free parameters in the model. The notation $\vec{\theta}$ refers to a vector of values for every free parameter of the model, including the semantic thresholds but also α_s, β_s, α_a, β_a and λ.

[5] The proportionality sign between the two sides of the equation means that they are to be equal up to the normalizing constant $\int P(\overline{\theta'}) \cdot P(D \mid \overline{\theta'}) d\overline{\theta'}$.

[6] The code is publicly available at `https://github.com/mic-he/ProbExp-PhD/tree/master/chapter5-6`.

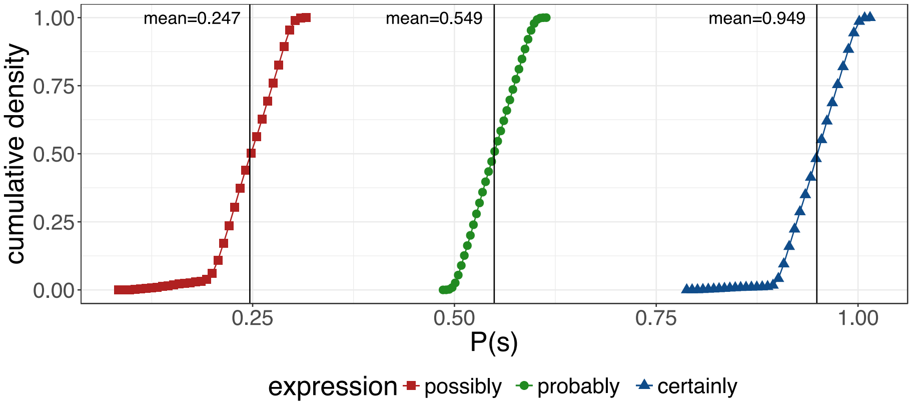


Figure 6.6: Posterior cumulative density distributions for the semantic threshold parameters.

and listener's probabilities (the linking function being the sampling from multinomial distributions mentioned above). In turn, these probabilities are connected to the nodes representing the remaining equations defining our model, up to the topmost nodes representing the free parameters. The data-generating procedure goes from top to bottom, whereas the Bayesian inference goes from bottom to top.

The inference was performed by collecting two MCMC chains of 2500 samples from the posterior distributions after an initial burn-in period of 2500 samples. We checked convergence via $\hat{R}$ (see footnote 6). Each sample is a vector containing one inferred value for each free parameter. Table 6.2 summarizes the results for the semantic threshold parameters ($\theta_{certainly}$, $\theta_{probably}$, $\theta_{possibly}$) in terms of mean inferred values and 95% HDIs (see footnote 6). Given the model and empirical data, these are the likely (or credible) values for the free parameters.[7]

	$\theta_{possibly}$	$\theta_{probably}$	$\theta_{certainly}$
lower	0.200	0.500	0.904
mean	0.247	0.549	0.949
upper	0.299	0.594	1.000

Table 6.2: Mean inferred values and HDIs of the semantic threshold parameters free in the model, given experimental data collected in Experiment 1.

Figure 6.6 displays posterior cumulative density distributions for the threshold parameters. These can be read as showing our (as modelers) posterior beliefs about semantic truth and falsity of simple uncertainty expressions, given model and data. These

[7] The results for the remaining free parameters are summarized in the following table:

	λ	α_s	β_s	α_a	β_a
lower	4.583	2.839	2.651	7.329	4.840
mean	4.873	3.251	3.050	10.601	6.950
upper	5.174	3.691	3.459	14.603	9.557

results showcase a nice feature of our model. Recall our discussion about the free parameters. We said that they build our uncertainty about some modeling choices into the model itself. We did not *assume* the values for the semantic thresholds from the beginning, limiting ourselves to fixing a reasonable range. Nonetheless, the model was able to *infer* plausible and intuitive values given the data, with a reasonably high degree of credibility.

In a sense, performing the described Bayesian inference taught us something new: starting from a situation of agnosticism about the literal meaning of *certainly*, *probably* and *possibly*, we can now reasonably conclude that participants' in Experiment 1 used and interpreted these expressions accordingly to the truth conditions displayed in Figure 6.6. This, of course, holds only if our pragmatic model of language use and interpretation is a good model. But is it?

Posterior predictive distributions. To assess model quality, we want to compare empirical data and model predictions. As already mentioned above, our Bayesian setting does not give us *one* set of predictions to directly compare with data, but rather a collection of 2500 samples of predictions, one for each possible vector of inferred parameter values. More in detail, model predictions have the form of samples of hypothetical repeat-data D_{rep} taken from the posterior predictive distribution:

$$P(D_{\text{rep}} \mid D_{\text{obs}}) = \int P(\vec{\theta} \mid D_{\text{obs}}) P(D_{\text{rep}} \mid \vec{\theta}) \mathrm{d}\vec{\theta}$$

A posterior predictive sample D_{rep} is obtained by taking a sample $\vec{\theta} \sim P(\vec{\theta} \mid D_{\text{obs}})$ from the posterior distribution over model parameters given the observed data D_{obs}, and then sampling a likely hypothetical alternative data point $D_{\text{rep}} \sim P(D_{\text{rep}} \mid \vec{\theta})$ from the model's likelihood function for parameters $\vec{\theta}$ (Gelman, Carlin, Stern, & Rubin, 2014; Kruschke, 2014).

For each of our 2500 samples from the posterior over $\vec{\theta}$, we generated one D_{rep} for each of the three relevant observed variables: speaker choices of expression (Equation 6.7), listener choices of state (Equation 6.9) and listener choices of access - observation pairs (Equation 6.10). In order to get an overall evaluation of the model we correlated each set of predictions with the corresponding set of experimental data, collecting the results in vectors of Pearson's correlation score. Table 6.3 summarizes the results in terms of mean correlation scores and HDIs. We observe that all the four means and HDIs are reassuringly high, suggesting that the model was overall able to capture regularities in the data.

	expression	*state*	*access*	*observation*
lower	0.824	0.666	0.736	0.766
mean	0.861	0.741	0.798	0.819
upper	0.902	0.824	0.858	0.868

Table 6.3: Mean Pearson's correlation scores and HDIs between model posterior predictive distributions and exerimental data collected in Experiment 1.

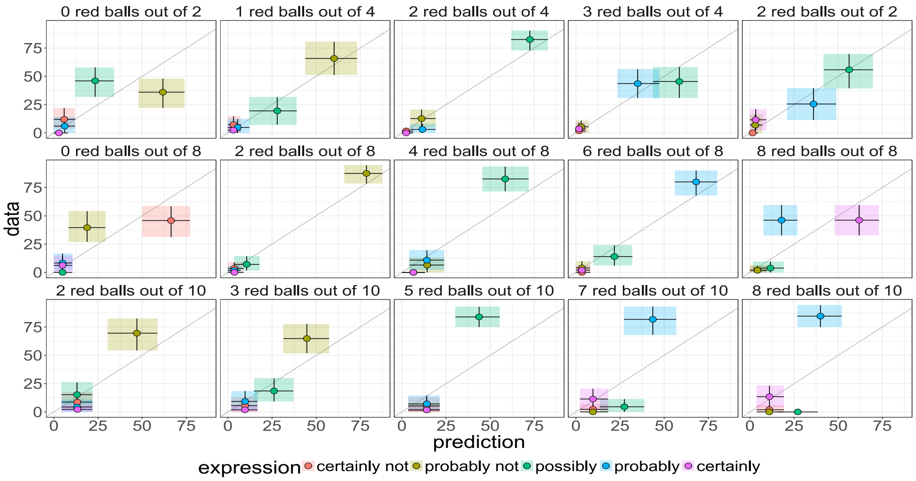


Figure 6.7: Percentages of expression choices in each observation condition compared to posterior predictive distributions. Rectangular “confidence areas" are bootstrapped 95% confidence intervals of the data against Bayesian 95% HDIs of the posterior predictive.

Generating repeat-data allows us to supplement correlation scores with a more detailed comparison between model predictions and experimental data via posterior predictive checks (PPCs). These are displayed in Figures 6.7 and 6.8. In particular we look at discrepancies between hypothetical and actual data, i.e., points in the plots where the “confidence areas”, obtained by plotting bootstrapped 95% confidence intervals of the data against Bayesian 95% HDIs of the posterior predictive, do not overlap with the diagonal. In these cases the model fails to predict those data points: the observed data is still unexpected or surprising, so to speak, in the light of the model trained on the data. Now, it is to be expected that the model will “fail” a PPC for some conditions due to multiple comparisons alone. What matters most is whether there are theoretically insightful patterns of systematic failure that might point to substantial conceptual shortcomings of the model.

Looking first at the production data and predictions (Figure 6.7), we can observe that the patterns displayed by the data seem to be captured quite well by the model. The most frequently chosen expression in each observation condition is always correctly predicted by the model (with the exception of one case: in the $^0/_2$ condition the model underpredicts *possibly* and overpredicts *probably not*). In the majority of conditions the model correctly predicts the second most frequently chosen expression too. Looking at the most glaring discrepancies in the plot we can observe that the model tends to underpredict *possibly* when the proportion is equal to $^1/_2$ (middle column, especially third and second row). Moreover, the model underpredicts *probably not* in favor of *possibly* with low proportions and no higher-order uncertainty (i.e., $o = 2$ or $o = 3$ and $a = 10$, bottom left corner); and symmetrically *probably* is underpredicted in favor of *possibly* with high proportions and no higher-order uncertainty (i.e., $o = 7$ or $o = 8$ and

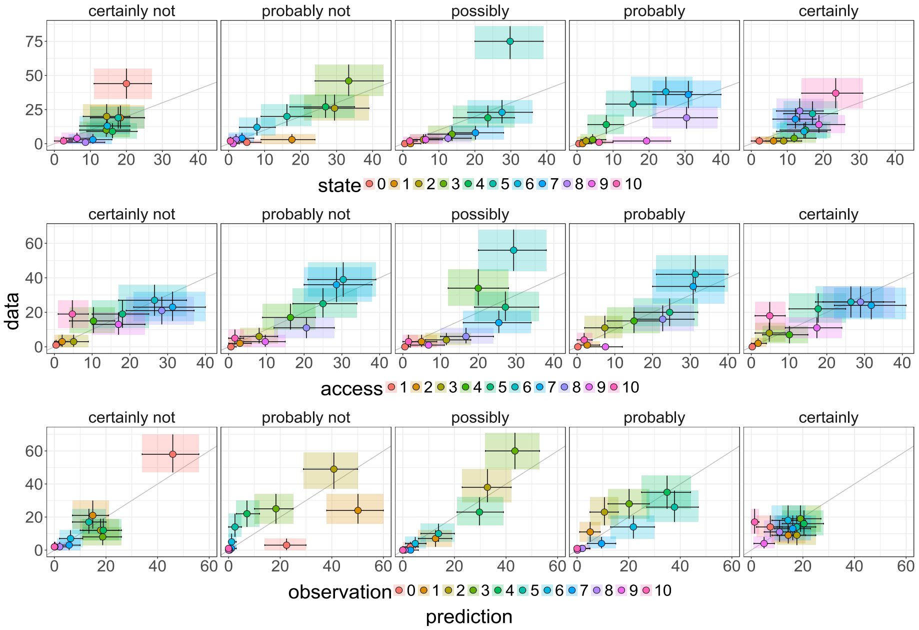


Figure 6.8: Counts of state, access and observation values choices in each expression condition compared to posterior predictive distributions. Rectangular "confidence areas" are bootstrapped 95% confidence intervals of the data against Bayesian 95% HDIs of the posterior predictive.

