How Many are *many*?
Exploring Context-Dependence of *few* and *many* with Probabilistic Computational Models

Dissertation
zur
Erlangung des akademischen Grades
Doktor der Philosophie
in der Philosophischen Fakultät
der Eberhard Karls Universität, Tübingen

vorgelegt von

Anthea Sofie Schöller

aus Ulm

2017

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen

Dekan:	Prof. Dr. Jürgen Leonhardt
Hauptberichterstatter:	Prof. Dr. Gerhard Jäger
Mitberichterstatter:	Prof. Dr. Friedrich Hamm

Tag der mündlichen Prüfung: 19. Juli 2017

Abstract

This dissertation investigates the context-dependence of the quantity words *few* and *many* and the interaction between their linguistic meaning and rich statistical world knowledge. Concretely, one theory by Fernando and Kamp (1996) is explored which makes precise predictions about how contextual information might be integrated in the semantics. This theory assumes that the "surprise reading" of *few* and *many* expresses that a number or a proportion is lower or higher than expected. Context-dependent prior expectations about a relevant quantity are formalized as probability distributions P_E over cardinalities and the cardinalities which count as *few* or *many* are determined by applying fixed, context-independent thresholds θ_{few} and θ_{many} to the cumulative density mass of these distributions. In other words, *few* and *many* comprise a stable core meaning, which explains why speakers and listeners manage to successfully communicate with these context-dependent expressions and how children can acquire proficiency in their use.

Fernando and Kamp's (1996) theory is tested by couching it in a probabilistic model of language use in which the threshold parameters are treated as latent parameters. Their values cannot be directly observed, but are estimated based on experimental data by applying Bayesian inference. In several series of experiments prior expectations are elicited and the production and interpretation of sentences with *few* and *many* are measured. In particular, the cardinal and the proportional reading of the quantity words is examined as well as the effect of overtly marking surprise with adverbs like *surprisingly* or *compared to* constructions.

Keywords: context-dependence, *few*, *many*, prior expectations, computational modeling, Bayesian inference

Zusammenfassung

Diese Dissertation untersucht die Kontextabhängigkeit der Wörter *few* ('wenige') und *many* ('viele') und die Interaktion zwischen deren linguistischer Bedeutung und statistischem Weltwissen. Konkret wird eine Theorie von Fernando und Kamp (1996) vorgestellt, die präzise Vorhersagen darüber macht, wie die Informationen aus dem Kontexts in the Semantik integriert werden. Diese Theorie nimmt an, dass die sogenannte "Überraschungs-Lesart" von *few* und *many* ausdrückt, dass eine Zahl oder Proportion kleiner oder größer ist als erwartet. A priori Erwartungen im jeweiligen Kontext werden als Wahrscheinlichkeitsverteilungen P_E über natürliche Zahlen oder Proportionen formalisiert. Welche Zahlenwerte als *few* oder *many* gelten, wird bestimmt indem feste, kontextunabhängige Grenzwerte θ_{few} and θ_{many} auf die kumulierten Wahrscheinlichkeiten dieser Verteilungen angewendet werden. Mit anderen Worten, *few* und *many* enthalten eine feste Kernbedeutung, die erklären kann, warum Sprecher und Hörer so erfolgreich mit diesen kontextabhängigen Begriffen kommunizieren und wie Kinder ihre Verwendung erlernen können.

Fernando und Kamps (1996) Theorie wird getestet indem sie in ein probabilistisches Modell übersetzt wird, welches die Grenzwerte als latente Parameter betrachtet. Deren Werte können nicht gemessen werden, sondern sie werden mit Bayesianischer Inferenz basierend auf experimentellen Daten geschätzt. In mehreren Versuchsreihen werden a priori Wahrscheinlichkeiten im jeweiligen Kontext und die Produktion und Interpretation von Sätzen mit *few* und *many* gemessen. Insbesondere wird die kardinale und die proportionale Lesart betrachtet, sowie die Modifikation von *few* und *many* mit *surprisingly* ('überraschend') und *compared to* ('im Vergleich zu') Konstruktionen, die die "Überraschungs-Lesart" explizit markieren.

Schlagwörter: Kontext-Abhängigkeit, *few*, *many*, a priori Erwartungen, probabilistische Modelle, Bayesianische Inferenz

Acknowledgements

This dissertation would not exist in its current state without the support of many[1] colleagues, friends and family members. I want to take the opportunity and say thank you.

First of all, I want to thank my advisor Michael Franke. This dissertation is as much your work as it is mine and I am grateful that I could learn so much from you. You introduced me to the power of computational models, you pushed me when it was necessary and you trusted in me in times when I didn't believe that I could succeed. And you made sure that I could focus on my work and was not distracted by other duties. I want to sincerely thank you for your trust and support and for making it a pleasure to be a PhD student in the ProComPrag project. Bayes rulez!

I am also very grateful to have Gerhard Jäger as my advisor. You immediately offered help and support when I asked for it and had my back so that I could focus on writing this dissertation. Furthermore, I would like to thank Fritz Hamm and Britta Stolterfoht for immediately agreeing to join my committee.

Apart from my supervisors, many other people helped me discover my passion for linguistics and especially for semantics. I want to thank Sonja Tiemann and Sigrid Beck for introducing me to the study of language in such an enjoyable, challenging and intellectually rewarding way. It was a pleasure to learn from you how science works.

The stages at which I received help to get this dissertation project started and finished are many. Thank you to Anna Howell and Josh Armstrong for your patience when I was asking for judgments. Thank you to Fabian Dablander, Judith Degen and Erin Benett for showing me how to code my experiments. Thanks to Cosi, Julia, Ann, Josh, Vera, Saskia and Markus for your valuable feedback and for finding the many typos in this dissertation.

I can honestly say that I wouldn't have completed this dissertation if it wasn't for the great company of Team Lambda: Nadine Bade, Uli Bausch, Sigrid Beck, Polina Berezovskaya, Julia Braun, Saskia Brockmann, Sarah Bußmann, Sonja Haas-Gruber, Verena Hehl, Vera Hohaus, Anna Howell, Ann-Cathrin Jurawel, Sabine Lohf, Konstantin Sachs, Cosima Schneider, Isil Senel, Sonja Tiemann and Alexander Wimmer. You made working at this university and learning about semantics great fun and I am honored to not only call you colleagues but friends. I love to remember our Gartenradtouren, Stocherkahnfahrten, Kaffeerunden, Lauftreffs, 100km Staffeln, Kirnbergläufe, X-Bar Trips, Hochzeiten, Promotionsfeiern, Tippspiele, Weihnachtssingen, Mittagessen in Mensa und Park and many many more fantastic events. I hope for many more to come.

[1] Quiz: How many uses of *many* can you spot on these two pages? Find the solution in the introductory chapter!

This dissertation profited greatly from the infrastructure and support of the SFB 833 and the Priority Program XPrag.de. I had the pleasure of becoming a member of a great community of linguists and could participate in retreats and conference trips to Bad Urach, Barcelona, Berlin, Göttingen, Heiligkreuztal, London, Stanford, Utrecht and Wroclaw. In particular, I want to thank the project ProComPrag with Fabian Dablander, Judith Degen, Michael Franke, Gerhard Jäger and Michele Herbstritt. Furthermore, I am grateful to Maribel Romero for inviting me to Konstanz and giving me the opportunity of giving my first invited talk and for discussing my work with me. A very special and wonderful person who can't be praised enough is the world's greatest secretary Beate Starke. You took care of for more things than I probably know and made the Nauklerstraße 35 a place that feels like home. Thank you!!

The project ProComPrag also gave me the opportunity to spend two months at Stanford University. I want to thank the CoCoLab, Judith Degen, Noah Goodman, Erin Bennett, Mike Frank, Lelia Glass, Masoud Jasbi, Justine Kao, Dan Lassiter, Prerna Nadathur, Chris Potts, Ciyang Qing, Greg Scontras and Michael Henry Tessler for welcoming me with open arms and for discussing my work with me. Thank you, Daniela Busse, for hosting me!

But even more people were around to make the three years that it took to complete this dissertation an enjoyable time. I'm thinking of my family, this large group of wonderful people that I can always rely on. Thank you for supporting me in more ways than you can image. Opa Paul, wie gerne hätte ich diesen Moment mit einem großen Eis mit dir gefeiert! And there are also my many friends in Tübingen (Geigerles, Mathematiker, Fitnessboxer, Mitbewohner, Tanzpartner, Stocherkahn Charlotte,...), at home in Weißenhorn, and all around the world. You know who you are and I want you to know that I love you and feel very lucky to have you by my side. Andy, mein allerliebster Andy, dir danke ich von ganzem Herzen für deine Liebe und deine Unterstützung. Jeder Augenblick mit dir an meiner Seite ist wunderbar.

Ich widme diese Arbeit meiner Familie, meiner Mama Edith, meinem Papa Dieter und meinem Bruder Christoph. Ich danke euch für eure bedingungslose Liebe und Unterstützung. Ihr habt mich zu dem Menschen gemacht der ich heute bin und ich kann mich immer auf euch verlassen. Das ist das größte Geschenk und ich danke euch dafür.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The Context-Dependence of *few* and *many*

Context-dependence is a key feature of the quantity words *few* and *many*. How many counts as "many" and how few counts as "few" can vary extremely across contexts. This is why "describing context-dependence can (needless to say) be a tricky matter" (Fernando and Kamp, 1996, 64). This is exemplified in the following sentences:

(1) a. Ben has many siblings.
    b. Chris' team scored many points in the last basketball match.

(2) a. Melanie owns few pairs of shoes.
    b. Few people watched the Olympics this time.

The number of Ben's siblings needed to make (1a) true is much lower than the number of points that are needed to make (1b) true. Similarly, the number of shoes Melanie needs to own for (2a) to be true is much lower than the number of viewers in (2b). Indeed, precise truth conditions seem to be impossible to determine. For this reason, it is a challenge for linguistic theory to explain how speakers and listeners successfully communicate with expressions so context-dependent and vague and how children can acquire proficiency in their use.

In this respect *few* and *many* share the properties of gradable adjectives like *short* and *tall* or *cheap* and *expensive*, which can be equally context-dependent in their positive form (Kennedy and McNally, 2005; von Stechow, 2009). What sets the quantity words apart from gradable adjectives, however, is their syntactic flexibility and broad distribution. They can occur in positions that are usually occupied by quantifiers (few/many people like punting), adjectives (the few/many bars were crowded), or numerals (many/few more punts participated), as pointed out by Solt (2009, 2015). Their occurrence in positions that could be called quantificational, predicative, attributive, and differential is exemplified below.

(3) a. **quantificational**: Many/few people in Tübingen like to go punting on the Neckar.

b. **predicative**: Josh's friends are many/few.

c. **attributive**: The many/few bars in Tübingen were crowded on Saturday night.

d. **differential**: Many more/few more/many fewer than 60 punts participated in the last punting race on the Neckar.

Another contrast between quantity words and gradable adjectives is that *few* and *many* exhibit several readings. The two most prominent readings are the cardinal and the proportional reading.

(4) Cardinal reading

a. Joe ate many burgers.

b. Sue has many friends on Facebook.

c. Andy drank few cups of coffee last week.

d. There are few Trump supporters in California.

(5) Proportional reading

a. Many women out of the 1000 participants who tested the new contraceptive method became pregnant.

b. Many of Mr. Smith's students passed the exam.

c. Few people voted for Trump in D.C. last year.

d. Few of Mr. Smith's students passed the exam.

The cardinal reading of *few* and *many* in (1), (2) and (4) describes a *number* as small or large whereas their proportional reading in (5) expresses that a *proportion* is small or large (Partee, 1989).

Both readings are equally context-dependent. The examples in (4) and (5) show that very different numbers or proportions can count as *few* or *many* depending on the context. (4a) is probably true if Joe ate more than four burgers, whereas Sue in (4b) needs to have more than, say, 700 friends on Facebook to make the sentence true. For cardinal *few*, (4d) can be truthfully uttered even though there are several million Trump supporters in California, but (4c) is only true if Andy drank less than, say, five cups of coffee last week. Similarly, the proportional reading of *many* in (5a) can describe a proportion as little as five per cent as being large, but in (5b) probably more than half of the students need to have passed the exam to count as *many*. Sentence (5c), on the other hand, is true for a proportion of, say, less than twenty per cent, but in a sentence like (5d) proportions up to sixty or seventy per cent could count as *few*.

Examples (5b) and (5d) show *few* and *many*'s context-dependence particularly clearly since - in adequate contexts - the same proportion can truthfully be described as both *few* and *many*.

(6) a. CONTEXT: Mr. Smith's class had to take a very difficult exam, which usually only a very small percentage of participants passes. Contrary to his expectations, 60% of his students passed the exam.

b. STATEMENT: Many of Mr.Smith's students passed the exam.
$\rightsquigarrow$ More students of Mr. Smith's than expected passed the exam.

(7) a. CONTEXT: Mr. Smith's students had to take an exam, for which they were very well prepared. Mr. Smith had expected all of his students to pass, but only 60% of the students passed the exam.

b. STATEMENT: Few of Mr.Smith's students passed the exam.
$\rightsquigarrow$ Fewer students of Mr. Smith's than expected passed the exam.

In both contexts in (6a) and (7a), the proportion of students who passed the exam is 60%. But the statement in (6b) describes this proportion as *many* whereas in (7b) it is described as *few*.

As pointed out above, an unsolved puzzle of the field of linguistics and cognitive science is how exactly context-dependent expressions receive their meaning in context. Numerous attempts have been made at assigning *few* and *many* a fixed semantic contribution, but, given their extreme vagueness and context-dependence, this undertaking turns out to be very difficult. At this point, semantic accounts tend to shift the load of determining the quantity words' concrete denotation to pragmatic theories and have some notion of context fix of what counts as "few" or "many" (more on this in the discussion of the semantics of positive form adjectives and quantity words in Section 2.2). In this dissertation we investigate an analysis of *few* and *many* that goes beyond simply assuming that a context simply outputs *few* and *many*'s meaning and makes more concrete predictions about the integration of the context.

The examples in (6) and (7) bring to light a concept that we assume to be a key factor determining the use of the quantity words *few* and *many*: quantitative expectations about events or cardinalities in the context. Following Clark (1991) and Fernando and Kamp (1996), we take it that the so-called "surprise reading" of *few* and *many* expresses that a cardinality or proportion is lower or higher than expected in the respective context. The idea of employing these prior expectations in the semantics of *few* and *many* seems particularly promising because it describes their relationship with the context using the mathematical concept of probabilities. Probabilities apply well to describe subtle differences in judgments and noisy empirical data, and have turned out to be fruitful in other domains of cognitive science

as well. Probabilistic models can formalize the underlying processes of coming to understand the world, for example in learning concepts, acquiring language, and grasping causal relations (Xu and Tenenbaum, 2007; Frank et al., 2009; Tenenbaum et al., 2011). Moreover, neural nets are used in deep learning algorithms, a branch of machine learning and artificial intelligence research, to simulate the human brain (Schulz, 2017). But also in semantics and pragmatics the role of probabilities is attracting increasing attention (see Goodman and Lassiter (2015), Franke and Jäger (2016) and references therein), as will become evident in the course of this dissertation.

One theory that makes concrete predictions about how the denotation of *few* and *many*'s surprise reading is determined in the context was first suggested by Clark (1991) and formally worked out by Fernando and Kamp (1996). These authors try to identify a *stable core meaning* of these expressions: a complex yet systematic function from contexts to precise denotations. According to this approach, a sentence of the form "Many As are B" is true if the actual number of $n = |\, A \cap B \,|$ is surprisingly high. More precisely, "Many As are B" is true if the actual cardinality $n = |\, A \cap B \,|$ exceeds a fixed threshold θ_{many} on a measure of surprise, which is derived from a contextually supplied measure of *a priori* expectations P_E about likely values of n. Even with a fixed and contextually-stable threshold for what counts as sufficiently surprising, whether a certain n counts as surprisingly high can still vary dramatically for numbers of siblings and points scored during a basketball match, because we may have dramatically different prior expectations P_E. This provides an explanation for the fact that context-dependence and vagueness can be possible despite a systematic, calculable and learnable stable core meaning.

While such a surprise-based semantics may seem like an appealing idea, it also raises methodological concerns. It becomes exceedingly hard to test the predictions of such an account because the precise nature of what counts as surprising is hard to assess based on solitary introspection. In this thesis, we set out to test Clark's (1991) and Fernando and Kamp's (1996) theory with modern methodology. Since neither prior expectations nor threshold values can be estimated based on introspection alone, data about the population's statistical world knowledge are elicited experimentally and the production and interpretation behavior of a large group of subjects will be measured. Based on these data, we seek to demonstrate how data-driven computational modeling can be a helpful addition to the linguist's toolbox, exactly where solitary introspection fails and the theory under scrutiny concerns *latent parameters* that are not directly observable, like a threshold θ_{many} on a measure of surprise. By Bayesian inference, we will estimate plausible values of latent parameters in probabilistic models. We will show how the semantic theories we want to test can be couched in a computational model and explore whether they make the

correct predictions to explain the context-dependence of *few* and *many* in its various uses. In particular, we make use of the models to investigate the cardinal and the proportional reading. Furthermore, we examine whether speaker behavior changes when prior expectations are overtly marked, either by the adverb *surprisingly* or by a *compared to* construction.