$a = 10$, bottom right corner). In general, these observations seem to point to a model which is a little more conservative or cautious, so to speak, than the participants.

Turning to the interpretation data displayed in Figure 6.8, PPCs show that the patterns displayed in the data are captured quite well by the model, with the exception of a few discrepancies, generally where the model seems once again to be more cautious than the participants.

In general, our comparison of the predictions of the model to the observed data in each experimental condition does not reveal any obvious systematic failure of the model.

6.5 Conclusion

The experimental data collected and analyzed in Chapter 5 suggested that subtle manipulations of higher-order uncertainty affect the use of simple uncertainty expressions and that listeners appear to draw systematic inferences about a speaker's higher-order uncertain belief state when interpreting simple uncertainty expressions.

In this chapter we formulated a probabilistic model of goal-oriented pragmatic communication within the RSA framework, revolving around the assumptions that interlocutors hold rational higher-order beliefs about the urn scenario and they desire to communicate these complex beliefs.

Conditioning on the empirical data, the model recovered values of latent semantic threshold parameters for simple uncertainty expressions, which are intuitive and in line with the relevant literature. Moreover, Bayesian model criticism revealed some mismatch between posterior model predictions and observed data (*all models are wrong*...), but no obvious systematic failure to capture particular patterns. We conclude that participants' communicative behavior in this uncertainty scenario is in line with a model of rational belief formation and goal-oriented pragmatic communication (*...but some models are useful*).

Chapter 7

Complex uncertainty expressions: data & model

We are not interested in falsifying our model for its own sake,
among other things, having built it ourselves,
we know all the shortcuts taken in doing so,
and can already be morally certain it is false.
(Gelman & Shalizi, 2013)

7.1 Introduction

Chapters 5 and 6 reported the encouraging results of our data-driven modeling and investigation of *simple* uncertainty expressions under higher-order uncertainty. As mentioned in the Introduction, however, *complex* uncertainty expressions are attested in English as well, although they are under-investigated in the literature.

In the last section of the Introduction we briefly discussed an observation inspired by an example proposed by Moss (2015), with the merit of showing that we tend to have clear pre-theoretic intuitions about the use and interpretation of complex uncertainty expressions. Let us go back to that example, this time in a form which is closer to the original given by Moss (2015). Imagine a person called Liem, very fond of green shirts. Liem's dad Eric has observed Liem wearing green on 500 of 800 consecutive days. Liem's friend Madeleine made a similar observation: she observed Liem wearing green 5 times out of 8 consecutive days. The proportion of green observations is exactly the same: 62.5%. However, as Moss (2015) argues convincingly, Eric is in the position to assert (1), whereas Madeleine should limit herself to (2):

(1) Liem is certainly likely to be wearing green.

(2) It might be likely that Liem is wearing green.

The simple expression embedded in both (1) and (2) is *likely*. Following our basic analysis of simple uncertainty expressions we can say that the sentence *It is likely that*

Liem is wearing green is literally true in those states where the objective chance of Liem wearing green is higher, say, than 50%. Now, given their respective observations, the best guess of Eric and Madeleine about this objective chance is the same, i.e. 62.5%. But then, why are not Eric and Madeleine intuitively able to truthfully assert the same set of sentences? This is the puzzle raised by Moss's example.

As we mentioned in the Introduction, our take on the matter is that, analogously to what happens with simple uncertainty expressions, complex ones are sensitive not only to the objective chance of the event under discussion, but also to the speaker's higher-order level of uncertainty. Because of the bigger sample size, Eric will be less uncertain about his guess about the objective chance than Madeleine can be about her own. This translates into the following intuitions about the truth-conditions of (1) and (2). *Relatively to Eric's belief*, and this is crucial, (1) is intuitively true. In fact, his belief state is not compatible with many situations where the objective chance is much different from 62.5%: Eric's uncertainty is so low that he can say with certainty that his guess allows him to predict that Liem will likely be wearing green. On the other hand, *relatively to Madeleine's belief* this does not hold. Madeleine shares the same guess about the objective chance but she is much more uncertain about it: her belief state is compatible with a number of situations where the objective chance is different from 62.5%, hence relatively to Madeleine's belief state only (2) is intuitively true.

This chapter is entirely dedicated to extending to complex uncertainty expressions the data-driven modeling approach adopted in Chapters 5 and 6. First, on the basis of our discussion of Moss's example, we develop a conservative extension of the simple model presented in 6, incorporating a uniform compositional semantics for both simple and complex uncertainty expressions. Next, we report on two experiments, analogous to the ones presented in Chapter 5, which we ran to collect empirical data on the use and interpretation of complex expressions in situations of higher-order uncertainty. Unlike Chapter 5 we do not approach empirical data with a set of precise hypotheses or intuitions in mind which we want to test. That is, we do not directly analyze our data with regression models or similar. Instead, we think of our pragmatic model as the theory —so to speak— in need of empirical validation (or falsification). Keeping this in mind, we use empirical data to train our model and subsequently evaluate it in terms of correlation scores and Bayesian PPCs.

7.2 Model

Compositional semantics. The basic setup of the model presented here (we call it "complex model") is the same as the one presented in Chapter 6 ("simple model"). The state space is the set of natural numbers $S = \{0, \ldots, 10\}$, where each $s \in S$ is a possible quantity of red balls in the urn; the speaker can access 1 to 10 balls, any number of which can be red. On the basis of her observation, the speaker forms a rational uncertain belief about the contents of the urn. As in the simple model, the speaker's beliefs are one of the main ingredients informing her choice of which message to send to the listener.

This brings us to the most obvious difference between the two models, i.e. the set of expressions available to the speaker. In this case, the speaker can choose from a set containing 3 simple expressions (*likely*, *possible*, *unlikely*) together 9 complex expressions obtained by combining the simple ones with 3 modifiers (*certainly*, *probably*, *might be*). The speaker sends messages of the following form:

It (is) [...] that the next ball will be red

where the gap is to be filled with one of the expressions in Table 7.1.

likely	*possible*	*unlikely*
certainly likely	*certainly possible*	*certainly unlikely*
probably likely	*probably possible*	*probably unlikely*
might be likely	*might be possible*	*might be unlikely*

Table 7.1: Complex uncertainty expressions.

The choice of these particular complex messages was dictated by our desire to cover a reasonably wide range of possibilities in a balanced way: the three simple expressions are on a scale from *likely* to *unlikely* and each of them appears nested under each of the modifiers. These can also be placed on a scale from *might* to *certainly*. According to our intuitions, the resulting 9 complex messages vary with respect to their naturalness, but they are all grammatical. An informal search on the Hansard corpus, containing speeches given in the British parliament from 1803-2005,[1] essentially confirmed our intuitions. For example, search results range from $\sim$ 20 occurrences of *might be unlikely* and *probably possible* to $\sim$ 200 of *certainly possible* and *might be likely*, up to $\sim$ 6700 of *might be possible*.[2]

In order to specify the literal meaning of simple and complex messages, we look back to the examples given in Chapter 6, repeated here as (3) and (4).[3]

(3) a. The next ball drawn from this urn will be red. (= RED)
b. $[\![\text{RED}]\!] = \{s \mid V_s(\text{RED}) = 1\}$
c. The set of all states in which the next draw is red.

(4) a. It is probable that the next ball drawn from this urn will be red.
b. $[\![\text{probably}_i\ (\text{RED})]\!] = \left\{s \mid \mu_{i,s}([\![\text{RED}]\!]) > \theta_{\text{probably}}\right\}$
c. The set of states in which agent i assigns a level of credence higher than the semantic threshold θ_{probably} to the proposition that the next draw will be red.

[1] https://www.hansard-corpus.org

[2] We should notice, however, that we cannot guarantee that all the tokens found in the corpus were used in the way we have in mind, i.e. as uncertainty expression. For example *possible* has a prominent ability-reading as well beside the epistemic one. More in general, Lassiter (2018) shows that nested uncertainty expressions naturally appearing in corpora can exhibit different readings than the one we are investigating here. As is to be expected, things can become more complicated than simply counting occurrences in a corpus. In order to keep our model as simple as possible, we abstract away from these complications here.

[3] Recall that the a-variant is the sentence to be analyzed, the b-variant is a formal characterization of its meaning and the c-variant a gloss of the b-variant in natural language. Moreover, V_s is a valuation function assigning truth values to propositional letters and $\mu_{i,s}$ is measure function assigning probability values to sets of states, such that if X is a proposition, then the value $\mu_{i,s}(X)$ represents the credence assigned by agent's i to X in s.

The idea behind these examples can be straightforwardly adapted to nested uncertainty expressions, as shown in the following example:

(5) a. It is certainly probable that the next ball drawn from this urn will be red.
b. $[\![\text{certainly}_i(\text{probably}_i\ (\text{RED}))]\!] = \{s \mid \mu_{i,s}([\![\text{probably}_i\ (\text{RED})]\!]) > \theta_{\text{certainly}}\}$
c. The set of states in which agent i assigns a level of credence higher than the semantic threshold $\theta_{\text{certainly}}$ to the proposition that the next draw will probably be red.

These definitions are uniform and compositional, nonetheless they imply a certain difference between simple and complex uncertainty expressions. When we evaluate the meaning of a simple sentence such as *probably RED* in (4) at a world s (from the point of view of agent i), the only thing that matters is the probability $\mu_{i,s}([\![\text{RED}]\!])$, i.e. the probability assigned by i to RED at s. Following the terminology adopted in Chapter 4, this is uncertainty of the first-order. In practice, for the evaluation of *probably RED*, we can lump together all worlds s which agree on $\mu_{i,s}([\![\text{RED}]\!])$. It is as if the sentence *probably RED* draws our attention to a partition of the set of states such that the states in each cell agree with each other in terms of the direct and fully resolving answers to the question *What is the probability of RED?*. On the other hand, when we evaluate a complex sentence such as *certainly likely RED* with the semantics in (5), what matters is a different aspect of each possible world s, namely agent i's higher-order beliefs $\mu_{i,w}([\![\text{probably}_i\ (\text{RED})]\!])$. The sentence draws attention to a different kind of partitioning of the same set of states. In this case the states in each cell of the partition agree with each other in terms of the answers to the question *What is the probability of 'probably RED'?*. This is higher-order uncertainty.