As well as learning about the production and interpretation of sentences with *few* and *many*, their semantics and in particular their compositional analysis will be discussed in detail. Based on Romero (2015, 2017), we discuss how a sentence's comparison class can be derived from the available semantic and pragmatic components.

1.2 The Structure of this Dissertation

The focus of this dissertation is on three series of experiments that were conducted to investigate the context-dependence of *few* and *many* and to test Fernando and Kamp's (1996) semantic theory. Before we delve into the details, however, an overview of previous linguistic and psychological work as well as an introduction into the applied methodology is provided. The structure of this dissertation is as follows.

Chapter 2 presents an overview of the many readings of *few* and *many* and discusses factors governing their availability. It presents three semantic analyses which reflect the variety of positions the quantity words can occupy, as exemplified in (3): *few* and *many* are treated on par with quantifiers (Romero, 2015), with adjectives (Romero, 2017) or as degree modifiers (Solt, 2009). After a brief discussion of the differences between *few* and *many*, the surprise-based semantics by Fernando and Kamp (1996) is introduced, which is to be tested experimentally in the following chapters. Before doing so, we present an attempt to derive the quantity words' "surprise reading" and comparison classes compositionally and propose an intensional version of the positive degree operator *POS*. To do so, we are building on Romero's (2015) semantic analysis and Fernando and Kamp's (1996) probabilistic approach to *few* and *many*'s "surprise reading".

Chapter 3 follows up on the linguistic background by presenting relevant psychological studies and experiments on the context-dependence of *many* and *few*. Clark (1991) proposes the use of probabilities to describe language. He suggests representing prior expectations as a probability distribution on which *few* and *many* impose a threshold. An extensive series of experimental studies was produced by Moxey and Sanford (2000) and Sanford et al. (1994), which also identifies prior expectations of the contexts as a factor which influences the use and interpretation of *few* and *many*. Finally, Newstead and Coventry (2000) and Coventry et al. (2005, 2010) investigate

the influence of visual cues on the use of *few* and *many*. One of their experiments will be replicated in Chapter 5 to investigate the role of prior expectations in a visually displayed context.

In Chapter 4, relevant terminology and concepts of computational modeling and Bayesian inference will be explained. In the subsequent chapters, we demonstrate how Bayesian inference in connection with data-driven computational modeling can be a helpful tool for learning about theories which are hard to grasp by solitary introspection and which concern latent parameters that are not directly observable.

Chapter 5 focuses on the cardinal reading of *few* and *many*. It sets off with a replication of Newstead and Coventry's (2000) experiment to investigate the influence of various visual factors on the use of the quantity words. We test whether these effects of visual presentation can be reduced to an explanation in terms of expectations. Next, we show how Fernando and Kamp's (1996) theory can be couched in a probabilistic, computational model. This demonstration builds on the proceedings paper by Schöller and Franke (2015) and was elaborated by Schöller and Franke (2017a). In three experiments we gather data on participants' prior expectations and their production and interpretation behavior of *few* and *many* in 14 contexts. This series of experiments was also presented in Schöller and Franke (2016) and Schöller and Franke (2017a). We demonstrate how the computational model and Bayesian inference can be applied to test the theory, which assumes as the semantic meaning of *few* and *many* a fixed pair of threshold values on a distribution representing prior expectations of the context.

Chapter 6 follows up on Chapter 5 by investigating constructions which mark the "cardinal surprise reading" of *few* and *many* overtly. We explore the role of the adverb *surprisingly* in combination with the quantity words and test whether it functions as an intensifier or just marks that surprise is expressed. *Surprisingly* is compared with the intensifying adverb *incredibly* and a *compared to* construction which openly addresses expectations. A judgment task is conducted and a variant of the computational model from the previous chapter is used to analyze the data. This chapter builds on a proceedings paper by Schöller and Franke (2017b). Additionally, we present an interpretation task as a follow-up experiment and complement the analysis of the data with further Bayesian methods.

The proportional reading of *few* and *many* is the subject of Chapter 7. An interpretation experiment tests whether proportional *few* and *many* can be accounted for by a fixed, context-independent threshold on proportions and finds that this is not the case. The proportional reading is influenced by expectations of the context, too. These findings go back to a proceedings paper by Schöller and Franke (2016). To further investigate this reading, a series of experiments in both real-world and very abstract contexts is conducted to collect information on prior expectations about

cardinalities and production data. We find that the proportional reading can both express that a proportion is numerically small or large or surprisingly small or large. We propose a computational model which is an extension of the model incorporating Fernando and Kamp's (1996) theory from Chapter 5. This model assumes that the contextual contribution required for the proportional reading is two-fold. The first is an uninformed, uniform belief about proportions and the second are informed prior expectations about likely proportions based on world knowledge. We propose a linear combination model which incorporates the assumption that the amount of world knowledge employed depends on its salience in the context. We test again whether a fixed pair of threshold values on prior expectations can explain the use of *few* and *many*.

In the final chapter, Chapter 8, the dissertation's main findings are summarized and we will discuss what they contribute to a theory of context-dependence. We conclude with open questions and interesting issues for future work.

Chapter 2

Linguistic Background of Quantity Words

The goal of this thesis is to explain the interaction between context-dependent expressions like *few* and *many* and the context. Before we start delving into the details of this undertaking, it is wise to get an overview of the general properties of the object of interest. A vast body of semantic literature on *few* and *many* has evolved which describes and explores the available readings of sentences with *few* and *many*, their semantic properties as well as their context-dependence. Building on the findings of previous research is inevitable because we can only learn about the pragmatics of *few* and *many*, how they are produced and interpreted across contexts, when we understand their semantic and syntactic properties.

Section 2.1 provides an overview of the many readings which a sentence with *few* and *many* can express and discusses the factors which allow or prevent their availability. The lexical semantics of *few* and *many* is discussed in Section 2.2. We will see that there is controversy in the literature about how to classify them. They have been claimed to share the semantic properties of quantifiers (Westerståhl, 1985; Partee, 1989), parametrized determiners (Hackl, 2000; Romero, 2015) and adjectives (Hackl, 2009; Dobrovie-Sorin, 2013) and also to be semantically empty gradable quantifiers over degrees (Rett, 2008; Solt, 2009). To avoid terminology which commits to one of the theories, *few* and *many* will be labeled "quantity words" from now on (cf. Rett, 2008). Sections 2.2.1, 2.2.2 and 2.2.3 introduce the approaches we consider to be most insightful, show how truth-conditions can be derived compositionally in the respective semantics and discuss strengths and weaknesses of each theory. Section 2.2.4 describes characteristics of *few*.

One issue that is left out in the cold by all of the competing semantic analyses from Section 2.2 is how exactly *few* and *many* interact with the context. Fernando and Kamp (1996) address this open issue and spell out a semantic theory which makes the truth conditions of *few* and *many* dependent on prior expectations. Their

approach is presented in Section 2.3 and will be tested experimentally and extended in Chapters 5, 6 and 7. A related open issue of the semantics of context-dependent expressions is how to derive their context-dependence, more concretely their comparison classes and the related prior expectations, compositionally. The difficulties of this undertaking and first steps towards a solution of the problem are presented in Section 2.4.

2.1 The Many Readings of *few* and *many*

Sentences with *few* and *many* can be ambiguous between different readings. In this section we introduce three readings which received a lot of attention in the linguistic literature. Section 2.1.1 discuses the cardinal and proportional reading of *few* and *many*, Section 2.1.2 the reverse proportional reading and Section 2.1.3 presents factors which influence which of the readings is dominant.

2.1.1 The Cardinal and the Proportional Reading

The most prominent readings were famously distinguished by Partee (1989) (and a long tradition thereafter) and labeled the *cardinal* and the *proportional* reading. They are exemplified in (8) and (10).

(8) a. There are few nightclubs in Tübingen.
b. Joe ate many burgers at the barbecue.

The sentences in (8) exhibit cardinal readings. (8a) expresses that only a small number of nightclubs exists in Tübingen. (8b), on the other hand, asserts that the number of burgers eaten by Joe is considered large.

Partee (1989) suggests that *many*'s cardinal reading has a meaning "like that of the cardinal numbers, *at least* $[x_{min}]$, with the vagueness located in the unspecified choice of $[x_{min}]$, it being part of the meaning of *many* that the value of $[x_{min}]$ must be one that counts as large in the given context. The cardinal reading of *few* is similar except that it means *at most* $[x_{max}]$, and $[x_{max}]$ is generally understood to be small" (Partee, 1989, 383)[1]. Simple truth conditions for a sentence of the form "Few/Many A are B" under a cardinal reading are given in (9) (Partee, 1989, 383).

(9) **Cardinal reading**
a. *Few*: $|A \cap B| \leq x_{\max}$
b. *Many*: $|A \cap B| \geq x_{\min}$

[1]Partee (1989) labels both threshold values as n. For consistency with the theory proposed in Section 2.3 we use x_{max} and x_{min} instead.

Under a semantics as in (9a), sentence (8a) is true if the number of nightclubs in Tübingen is smaller than x_{max}. (8b) is true if the number of burgers eaten by Joe is larger than x_{min}.

An example of the proportional reading of *few* and *many* is given in (10).

(10) a. Few children like spinach.

b. Many of my students passed the exam.

Sentence (10a) states that the proportion of children who like spinach is low whereas (10b) expresses that the proportion of the speaker's students who passed the exam is large.

Partee (1989) describes the threshold of the proportional reading "as a fraction between 0 and 1 or as a percentage" (Partee, 1989, 384). Truth conditions of "Few/Many A are B" under a proportional reading could look as in (11).

(11) **Proportional reading**

a. *Few*: $|A \cap B| : |A| \leq k_{\max}$

b. *Many*: $|A \cap B| : |A| \geq k_{\min}$

For *few*, sentence (10a) is true if the proportion of children who like spinach is not greater than $k_{\max}$. The sentence in (10b) is true if the proportion of the speaker's students who passed the exam is at least $k_{\min}$.

Partee (1989) suggested that the cardinal and proportional reading correspond to two lexically distinct meanings of *few* and *many*. The *lexical ambiguity theory* is supported by a scenario as in (12a) and sentences (12b) and (12c), where we find a truth-conditional difference between the readings (cf. Partee, 1989; Romero, 2015).

(12) a. Scenario: All the faculty children attended the 1980 picnic, but there were few faculty children back then. Almost all faculty children had a good time.

b. Few faculty children attended the 1980 faculty picnic.

c. Many (of the) faculty children had a good time.

In this scenario, (12b) expresses that the number of faculty children who were present at the picnic is small, regardless of the fact that all children attended. The sentence is true under a cardinal but not under a proportional reading because proportional *few* "certainly never means 'all'... The cardinal reading, on the other hand, is quite compatible with few being all, since it asserts the number of [children] that satisfy the predicate is small without saying anything about what proportion of the set of [children] that is" (Partee, 1989, 391). Partee's (1989) line of reasoning can be made clearer when spelling out the truth conditions of the sentences in terms of the semantics of *few* in (9) and (11). The cardinal reading of *few* predicts the

sentence to be true if the *number* of kids attending is smaller than a contextually provided threshold $x_{\max}$: $|\,\text{faculty kids at picnic}\,| \leq x_{\max}$. According to the context in (12a), these truth conditions are fulfilled, rendering the sentence true under a cardinal reading. For the proportional reading of *few*, however, the *proportion* of kids attending must be smaller than a certain threshold $k_{\max}$. But the proportion of attending faculty kids of all faculty kids does certainly not fulfill this condition, since all faculty kids attended; $|\,\text{faculty kids at picnic}\,| : |\,\text{faculty kids}\,| = 1$. Consequently, sentence (12b) is not true under a proportional reading of *few*, but true under a cardinal reading. Partee (1989) concludes that these different truth conditions require two distinct lexical entries. In contrast, (12c) is true under a proportional but not under a cardinal reading because even though a large proportion of the children had fun, their overall number was small. The truth-conditional difference between the cardinal and the proportional reading in this scenario supports Partee's (1989) ambiguity hypothesis of *few* and *many*, which is advocated also by Westerståhl (1985); Cohen (2001) and Krasikova (2011). Further factors which differentiate between a cardinal and proportional interpretation of *few* and *many* are presented in Section 2.1.3.

2.1.2 The Reverse Proportional Reading

Besides the cardinal and proportional readings, Westerståhl (1985) claims that there is an additional reading of *few* and *many*, the *reverse proportional reading*.

(13) a. Scenario: 14 out of a total of 81 winners of the Nobel Prize in literature come from Scandinavia.

b. Many Scandinavians have won the Nobel Prize in literature.

c. Paraphrase: Many winners of the Nobel Prize in literature are Scandinavians.

Westerståhl (1985) suggests that sentence (13b) has a reading paraphrasable as (13c). This reading is not accounted for by the proportional semantics of *many* in (11b) because 14 Scandinavians are not enough to constitute a large proportion of the roughly 15 million Scandinavians. *Many*'s arguments need to be reversed to arrive at the desired truth conditions.

The same argument has been made for *few* by Herburger (1997) and Cohen (2001). Herburger (1997) makes the strong claim that for a scenario as in (14a), both a proportional reading and a cardinal reading are ruled out, because neither the total number of applicants nor the total number of cooks nor the total number of cooks who applied is known.

(14) a. Scenario: "The fellowship committee is sorting through the applications for travel funding to Paris. Without knowing how many applicants there are, at an early point during the review process they observe that on average only every twentieth application was sent in by a cook, which is a much lower percentage than they had anticipated." (Herburger, 1997, 61f)

b. Few cooks applied.

c. Paraphrase: Few applicants were cooks.

Truth conditions which account for the reverse proportional are given below:

(15) **Reverse proportional reading**

a. *Few*: $|A \cap B| : |B| \leq k_{\max}$

b. *Many*: $|A \cap B| : |B| \geq k_{\min}$

This third reading is problematic for semantic theory no matter whether *few* and *many* are classified as a quantifier or as an adjective (Romero, 2017). These problems and a solution proposed by Romero (2015, 2017) are discussed in more detail in Sections 2.2.1 and 2.2.2, which present the respective semantic theories.

2.1.3 Factors Influencing the Availability of the Readings

Before we move on to review the competing semantic analyses, we want to inspect the factors which influence the availability of the readings of *few* and *many*. From a pragmatic point of view, the context has to provide enough information about the cardinalities which *few* and *many* are meant to describe. What marks the proportional reading off from the cardinal reading is a difference in scale structure. The key characteristic of the proportional reading is the existence of an upper bound on $|A|$ or $|B|$.

In terms of the semantic environment, Partee (1989) and Herburger (1997) follow Milsark (1977) in relating cardinal and proportional *few* and *many* to general properties of the distribution of determiners. "Milsark argues that some restrictions normally posed in terms of definiteness could be better explained on the basis of a classification of determiners as 'weak' or 'strong' " (Partee, 1989, 387). Weak determiners include the indefinite determiner *a*, unstressed *some*, cardinal numbers, *a few*, and *no* and fulfill the symmetry property (see below). Strong determiners include the definite determiner *the*, *all*, *most*, and *neither* (Barwise and Cooper, 1981; Partee, 1989). In contrast to other determiners, *few* and *many* can be attributed to both determiner classes, depending on their reading. Cardinal *few* and *many* are classified as weak determiners whereas proportional *few* and *many* count as strong.

Only weak determiners can occur in *there*-existentials.

(16) a. **weak**: There are few/many/some[2]/no children in the garden.

b. **strong**: *There is/are few (of the)/many (of the)/every/each/all children/child in the garden[3].

Furthermore, when *few* and *many* are used in explicitly adjectival positions, in which they are preceded by the definite determiner, for example, only the cardinal reading is available (Partee, 1989). Cardinal *many*, for example, can be taken to mean "large in number" and rather patterns with an adjective (compare (17b) and (17c)). It characterizes the NP, but it does not contribute quantificational force (Herburger, 1997).

(17) a. The few women at the party had lots of fun.

b. The many children in the park enjoyed the sunshine.

c. The numerous children in the park enjoyed the sunshine.

d. *The few/many of the guests are from Bavaria.

Strong DPs pattern with definites in that they denote a small percentage of the NP (Herburger, 1997, 55). The denotation of strong DPs is quantificational and not adjectival because they do not express an intersective property. Instead, they contribute how two sets are combined with each other. This can be exemplified with the symmetry property, as spelled out by Solt (2009, 7) (first formally, then more intuitively).

(18) a. A determiner Det is symmetric iff for all $A, B: \ B \in Det(A)$ iff $A \in Det(B)$.

b. DetAs are B iff DetBs are A.

This property holds for weak, cardinal *few* and *many* and also for number words as in (19a) and (19b) but not for strong, proportional *few* and *many* or for *all* as in (19c) and (19d).

(19) a. $\text{Many}_{\text{card}}$ women were at the party. $\Leftrightarrow$ $\text{Many}_{\text{card}}$ guests at the party were women.

b. Five women were at the party. $\Leftrightarrow$ Five guests at the party were women.

c. Few_{prop} women are great-grandmothers. $\not\Leftrightarrow$ Few_{prop} great-grandmothers are women.

d. All great-grandmothers are women. $\not\Leftrightarrow$ All women are great-grandmothers.

Strong determiners are not symmetrical since their semantics specifies exactly how the two sets they quantify over are related with each other. Proportional *few* and *many* are thus classified as strong.

[2] the weak, unstressed version of *some*.

[3] Ungrammatical or marked sentences are marked with * or ??.

	individual level predicate	**stage level predicate**
example	Few Chinese have blue eyes.	Few students were hungry for lunch.
	permanent	temporary
reading	proportional	proportional and cardinal

Table 2.1: Readings of *few* and *many* with ILPs and SLPs

In terms of the types of predicates *few* and *many* combine with, only the proportional reading is available when *few* and *many* occur in the subject position of individual level predicates (ILP), as in (20).

(20) a. Many of my friends are smokers.

b. Few Chinese have blue eyes.

(20a) exemplifies another property of proportional *few* and *many*: they can be accompanied by the partitive construction *of the* which makes the subset-superset relation overt. The subject position of a stage-level predicate (SLP) allows both readings.

(21) Few kids were hungry for lunch because of the big breakfast they had.

Table 2.1 presents a brief overview. For a more formal characterization of weak and strong determiners, I refer the reader to Barwise and Cooper (1981) or Partee (1989).

Which reading is available is also influenced by information structure. Focus influences the *comparison class* in relation to which *few* and *many* receive their meaning. A comparison class is a set of objects that are similar in some way to whatever is being discussed. "In many cases, the comparison class is just the set of things that the participants in a conversation happen to be talking about at a given time. In formal terms, a comparison class is a subset of the universe of discourse which is picked out relative to a context of use" (Klein, 1980, 13). In (22a), focus on Joe has the effect that the number of burgers consumed by Joe is compared to the number of burgers eaten by *other relevant people*, like other guests at the barbecue. In contrast, (22b) expresses that the number of burgers consumed by Joe is large as compared to *other food* that he might have eaten, like hot dogs, sandwiches or muffins.

(22) a. JOE ate many burgers at the barbecue.

b. Joe ate many BURGERS at the barbecue.

For our purposes, *focus* is used to mark a constituent off from relevant alternatives (Schwarzschild, 1997). The next section introduces how focus marking is analyzed semantically. In the compositional analyses we will treat focus marking on par with contrastive topic marking. Partee (2010) illustrates these two concepts with

an example and points out that not everything with "intonational prominence" is focus.

(23) a. Where do your sons live?

b. Well, [my oldest son]$_{\text{CT}}$ lives in [Massachusetts]$_{\text{F}}$, [my middle son]$_{\text{CT}}$ lives in [Alaska]$_{\text{F}}$, and [my youngest son]$_{\text{CT}}$ lives in [Salt Lake City, Utah]$_{\text{F}}$.

The answer's *topic* is 'my sons', and the subject of each clause is a *contrastive topic*, because different answers are necessary for different sons. If they all lived in Massachusetts, the answer could begin with 'My sons live...', using a simple topic phrase with no intonational prominence (Partee, 2010, 4).

Moreover, Herburger (1997) claims that the reverse proportional reading is only available if the quantity word's first argument is focused. She calls these readings "focus-affected" because she claims that in cases like (24) it is focus rather than syntax that determines the order in which the quantity word applied to its arguments.

(24) a. Many SCANDINAVIANS have won the Nobel Prize in literature.

b. Few COOKS applied.

Finally, another linguistic construction that influences the readings of *few* and *many* and especially their comparison class are frame setters like *for-* and *compared to*-phrases. These phrases are called frame setters because they "set the frame" for the matrix clause and contribute its comparison class. They denote comparison classes which "affect the standard involved in the semantics of positive forms of gradable adjectives" (Bylinina, 2014, 143) and they presuppose that the subject of the gradable predicate (a gradable adjective or a quantity word) is included in the comparison class set (Kennedy, 2007; Schwarz, 2010). [4]

(25) a. There are few cars in the car park for a Monday evening.

b. Compared to the other kids, Jimmy ate many muffins.

c. *For a small dog, our cat Billy catches many mice.

The *for*-phrase in (25a) evokes a comparison between the number of cars on today's Monday evening and the number of cars typically present on Monday evenings. Furthermore, it has the effect that the sentence can be uttered meaningfully only on a Monday evening, otherwise we run into a presupposition failure. Similarly, (25b) triggers a comparison with a group of relevant kids. (25c) exemplifies a presupposition failure because the cat Billy is not a member of the comparison class consisting of small dogs. In the experiments presented in Chapters 5, 6 and 7, frame setters will be used to make sure that participants produce and interpret the experimental

[4] A gradable adjective is an adjective that occurs in comparison constructions (for example *tall, taller, tallest*). It is semantically treated as making reference to degrees on a scale of a particular dimension like *height, weight, beauty or cardinality* (cf. Beck, 2011).

test sentences with respect to the same comparison class. In particular Chapter 6 will compare *for*- and *compared to*-phrases.

2.2 The Semantics of *few* and *many*

As already pointed out in the introduction to this chapter, there is controversy about how to classify *few* and *many* semantically. There are by now three prominent semantic treatments. *Few* and *many* have been attributed to share the semantic properties of a

- **quantifier** (Barwise and Cooper, 1981; Westerståhl, 1985; Partee, 1989; Herburger, 1997; Heim and Kratzer, 1998; Hackl, 2000; Romero, 2015)
- **adjective** (Partee, 1989; Hackl, 2009; Krasikova, 2011; Dobrovie-Sorin, 2013; Romero, 2017)
- semantically empty **degree quantifier** (Rett, 2008, 2016; Solt, 2009, 2015)

These accounts and their characteristic features will be summarized in the following sections. We will show that each of them generates the desired truth conditions, but comes with different problems, as pointed out above. Furthermore, we will see that none of them commits to how the standard of comparison or threshold is determined in relation to the context. To anticipate the remainder of this chapter, Romero's (2015) quantifier semantic account in Section 2.2.1 (which is formally very similar in spirit to Romero's (2017) adjectival semantics in Section 2.2.2) will be picked up again in Section 2.4 in which we risk a first attempt to systematically derive prior expectations and to formalize their interaction with the sentence's comparison class. Solt's (2009) is introduced to present an account which can uniformly account for the quantity words' many occurrences and positions, as pointed out in the introduction of this thesis and repeated in Section 2.2.3.

What all of the competing analyses share, however, is the essential meaning contribution: *few* expresses a small cardinality, *many* a large cardinality. A degree-semantic framework has been developed to formalize the intuition that cardinalities come about by counting, can be compared with each other and therefore be ordered on a scale. In what Beck (2011) calls the "standard theory" of comparison constructions, gradable adjectives like *tall* are taken to relate individuals to degrees on a scale (more on scales below).

(26) ⟦tall⟧: x is tall to degree d

The simplified semantics of *tall* in (26) then expresses that an individual has the property of reaching a certain degree on a height scale. This degree can be compared

to other degrees on the scale, for example with the comparative operator *-er* (Beck, 2011, 5).

(27) a. Andy is taller than Anthea.

b. ⟦-er⟧: the degree matrix clause > the degree than-clause
⇝ Andy's height > Anthea's height

The degree semantics literature assumes that not only comparatives are composed of a gradable predicate and a comparison operator *-er*, but that also other comparison constructions are made up of a gradable predicate and a comparison operator (von Stechow, 1984; Heim, 1999; Beck, 2011). For example, superlatives decompose into a gradable predicate and the superlative operator *-est* and equatives come with the equative operator *as...as*. Interestingly, these comparison operators do not always have to be overt. The frequently used positive form of the adjective does not immediately suggest that a comparison is made at all.

(28) a. Andy is tall.

b. Michael's friends are many.

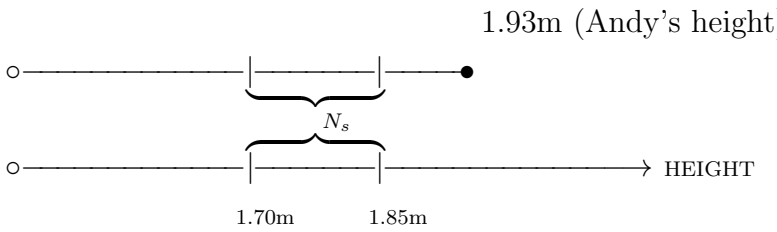
Nevertheless, even in cases as in (28), the height described by *tall* or the cardinality described by *many* are related to other relevant degrees in the context and compared to what counts as normal in this case. For example, (28a) is interpreted to express that Andy's height is taller than the average height and (28b) expresses that Michael has more friends than other people. This comparison to a "standard value" on the scale is assumed to be contributed by a null morpheme *POS* which binds the gradable predicate's degree argument and makes reference to a neutral interval on the respective scale (von Stechow, 1984). A scale S, in turn, is a triplet $\langle D, <, DIM\rangle$ consisting of a set of degrees D, an ordering relation on that set $<$ and a dimension of measurement DIM. Dimensions include cardinality, length, height, duration, volume, and weight, for example (Solt, 2015, 231f).

(29) $[\![POS]\!] = \lambda \mathrm{C}_{\langle dt,t\rangle}.\lambda D_{\langle dt,t\rangle}.L_{\langle\langle dt,t\rangle,\langle d,t\rangle\rangle}(\mathrm{C}) \subseteq D$

The *POS* operator is a quantifier over degrees. The variant used here (cf. Romero, 2015) takes as its first argument a set of sets of degrees C which corresponds to the comparison class. The comparison class contains the sets of degrees of focus/contrastive topic alternatives which are evaluated by the $\sim$ operator (more below). These alternatives are also called the focus/contrastive topic associates of *POS*. For our purposes it does not make a difference which kind of marking the constituent carries (see Romero, 2015). C is input to the function L which returns the so-called neutral interval $N_s = L([\![\mathrm{C}]\!])$ of the comparison class. For a sentence as in (28a) and focus on Andy, the comparison class contains the heights of other relevant people, say men in Germany. The neutral interval returned by L would

then contain those degrees on the height scale which count as neither tall nor short. Andy's height formalized as a set of degrees as well (not as a single degree); it is the set of all degrees below his actual, maximal height. The idea is illustrated in the example below.

(30)



The relevant scale is the height scale and the corresponding degrees are indicated in meters and centimeters. The height scale is illustrated by the bottom line in the graphic. Let's assume for the sake of the example that the neutral interval of the heights of adult men in Germany is $[1.70, 1.85]$. Sentence (28a) is mapped to true if Andy's height is greater than those heights which count as neutral (or which are the endpoints of the interval), that means when the neutral interval is fully contained in the interval of Andy's heights (the top line). In other words, Andy can be called "tall" in this context if he is at least 1.86m tall. Since Andy's height is 1.93m, the sentence can be truthfully uttered in this context. See Beck (2011), von Stechow (2009), Kennedy (2007) and references therein for a more thorough introduction into degree semantics and the at times elusive semantics of the positive operator.

The decomposition of the positive form of gradable adjectives into their stem and the degree operator *POS* is transferred to *few* and *many* (Solt, 2009; Hackl, 2009; Romero, 2015, 2017):

(31) a. *few* = FEW + *POS*

b. *many* = MANY + *POS*

POS is responsible for the quantity words' interaction with the context and will be central to the compositional analysis in all three semantic theories of *few* and *many*.

We mentioned above that *few* and *many* are evaluated relative to a comparison class. The comparison class contains the sets of degrees of focus/contrastive topic alternatives which are evaluated by the $\sim$ operator. This idea goes back to the *theory of focus* by Rooth (1992) and neatly summarized by Beck (2006). In the two examples below, focus is used to mark contrast. Contrastive focus can occur in a discourse in which one speaker objects the other. Semantically, focus introduces alternatives. We will assume the same for constituent marked as a contrastive topic.

(32) a. Hayley invited Ross.

b. RYAN$_F$ invited Ross.

(33) a. Ryan invited Harriet.

b. Ryan invited ROSS_F.

Sentences (32b) and (33b) are associated with the structures in (34a) and (34b). According to Rooth (1992), an operator $\sim$ (the squiggle operator) and a variable C_i adjoin to the structure labeled α. C_i is called the focus anaphor and will eventually contain the relevant focus alternatives. In the following sections, this variable will be the semantic representation of the sentence's comparison class.

(34) a. $[[_\alpha\, [_F$ RYAN] invited ROSS] $\sim$C1]

b. $[[_\alpha$ Ryan invited $[_F$ROSS]] $\sim$C2]

Rooth (1992) associates two different semantic objects with α, its ordinary semantic value $[\![\alpha]\!]_0$ and its focus semantic values $[\![\alpha]\!]_f$. Note that the sentences (32b) and (33b) have the same ordinary semantic value, but differ in their focus semantic value, due to the different focus marking (Beck, 2006).

(35) a. $[[_\alpha\, [_F$ RYAN] invited Ross] $\sim$C1].

b.
$$\begin{aligned} [\![\alpha]\!]_0 &= \lambda w.\ \text{Ryan invited Ross in w.} \\ &= \text{that Ryan invited Ross.} \\ [\![\alpha]\!]_f &= \lambda w.\ \text{x invited Ross in w} \mid x \in \text{D} \\ &= \{\text{that Ryan invited Ross, that Hayley invited Ross, ...}\} \end{aligned}$$

(36) a. $[[_\alpha$ Ryan invited $[_F$ROSS]] $\sim$C2].

b.
$$\begin{aligned} [\![\alpha]\!]_0 &= \lambda w.\ \text{Ryan invited Ross in w.} \\ &= \text{that Ryan invited Ross.} \\ [\![\alpha]\!]_f &= \lambda w.\ \text{Ryan invited y in w} \mid y \in \text{D} \\ &= \{\text{that Ryan invited Ross, that Ryan invited Harriet, ...}\} \end{aligned}$$

The $\sim$ operator introduces a presupposition requiring that the context provides at least one proper focus alternative to the proposition that is asserted, ie. an element differing from the ordinary semantic value of the focused phrase with respect to the accented item (Umbach, 2001). The $\sim$ operator "does not determine the interpretation of the variable C_i uniquely, but it does constrain it. It basically says that whenever you have a sentence with something focused in it, its presupposed that there is some relevant set of alternatives in the context" (Partee, 2010).

(37) $[\![\ [\alpha \sim \text{C}_i]]\!]^g_0$ is only defined if $[\![\text{C}_i]\!] \subseteq [\![\alpha]\!]^g_f$ & $[\![\text{C}_i]\!] \neq [\![\alpha]\!]^g_0$.
If defined, $[\![\ [\alpha \sim \text{C}_i]]\!]^g_0 = [\![\alpha]\!]^g_0$

For the examples in (34a) and (34b), C_1 and C_2 differ in whether the relevant alternatives are inviters or invitees.

In the following three sections, three standard semantic treatments of *few* and *many* are introduced under the decomposition assumption in (31). The lexical

entries as well as positive and negative features of each account are summarized in Table 2.2. We assume the compositional rules of Functional Application (FA), Predicate Modification (PM) and Predicate Abstraction (PA) by Heim and Kratzer (1998). Where additional rules are necessary, we point them out explicitly. Section 2.2.4 will briefly review previous work on the negativity of *few*. Section 2.3 introduces the surprise-based semantics of Fernando and Kamp (1996), which makes a concrete proposal of how to calculate the neutral interval across contexts. The challenge of a formal integration of prior expectations and comparison classes into the compositional analysis is subject to Section 2.4.

2.2.1 Quantifier Semantics

The quantifier account is probably the most common analysis of *few* and *many*. The quantificational analysis in its most standard form goes back to the Generalized Quantifier Theory by Barwise and Cooper (1981). *Few* and *many* are treated as "quantifying determiners" that express relationships between two sets of individuals (type $\langle et, \langle et, t\rangle\rangle$), similar to *some, lots of* and *all*. They specify "that the cardinality of their intersection exceeds (*many*) or falls short of (*few*) some standard determined by the context" (Solt, 2015, 225). Following Partee (1989) as in the previous section, the semantics of *few* and *many* look as in (38) and (39).