Having clarified this, we can finally move to the definition of the literal meaning of our simple and complex messages in the way in which they will enter the computational model. We begin with the simple messages, whose meaning is defined along the same lines as the messages in Chapter 6. As seen in Chapter 6, the meaning of a sentence of the form $X(p)$, where X is a simple uncertainty expression, is obtained by collecting the states where agent i assigns a probability value higher than a certain threshold θ_X to the proposition $[\![p]\!]$. As we said above, what matters for the evaluation of a simple sentence $X(p)$ is only the probability assigned to p. Moreover, the only possible states in our urn-based scenario are those corresponding to different quantity of red balls in the urn. These considerations allow us to give the following simplified definitions:

$$
\begin{aligned}
&[\![\text{likely}(p)]\!] = \{s \in S \mid {}^{s}\!/_{10} > \theta_{likely}\} \\
&[\![\text{possible}(p)]\!] = \{s \in S \mid {}^{s}\!/_{10} > \theta_{possible}\} \\
&[\![\text{unlikely}(p)]\!] = \{s \in S \mid {}^{s}\!/_{10} < 1 - \theta_{likely}\}
\end{aligned}
$$

where θ_{likely} and $\theta_{possible}$ are free parameters in the model.[4]

[4] Notice that we assume here that *unlikely* is essentially interpreted as the logical negation of *likely*. This assumption is not uncontroversial, but it is at least compatible with the empirical results of Tessler and Franke (2018) who found that expressions like *unhappy* are interpreted like *not happy* when presented in isolation. Only when listeners interpret multiple utterances from the same speaker, also including expressions like *not unhappy*, the interpretation assigned to *unhappy* is more negative than that of *not happy*. Tessler and Franke give an account of these experimental results with an RSA model that includes the listener's uncertain reasoning about the speaker's use of negation markers like *un-*. For simplicity, we chose not to include this

Moving to complex messages, the example in (5) shows that nesting a simple sentence such as $X(p)$ inside another uncertainty expression makes thing slightly more convoluted but results in the same kind of meaning definition: the meaning of $Y(X(p))$, where Y, X are uncertainty expressions, is obtained by collecting the states where agent i assigns a probability value higher than a certain threshold θ_Y to the proposition $[\![X(p)]\!]$, which in turn is obtained as before. As we said above, what matters for the evaluation of $Y(X(p))$ is the agent's probabilistic belief about $X(p)$. Now, similarly to what happens for simple messages, our urn-based scenario imposes constraints on the set of possible interpretations of complex messages as well. In fact, not every probabilistic belief about $X(\mathrm{p})$ is compatible with a rational belief obtained from a partial observation of the urn's contents. In our setting, all the possible rational beliefs are fully and uniquely defined in terms of any given pairs $\langle o, a\rangle$ of o observed red balls out of a balls drawn from the urn. As a consequence, we can partition the state space by lumping together all the states which agree on the speaker's contextually-relevant higher-order beliefs, which yield the following compact definition of the meaning of a sentence such as $Y(X(p))$ in terms of pairs $\langle o, a\rangle$:

$$[\![Y(X((p))]\!] = \{\langle o, a\rangle \mid \sum_{s \in [\![X(p)]\!]} \text{rat.bel}(s|o,a) > \theta_Y\} \tag{7.1}$$

Equation 7.1 defines the literal meaning of complex messages on top of the simple ones. First, the meaning of $X(p)$ is computed, where X is nothing but a placeholder for one of the simple messages in Table 7.1. Then, each state in $[\![X(p)]\!]$ is associated with a certain probability mass according to the rational belief induced in the speaker by each observation $\langle o, a\rangle$. Finally, the meaning of the complex message is computed as the set of pairs $\langle o, a\rangle$ where the total probability mass of the states in $[\![X(p)]\!]$ is greater than the semantic threshold θ_Y of the embedding expression.

As an illustration, suppose we draw 4 balls, and observe that 3 of them are red. Can we say that it is *certainly likely* that a randomly drawn ball will be red? Alternatively, suppose we draw 8 balls and observe that 6 are red. How about now? Our intuition is that in the latter case it is more natural to say that a random ball is *certainly likely* to be red. Let us illustrate Equation 7.1 by evaluating the complex message *It is certainly likely that the ball will be red* in the mentioned situations. First, the embedded simple expression is *likely*. Hence, fixing the threshold $\theta_{\text{likely}} = 0.5$, the literal meaning of *likely*$($*red*$)$ is the set of states $\{6, 7, 8, 9, 10\}$. Next, for each of these states, we compute its probability mass under the belief distributions induced by the observations 3/4 and 6/8. These are represented in Figure 7.1.

Finally, for each of the two situations, we sum the probability mass of each "surviving" state and compare the result with the threshold for the modifier, in this case *certainly*. Let us assume $\theta_{\text{certainly}} = 0.999$. Summing the probabilities obtained in the 6/8 situation (right hand side of Figure 7.1), we obviously get 1, which is greater than 0.999. Hence, the message *It is certainly likely that the ball will be red* is predicted to be literally true. On the other hand, summing the probabilities obtained in the 3/4 situation (left hand side of Figure 7.1) we get approximately 0.825, which is smaller than 0.999. Therefore, the message is predicted to be literally false. The definition in

potential level of listener uncertainty in the present model.

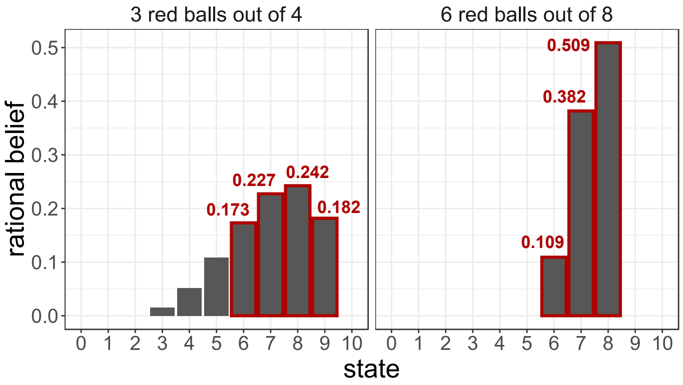


Figure 7.1: Rational belief distributions given observations 3/4 and 6/8. States which make *likely* true have been highlighted in red. Red numbers near the bars indicate the (rounded) probability mass of the corresponding state, assuming flat prior beliefs.

Equation 7.1 repeats this process for each logically possible pair $\langle o,a\rangle$ and collects the sets of pairs where each complex message results to be literally true.

Complex beliefs. The literal meaning of messages enters the model at the level of the naive literal listener, who is taken to update his belief about the urn exclusively on the basis of the literal semantics of the received message. This was easily formalized for simple messages in Equation 6.5, Chapter 6: the literal listener's beliefs are his prior beliefs restricted to the states where the message is literally true, and re-normalized. The idea behind this definition is at work in the complex model too: the literal listener updates her prior beliefs by ruling out states that are incompatible with the literal meaning of the observed message. However, things are slightly more complicated here because of the way in which we defined the meaning of complex uncertainty expression in 7.1. The difference is that the denotation of a simple message $X(p)$ is a set of states $[\![X(p)]\!] \subseteq S$ (or equivalently a set of probabilities assigned to the truth of p), whereas the denotation of a complex expression $Y(X(p))$ is a set of observation-access pairs $[\![Y(X(p))]\!] \subseteq O \times A$ (or equivalently a set of rational belief distributions). Consequently, we formulate the literal listener's beliefs after receiving a message in terms of state distinctions that are relevant to the given urn scenario: states for simple messages, $\langle o,a\rangle$ pairs for complex ones:

$$\begin{aligned} \text{lit.bel}(s \mid X(p)) &\propto \delta_{s\in[\![X(p)]\!]} \cdot \text{prior}(s) \\ \text{lit.bel}(o,a \mid Y(X(p))) &\propto \delta_{\langle o,a\rangle\in[\![Y(X(p))]\!]} \cdot \text{prior}(o,a) \end{aligned} \tag{7.2}$$

where prior(s) and prior(a) are the literal listener's priors over states and access values, ad the priors over $\langle o,a\rangle$ pairs are computed as follows:

$$\text{prior}(o,a) = \text{prior}(a) \cdot \sum_s \text{prior}(s) \cdot \text{Hypergeometric}(o \mid a,s,10) \tag{7.3}$$

Generally speaking, the interpretations of simple and complex expressions are parallel: both rule out semantically incompatible possible worlds from the literal listener's beliefs. However, a difference should be highlighted. While simple expressions are semantically about first-order uncertainty, complex expressions about higher-order uncertainty. The literal listener's beliefs reflect this distinction in terms of a partitioning of the state space into distinctions relevant to the kind of the received message.

Speaker's behavior. The last difference between the simple model and the complex one regards how we define the EU of the messages in the complex model. This is mostly a technical issue, deriving from our different definition of literal beliefs in Equation 7.2. Recall that in Chapter 6 we computed the EU of each message in each situation by directly comparing the probability distributions expressing the speaker's beliefs over S (i.e. about the probability of RED) in the given situation and the naive listener's beliefs over S induced by the message, using negative Hellinger distance. In the case of simple expressions, everything remains as in Chapter 6. Moving to complex expressions, by virtue of their semantics, they induce in the literal listener a higher-order probabilistic belief, i.e. a probability distribution over rational belief states (represented by $\langle o, a\rangle$ pairs). In order to define the EU of complex messages we imagine that the literal listener will sample one interpretation from the set of rational belief states with probability given by the higher-order belief induced by the message. The result in Equation 7.4 is a conservative extension of the utility function defined for the simple case, where ω ranges over the set of relevant rational beliefs and $\text{lit.bel}'$ is just an alternative compact notation for the literal listener's beliefs, defined below in Equation 7.5:

$$\text{EU}(m; o, a) = -\sum_{\omega} \text{lit.bel}'(\omega \mid m) \cdot \text{HD}(\text{rat.bel}(\cdot|o, a), \omega) \tag{7.4}$$

$$\text{lit.bel}'(\omega \mid m) = \begin{cases} 1 & \text{if } m \text{ is simple and } \omega = \text{lit.bel}(\cdot \mid m) \\ \text{lit.bel}(o, a \mid m) & \text{if } m \text{ is complex and there is an } a \text{ and } o \\ & \quad \text{such that } \omega = \text{rat.bel}(\cdot \mid o, a) \end{cases} \tag{7.5}$$

The reader can easily verify that, if m is a simple expression, Equation 7.4 is reduced to Equation 6.6 from Chapter 6. In the case of complex expressions, the expected utility of a message m in the situation resulting from drawing a balls and observing that o of them are red is computed as the weighted average of the negative Hellinger Distance between the rational belief induced in the speaker by said observation and each of the beliefs induced by m in the literal listener.

Having defined the EU of each message given each observation we can finally turn to modeling the pragmatic speaker's and listener's behavior. Here the complex model does not diverge in any way from the simple one:

$$\text{speak.prob}(m|o, a) \propto e^{\lambda \cdot \text{EU}(m; o, a)} \tag{7.6}$$

$$\begin{aligned} \text{listen.prob}(s, o, a|m) \propto\ & \text{speak.prob}(m|o, a) \cdot \text{Hypergeometric}(o|a, s, 10) \cdot \\ & \text{prior}(a) \cdot \text{prior}(s) \end{aligned} \tag{7.7}$$

As before, we derive listen.prob$(s|m)$ and listen.prob$(o,a|m)$ from Equation 7.7 by marginalization.