(38) Cardinal reading

a. $[\![\text{few}]\!]_{\langle et, \langle et, t\rangle\rangle} = \lambda P_{\langle e,t\rangle}.\lambda Q_{\langle e,t\rangle}.\,|\,P \cap Q\,| \leq x_{\max}$

b. $[\![\text{many}]\!]_{\langle et, \langle et, t\rangle\rangle} = \lambda P_{\langle e,t\rangle}.\lambda Q_{\langle e,t\rangle}.\,|\,P \cap Q\,| \geq x_{\min}$

(39) Proportional reading

a. $[\![\text{few}]\!]_{\langle et, \langle et, t\rangle\rangle} = \lambda P_{\langle e,t\rangle}.\lambda Q_{\langle e,t\rangle}.\,|\,P \cap Q\,| : |\,P\,| \leq k_{\max}$

b. $[\![\text{many}]\!]_{\langle et, \langle et, t\rangle\rangle} = \lambda P_{\langle e,t\rangle}.\lambda Q_{\langle e,t\rangle}.\,|\,P \cap Q\,| : |\,P\,| \geq k_{\min}$

In the previous section, we already pointed out that there is a third attested reading of *few* and *many*. The "reverse proportional reading" reverses the order of the arguments of the quantity words.

(40) Reverse proportional reading

a. $[\![\text{few}]\!]_{\langle et, \langle et, t\rangle\rangle} = \lambda P_{\langle e,t\rangle}.\lambda Q_{\langle e,t\rangle}.\,|\,P \cap Q\,| : |\,Q\,| \leq k_{\max}$

b. $[\![\text{many}]\!]_{\langle et, \langle et, t\rangle\rangle} = \lambda P_{\langle e,t\rangle}.\lambda Q_{\langle e,t\rangle}.\,|\,P \cap Q\,| : |\,Q\,| \geq k_{\min}$

The reverse proportional reading poses a problem for semantic theory because it violates a universal property that the Generalized Quantifier Theory ascribes to all quantifiers: reverse proportional *few* and *many* with a semantics as in (40) are not conservative. Conservativity is defined in (2.2.1) (Barwise and Cooper, 1981; Keenan and Stavi, 1986).

(41) A function $f \in D_{\langle et, \langle et, t\rangle\rangle}$ is conservative iff for any P and $Q \in D_{\langle e, t\rangle}$:
$f(P)(Q) = 1$ iff $f(P)(P \cap Q) = 1$

For example, *some* is a conservative quantifier because "some books are new" iff "some books are new books". This does not hold for reverse proportional *few* and *many* as exemplified in (14) and (13). It is not the case that "Few cooks applied" in (14) (in the sense that few applicants are cooks) iff "Few cooks are applying cooks". Equally, it is not true that "Many Scandinavians won a Nobel Prize in Literature" in (13) (with the reverse proportional reading that many winners of a Nobel Prize in literature are Scandinavian) iff "Many Scandinavians are Nobel Prize winning Scandinavians".

A solution for the problem is brought forward by Romero (2015), who builds on Hackl (2000), Heim (1999) and Schwarz (2010). She analyzes *few* and *many* as parametrized determiners (type $\langle d, \langle et, \langle et, t\rangle\rangle\rangle$), which have a degree argument and decompose into the stem and *POS* (as illustrated in example (44) below). This semantics allows her to derive the truth conditions of all three readings in a compositional way while preserving conservativity. There is only one proportional lexical entry and the difference between the regular proportional and the reverse proportional reading is due to a scopally-mobile *POS* which can associate with material external or internal to the original host NP (cf. Heim, 1999; Schwarz, 2010). Romero's (2015) lexical entries for *few* and *many* are defined as follows:

(42) Romero's (2015) cardinal reading

a. $[\![\text{few}_{card}]\!]_{\langle d, \langle et, \langle et, t\rangle\rangle\rangle} = \lambda d_d.\lambda P_{\langle e, t\rangle}.\lambda Q_{\langle e, t\rangle}.\, |\, P \cap Q\, | \leq d$

b. $[\![\text{many}_{card}]\!]_{\langle d, \langle et, \langle et, t\rangle\rangle\rangle} = \lambda d_d.\lambda P_{\langle e, t\rangle}.\lambda Q_{\langle e, t\rangle}.\, |\, P \cap Q\, | \geq d$

(43) Romero's (2015) proportional reading

a. $[\![\text{few}_{prop}]\!]_{\langle d, \langle et, \langle et, t\rangle\rangle\rangle} = \lambda d_d.\lambda P_{\langle e, t\rangle}.\lambda Q_{\langle e, t\rangle}.\, |\, P \cap Q\, | : |\, P\, | \leq d$

b. $[\![\text{many}_{prop}]\!]_{\langle d, \langle et, \langle et, t\rangle\rangle\rangle} = \lambda d_d.\lambda P_{\langle e, t\rangle}.\lambda Q_{\langle e, t\rangle}.\, |\, P \cap Q\, | : |\, P\, | \geq d$

The relationship between *many* and the context is established by having the positive operator *POS* bind *many*'s degree argument. We assume here that the associate of *POS* bears focal stress or functions as a contrastive topic. *POS* can associate with a constituent internal or external to the host NP, thereby deriving the different readings (cf. Romero, 2015). An external associate results in the regular reading, an internal associate in the reverse reading.

To demonstrate these semantics at work, a compositional analysis of each reading is carried out with the semantics of *many* from above. An LF for each sentence as well as the most important steps in the calculation are provided. First, the analysis of the cardinal reading of *many* is demonstrated. Too keep the analysis concise, only the most important composition steps are spelled out. The respective nodes

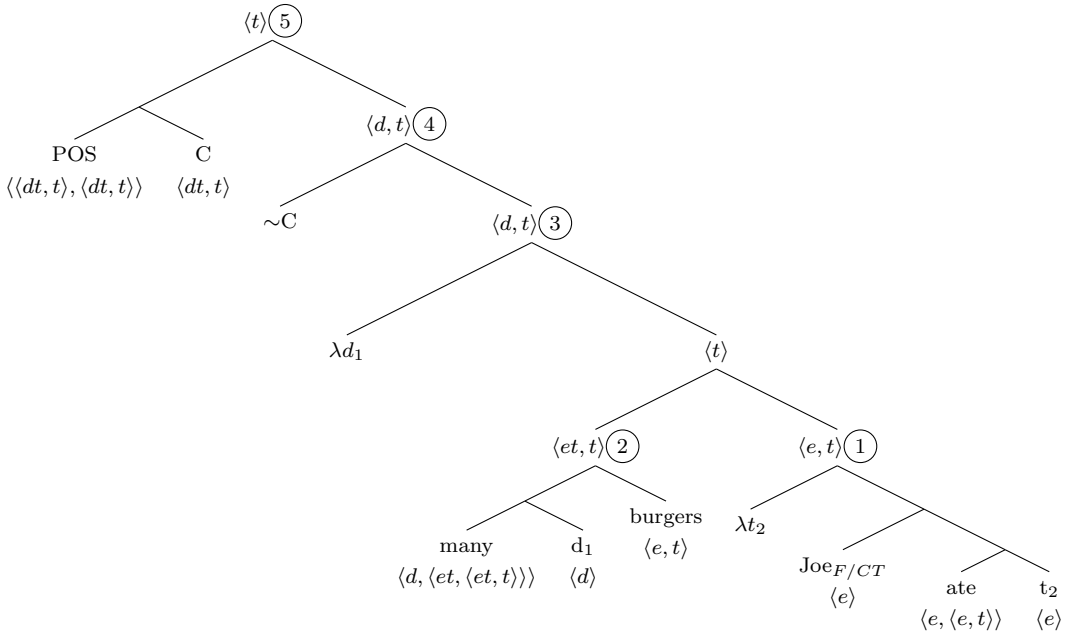
	quantificational semantics	**adjectival semantics**	**degree quantifier semantics**
author	Romero (2015)	Romero (2017)	Solt (2009)
cardinal	$[\![\text{few}]\!] = \lambda d_d.\lambda P_{\langle e,t\rangle}.\lambda Q_{\langle e,t\rangle}.\mid P \cap Q \mid \leq d$ $[\![\text{many}]\!] = \lambda d_d.\lambda P_{\langle e,t\rangle}.\lambda Q_{\langle e,t\rangle}.\mid P \cap Q \mid \geq d$	$[\![\text{few}]\!] = \lambda d_d.\lambda x_{\langle e\rangle}.\mid x \mid \leq d$ $[\![\text{few}]\!] = \lambda d_d.\lambda P_{\langle e,t\rangle}.\lambda x_{\langle e\rangle}.P(x) \wedge \mid x \mid \leq d$ $[\![\text{many}]\!] = \lambda d_d.\lambda x_{\langle e\rangle}.\mid x \mid \geq d$ $[\![\text{many}]\!] = \lambda d_d.\lambda P_{\langle e,t\rangle}.\lambda x_{\langle e\rangle}.P(x) \wedge \mid x \mid \geq d$	$[\![\text{few}]\!] = \lambda d_d.\lambda I_{\#} \in D_{\langle d,t\rangle}.d \notin I$ $[\![\text{many}]\!] = \lambda d_d.\lambda I_{\#} \in D_{\langle d,t\rangle}.d \in I$ $[\![\text{Meas}]\!] = \lambda x_e.\lambda d_d.\mu_{DIM}(x) \geq d$
proportional	$[\![\text{few}]\!] = \lambda d_d.\lambda P_{\langle e,t\rangle}.\lambda Q_{\langle e,t\rangle}.$ $\mid P \cap Q \mid : \mid P \mid \leq d$ $[\![\text{many}]\!] = \lambda d_d.\lambda P_{\langle e,t\rangle}.\lambda Q_{\langle e,t\rangle}.$ $\mid P \cap Q \mid : \mid P \mid \geq d$	$[\![\text{few}]\!] = \lambda d_d.\lambda P_{\langle e,t\rangle}.\lambda x_e.P(x) \wedge$ $\mid x \mid : \mid P_{Atomic} \mid \leq d$ $[\![\text{many}]\!] = \lambda d_d.\lambda P_{\langle e,t\rangle}.\lambda x_e.P(x) \wedge$ $\mid x \mid : \mid P_{Atomic} \mid \geq d$	
⊕	conservative most straightforward account for the proportional reading and *few* and *many* in determiner position no further covert semantic machinery needed	compositional most straightforward account for the cardinal reading can also account for adjectival uses (combination with definite determiner)	unified treatment of all syntactic positions and in parallel with *much* and *little*
⊖	cannot account for adjectival or differential uses	cannot account for differential uses requires existential closure in determiner-like uses	several covert operators are necessary availability of the proportional reading is only influenced by the discourse, not by syntactic or semantic properties of the sentence

Table 2.2: Standard semantics accounts of *few* and *many*

are marked with a circled number in the tree and their denotations are provided in the list underneath the tree.

Focus/topic-marking on Joe results in a comparison between Joe and other relevant people (other guests at the barbecue, for example). Plural individuals are marked with by the * operator (see Hackl, 2001). The variable's domain is only given explicitly where considered necessary (for example, λx. ... instead of $\lambda x \in D_e$. ...)

(44) Joe$_{F/CT}$ ate many burgers.



(45) a. (1) $= \lambda x$. Joe ate x

b. $[\![\text{many}_{card}]\!] = \lambda d_d.\lambda P_{\langle e,t\rangle}.\lambda Q_{\langle e,t\rangle}.\,|\,P \cap Q\,| \geq d$

c. (2) $= \lambda Q_{\langle e,t\rangle}.\,|\,\{x : *\text{burgers}(x)\} \cap Q\,| \geq \text{d}_1$

d. (3) $= \lambda d.\,|\,\{x : *\text{burgers}(x)\} \cap \{x : \text{Joe ate } x\}\,| \geq d$

e. (4) is defined iff $[\![\text{C}]\!] \subseteq \{\lambda d'.\,|\,\{x : *\text{burgers}(x)\} \cap \{x : \text{Joe ate } x\}\,| \geq d',$
$\lambda d'.\,|\,\{x : *\text{burgers}(x)\} \cap \{x : \text{Max ate } x\}\,| \geq d',$
$\lambda d'.\,|\,\{x : *\text{burgers}(x)\} \cap \{x : \text{Sue ate } x\}\,| \geq d', ...\}$

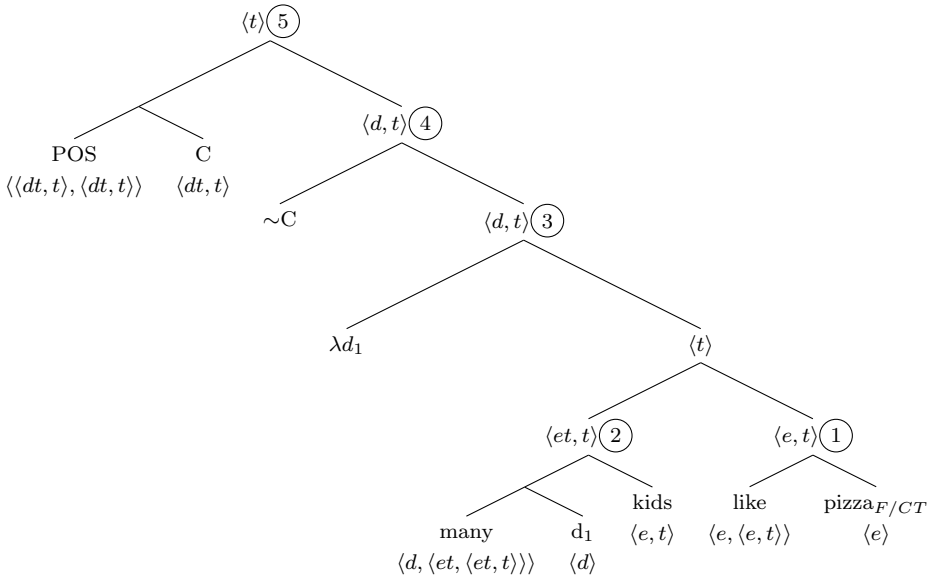
f. $[\![POS]\!] = \lambda \text{C}_{<dt,t>}.\lambda D_{<d,t>}.L_{\langle\langle dt,t\rangle,\langle d,t\rangle\rangle}(\text{C}) \subseteq D$

g. (5) $= 1$ iff $L([\![\text{C}]\!]) \subseteq \lambda d.\,|\,\{x : *\text{burgers}(x)\} \cap \{x : \text{Joe ate } x\}\,| \geq d$

Sentence (44) is true if Joe ate many burgers, where many is evaluated relative to the number of burgers that other people ate.

Second, the truth conditions of the regular proportional reading are derived. Romero (2015) analyzes this reading as a result of *POS* associating with an element *external* to the host NP. In sentence (46), "pizza", *many*'s second argument, is focus-marked and triggers the alternatives contained in C.

(46) Many (of the) kids like pizza$_{F/CT}$.



(47) a. (1) = $\lambda x.\, x$ like pizza

b. $[\![\text{many}_{prop}]\!] = \lambda d_d.\lambda P_{\langle e,\, t\rangle}.\lambda Q_{\langle e,\, t\rangle}.\,|\, P \cap Q \,|\, :\, |\, P \,| \geq d$

c. (2) = $\lambda Q_{\langle e,\, t\rangle}.\,|\, \{x : {}^*\text{kids}(x)\} \cap Q \,|\, :\, |\, \{x : {}^*\text{kids}(x)\} \,| \geq \text{d}_1$

d. (3) = $\lambda d.\,|\, \{x : {}^*\text{kids}(x)\} \cap \{x : x \text{ like pizza}\} \,|\, :\, |\, \{x : {}^*\text{kids}(x)\} \,| \geq d$

e. (4) is defined iff
$[\![\text{C}]\!] \subseteq \{\lambda d'.\,|\, \{x : {}^*\text{kids}(x)\} \cap \{x : x \text{ like pizza}\} \,|\, :\, |\, \{x : {}^*\text{kids}(x)\} \,| \geq d',$
$\lambda d'.\,|\, \{x : {}^*\text{kids}(x)\} \cap \{x : x \text{ like spinach}\} \,|\, :\, |\, \{x : {}^*\text{kids}(x)\} \,| \geq d',$
$\lambda d'.\,|\, \{x : {}^*\text{kids}(x)\} \cap \{x : x \text{ like cherries}\} \,|\, :\, |\, \{x : {}^*\text{kids}(x)\} \,| \geq d', ...\}$

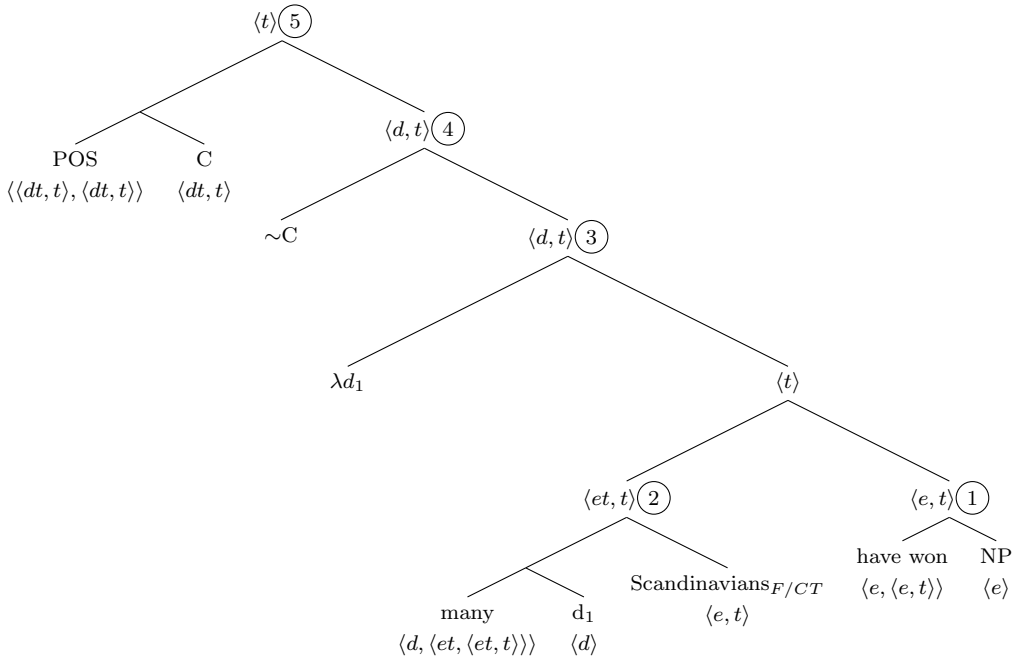
f. $[\![POS]\!] = \lambda \text{C}_{<dt,t>}.\lambda D_{<d,t>}.L_{\langle\langle dt,\, t\rangle,\, \langle d,\, t\rangle\rangle}(\text{C}) \subseteq D$

g. (5) = 1 iff
$L([\![\text{C}]\!]) \subseteq \lambda d.\,|\, \{x : {}^*\text{kids}(x)\} \cap \{x : x \text{ like pizza}\} \,|\, :\, |\, \{x : {}^*\text{kids}(x)\} \,| \geq d$

Sentence (46) is true if many kids like pizza, where many is evaluated relative to the proportion of kids who like other kinds of food.

Third, we address the reverse proportional reading. Romero's (2015) achievement is that she can derive the truth conditions of this reading with only one proportional determiner *many*$_{prop}$ and *few*$_{prop}$, which are both conservative. The reverse proportional reading is obtained when *POS* is associated with an element *internal* to the host NP. In example (48), repeated from above, Scandinavians, *many*'s first argument, is focused which triggers the reversed reading that many winners of a Nobel Prize in literature are Scandinavians (cf. Herburger, 1997). See Romero (2015) for a more thorough discussion of the truth conditions of the reverse proportional reading and its comparison class. Note that in the calculation "the Nobel Prize in literature" will be abbreviated to "NP" of type $\langle e \rangle$.

(48) Many Scandinavians$_{F/CT}$ have won the Nobel Prize in literature.



(49) a. (1) = $\lambda x.x$ have won NP

b. $[\![\text{many}_{prop}]\!] = \lambda d_d.\lambda P_{\langle e,\,t\rangle}.\lambda Q_{\langle e,\,t\rangle}.\,|\,P \cap Q\,| : |\,P\,| \geq d$

c. (2) = $\lambda Q_{\langle e,\,t\rangle}.\,|\,\{x : {*}\text{Scand}(x)\} \cap Q\,| : |\,\{x : {*}\text{Scand}(x)\}\,| \geq \text{d}_1$

d. (3) = $\lambda d.\,|\,\{x : {*}\text{Scand}(x)\} \cap \{x : \text{NP-winner}(x)\}\,| : |\,\{x : {*}\text{Scand}(x)\}\,| \geq d$

e. (4) is defined iff $[\![\text{C}]\!] \subseteq$
$\{\lambda d'.\,|\,\{x : {*}\text{Scand}(x)\} \cap \{x : \text{NP-winner}(x)\}\,| : |\,\{x : {*}\text{Scand}(x)\}\,| \geq d',$
$\lambda d'.\,|\,\{x : \text{Mediterr}^*(x)\} \cap \{x : \text{NP-winner}(x)\}\,| : |\,\{x : \text{Med}^*(x)\}\,| \geq d',$
$\lambda d'.\,|\,\{x : \text{M.East}^*(x)\} \cap \{x : \text{NP-winner}(x)\}\,| : |\,\{x : \text{M.East}^*(x)\}\,| \geq d',$
$...\}$

f. $[\![POS]\!] = \lambda \text{C}_{<dt,t>}.\lambda D_{<d,t>}.L_{\langle\langle dt,\,t\rangle,\,\langle d,\,t\rangle\rangle}(\text{C}C) \subseteq D$

g. (5) = 1 iff $L([\![\text{C}]\!]) \subseteq$
$\lambda d.\,|\,\{x : {*}\text{Scand}(x)\} \cap \{x : \text{NP-winner}(x)\}\,| : |\,\{x : {*}\text{Scand}(x)\}\,| \geq d$

Sentence (48) is true under a reverse proportional reading if "the proportion of Scandinavians that have won the Nobel Prize in literature is large compared to a threshold based on the proportions of inhabitants of other world regions that have won the Nobel Prize in literature" (Romero, 2015, 23).

After having seen the cardinal reading, the regular proportional and the reverse proportional reading, the question might arise whether there is also such a thing as a "reverse cardinal" reading. This is actually possible when cardinal *many*'s first argument is focused as in

(50) Joe ate many burgers$_{F/CT}$.

Focus on "burgers" triggers the comparison class of other things Joe might have eaten and the sentence is true if Joe ate more burgers than, say, sandwiches, hot dogs and muffins. Since the cardinal reading is symmetrical anyway, this reading is not considered particularly interesting or problematic. We leave it to the reader to derive this reading compositionally.

In a nutshell, we see the strength of the quantificational approach in its simplicity. It does not need further covert semantic machinery and its application is straightforward. Moreover, it fits the proportional reading well by intersecting two sets of individuals. We have shown that even the reverse proportional reading can be captured by a single proportional lexical entry while preserving the conservativity feature (Cohen, 2001; Romero, 2015). On the downside, the quantificational account does not capture adjectival uses in which *few* and *many* are preceded by the definite determiner. Quantifiers and the definite determiner are supposed to be in complimentary distribution and thus the quantificational analysis would rule out this well-attested use. Such a case is better accounted for by an adjectival semantics.

2.2.2 Adjectival Semantics

Few and *many*'s similarity to gradable adjectives has led to the development of an adjectival semantics similar to cardinality predicates like *numerous*, which expresses the property "small/large in number relative to the average number in the given context" (cf. Partee, 1989; Hackl, 2009; Krasikova, 2011; Dobrovie-Sorin, 2013; Romero, 2017). Arguments for treating *few* and *many* on par with adjectives come from their parallel behavior in comparative constructions. Both quantity words and gradable adjectives are available in the positive, comparative and superlative form:

(51) a. many, more, most
b. few, fewer, fewest
c. tall, taller, tallest

The standard semantic analysis of these three comparison constructions is to decompose them into a stem and functional operators (cf. von Stechow, 1984; Kennedy and McNally, 2005; Beck, 2011). This analysis has also been applied to quantity words.

(52) a. tall = TALL + *POS*
taller = TALL + *-er*
tallest = TALL + *-est*

b. many = MANY + *POS* (Romero, 2015, 2017)
more = MANY + *-er* (Hackl, 2000)
most = MANY + *-est* (Hackl, 2009)

Furthermore, only adjectives, but not quantifiers, can occur in explicitly adjectival positions and be preceded by the definite determiner, similar to example (17) above.

(53) a. The many construction sites brought the traffic to a standstill.
b. The few shops that were still open didn't sell the shoes I wanted.
c. The *some/*no/*all students were late.

An adjectival semantics of *few* and *many* can be straightforwardly transferred from gradable adjectives. The cardinal semantics in (54) provides two lexical entries, depending on whether the quantity word appears as an intersective adjective or combines directly with a plural individual whose atoms are counted. The intersective version of type $\langle e, t\rangle$ would combine with its plural noun sister via Predicate Modification and thereby pattern with other adjectives, whereas the entry in the second line takes its sister as an argument via applying Functional Application (Heim and Kratzer, 1998).

(54) Cardinal reading

a. $[\![\text{few}]\!]_{\langle e,t\rangle} = \lambda x_{\langle e\rangle}.\,|\,x\,| \leq x_{\max}$
$[\![\text{few}]\!]_{\langle et,\langle e,t\rangle\rangle} = \lambda P_{\langle e,t\rangle}.\lambda x_{\langle e\rangle}.P(x) \wedge |\,x\,| \leq x_{\max}$

b. $[\![\text{many}]\!]_{\langle e,t\rangle} = \lambda x_{\langle e\rangle}.\,|\,x\,| \geq x_{\min}$
$[\![\text{many}]\!]_{\langle et,\langle e,t\rangle\rangle} = \lambda P_{\langle e,t\rangle}.\lambda x_{\langle e\rangle}.P(x) \wedge |\,x\,| \geq x_{\min}$

(55) Proportional reading

a. $[\![\text{few}]\!]_{\langle et,\langle e,t\rangle\rangle} = \lambda P_{\langle e,t\rangle}.\lambda x_{\langle e\rangle}.P(x) \wedge |\,x\,| : |\,P\,| \leq k_{\max}$

b. $[\![\text{many}]\!]_{\langle et,\langle e,t\rangle\rangle} = \lambda P_{\langle e,t\rangle}.\lambda x_{\langle e\rangle}.P(x) \wedge |\,x\,| : |\,P\,| \geq k_{\min}$

The cardinal and the proportional reading again seem unproblematic, but once we turn to the reverse proportional reading, we run into a compositionality problem.

(56) Reverse proportional reading

a. $[\![\text{few}]\!]_{\langle et,\langle e,t\rangle\rangle} = \lambda P_{\langle e,t\rangle}.\lambda x_{\langle e\rangle}.P(x) \wedge |\,x\,| : |\,Q\,| \leq k_{\max}$

b. $[\![\text{many}]\!]_{\langle et,\langle e,t\rangle\rangle} = \lambda P_{\langle e,t\rangle}.\lambda x_{\langle e\rangle}.P(x) \wedge |\,x\,| : |\,Q\,| \geq k_{\min}$

It is a property of adjectives that they only "see" their sister in the tree, be it an individual or a noun. They cannot, however, raise out of the NP they are contained in and take higher scope. This is why a semantics for the reverse proportional reading as in (56) is not compositional. The quantity word needs to calculate a proportion over $|\,Q\,|$, but does not have a λQ-argument (Romero, 2017).

Romero (2017) sets out to solve the compositionality problem. Regular and reverse proportional *few* and *many* are again fused into one lexical entry, which

however, does not only look at its $\langle e, t\rangle$ complement as a whole but also at its atomic subparts. Both the cardinal and the proportional variant receive a degree argument which is bound by the null morpheme *POS*. *POS* once more associates with a focus- or contrastive topic-marked constituent. Associates external to the host DP trigger the regular reading, internal associates a reverse reading.

(57) Romero's (2017) cardinal reading

a. $[\![\text{few}_{card}]\!]\langle d, \langle e, t\rangle\rangle = \lambda d_{\langle d\rangle}.\lambda x_{\langle e\rangle}.\,|\,x\,| \leq d$
$[\![\text{few}_{card}]\!]\langle d, \langle et, \langle e, t\rangle\rangle\rangle = \lambda d_{\langle d\rangle}.\lambda P_{\langle e,\,t\rangle}.\lambda x_{\langle e\rangle}.P(x) \wedge |\,x\,| \leq d$

b. $[\![\text{many}_{card}]\!]\langle d, \langle e, t\rangle\rangle = \lambda d_{\langle d\rangle}.\lambda x_{\langle e\rangle}.\,|\,x\,| \geq d$
$[\![\text{many}_{card}]\!]\langle d, \langle et, \langle e, t\rangle\rangle\rangle = \lambda d_{\langle d\rangle}.\lambda P_{\langle e,\,t\rangle}.\lambda x_{\langle e\rangle}.P(x) \wedge |\,x\,| \geq d$

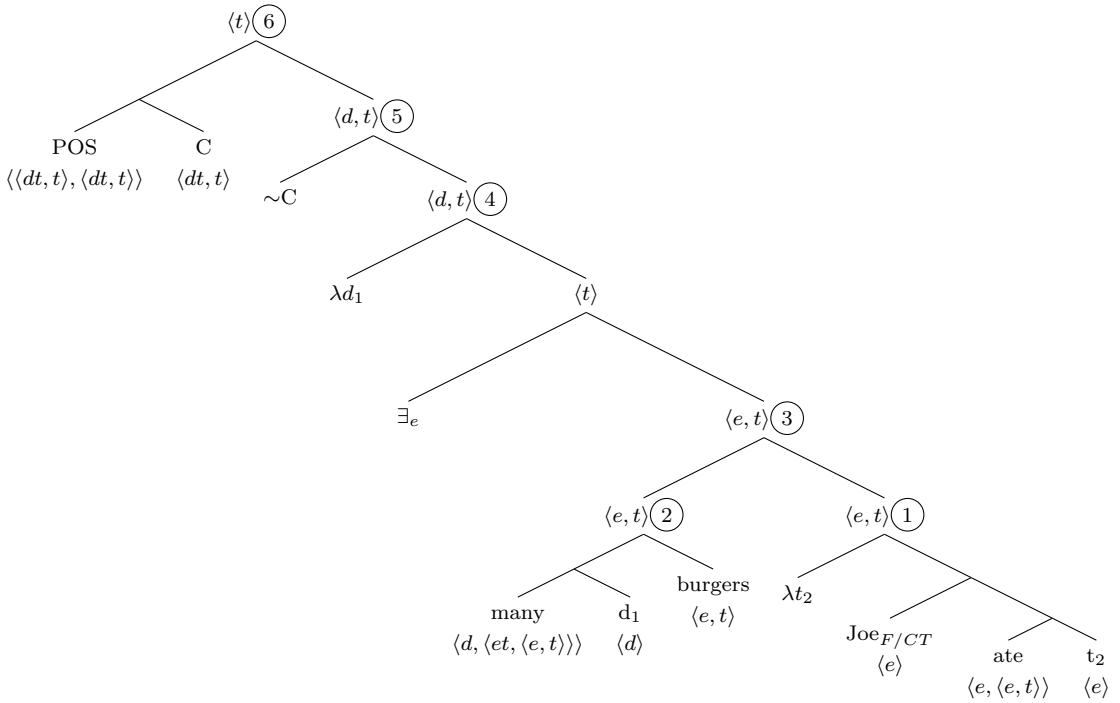
(58) Romero's (2017) proportional reading

a. $[\![\text{few}_{prop}]\!]_{\langle d,\,\langle et,\,\langle e,\,t\rangle\rangle\rangle} = \lambda d_{\langle d\rangle}.\lambda P_{\langle e,\,t\rangle}.\lambda x_{\langle e\rangle}.P(x) \wedge |\,x\,| : |\,P_{Atomic}\,| \leq d$

b. $[\![\text{many}_{prop}]\!]_{\langle d,\,\langle et,\,\langle e,\,t\rangle\rangle\rangle} = \lambda d_{\langle d\rangle}.\lambda P_{\langle e,\,t\rangle}.\lambda x_{\langle e\rangle}.P(x) \wedge |\,x\,| : |\,P_{Atomic}\,| \geq d$

The adjectival semantics is now to be applied to the sentences from the previous section. A necessary assumption for the composition to work, however, is that *many*'s type $\langle e\rangle$ argument is existentially bound by the covert operator *existential closure* $\exists$ if no overt determiner is present. Furthermore, *many*'s host NP and the VP will be combined via Predicate Modification, not by Functional Application as above. We will see that the result of the compositional analysis will be the same as for a quantifier analysis of *few* and *many*. The cardinal reading of *many* of the example sentence repeated from above is presented first.

(44) Joe$_{F/CT}$ ate many burgers.