7.3 Experiment 4: complex uncertainty expressions

Goal. We collected empirical data on the production and interpretation of complex uncertainty expressions in situations of higher-order uncertainty. As already mentioned in the introductory section of this chapter, our approach to empirical data here is slightly different from Chapter 5. There, we had begun our investigation from a set of rather precise hypotheses about the influence of different levels of higher-order uncertainty on speakers' use of uncertainty expressions and about the ability of listeners to infer information about the speakers' uncertain belief state, upon hearing uncertainty expressions. For this reason, a large part of Chapter 5 was devoted to a rather detailed regression analysis testing the mentioned hypotheses on the empirical data. Subsequently, we also used the data to train and evaluate the pragmatic model developed in Chapter 6, with promising results.

Our approach is different here. Even though we still have intuitions about the behavior of complex uncertainty expressions, some of which also informed the development of the pragmatic model in the previous section, we will not attempt a thorough statistical analysis of the collected data with the goal of testing those intuitions. Instead, we take the model as our starting point. In a sense, we think of the model as our (first attempt to a) theory of the pragmatic production and interpretation of complex uncertainty expressions, and we will use the data to train and, most importantly, evaluate and criticize the model.

Participants. 255 self-reported English native speakers with IP addresses located in the USA were recruited on Amazon's Mechanical Turk. 104 participants took part in the production task (Experiment 4a) and 151 participants took part in the interpretation task (Experiment 4b). The workers were paid 1 USD for their participation, amounting to an average hourly wage of approximately 10 USD.[5]

Materials and procedure. For the most part, the experiment had the same structure and contents as Experiment 3 (Chapter 5). Participants in either task completed an introductory phase similar to the one described for Experiment 3.[6] The experimental conditions of the production task were the same 15 observations summarized in Table 5.1, repeated here as Table 7.2. Participants in the production task completed 12 trials, one for each of 12 unique conditions, in random order.

[5] Code and data are publicly available at `https://github.com/mic-he/ProbExp-PhD/tree/master/chapter7`.

[6] In more detail, participants of the production task completed a training in which they played 2 fixed rounds in the role of receivers, reporting their intuitions about the contents of red balls in the urn having received, respectively, the message "It's possible that the next ball will be red" and "It's certainly likely that the next ball will be red". Participants of the interpretation task played 3 fixed rounds as sender, observing the conditions 3/6, 1/2 and 3/8, respectively, and choosing a message to send (see main text for the options).

high	0/2	1/4	2/4	3/4	2/2
low	0/8	2/8	4/8	6/8	8/8
none	2/10	3/10	5/10	7/10	8/10

Table 7.2: Experimental conditions in Experiment 4a. The fractions represent observations of the urn. The labels on the left refer to levels of higher-order uncertainty.

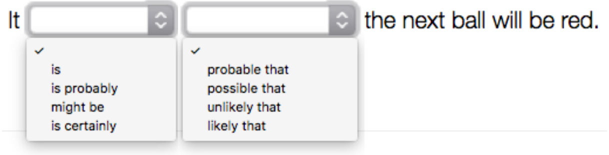


Figure 7.2: Input menus in the production task.

In each trial participants looked a the pictures corresponding to the selected condition (see Figure 5.1, Chapter 5) and made a prediction about the color of a randomly drawn ball. As per Experiment 3, the prediction must be expressed by completing a message which the participants would send to the receiver. In this case, the message had the form

It [...] [...] the next ball will be red

where the gaps must be filled in with the most appropriate combination of auxiliary/modifier and simple uncertainty expressions selected from two drop-down menus (see Figure 7.2). Notice that the choices in the second menu in Figure 7.2 included *probable* alongside *likely*. This was to offer participants their favorite pick between *probable* and *likely*. However, for simplicity, all analyses and modeling reported in this chapter treat *probable* and *likely* as synonymous and participants' choices of either as belonging to the same category, which we simply refer to as a *likely*.[7]

Moving to the interpretation task, the experimental conditions were the 12 message combinations obtained combining the 4 auxiliaries/modifiers *is*, *is probably*, *might be*, *is certainly* with the 3 simple expressions *possible*, *likely*, *unlikely*. Participants completed 24 trials, alternating state trials and observation trials for each of the 12 expressions in random order. The input sliders and recorded measures were exactly the same as in Experiment 3 (for example, see Figure 5.3, Chapter 5).

Results. We discarded the answers given by 2 participants in the production study who explicitly admitted to have had a poor understanding of the task in our final survey. Moreover, we discarded the answers given by one participant in the interpretation task who selected both access and observation values equal to 0 for at least one expression.[8]

[7] This is in line with usual assumptions in the literature, although it is worth noticing that recent work found experimental evidence of subtle differences between *probable* and *likely* (Lassiter & Goodman, 2015a). We leave the investigation of this issue to future work.

[8] While such a selection was logically possible, it explicitly contradicted the instructions given to the participants at the beginning of the task.

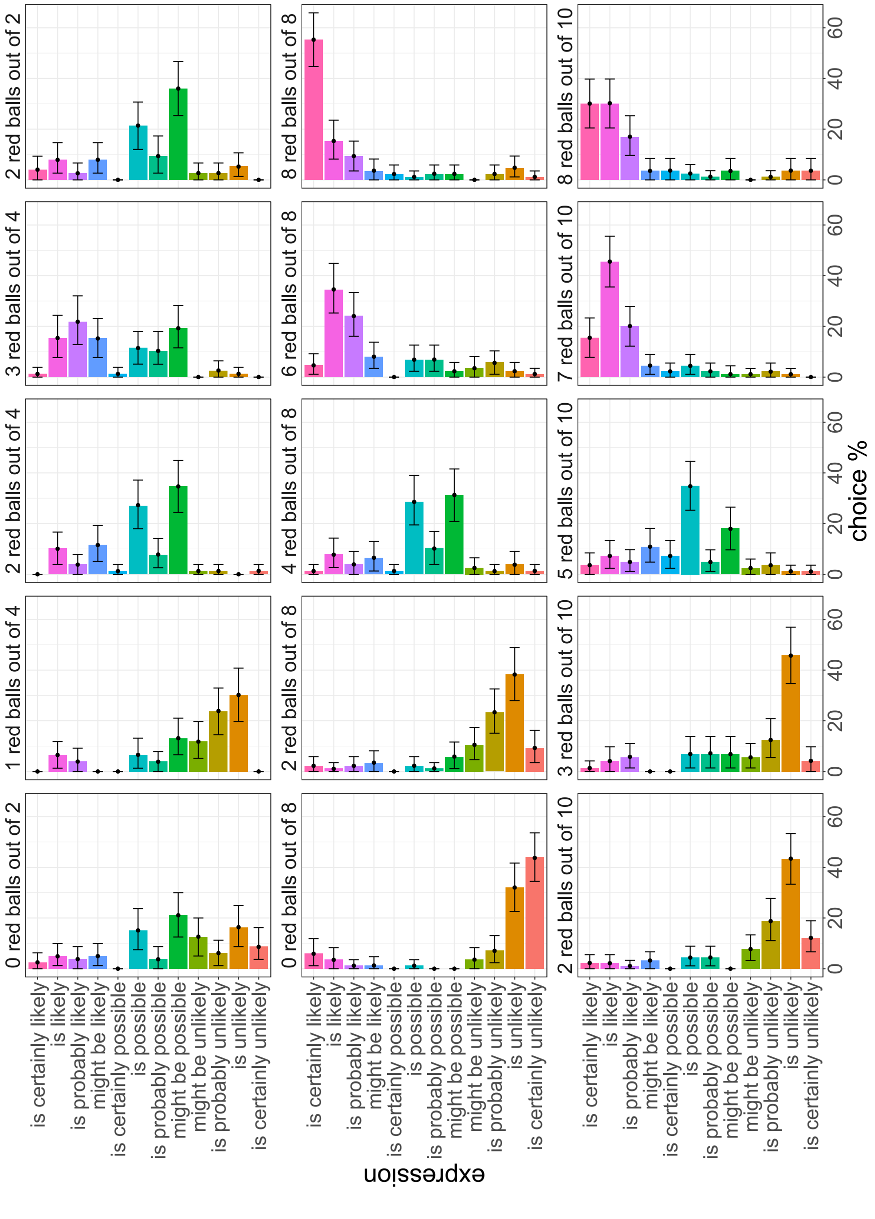


Figure 7.3: Percentages of expression choices in each observation condition, together with bootstrapped 95% confidence intervals (black bars).

Figure 7.3 displays percentages of 102 participants' expression choices in each observation condition. We highlight three interesting features of the data. First, we observe that participants selected complex expressions in a systematic manner: in fact, five out of the nine complex expressions are the modal choice in at least one experimental condition, as are all three simple expressions.

Second, by visual inspection, we can already observe that the data seems to be in line with the general predictions we would expect from the pragmatic model. The middle column of Figure 7.3 represents situations in which the speaker sees an equal number of red and blue balls. In these situations $s = 5$ is the most likely state under the relevant rational beliefs (assuming unbiased priors). As a consequence, we expect to see frequent choices of expressions which include *possible*, which indeed we can observe in the data. In the two columns on the right we find epistemic states where a higher proportion of red balls is subjectively more likely than a lower. As a consequence, we expect more choices of expressions with *likely*. The reverse is the case in the two columns on the left, where we expect more choices with *unlikely*. Again, we can indeed observe these general regularities in the data. Moreover, going through Figure 7.3 row-wise from top to botton, speakers have increasing access, so less higher-order uncertainty. With less higher-order uncertainty we expect (and observe) a general trend of increasing use of *certainly* or *is* in connection with *unlikely* (in the two left-most columns) and *likely* (in the two right-most columns).

Finally, we observe that *certainly* appears to behave like an intensifier, compared to *is*. Choice counts of *certainly likely* increase from condition 6/8 to 8/8 and from 7/10 to 8/10, whereas choice counts of *is likely* decrease. Similarly, choice counts of *certainly unlikely* increase from condition 2/8 to 0/8 and from 3/10 to 2/10, whereas choice counts of *is unlikely* decrease. Similarly, *might be* appears to be something like a 'downtoner.' These patterns are also reflected in the interpretation data, to which we turn next.

Moving to interpretation, Figure 7.4 displays counts of 150 participants' choices of state, access and observation values in each expression condition. Our model predicts that interpretation, by Bayes rule, follows the likelihood of production choices. As a consequence, we would expect to observe patterns similar to those observed in the production data (ignoring strong prior effects). This is indeed the case: modal and mean interpretation choices of state s are lowest for expression choices with *unlikely*, higher for *possible* and highest for expressions with *likely*. Moreover, participants' interpretation choices for the number of observed red balls are very well behaved and appear to line up with the interpretation of the state component under a rational belief model.