(59) a. ① $= \lambda x$. Joe ate x

b. $[\![\text{many}_{card}]\!] = \lambda d_{\langle d \rangle}.\lambda P_{\langle e,t \rangle}.\lambda x_{\langle e \rangle}.P(x) \wedge |\, x \,| \geq d$

c. ② $= \lambda x_{\langle e \rangle}$. *burgers$(x) \wedge |\, x \,| \geq \text{d}_1$

d. ③ $= \lambda x_{\langle e \rangle}$.*burgers$(x) \wedge |\, x \,| \geq \text{d}_1 \wedge \text{ate}(x)(\text{Joe})$

e. ④ $= \lambda d.\exists x[$*burgers$(x) \wedge |\, x \,| \geq d \wedge \text{ate}(x)(\text{Joe})]$

f. ⑤ is defined iff $[\![\text{C}]\!] \subseteq \{\lambda d'.\exists x[$*burgers$(x) \wedge |\, x \,| \geq d' \wedge \text{ate}(x)(\text{Joe})]$,
$\lambda d'.\exists x[$*burgers$(x) \wedge |\, x \,| \geq d' \wedge \text{ate}(x)(\text{Max})]$,
$\lambda d'.\exists x[$*burgers$(x) \wedge |\, x \,| \geq d' \wedge \text{ate}(x)(\text{Sue})], ...\}$

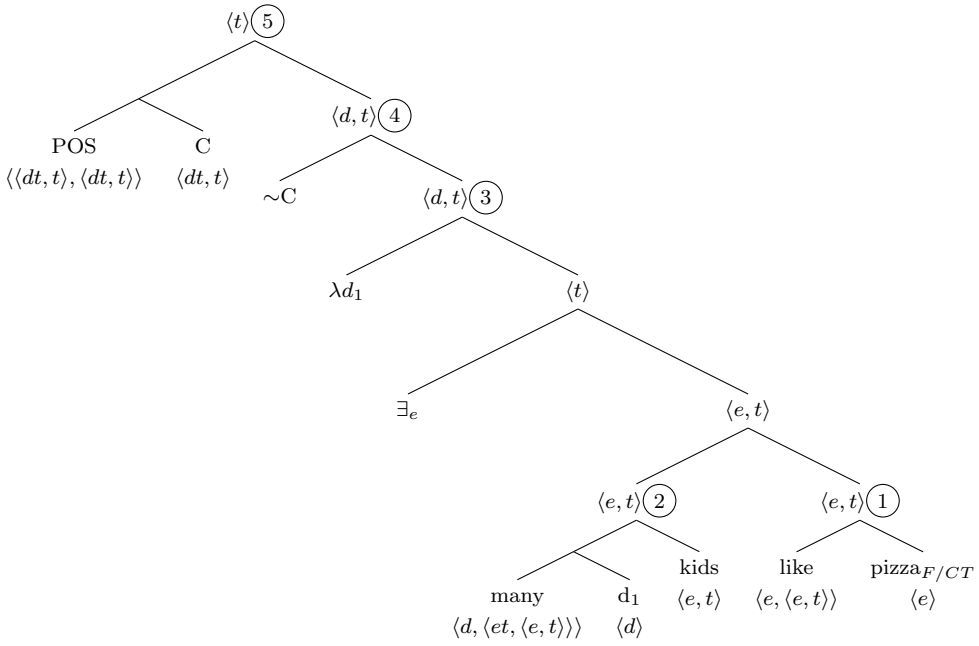
g. $[\![POS]\!] = \lambda \text{C}_{<dt,t>}.\lambda D_{<d,t>}.L_{\langle\langle dt,t\rangle,\langle d,t\rangle\rangle}(\text{C}) \subseteq D$

h. ⑥ $= 1$ iff $L([\![\text{C}]\!]) \subseteq \lambda d.\exists x[$*burgers$(x) \wedge |\, x \,| \geq d \wedge \text{ate}(x)(\text{Joe})]$

Sentence (44) is true if Joe ate many burgers, where many is evaluated relative to the number of burgers that other people ate.

Next, the regular proportional reading is analyzed. Again, this reading arises if the focused constituent is external to the host NP, see the repeated example.

(46) Many (of the) kids like pizza$_{F/CT}$.



(60) a. ① $= \lambda x.x$ like pizza

b. $[\![\text{many}_{prop}]\!] = \lambda d_{\langle d \rangle}.\lambda P_{\langle e,t \rangle}.\lambda x_{\langle e \rangle}.P(x) \wedge |\, x \,| : |\, P_{Atomic} \,| \geq d$

c. ② $= \lambda x_{\langle e \rangle}$. *kids$(x) \wedge |\, x \,| : |\, \{z :$ *kids$(z)\} \,| \geq \text{d}_1$

d. ③ $= \lambda d.\exists x[$*kids$(x) \wedge |\, x \,| : |\, \{z :$ *kids$(z)\} \,| \geq d \wedge \text{like}(\text{pizza})(x)]$

e. ④ is defined iff
$[\![\text{C}]\!] \subseteq \{\lambda d'.\exists x[$*kids$(x) \wedge |\, x \,| : |\, \{z :$ *kids$(z)\} \,| \geq d' \wedge \text{like}(\text{pizza})(x)]$,

$\lambda d'.\exists x[*\text{kids}(x) \wedge |\, x \,| : |\, \{z : *\text{kids}(z)\} \,| \geq d' \wedge \text{like}(\text{spinach})(x)]$,
$\lambda d'.\exists x[*\text{kids}(x) \wedge |\, x \,| : |\, \{z : *\text{kids}(z)\} \,| \geq d' \wedge \text{like}(\text{cherries})(x)], ...\}$

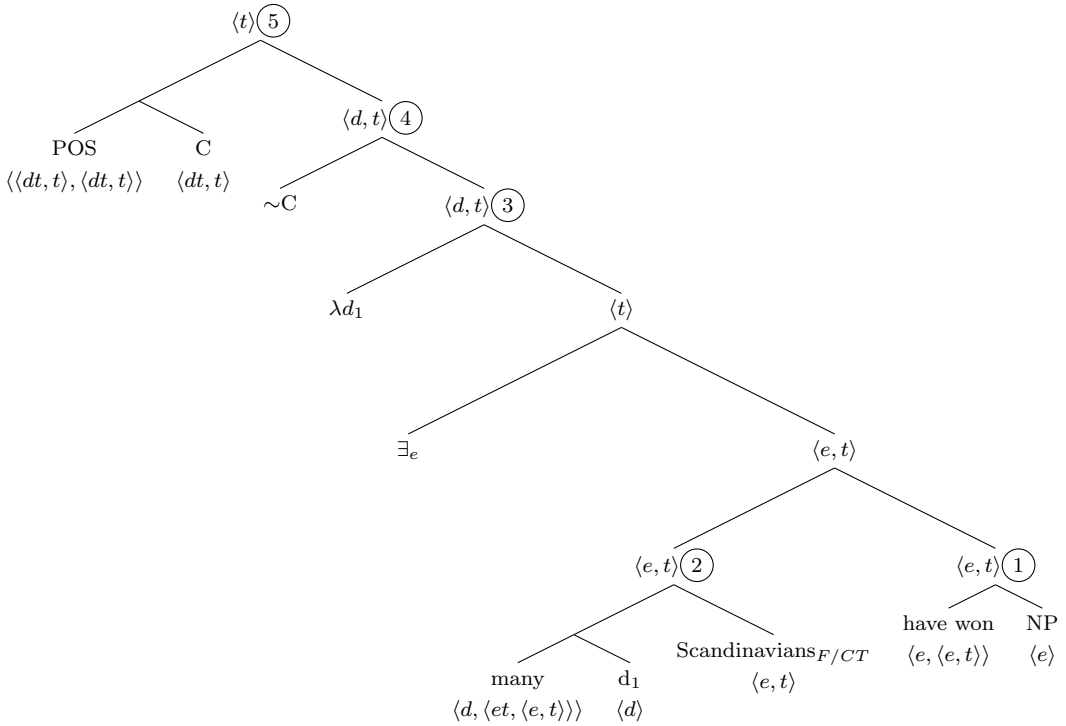
f. $[\![POS]\!] = \lambda \text{C}_{<dt,t>}.\lambda D_{<d,t>}.L_{\langle\langle dt,t\rangle,\langle d,t\rangle\rangle}(\text{C}) \subseteq D$

g. (5) = 1 iff
$L([\![\text{C}]\!]) \subseteq \lambda d.\exists x[*\text{kids}(x) \wedge |\, x \,| : |\, \{z : *\text{kids}(z)\} \,| \geq d \wedge \text{like}(\text{pizza})(x)]$

Sentence (46) is true if many kids like pizza, where many is evaluated relative to the proportion of kids who like other kinds of food.

The last compositional analysis to be presented is of the reverse proportional reading. Again, the definite description "the Nobel Prize in literature" is abbreviated by "NP".

(48) Many Scandinavians$_{F/CT}$ have won the Nobel Prize in literature.



(61) a. (1) = $\lambda x.x$ have won NP

b. $[\![\text{many}_{prop}]\!] = \lambda d_{\langle d\rangle}.\lambda P_{\langle e,t\rangle}.\lambda x_{\langle e\rangle}.P(x) \wedge |\, x \,| : |\, P_{Atomic} \,| \geq d$

c. (2) = $\lambda x_{\langle e\rangle}.*\text{Scand}(x) \wedge |\, x \,| : |\, \{z : *\text{Scand}(z)\} \,| \geq \text{d}_1$

d. (3) = $\lambda d.\exists x[*\text{Scand}(x) \wedge |\, x \,| : |\, \{z : *\text{Scand}(z)\} \,| \geq d \wedge \text{NP-winner}(x)]$

e. (4) is defined iff
$[\![\text{C}]\!] \subseteq \{\lambda d'.\exists x[*\text{Scand}(x) \wedge |\, x \,| : |\, \{z : *\text{Scand}(z)\} \,| \geq d' \wedge \text{NP-winner}(x)]$,
$\lambda d'.\exists x[*\text{Mediterr}(x) \wedge |\, x \,| : |\, \{z : *\text{Mediterr}(z)\} \,| \geq d' \wedge \text{NP-winner}(x)]$,
$\lambda d'.\exists x[*\text{M.Eastern}(x) \wedge |\, x \,| : |\, \{z : *\text{M.Eastern}(z)\} \,| \geq d' \wedge \text{NP-winner}(x)]$,
$...\}$

f. $[\![POS]\!] = \lambda \text{C}_{<dt,t>}.\lambda D_{<d,t>}.L_{\langle\langle dt,t\rangle,\langle d,t\rangle\rangle}(\text{C}) \subseteq D$

g. $\textcircled{5} = 1$ iff
$L([\![\text{C}]\!]) \subseteq \lambda d.\exists x[\text{*Scand}(x) \wedge |\, x \,| : |\, \{z : \text{*Scand}(z)\} \,| \geq d \wedge \text{NP-winner}(x)]$

Sentence (48) is true under a reverse proportional reading if "the proportion of Scandinavians that have won the Nobel Prize in literature is large compared to a threshold based on the proportions of inhabitants of other world regions that have won the Nobel Prize in literature" (Romero, 2015, 23).

In sum, we find that the adjectival account can derive the same truth conditions under the assumption that the degree variable is bound by *POS* whose associates are prosodically marked. In contrast to a quantifier semantics, the adjective semantics needs an additional silent operator $\exists$ to bind the pronoun of type $\langle e \rangle$. The adjectival account is particularly elegant to derive cardinal readings and quantity words in combination with the definite determiner.

2.2.3 Degree Quantifier

A third account was proposed by Solt (2009, 2015) and Rett (2008, 2016). They analyze *few* and *many* as a semantically empty, gradable quantifier over degrees. Further semantic machinery, like a measure function, is contributed by covert semantic operators. This summary focuses on Solt's work, who treats *few* and *many* in parallel with *little* and *much* and aims to account not only for their quantificational and adjectival uses, but also for differential uses in a unified semantics.

(62) Solt (2015, 222)

a. **quantificational**: Many/few students attended the lecture.

b. **predicative**: John's friends are many/few.

c. **attributive**: The many/few students who attended enjoyed the lecture.

d. **differential**: Many more/few more/many fewer than 100 students attended the lecture.

Instead of treating *few* and *many* as quantifiers like *every* or adjectives like *tall* Solt (2009, 2015) suggests that they are gradable predicates of intervals (sets of degrees) on some dimension of measurement. She claims that only such a semantics can capture all of the various uses exemplified in (62). In her decompositional account she strips *few, many, little* and *much* of most of their semantics content which is instead contributed by a series of null functional elements (see Solt, 2009, Chapter 3). In a compositional analysis of a sentence with quantificational *few* or *many*, the truth conditions are the same as under a quantificational semantics in which all of these operators are already contained in the semantics of *many*. This is why her proposal looks unnecessarily complicated, at least for quantificational *few* and *many*. Solt, however, claims that her reduced semantics is necessary to account

for the various uses of quantity words in a unified way. The syntactic form she assumes and the lexical entries she proposes are given below.

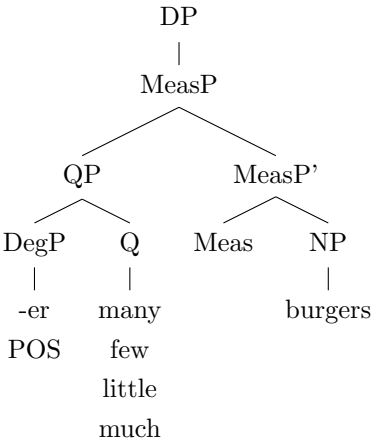
(63) a. $[\![\text{many}]\!] = \lambda d_d.\lambda I_{\#} \in D_{\langle d,\, t\rangle}.d \in I$

b. $[\![\text{few}]\!] = \lambda d_d.\lambda I_{\#} \in D_{\langle d,\, t\rangle}.d \notin I$

c. $[\![\text{Meas}]\!] = \lambda x_e.\lambda d_d.\mu_{DIM}(x) \geq d$

d. $\exists$ (Existential Closure)

e. $[\![\text{POS}]\!] = \lambda I_{\langle d,\, t\rangle}.\forall d \in N_s[d \in I]$

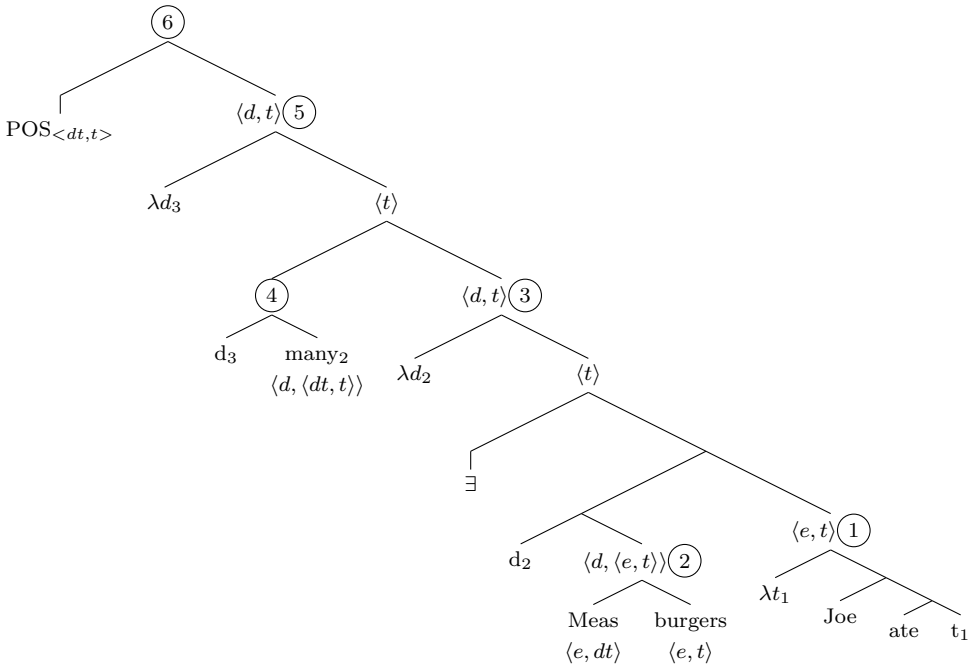


According to Solt's theory, the only semantic contribution of *many* and *few* is to check whether a degree is contained in an interval. The remainder of their meaning is contributed by *Meas* and *POS*. In terms of the LF, Solt (2009, 2015) proposes that the quantity words are located in the specifier position of a DP-internal functional head *Meas*, which introduces a degree argument. Then, they undergo Quantifier Raising (Heim and Kratzer, 1998) and take sentential scope. They take as their argument a set of degrees (see the LF structure below). Consequently, the quantity words are not assumed to combine with their NP restrictor directly. They are taken to be embedded in a measure phrase MeasP. Note that "Meas does not encode a specific dimension. Rather, the dimension in question is 'filled in' on the basis of the NP denotation, the nature of the degree expression that it combines with, and the context of interpretation" (Solt, 2009, p. 105). The scale's dimension is underspecified in the semantics of *Meas* because Solt (2009, 2015) aims to apply the operator to *few* and *many* as well as to *much* and *little*. In contrast to *much* and *little*, which are more free in the choice of their scale and can operate on scales of volume, weight etc., *many* and *few* require that the interval is of dimension cardinality. The degrees on the cardinality scale must be countable (indicated by the # subscript). A last remarkable feature of Solt's (2009) proposal is that instead of attributing quantificational force over individuals to Q-adjectives themselves, she opts for existential closure in order to be able to also account for their predicative uses.