If we look at mean interpretation choices for the state dimension (top row, Figure 7.5) we can observe patterns along the same lines of the "intensifying/downtoning effects" of outer expressions *certainly* and *might be*, respectively, observed in the production data. Listeners estimate the true number of red balls to be higher when they hear *certainly likely* (mean interpretation of state 7.05 and bootstrapped 95% CI [6.76; 7.32]) than when they hear *is likely* (mean 6.39, [6.15; 6.65]). If we compare the first to the last row (especially *is (un-)likely* to *might be (un-)likely*) we can observe a tendency to a similar "downtoning effect." For example, the interpretation of *might be likely* gets a mean of 5.35 ([5.13; 5.59]).

Finally, we look at mean interpretation choices along the access dimension (bottom

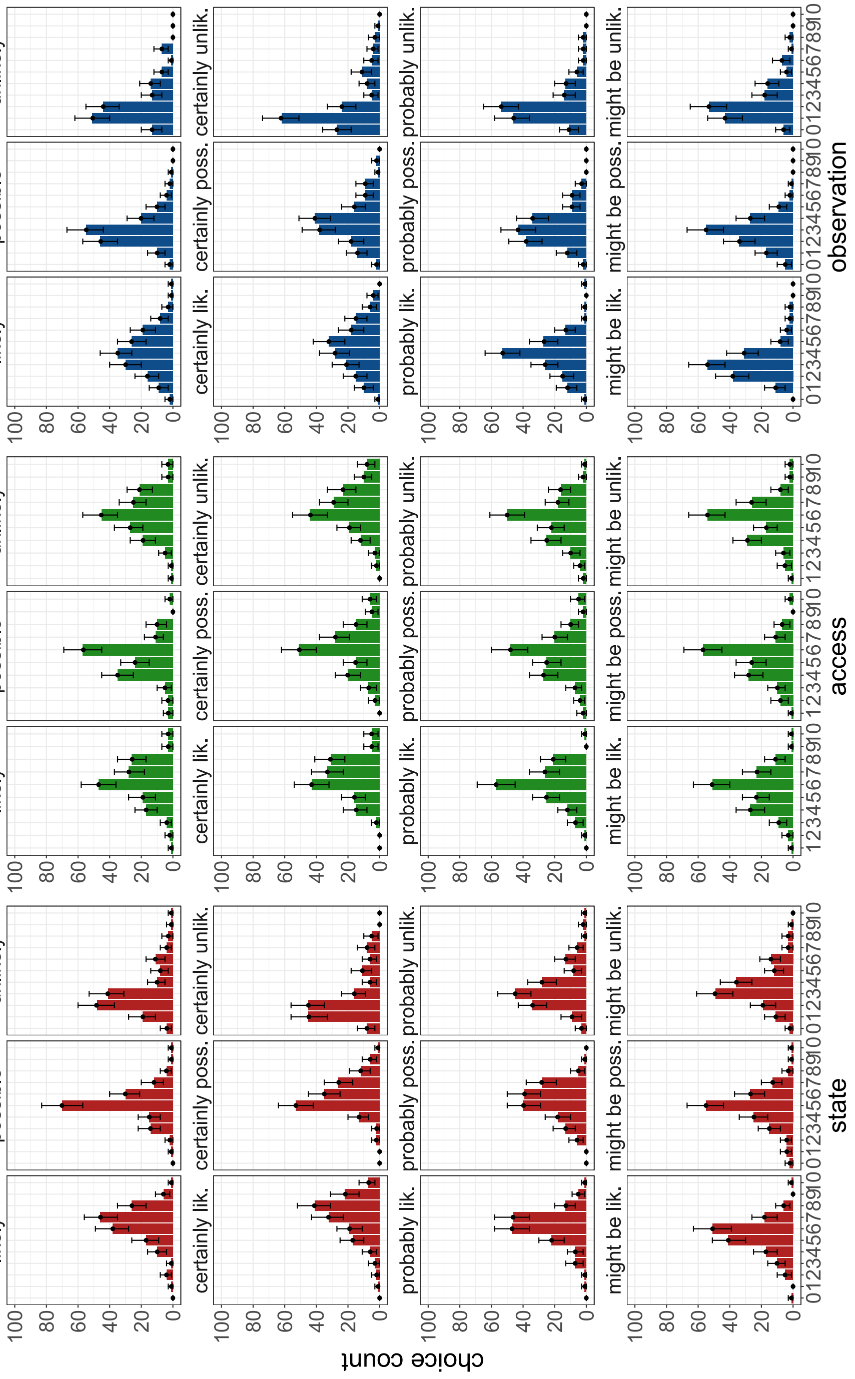


Figure 7.4: Counts of state, access and observation value choices in each expression condition, together with bootstrapped 95% confidence intervals (black bars).

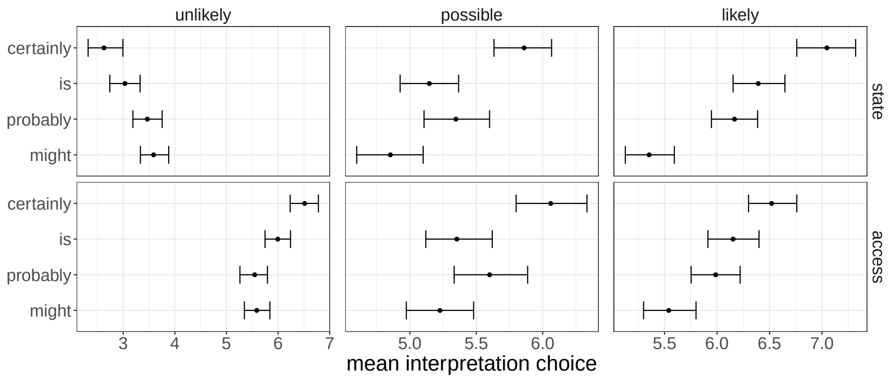


Figure 7.5: Means of choices in the state and access interpretation conditions, with bootstapped 95% confidence intervals.

row, Figure 7.5), which shows how knowledgeable the speaker is estimated to be after hearing certain messages. Again, we can observe a trend to estimate the speaker to be more informed (higher access) when an expression is modified with *certainly* than when *is* is used, and even less competent when *might be* is used. For example, mean interpretation of access for *certainly likely* is 6.52 ([6.3; 6.76]), for *is likely* it is 6.15 ([5.91; 6.4]) and for *might be likely* it is 5.54 ([5.3; 5.8]).

7.4 Inference and model evaluation

We followed the same procedure described in Chapter 6. First, we inferred credible values for the free parameters of the model given the data. That is, we implemented the model in JAGS and we estimated the posterior distribution over values for the free parameters conditioning on the data, under the assumption that the observed choice counts of messages and of state, access and observation values are modeled as samples from appropriate multinomial distributions with weights equal to the probabilities computed by the pragmatic model (in particular Equation 7.6 and 7.7).[9]

The results obtained for the semantic threshold parameters are summarized in Table 7.3 in terms of mean and HDIs.[10] Notice that the values are similar to the ones inferred in Section 6.4, Chapter 6, which is encouraging: the 95% HDIs of the earlier inference overlap those of the current inference for all three threshold parameters.[11] Another

[9] The code is publicly available at `https://github.com/mic-he/ProbExp-PhD/tree/master/chapter7`.

[10] The results for the remaining free parameters are summarized in the following table:

	λ	α_s	β_s	α_a	β_a
lower	4.639	4.932	4.882	27.956	20.379
mean	4.846	5.546	5.488	68.835	49.43
upper	5.075	6.125	6.07	128.255	91.733

[11] Notice that $\theta_{probably}$ from the simpler model should be mapped onto θ_{likely} in Table 7.3 because the latter represents the threshold of the inner expressions *likely/probably* and thus correspond to the simple expression

	$\theta_{possible}$	θ_{might}	θ_{likely}	$\theta_{probably}$	$\theta_{certainly}$
lower	0.200	0.328	0.506	0.629	0.964
mean	0.251	0.332	0.530	0.690	0.979
upper	0.295	0.336	0.548	0.745	0.996

Table 7.3: Mean inferred values and HDIs of the semantic threshold parameters free in the model given experimental data collected in Experiment 2.

interesting observation is that uncertainty expressions in outer position are estimated to have a higher semantic threshold than related expressions in inner position: for example, the threshold of *might be* (a nesting expression) is estimated to be higher than that of *possible* (a nested expression) and similarly for *probably* and *likely*.

Next, we computed credible ranges of Pearson's correlation scores between the model's posterior predictive distributions and experimental data, as summarized in Table 7.4 in terms of mean and HDIs. These results are encouraging. Comparing the results with those reported in Section 6.4 of Chapter 6, we can observe two facts. Starting from the the correlations between the data obtained in the interpretation task and the predictions of listener's model, we notice that the average scores are slightly higher for the complex model. Moving to production data and speaker's model, we can observe that the range of correlations, while still being reasonably high, is noticeably lower in the case of complex expressions. However, this fact does not seem especially worrying or surprising to us: while the simple model is trying to predict observed choices of five different categories, the complex model is trying to predict choices of twelve categories, hence more noise is to be expected.

	expression	*state*	*access*	*observation*
lower	0.596	0.812	0.818	0.915
mean	0.654	0.849	0.853	0.934
upper	0.719	0.883	0.888	0.952

Table 7.4: Mean Pearson's correlation scores and HDIs between model posterior predictive distributions and exerimental data collected in Experiment 4.

Finally, we compared model predictions and experimental data in more detail by looking at posterior predictive distributions, displayed in Figures 7.6 and 7.7. We begin with production. In general, the posterior predictive distributions appear to support the basic observations we made for the production data above. Going from left to right, we can observe that the model indeed predicts that the conditions in the two leftmost columns are mostly associated with the use of expressions containing *unlikely*; the conditions in the middle column mostly associate with expressions containing *possible*; and the two rightmost columns show a dominant ending in *likely*, especially in conditions with low higher-order uncertainty. Moreover, the posterior predictive distributions also support the observation that *certainly* seems to behave like an intensifier. For example, the model predicts that *certainly likely* is less frequent in condition 6/8

from the first model.

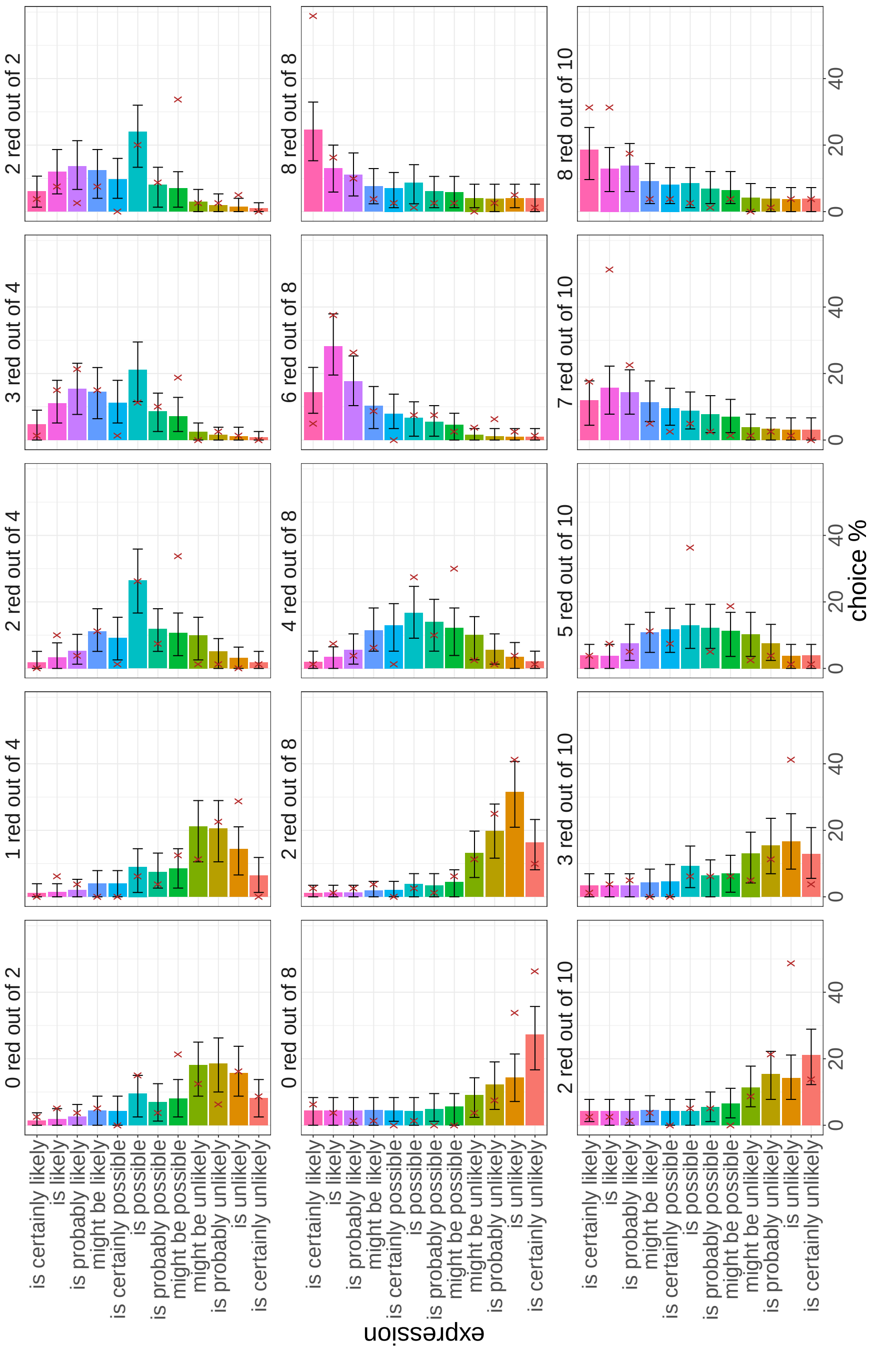


Figure 7.6: Mean predicted percentages of expression choices in each observation condition, together with Bayesian 95% HDIs (black bars). Red crosses mark the empirically observed counts for comparison (see also Figure 7.3).

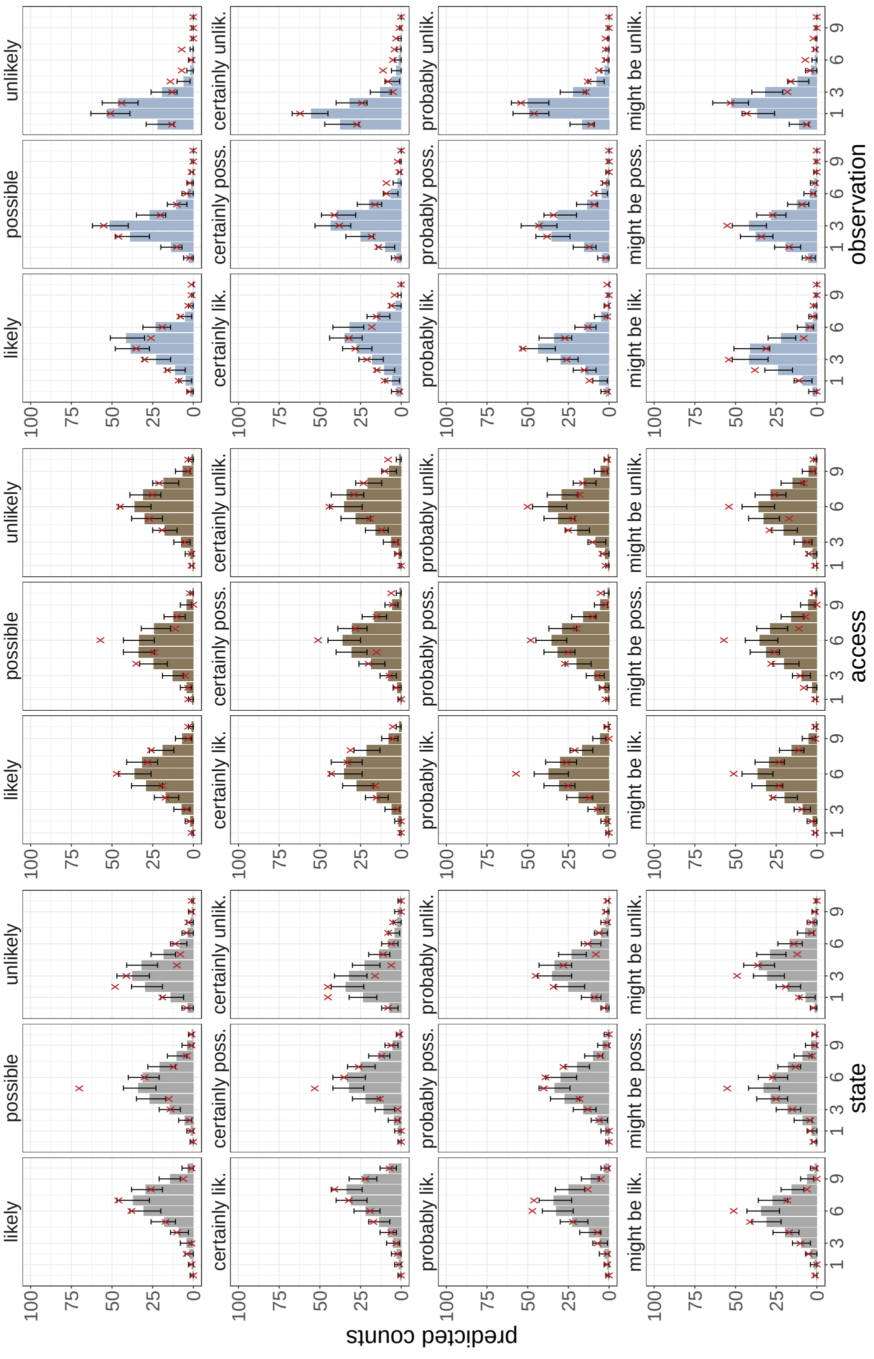


Figure 7.7: Means of the predicted counts of state, access and observation value choices in each expression condition, together with Bayesian 95% HDIs (black bars). Red crosses mark the empirically observed counts for comparison (see also Figure 7.4).

than in 8/8, whereas *is likely* is more frequent in 6/8 than in 8/8. A similar pattern is predicted for *certainly unlikely*.

However, we can also observe a number of discrepancies between experimental data and the model's posterior predictive distributions. These are points in the plots where the red crosses representing observed count data do not fall inside the range delimited by the black bars representing the predictions' HDIs. If the discrepancies are systematic, then they might suggest that the model does not capture some aspects of the experimental data. For example, the model constantly underpredicts choice frequencies of *might be possible* in the conditions where the expression would be appropriate (first rows in Figures 7.3 and 7.6). This is likely due to the fact that the model, in general, prefers logically stronger expressions, but *might be possible* is very weak. It is an interesting question why participants chose *might be possible* more than we would expect (perhaps they did not want to commit too much?). Another interesting observation is that the discrepancies concerning simple expressions always go in the same direction: the model underpredicts choice frequencies of simple expressions. This suggests that an important component might be missing from the model, for example a baseline preference for some expressions over others, possibly defined in terms of cost, efficiency, or frequency of occurrence.

Turning to interpretation, visual inspection of Figure 7.7 reveals that the patterns displayed in the data are captured rather well by the model. However, we can highlight a number of discrepancies here as well. For example, if we look at the state interpretation for *possible* and complex messages containing it (left panel, middle column of Figure 7.7) we can see that the model often underpredicts participants' choices of the middle state 5. This is probably the most systematic of the discrepancies revealed by the comparison, and it could be an effect of the salience of option 5, which participants might have perceived as the least committing "default" option on the slider. Other shortcomings of the model include *unlikely* and *certainly unlikely* (left panel, right column), where we can observe that the predictions are visibly shifted to the right compared to the data: the model underpredicts lower state values (1,2,3) in favor of middle values (4,5,6). Looking at the access interpretation, another feature that looks unexpected in the light of the model is the low counts of access choices of 5 (compared to 4 and 6) for several expressions (middle panel).

Finally, let us zoom in on the model's posterior predictions on mean values for the interpretation of state and access (Figure 7.8). The model seems to capture the tendency towards an intensifying effect on the state interpretation of *certainly* and the downtoning effect of *might be* when these expressions embed *unlikely* and *likely* (top row, leftmost and rightmost columns respectively). Similarly, the model supports the conclusion that the speaker is perceived as more knowledgeable (higher access interpretation) when the embedding expression is *certainly* than when it is simply *is* (bottom row). However, the model visibly fails to predict the observed access interpretation of *might be possible* (bottom row, middle column). According to the observed data, the speaker is taken to be roughly as informed after hearing *might be possible* as after hearing *is possible*, whereas the model predicts a value of access for *might be possible* which is too high.

The importance of the observed discrepancies should not be underestimated: these are data points which are still surprising —so to speak— for a model which was trained

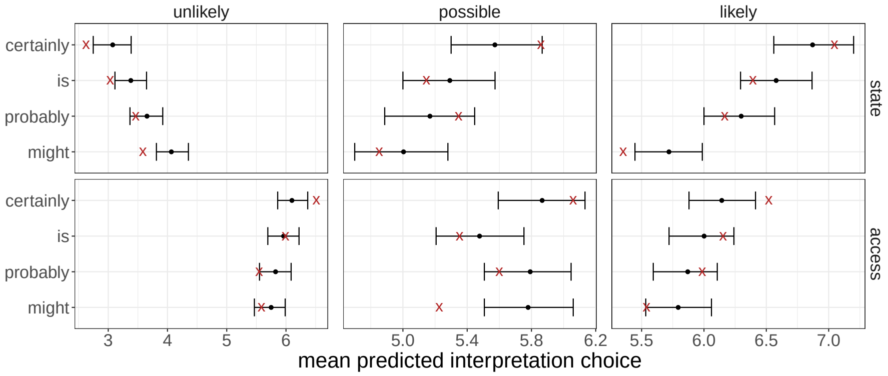


Figure 7.8: Posterior predictive of the means of choices in the state and access interpretation conditions. Error bars are 95% HDIs. Red crosses are the empirically observed means (see also Figure 7.5).

on the data. Moreover, it is not easy to identify an obvious conceptual shortcoming of the model which can explain the discrepancies away. However, we believe that in general the tendencies displayed in the data are captured by the model, which is encouraging.