In the positive form, the degree argument of *many* is bound by the null morpheme POS. POS establishes a comparison with a neutral interval provided by the context. In the following, the three example sentences from above are analyzed. We start out again with the cardinal reading of *many* in its quantificational use. An LF structure and a compositional analysis are presented. Note that it seems inconvenient to apply Quantifier Raising three times and Existential Closure only in order to avoid type

mismatches. Furthermore, the measure function cannot be included via Function Application but needs an own composition rule. The rule Variable Identification is based on Kratzer's (1996) Event Identification rule and combines the *Meas* Node with its nominal complement. Note that in contrast to Romero (2015, 2017), the *POS* operator used here, does not take the comparison class as its argument and thus does not build on focus/contrastive topic alternatives.

(44) Joe ate many burgers.



(64) a. $\textcircled{1} = \lambda x.$ Joe ate x

b. $[\![\text{Meas}]\!] = \lambda x.\lambda d.\mu_{DIM}(x) \geq d$

c. $\textcircled{2} = \lambda d.\lambda x.^*\text{burgers}(x) \wedge \mu_{DIM}(x) \geq d$

d. $\textcircled{3} = \lambda d.\exists x[^*\text{burgers}(x) \wedge \mu_{DIM}(x) \geq d \wedge \text{ate}(x)(\text{Joe})]$

e. $[\![\text{many}]\!] = \lambda d_d.\lambda I_{\#} \in D_{<d,t>}.d \in I$

f. $\textcircled{4} = \lambda I_{\#} \in D_{<d,t>}.d_3 \in I$

g. $\textcircled{5} = \lambda d.\exists x[^*\text{burgers}(x) \wedge \mu_{\#}(x) \geq d \wedge \text{ate}(x)(\text{Joe})]$

h. $[\![\text{POS}]\!] = \lambda I_{<d,t>}.\forall d \in N_s[d \in I]$

i. $\textcircled{6} = 1$ iff $\forall d \in N_s\, \exists x[^*\text{burgers}(x) \wedge \mu_{\#}(x) \geq d \wedge \text{ate}(x)(\text{Joe})]$

Sentence (44) is true if Joe ate many burgers, where many is evaluated relative to the neutral interval on the cardinality scale.

For the proportional reading, Solt (2009) takes a different path than the approaches in the previous sections. Instead of a lexical ambiguity, she claims that the ambiguity is caused by a difference in scale structure. "The proportional reading arises when an upper bound to the scale is assumed, whereas the cardinal reading

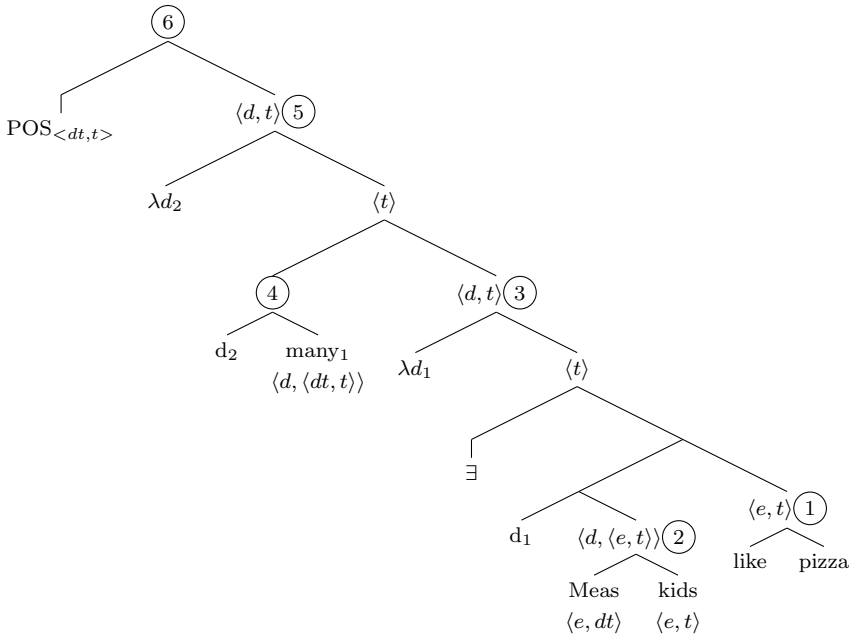
arises when there is no salient upper bound" (Solt, 2009, p.209). Two environments force a proportional reading: the subject position of an individual-level predicate and the partitive construction *of the*. Solt (2009) suggests that they share the "totalizing" property. The concepts of totalizing and individualizing can be explained with the determiners *every* and *each*. *Every* is totalizing whereas *each* is individualizing.

(65) a. Every cake is tasty.

b. Each cake is tasty.

"The totalizing effect comes not from the [quantity word] itself but from the construction in which it occurs... Totalizing has a consequence for measurement, in that the measurement scale introduced by the functional head Meas is restricted to measuring the extent of that totality" (Solt, 2009, p.218). A plural NP in the subject position of an individual-level predicate "is interpreted by first pulling aside the totality of individuals in its extension, and then subjecting them to the predicate" (Solt, 2009, p.222). Based on these observations, Solt (2009) adjusts the semantics of the NP itself by adding a supremum operator. Find my interpretation of her analysis applied to example (46) below.

(46) Many (of the) kids like pizza$_{F/CT}$.



(66) a. (1) = $\lambda x.\, x$ like pizza

b. $[\![\text{kids}]\!]_{\langle e\rangle} = \sup(\lambda x.\ ^*\text{kids}(x))$
$\Rightarrow$ shift from group to set type, PSP as domain restriction on Meas
$[\![\text{kids}]\!]_{\langle e,t\rangle} = \lambda y : y \subseteq \sup(\lambda x.\ ^*\text{kids}(x)).\ ^*\text{kids}(y)$

c. $[\![\text{Meas}]\!] = \lambda x.\lambda d.\mu_{DIM}(x) \geq d$

d. ② $= \lambda d : d \leq \mu_{DIM}(\sup(\lambda x.\ *\text{kids}(x))).\lambda y : y \subseteq \sup(\lambda x.\ *\text{kids}(x)).$
*kids$(y) \wedge \mu_{DIM}(y) \geq d$
$\Rightarrow$ scale is bounded on the upper end

e. ③ $= \lambda d : d \leq \mu_{DIM}(\sup(\lambda x.\ *\text{kids}(x))).\exists y\,[\ *\text{kids}(y) \wedge$
$\mu_{DIM}(y) \geq d \wedge \text{like}(\text{pizza})(y)]$
③ is defined iff $y \subseteq \sup(\lambda x.\ *\text{kids}(x))$ and $\text{d}_1 \leq \mu_{DIM}(\sup(\lambda x.\ *\text{kids}(x)))$

f. $[\![\text{many}]\!] = \lambda d_d.\lambda I_{\#} \in D_{<d,t>}.d \in I$

g. ④ $= \lambda I_{\#} \in D_{<d,t>}.d_3 \in I$

h. ⑤ $= \lambda d : d \leq \mu_{\#}(\sup(\lambda x.\ *\text{kids}(x))).\exists y\,[\ *\text{kids}(y) \wedge \mu_{\#}(y) \geq d \wedge$
$\text{like}(\text{pizza})(y)]$

i. $[\![\text{POS}]\!] = \lambda I_{<d,t>}.\forall d \in N_s[d \in I]$

j. ⑥ $= 1$ iff $\forall d \in N_s\ \exists y\,[\ *\text{kids}(y) \wedge \mu_{\#}(y) \geq d \wedge \text{like}(\text{pizza})(y)]$
where the domain of degrees d is restricted to $d \leq \mu_{\#}(\sup(\lambda x.\ *\text{kids}(x)))$

Sentence (46) is predicted to be true if many kids like pizza where the number denoted by "many" is bounded by the totality of kids.

We see that it is possible to derive the truth conditions of the proportional reading with Solt's (2009) decompositional account. The question that arises and remains unanswered, however, is what in the semantics triggers the totalizing effect and restricts the scale. Following Solt (2009), this should be triggered by the semantics of an individual-level predicate or the partitive construction, but Solt (2009) does not show how this should work in the composition. An even more obvious question is how Solt (2009) would account for the proportional reading of stage-level predicates. Sentences containing *few* and *many* and a stage-level predicate can be interpreted both proportionally or cardinally. The only explanation offered by Solt (2009) is that discourse cues either impose a boundary (resulting in a proportional) reading or they do not (resulting in a cardinal reading). The lexical ambiguity theory assumed in the quantificational and adjectival account is by far more straightforward in this respect. Moreover, it is not clear how Solt's (2009) account of the proportional reading should be extended to the reverse proportional reading. For a sentence like (48), the scale would have to be restricted by the VP, with which *Meas* does not combine directly. Whether restricting the scale and the totalizing effect can take place randomly, detached from syntactic and semantic properties, is questionable.

All in all, Solt's (2009) theory is a well-motivated attempt to give a unified semantics of the various uses of *few* and *many*. This attempt works out well compositionally for the cardinal reading in a quantificational position. Nevertheless, the semantics is blown up by a number of covert operators and, at least for the quantificational use, the same result can be achieved with a simpler (be it quantificational

or adjectival) semantics. The derivation of the proportional reading is equally complicated and leaves the reader with several open questions. For the remainder of this thesis we will therefore adhere to the former two accounts.

2.2.4 The Semantics of *few*

After having intensively discussed the semantics of *many*, we notice that most theories put *few* in second place. This might be due to the complex issue of the semantics of antonyms. In this section we only provide a brief overview over the semantics of *few*. Further, non-semantic properties of *few*, like referents it prefers bring into focus, are also discussed in the next chapter, where psychological work is introduced. The difference between *few* and *many* is picked up again in Chapter 6 and especially in Section 6.7, where unexpected results of the interaction between *surprisingly* and *few* are discussed.

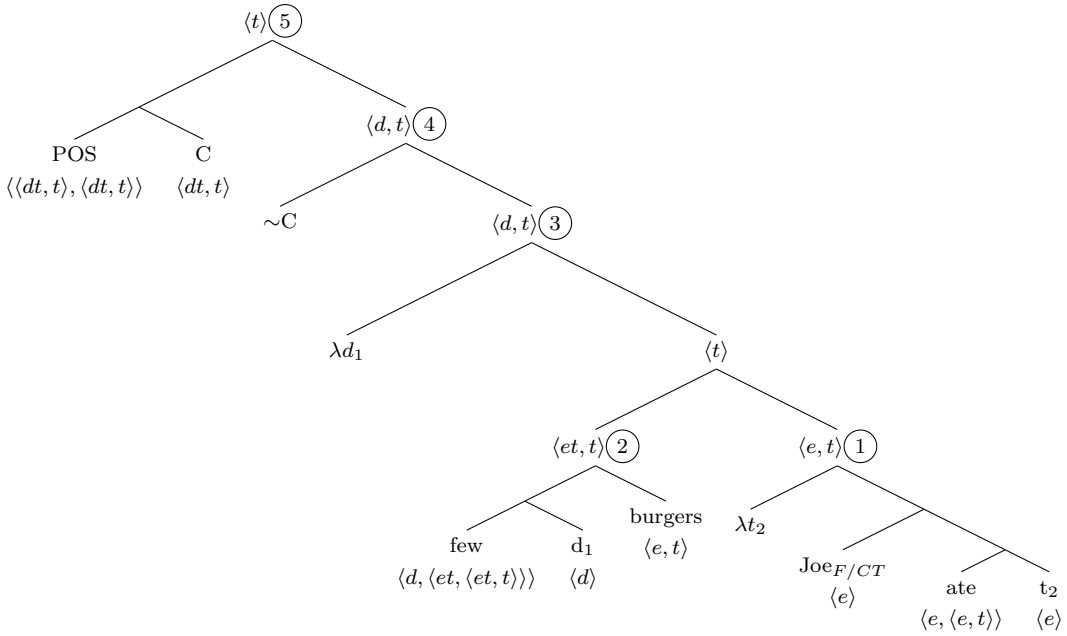
The literature agrees in that the negative member of an adjective pair like *short - long, young - old, slow - fast, few - many* involves negation. There is a debate, however, whether the negation is lexicalized as in (67a) or whether the antonym is split in the syntax into a negation operator *little* and the positive adjective as in (67b) (Kennedy and McNally, 2005; Heim, 2006, 2008; Beck, 2011). See an illustration of the two theories below:

(67) a. $[\![\text{few}]\!] = \lambda\ \text{A}.\ \lambda\ \text{B}.\ |\,\text{A} \cup \text{B}\,| \leq x_{\max}$

b. $[\![\text{few}]\!] = [\![\text{little}]\!]\ ([\![\text{many}]\!])$

A lexicalized negation is what we have assumed in the previous sections for the sake of simplicity. It is just intuitive that *few*, which expresses that a cardinality is small, denotes the *smaller than* relation $\leq$ between a cardinality and a degree. The example sentences (44), (46) and (48) originally containing *many* can be analyzed with *few* in a straightforward way. We exemplify the cardinal reading under a quantifier semantics.

(68) $\text{Joe}_{F/CT}$ ate few burgers.



(69) a. (1) = λx. Joe ate x

b. $[\![\text{few}_{card}]\!] = \lambda d_d.\lambda P_{\langle e,t\rangle}.\lambda Q_{\langle e,t\rangle}.\,|\,P \cap Q\,| \leq d$

c. (2) = $\lambda Q_{\langle e,t\rangle}.\,|\,\{x : *\text{burgers}(x)\} \cap Q\,| \leq \text{d}_1$

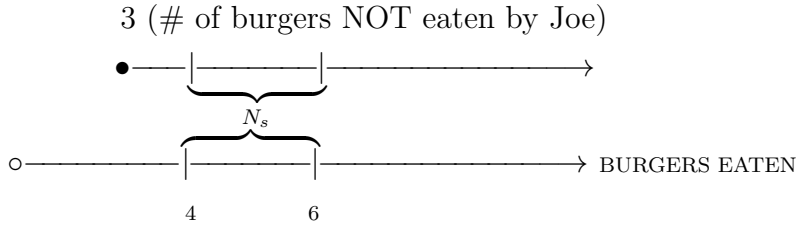
d. (3) = $\lambda d.\,|\,\{x : *\text{burgers}(x)\} \cap \{x : \text{Joe ate } x\}\,| \leq d$

e. (4) is defined iff $[\![\text{C}]\!] \subseteq \{\lambda d'.\,|\,\{x : *\text{burgers}(x)\} \cap \{x : \text{Joe ate } x\}\,| \leq d',$
$\lambda d'.\,|\,\{x : *\text{burgers}(x)\} \cap \{x : \text{Max ate } x\}\,| \leq d',$
$\lambda d'.\,|\,\{x : *\text{burgers}(x)\} \cap \{x : \text{Sue ate } x\}\,| \leq d', ...\}$

f. $[\![POS]\!] = \lambda \text{C}_{<dt,t>}.\lambda D_{<d,t>}.L_{\langle\langle dt,t\rangle,\langle d,t\rangle\rangle}(\text{C}) \subseteq D$

g. (5) = 1 iff $L([\![\text{C}]\!]) \subseteq \lambda d.\,|\,\{x : *\text{burgers}(x)\} \cap \{x : \text{Joe ate } x\}\,| \leq d$

(70)



The truth conditions are illustrated in the graphic in (70). Node (3) in (2.2.4) denotes the set of degrees such that the number of burgers eaten by Joe is *smaller*. This is equivalent to the interval of the numbers of burger that Joe *did not* eat. If Joe ate, say, only 2 burgers, (3) would denote the interval [3 , ∞]. Further assuming that the neutral interval $N_s = L([\![\text{C}]\!]) = [4\ ,\ 6]$, (3) is fully contained in the number of burgers not eaten by Joe. This results in the number of burgers eaten by Joe, namely 2, counting as *few* in this context. The truth conditions in (5) are met.

We see that a lexicalized analysis of the negation contained in *few* is easily derived compositionally. However, since there is only one lexical item, this analysis cannot account for the ambiguity that sentences with *few* exhibit, but not sentences with *many* (Heim, 2006, 2008; Solt, 2009). That sentences with *few* carry one more scope-taking element than sentences with *many* can be observed in the following example from Solt (2009, 45f.). The difference is explained with the claim that sentences with *few* can be ambiguous because the negation contributed by *few* can take variable scope, either above or below the modal *can* ($\Diamond$). The differences in truth conditions are exemplified by the mutually exclusive continuations. Sentences with *many* are not ambiguous.

(71) The students can take few advanced classes...

a. '... because not many courses are offered.'
$\rightsquigarrow$ *little* $>$ $\Diamond$ $>$ *many*
It is not possible for students to take a large number of classes.

b. '... and still get their degree.'
$\rightsquigarrow$ $\Diamond$ $>$ *little* $>$ *many*
It is possible for students to not take a large number of classes.