7.5 Conclusion

In this chapter we extended our data-driven modeling approach to account for complex uncertainty expressions. Once again, our working hypothesis was that higher-order levels of uncertainty play a role in the production and interpretation of complex uncertainty expressions. The data collected in our experimental setting showed that complex expressions are indeed systematically used to communicate higher-order uncertain information and that listeners are able to draw inferences about the speaker's uncertain beliefs from complex expressions.

We developed our RSA model of complex expressions with the goal of keeping it as conservative as possible with respect to the simple model proposed in Chapter 6. In doing so we faced the challenge how to define the literal semantics of complex messages in a compositional way and, as a consequence, how to adapt the literal listener's belief update rule to the new semantics. The approaches we proposed to tackle these challenges were in line with recent literature on the topic (Moss, 2015).

Despite not being flawless, the behavior of our model with respect to the experimental data appears to be insightful and promising. First of all, the basic assumptions which we made (compositional semantics) and which are built-in into RSA (Gricean cooperative reasoning in communication) have not been outright refuted by the data. Moreover, the model developed on the basis of these assumptions is indeed able to capture and explain interesting patterns of choice preferences observed in the data.

Summing up, this chapter presents, to the best of our knowledge, the first attempt at a systematic analysis of complex uncertainty expressions, grounded in empirical data and deriving quantitative predictions about use and interpretation of such expressions in a specific communicative context.

Conclusion

In the following paragraphs we summarize what we believe to be the most important take-home messages of this dissertation. While doing so, we also try to highlight some of the possible shortcomings of our approach as well as open issues and outlook for future research.

First of all, the research reported in this dissertation strengthens the case in favor of adopting a semantic analysis of uncertainty expressions which incorporates a quantitative measure of uncertainty, such as probability, or an equally rich model-theoretic structure. As we have seen, this idea has been put forward by authors belonging to the logico-philosophical tradition (Swanson, 2006, 2016; Yalcin, 2010; Moss, 2015; Lassiter, 2011a, 2017) as a reaction to the predominant qualitative paradigm dating back to Kratzer (1991). In this work we have not directly argued in favor of the quantitative approach to uncertainty expressions. The goal of our work was not to reconstruct the debate and/or evaluate the different concrete proposals available on the market. However, we do believe that the analyses and arguments put forward by the mentioned authors in favor of the quantitative approach are mostly convincing, and for this reason we have chosen to adopt (some variations on the theme of) a probabilistic semantics for uncertainty expressions as a given. On the basis of our choice, we designed the experiment reported in Chapter 2 (about context dependency and QUD sensitivity); we extended Roberts' context-sensitive semantics to account for the differences between uncertainty adverbs and adjectives in Chapter 3; and we developed our probabilistic model of the use and interpretation of simple and complex uncertainty expressions in Chapters 6 and 7. We hope to have shown that the research reported in each of these chapters gave very promising results, which indirectly speaks in favor of the probabilistic approach to the semantics of uncertainty expressions. This is the first take-home message of the dissertation.

The second take-home message is that, on top of the assumed probabilistic semantics, the general picture of uncertainty expressions emerging from our work is a picture of context sensitivity, QUD sensitivity and pragmatics. Even if we could not experimentally elicitate a clear QUD-effect in Chapter 2, we were able to replicate the well-known Alternative Outcome effect, which corroborates the importance of contextual information for the production of uncertainty expressions. The role played by QUD sensitivity with respect to uncertainty expressions is apparent in Chapter 3, in which our explanation of the differences between uncertainty adverbs and adjectives crucially revolves around the assumption that the two kinds of uncertainty expressions are typically used to address different kinds of QUDs.

Finally, the notion of QUD played a role in our probabilistic models of simple and complex uncertainty expressions reported in Chapters 5-7. Specifically, we made the simplifying assumption that simple uncertainty expressions (*possible*, *probably*) are typically interpreted as if they are used to answer a direct question about the probability of an event (*How likely is X?*), whereas complex uncertainty expressions (*certainly possible*, *might be likely*) are typically taken to express an answer to a question regarding higher levels of uncertainty. It goes without saying that things are certainly more complicated than this. More realistically, the listener would need to reason about and infer which among these (and possibly other) QUDs the speaker might have had in mind.[12] However, reasoning about possible alternative QUDs is relatively easy in an artificial set-up such as ours, where the urn-based modeling scenario made it easy for us to specify different possible partitions of the state space. On the other hand, if we look at real life conversation, such a precise and abstract apparatus might not be apparent to the speakers, or it might not be present at all. This brings us to the open question, which we believe can lead to interesting possibilities for future empirical and computational investigations, of how speakers and listeners flexibly coordinate on the contextual structure in which they use and interpret uncertainty expressions, and more generally, other context-sensitive linguistic constructions.

Next, we mentioned in the Introduction that one of the goals of our work was to bring closer together the psychological and the philosophical approaches to uncertainty expressions, and we believe that we achieved our goal, at least to some extent. First of all, both Chapter 2 and Chapter 3, each in its own way, investigate issues stemming from the interplay between formal semantic theories of uncertainty expressions and observations about empirical phenomena. Chapter 2 follows the lead of the works by Yalcin (2010) and Lassiter (2011a) and it investigates experimentally whether well-known psychological results about the context-dependency of uncertainty expressions (Teigen, 1988; Windschitl & Wells, 1998) can be reconciled with the formal probabilistic semantics proposed by authors in the logico-philosophical tradition. For this reason, Chapter 2 perfectly exemplifies what we have in mind when we talk about bringing the two traditions closer together.

From this point of view, our work on uncertainty adverbs and adjectives reported in Chapter 3 can be seen as a first step towards a more systematic investigations of the differences between the two categories of uncertainty expressions, which would naturally need to include experimental data as well. In fact, we based our theory on a set of empirical phenomena observed and collected by other authors, but our goal was to test how far we could go in explaining these phenomena without assuming Wolf's distinction between adverbs and adjectives. So crucially, our job has not been to thoroughly question the empirical validity of the phenomena themselves, which were taken —for the most part— as given. And, whenever we did assess empirical phenomena (either to criticize Wolf's theory or to support or own), we did so on the basis of our own (or other authors') intuitions or naive Google searches. This (perfectly acceptable) approach situates our work within the logico-philosophical tradition, and it leaves the door wide open for possible directions of future empirical work. In particular, the following two questions seem more pressing: Are the distributional differences between adjectives

[12] see Kao, Wu, et al. (2014)

and adverbs collected by Wolf corroborated by a systematic corpus study? How solid are (our and other authors') intuitions about the semantic/pragmatic/conversational differences, i.e. to what extent can they be observed in behavioral experiments?

Finally, let us go back to the focal point of the dissertation, which is the data-driven modeling reported in Chapters 4-7. Here, too, the two traditions converge. Following the research trend of experimental probabilistic pragmatics (exemplified, among others, by RSA and RSA-like modeling of pragmatic phenomena) we bring together formal semantics, rationalistic pragmatics and computational simulations into a model of the use and interpretation of uncertainty expressions whose quantitative (posterior) predictions are compared to experimentally observed human data. We believe that the promising results obtained in the research reported in Chapters 4-7 can be taken as the foundations of a systematic investigation of simple and complex uncertainty expressions. From this point of view, we have shown that higher-order uncertainty will likely have to play a prominent role in any such investigation. This is another take-home message of the dissertation. Clearly, there will likely be situations in which taking into account several layers of uncertainty is not necessary and can be abstracted away with, or situations in which we will need a different conceptualization of higher-order uncertainty than the one proposed here. In any case, our results have helped establishing the fact that uncertainty expressions are sensitive to different layers of uncertainty. As a consequence, *some* reference to the notion of higher-order uncertainty cannot be avoided, be it to further investigate complex uncertainty expressions following our work, or to argue that the notion of higher-order uncertainty is superfluous in a given context or that it should be revised in some way.

More in general, our work has helped establish that communication under uncertainty and about uncertainty is a matter of reasoning about our own and other agents' partial information and uncertain beliefs, cooperation and coordination, and context sensitivity. For this reason, we believe that the right tools to investigate this domain are the ones which integrate rationalistic approaches to communication with computational cognitive modeling and experimental psycholinguistics, on the background of open and reproducible scientific method.

References

Bellert, I. (1977). On semantic and distributional properties of sentential adverbs. *Linguistic Inquiry*, *8*, 337–351.

Benz, A., Jäger, G., & Van Rooij, R. (2005). *Game theory and pragmatics*. Basingstoke, UK: Palgrave Macmillan.

Beyth-Marom, R. (1982). How probable is probable? a numerical translation of verbal probability expressions. *Journal of Forecasting*, *1*(3), 257–269.

Box, G. E. (1979). Robustness in the strategy of scientific model building. In *Robustness in statistics* (pp. 201–236). Academic Press.

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*, *41*(3), 390–404.

Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Carnap, R. (1947). *Meaning and necessity*. Chicago, IL: University of Chicago Press.

Clark, D. A. (1990). Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology*, *9*(3), 203–235.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In D. C. Noelle et al. (Eds.), *Proceedings of CogSci37* (pp. 548–553).

Ernst, T. (2009). Speaker-oriented adverbs. *Natural Language & Linguistic Theory*, *27*(3), 497–544.

Fagin, R., & Halpern, J. Y. (1994). Reasoning about knowledge and probability. *Journal of the Association of Computing Machinery*, *340–367*.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Franke, M. (2014). Typical use of quantifiers: A probabilistic speaker model. In P. Bello (Ed.), *Proceedings of the 36th annual meeting of the Cognitive Science Society* (pp. 487–492).

Franke, M. (2017). Game theory in pragmatics: Evolution, rationality & reasoning. In *Oxford research encyclopedia of linguistics.* Oxford University Press.

Franke, M., Dablander, F., Schöller, A., Bennett, E., Degen, J., Tessler, M. H., ... Goodman, N. D. (2016). What does the crowd believe? a hierarchical approach to estimating subjective beliefs from empirical data. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th annual conference of the Cognitive Science Society* (pp. 2669–2674).

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why bayes rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*(1), 3–44.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd edition ed.). Boca Raton: Chapman and Hall.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–472.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38.

Geurts, B. (2010). *Quantity implicatures*. Cambridge, UK: Cambridge University Press.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

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

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Grice, P. (1975). Logic and conversation. *Syntax and semantics*, *3*, 41–58.

Groenendijk, J., & Stokhof, M. (1997). Questions. In J. van Benthem & A. ter Meulen (Eds.), *Handbook of logic and language, first edition* (pp. 1055–1124). Elsevier.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016, Sep 01). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842.

Hamblin, C. L. (1959). The modal 'probably'. *Mind*, *68*(270), 234–240.

Herbstritt, M. (2015). Experimental investigations of probability expressions: a first step in the (probably) right direction. In M. Kaeshammer & P. Schulz (Eds.), *Proceedings of ESSLLI 2015 Student Session* (pp. 77–88).

Herbstritt, M., & Franke, M. (2016). Definitely maybe and possibly even probably: efficient communication of higher-order uncertainty. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th annual conference of the Cognitive Science Society* (pp. 2639–2644).

Herbstritt, M., & Franke, M. (2017). Modeling transfer of high-order uncertain information. In *Proceedings of the 39th annual conference of the Cognitive Science Society* (pp. 507–512).

Herbstritt, M., & Franke, M. (2019). Complex probability expressions & higher-order uncertainty: Compositional semantics, probabilistic pragmatics & experimental data. *Cognition*, *186*, 50–71.

Hintikka, J. (1961). Modality and quantification. *Theoria*, *27*(3), 119–128.

Jackendoff, R. S. (1972). *Semantic interpretation in generative grammar.* Cambridge, MA: The MIT Press.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*, 169–188.

Kao, J. T., Bergen, L., & Goodman, N. (2014). Formalizing the pragmatics of metaphor understanding. In P. Bello (Ed.), *Proceedings of the 36th annual meeting of the Cognitive Science Society* (pp. 719–724).

Kao, J. T., & Goodman, N. D. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual meeting of the Cognitive Science Society* (pp. 1051–1056).

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *PNAS*, *111*(33), 12002–12007.

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

Kratzer, A. (1977). What 'must'and 'can'must and can mean. *Linguistics and philosophy*, *1*(3), 337–355.

Kratzer, A. (1991). Modality. In A. von Stechow & D. Wunderlich (Eds.), *Semantics: An international handbook of contemporary research* (pp. 639–650). Berlin, DE: de Gruyter.

Kripke, S. A. (1980). *Naming and necessity*. Cambridge, MA: Harvard University Press.

Kruschke, J. (2014). *Doing Bayesian Data Analysis, 2nd Edition: A tutorial with R, JAGS, and Stan*. Academic Press.

Lassiter, D. (2010). Gradable epistemic modals, probability, and scale structure. In L. N. & D. Lutz (Eds.), *Proceedings of SALT20* (pp. 197–215).

Lassiter, D. (2011a). *Measurement and modality: the scalar basis of modal semantics* (Unpublished doctoral dissertation). NYU Linguistics.

Lassiter, D. (2011b). *Measurement and modality: the scalar basis of modal semantics* (Unpublished doctoral dissertation). NYU Linguistics.

Lassiter, D. (2014). Epistemic comparison, models of uncertainty, and the disjunction puzzle. *Journal of Semantics*, *32*(4), 649–684.

Lassiter, D. (2017). *Graded modality: Qualitative and quantitative perspectives*. Oxford, UK: Oxford University Press.

Lassiter, D. (2018). Talking about (quasi-)higher-order uncertainty. In C. Condoravdi & T. H. King (Eds.), *Tokens of meaning: Papers in honor of Lauri Karttunen.* CSLI Publications.

Lassiter, D., & Goodman, N. D. (2015a). How many kinds of reasoning? inference, probability, and natural language semantics. *Cognition*, *136*(0), 123 – 34.

Lassiter, D., & Goodman, N. D. (2015b). *Nested and informative epistemics in a graphical models framework.* Talk given at Bridging Logical and Probabilistic Approaches to Language and Cognition, ESSLLI 2015, Barcelona.

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

Lee, M., & Wagenmakers, E. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press.

Levinson, S. C. (1983). *Pragmatics*. Cambridge, UK: Cambridge University Press.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science*, *9*(10), 563–564.

MacFarlane, J. (2009). Epistemic modals are assessment-sensitive. In B. Weatherson & A. Egan (Eds.), *Epistemic modality.* Oxford University Press.

MacFarlane, J. (2010). *Epistemic modals: Relativism vs cloudy contextualism.* Retrieved from `http://johnmacfarlane.net/cloudy.pdf` (Presented at Chambers Philosophy Conference on Epistemic Modal, University of Nebraska)

Moss, S. (2015). On the semantics and pragmatics of epistemic vocabulary. *Semantics and Pragmatics*, *8*(5), 1–81.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., … Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.

Nilsen, Ø. (2004). Domains for adverbs. *Lingua*, *6*, 809-847.

Nuyts, J. (2001a). *Epistemic modality, language, and conceptualization: A cognitive-pragmatic perspective* (Vol. 5). Amsterdam, NL: John Benjamins Publishing.

Nuyts, J. (2001b). Subjectivity as an evidential dimension in epistemic modal expressions. *Journal of pragmatics*, *33*(3), 383–400.

Papafragou, A. (2006). Epistemic modality and truth conditions. *Lingua*, *116*(10), 1688–1702.

Peskun, P. (2016). *Some relationships and properties of the hypergeometric distribution.* Retrieved from `https://arxiv.org/abs/1610.07554v1`

Piñón. (2006). *Modal adverbs again.* (Honoring Anita Mittwoch on her 80th birthday, Syntax, Lexicon, and Event Structure, The Hebrew University of Jerusalem, 4–6 July)

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of DSC3* (Vol. 124, p. 125).

Potts, C. (2007). The expressive dimension. *Theoretical linguistics*, *33*(2), 165–198.

Qing, C., & Franke, M. (2014). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In T. Snider (Ed.), *Proceedings of SALT24* (pp. 23–41).

Qing, C., & Franke, M. (2015). Variations on a Bayesian theme: Comparing Bayesian models of referential reasoning. In H. Zeevat & H.-C. Schmitz (Eds.), *Bayesian natural language semantics and pragmatics* (pp. 201–220). Berlin: Springer.

Roberts, C. (1996). Information structure in discourse. In J. Yoon & A. Kathol (Eds.), *Osu working papers in linguistics* (Vol. 49, pp. 91–136). Ohio State University.

Roberts, C. (2012, December). Information structure: Afterword. *Semantics and Pragmatics*, *5*(7), 1–19.

Roberts, C. (2015). *The character of epistemic modality: Evidentiality, indexicality, and what's at issue.* (Manuscript available at http://www.ling.ohio-state.edu/ roberts.21/Roberts.EpistemicModality.pdf)

Roberts, C. (2017, June). *Agreeing and assessing: Epistemic modals and the question under discussion.* (Manuscript available at http://www.ling.ohio-state.edu/ roberts.21/Roberts.A&A.pdf)

Roberts, C., Simons, M., Beaver, D., & Tonhauser, J. (2009). Presupposition, conventional implicature, and beyond: A unified account of projection. In N. Klinedist & D. Rothschild (Eds.), *Proceedings of the esslli 2009 workshop new directions in the theory of presupposition.*

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.

Schöller, A., & Franke, M. (2015). Semantic values as latent parameters: Surprising few & many. In S. D'Antonio, M. Moroney, & C.-R. Little (Eds.), *Proceedings of SALT25* (pp. 143–162).

Schöller, A., & Franke, M. (2017). Semantic values as latent parameters: Testing a fixed threshold hypothesis for cardinal readings of *few* & *many*. *Linguistic Vanguard*, *3*(1).

Sheldrake, R. (2004). The need for open-minded scepticism: A reply to David Marks. *The Skeptic*, *16*(4), 8–13.

Simons, M., Tonhauser, J., Beaver, D., & Roberts, C. (2010). What projects and why. In *Semantics and linguistic theory* (Vol. 20, pp. 309–327).

Stephenson, T. (2007). Judge dependence, epistemic modals, and predicates of personal taste. *Linguistics and Philosophy*, *30*(4), 487–525.

Swanson, E. (2006). *Interactions with context* (Unpublished doctoral dissertation). Massachusetts Institute of Technology, Cambridge MA.

Swanson, E. (2016). The application of constraint semantics to the language of subjective uncertainty. *Journal of Philosophical Logic*, *45*(2), 121–146.

Teigen, K. H. (1988). When are low-probability events judged to be 'probable'? effects of outcome-set characteristics on verbal probability estimates. *Acta Psychologica*, *6*(2), 157–174.

Tessler, M. H., & Franke, M. (2018). Not unreasonable: Carving vague dimensions with contraries and contradictions. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Proceedings of the 40th annual meeting of the Cognitive Science Society* (pp. 1108–1113).

Tessler, M. H., & Goodman, N. D. (2019). The language of generalization. *Psychological Review*, *126*(3), 395–436.

Tessler, M. H., Lopez-Brau, M., & Goodman, N. D. (2017). Warm (for winter): Comparison class understanding in vague language. In *Proceedings of the 39th annual conference of the Cognitive Science Society* (pp. 1181–1186).

Tversky, A., & Kahnemann, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Ülkümen, G., Fox, C. R., & Malle, B. F. (2016). Two dimensions of subjective uncertainty: Clues from natural language. *Journal of Experimental Psychology: General*, *145*(10), 1280.

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*.

von Fintel, K., & Gillies, A. (2008). Cia leaks. *Philosophical Review*, *117*(1), 77–98.

von Fintel, K., & Gillies, A. (2009). 'might' made right. In B. Weatherson & A. Egan (Eds.), *Epistemic modality.* Oxford University Press.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive psychology*, *60*(3), 158–189.

Wallsten, T. S., Fillenbaum, S., & Cox, J. A. (1986). Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language*, *25*(5), 571–587.

Weatherson, B., & Egan, A. (2009). Introduction: Epistemic modals and epistemic modality. In B. Weatherson & A. Egan (Eds.), *Epistemic modality.* Oxford University Press.

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, *2*(4), 343.

Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology*, *75*(6), 1411–1423.

Wittgenstein, L. (2010). *Philosophical investigations*. UK: Wiley-Blackwell.

Wolf, L. (2014). *Degrees of assertion* (Unpublished doctoral dissertation). Ben Gurion University of the Negev, Faculty of Humanities and Social Sciences, Department of

Foreign Literatures and Linguistics.

Wolf, L. (2015). It's probably certain. In N. Melnik (Ed.), *Proceedings of IATL 30* (pp. 115–142).

Wolf, L., & Cohen, A. (2009). Modal adverbs as negotiation chips. *Sprache und Datenverarbeitung*, *33*(1-2), 169–177.

Wolf, L., Cohen, A., & Simchon, A. (2015). *An experimental investigation of epistemic modal adverbs and adjectives.* (Manuscript available at http://semanticsarchive.net/sub2015/SeparateArticles/Wolf-Cohen-Simchon-SuB20.pdf)

Yalcin, S. (2007). Epistemic modals. *Mind*, *116*(464), 983-1026.

Yalcin, S. (2009). *The language of probability.* (Talk given at U.C. Berkeley)

Yalcin, S. (2010). Probability operators. *Philosophy Compass*, *5*(11), 916–37.

Yanovich, I. (2014). Standard contextualism strikes back. *Journal of Semantics*, *31*(1), 67-114.