(72) The students can take many advanced classes.
It is possible for students to take a large number of classes.

To account for the ambiguity, Heim (2006, 2008) and Büring (2007a,b) suggest a decomposition of the negative antonym into *little* and the positive antonym. *Little* contributes (adjectival) negation and is scopally mobile. Büring (2007b) suggests the following semantics:

(73) for any gradable adjective A, $[\![\text{little A}]\!] = \lambda d_{\langle d \rangle}.\lambda x_{\langle e \rangle}.\neg[[\![\text{A}]\!](d)(x)]$

This adjectival negation is immediately compatible with the adjectival version of *many* in (57).

However, the decomposition analysis is not an uncontroversial approach to the semantics of antonyms. Heim (2008) points out that there is evidence for and against it, and "the dilemma that results defies a simple solution". I refer the reader to Heim (2006, 2008), Büring (2007a,b) and Beck (2012) for a more thorough introduction into the elusive semantics of antonyms and *little* in particular as well as a discussion of the restrictions on the scope positions of the negation. The decomposition account of *few* as well as the apparent difference between *few* and *many* will be brought up again in Sections 6.7 and 7.7.

2.3 A Surprise-Based Semantics for *few* and *many*

In the last section, we have seen three semantic theories of *few* and *many*. Even though they differ in the semantic properties they ascribe to the quantity words, the predicted truth conditions of sentences with *few* and *many* are essentially the same: for a sentence with *few*, the described cardinality needs to be smaller than a certain threshold value (i.e. the respective boundary of the neutral interval), for *many* it needs to be greater. And these conditions lead directly to the open issue that all three theories leave unanswered. None commits to how these threshold values are derived for the respective comparison class. Romero (2015, 2017) derives at least the superset of the comparison class in a compositional way, but she does not specify how the function L determines the boundaries of the neutral intervals N_s. A more fine-grained variant to set the standard of comparison is suggested in Solt (2011a), where the neutral interval, or standard range R_{Std}, is constructed around the median of the comparison class. "The standard range R_{Std} can be defined as a central range whose width is dependent on the degree of dispersion in the comparison class" (Solt, 2011a, 194).

(74) $[\![\text{Fred is short for a jockey}]\!] = 1$ iff $\text{HEIGHT(Fred)} < \mathrm{R}_{Std}$,
where $\mathrm{R}_{Std} = \text{median}_{x:jockey(x)}(\text{HEIGHT(x)}) \pm \text{n}$; for some value n.

The proposal is, however, only spelled out for sentences in which the comparison class is made overt by a *for*-phrase and a compositional analysis is not presented. This is why it is not clear how and in which form the contextual information is integrated on the basis of which the standard of comparison (i.e. the threshold value) is derived. And even though the derivation of the standard is here described in more detail than in other approaches, Solt (2011a) does not further specify how the deviation n from the median is to be determined for each comparison class. And exactly this is an aspect of Solt's (2011a) proposal that could turn out to be problematic: it commits to an equal distance n below and above the median. Especially in contexts in which the value's distribution is very left- or right-skewed, the equal distance from the median makes wrong predictions. For example, in a comparison class of jockeys, the left-skewed distribution of heights shows a very small dispersion at the left end of the scale, but a larger degree of dispersion towards the upper end. See Figure 2.1 for an illustration. Let's assume that the distribution's median is 153cm. In such a comparison class, a jockey does not need to be much shorter than the median to count as short, say 149cm and below counts as short. This would result in a value of $n = 3$cm. An upper bound of R_{Std} of 153cm + 3cm = 156cm would be too low, however, since a significant group of jockeys is tall up to at least 160cm. Given this example, we suggest to drop the standard range's

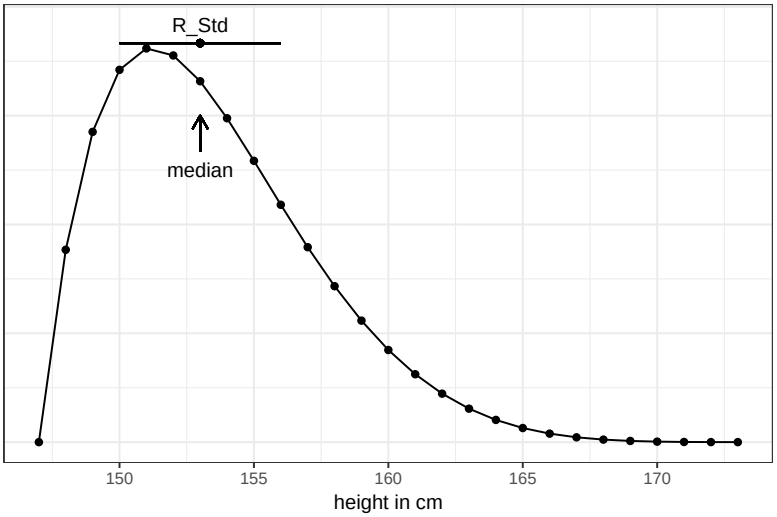


Figure 2.1: Sample distribution for jockey's heights

symmetry assumption. For a definite verdict, the proposal would have to be tested experimentally though.

A further complication in the undertaking of deriving *few* and *many*'s threshold values is that in many cases extensional alternatives as in the compositional analyses (45), (47) or (49) are not sufficient. With a well-known example Keenan and Stavi (1986) demonstrate the variance in *many*'s denotation and how to derive it. The use of *many* cannot even be predicted in a reliable manner when it refers to the same group of individuals. Nouwen (2010) sums up their example:

(75) a. Imagine a conference of lawyers and policemen where normally 60 lawyers and 40 policemen attend. Also, on average, only 10 attendants are women. This year, there are only 20 lawyers, but a staggering 80 policemen. Strikingly, all the lawyers happen to be women and all the policemen are men. (Nouwen, 2010, 238)

b. Many lawyers attend the meeting this year.

c. Many women attended the meeting this year.

Given the context in (75a), (75b) is probably judged false whereas (75c) tends to be accepted even though the set of lawyers and the set of women is exactly the same! "This shows that if the context and the number of relevant individuals are both fixed, *many* still gives rise to different meanings" (Nouwen, 2010, 238). Building on the example above Keenan and Stavi (1986) and Lappin (2000) conclude that *few* and *many* cannot be treated extensionally, because the context, alternatives, and expectations and desires about them play a crucial role. Bastiaanse (2014) arrives at the same conclusion and suggests an intensional treatment, too, but bases his argumentation on Barwise and Cooper's (1981) Generalized Quantifier Theory.

One theory that sets out to resolve the open issue of the derivation of threshold values while allowing for intensionality is the surprise-based semantics proposed by Fernando and Kamp (1996). The intensionality that Keenan and Stavi (1986), Lappin (2000), and Bastiaanse (2014) demand can be accounted for by a systematic incorporation of a measure of expectations into the semantics. In the case of *few* and *many*, prior expectations capture which cardinalities a speaker considers likely or unlikely in a certain context. For example, given the scenario in (75a), a speaker might expect that between 8 and 12 women attend the conference, whereas numbers of women of 7 and lower or of 13 and higher are considered increasingly unlikely. On the other hand, the speaker might consider it likely that roughly between 50 and 70 lawyers attend the conference, but higher or lower numbers are ascribed an increasingly lower possibility. For this reason the same number can count as *many* when compared to expectations about women but not when compared to expectations about lawyers.

Examples like (75) bring up the idea that sentences with *few* and *many* exhibit what we will call "surprise readings". The cardinality described by the quantity word is compared with cardinalities that are considered likely in the situation. Let us for now have a look at a simpler example. The sentence in (76) would then express that the number of cups of coffee drunk by Andy is lower or higher than expected.

(76) Andy drank few / many cups of coffee last week.
$\rightsquigarrow$ Andy drank less / more cups of coffee than expected.

In the context of coffee consumption, the contextual contribution based on which *few* and *many* receive their meaning would be expectations about the number of cups of coffee that Andy or people with Andy's coffee drinking habits might have drunk last week. Some cardinalities might obviously be considered more likely than others in this context. For example, numbers higher than 40 are probably negligible. The possibilities a speaker ascribes to each cardinality can be formalized in a probability distribution P_E and then be the contextual input for the semantics.

The idea that probability distributions play a role in the semantics of vague and context-dependent expressions has already been brought forward by Clark (1991), who builds on early work by Hörmann (1983). Clark's (1991) account of Hörmann's (1983) observation that context-depence is closely related with expectations is discussed in more detail in Section 3.1 and only briefly summarized here. Clark suggests that *few* could rather be taken to denote "the 25th percentile (range: 10th to 40th percentile) on the distribution of items inferred possible in [the current] situation" (Clark, 1991, 271). This approach explains the "cardinal surprise reading" of *few* and *many* in sentences like (76) as intensional, comparing the actual number of

cups of coffee that Andy drank last week to a probabilistic belief P_E which captures our expectations about the number of cups of coffee that Andy drank. This proposal was formally worked out by Fernando and Kamp (1996). We will call it the Clark-Fernando-Kamp (CFK) semantics.

The core idea of the CFK semantics is that *few* and *many*'s lexical meaning is a stable, context-independent function which contains a fixed threshold value θ. This complex yet systematic function maps contextual input to precise denotations by taking as input the cumulative density mass of P_E and cutting it off at a fixed percentage θ. Cardinalities higher than the cut-off would then count as *many*, for example. Here, we focus on the extent to which this approach can explain in particular unstressed cardinal readings as in example (76) (see Section 5.6 for further discussion).

Coming back to the example in (76), the prior expectation P_E is highly context-dependent. It assigns a measure of relative probability to each number n, or more precisely, to each proposition 'Andy drank n cups of coffee last week'. If Andy is a close friend, expectations could be very specific about Andy and his coffee drinking habits. If Andy is a complete stranger, expectations are more likely very general and derive from what one would normally expect from a person like Andy (with fuzziness in what counts as relevantly being like Andy). In both cases, we may think of P_E as reflecting the relevant properties of a *comparison class* in a, perhaps, loose sense of the term. Explicit lexical material or intonational and contextual cues may guide the inference of the relevant P_E. Usually, some uncertainty about the exact properties that form the relevant comparison class will remain.

In contrast to the elusive parameter P_E, there is also a context-independent lexical meaning of *few* and *many*, namely a pair of fixed thresholds θ_{few} and θ_{many} on the cumulative distribution of P_E. Truth conditions of the CFK semantics for sentences as in (76) are given in (77)[5].

(77) **CFK Semantics**

a. $[\![\text{Few As are B}]\!] = 1$ iff $|\,\text{A} \cap \text{B}\,| \leq x_{\max}$

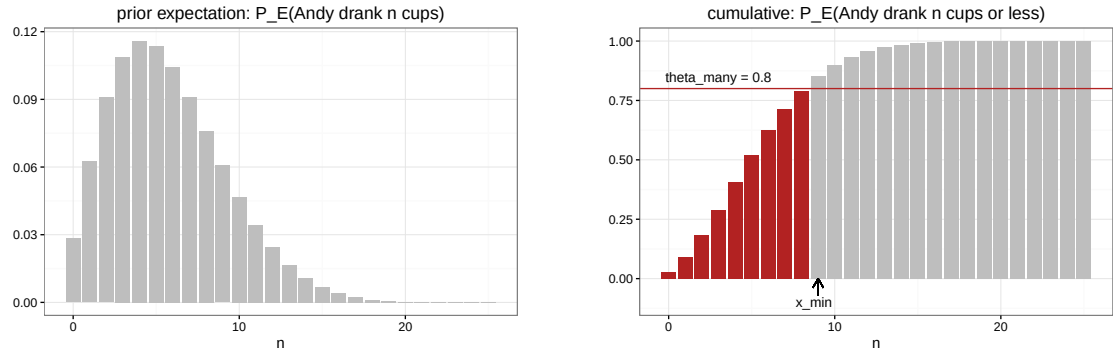
$$\text{where } x_{\max} = \max\{n \in \mathbb{N} \mid P_E(|\,\text{A} \cap \text{B}\,| \leq n) < \theta_{\text{few}}\}$$

b. $[\![\text{Many As are B}]\!] = 1$ iff $|\,\text{A} \cap \text{B}\,| \geq x_{\min}$

$$\text{where } x_{\min} = \min\{n \in \mathbb{N} \mid P_E(|\,\text{A} \cap \text{B}\,| \leq n) > \theta_{\text{many}}\}$$

From (77b), the sentence "Many As are B" is true if the number $n = |\,\text{A} \cap \text{B}\,|$ is no smaller than $x_{\min}$. In turn, $x_{\min}$ is specified as the lowest number for which the

[5]Fernando and Kamp (1996) spell out their semantics in terms of possible worlds. To illustrate the basic idea we opt for a simpler extensional version here, also because we do not find a contradiction to the expectation-based comparison classes we assume. An intensional semantics will be discussed in more detail in the next section.



(a) Prior (b) Cumulative

Figure 2.2: Illustration of the CFK-semantics

cumulative density mass of our prior expectation P_E about the number of As with property B is higher than the semantically fixed threshold θ_{many}. As a result, "Many As are B" is true if the actual number of As with property B is sufficiently surprising, where surprise is relative to contextually-variable P_E and what is sufficient surprise is encoded in contextually-stable θ_{many}.

To illustrate, consider the example in Figure 2.2 for the *many*-sentence in (76). Prior expectations P_E could look like in Figure 2.2a: they would assign a probability to any natural number n, indicating how likely we think it is that Andy drank n cups of coffee last week. Figure 2.2b shows the cumulative distribution of the distribution in Figure 2.2a. If θ_{many} was fixed to, say, 0.8, then the CFK-semantics would identify x_{min} to be 8. Accordingly, for this P_E, the *many*-sentence in (76) would be false for any $n < 8$ and true for any $n \geq 8$.

While such a surprise-based semantics may seem like an appealing idea, it also raises methodological concerns. Since the precise nature of what counts as surprising is hard to assess based on solitary introspection, it becomes exceedingly hard to test the predictions of such an account. The main contribution of Chapter 5 is therefore methodological. We seek to demonstrate how data-driven computational modeling can be a helpful addition to the linguists' toolbox, exactly where solitary introspection fails and the theory under scrutiny concerns *latent parameters* that are not directly observable, like a threshold on a measure of surprise. In other words, we argue here, by means of a case study on the meaning of *many* and *few*, for the usefulness of a particular approach to theoretically inspired statistical modeling of empirical data.

2.4 Comparison Classes and Prior Expectations

Throughout this chapter we have seen various semantic accounts of *few* and *many*. All of them formulate their truth conditions in terms of threshold values, which

determine the applicability of the quantity words. These threshold values in turn are dependent on a comparison class which "in some way serves to provide a frame of reference or standard of comparison" (Solt, 2011a, 190). The comparison class is closely linked to prior expectations, because it determines which information is taken into account when speaker or listener reason about the prior distribution P_E they have in mind. To our knowledge, how a sentence's semantic meaning and comparison classes constrain reasonable prior expectations has not yet been formalized. For example, Fara (2000), Kennedy (2007), Solt (2011a), or Bylinina (2014) make reference to comparison classes, but do not further formalize or formally integrate expectations. We set out to discuss the interesting relationship between the speaker's and listener's epistemic state, their prior expectations and the sentence's semantics.

In the following, we will propose a "moderately radical pragmatic" account of prior expectations, which in concert with a formal semantic analysis of the sentence's contribution explains how speaker and listener arrive at *few* or *many*'s denotation in context. This approach is "radical" because it allows for a lot of freedom in how to obtain P_E. But it is "moderate" in the sense that it requires P_E to be natural and inferable. It is also "moderate" in that it allows the linguistic material to inform the inference of P_E. The pragmatic proposal will be complemented by a formal semantic analysis of sentences containing *few* or *many* and a formalization of prior expectations. Instead of simply assuming that P_E is magically fixed at some point to be able to proceed with the semantic analysis, we will formally derive it from the semantic and pragmatic contributions of the utterance. We propose a modified version of the positive operator *POS*, an intensional degree operator POS^{surp}, to compositionally derive the truth conditions of the surprise reading of *few* and *many*.

2.4.1 A "Moderately Radical Account" of Prior Expectations

P_E is treated as a contextually free variable in the sense that the speaker has a concrete P_E in mind when uttering a sentence with a context-dependent expression, but the listener may have to infer it in order to interpret the utterance. Let us expand on this by assuming that a speaker wants to express with (44) from above that Joe ate more burgers than she had expected him to eat.

(44) $\text{Joe}_{F/CT}$ ate many burgers.

What the speaker does is compare the actual number n of burgers eaten by Joe with her probabilistic belief P_E about the number of consumed burgers. To be clear, the distribution P_E provides the prior probability of Joe eating a certain number of

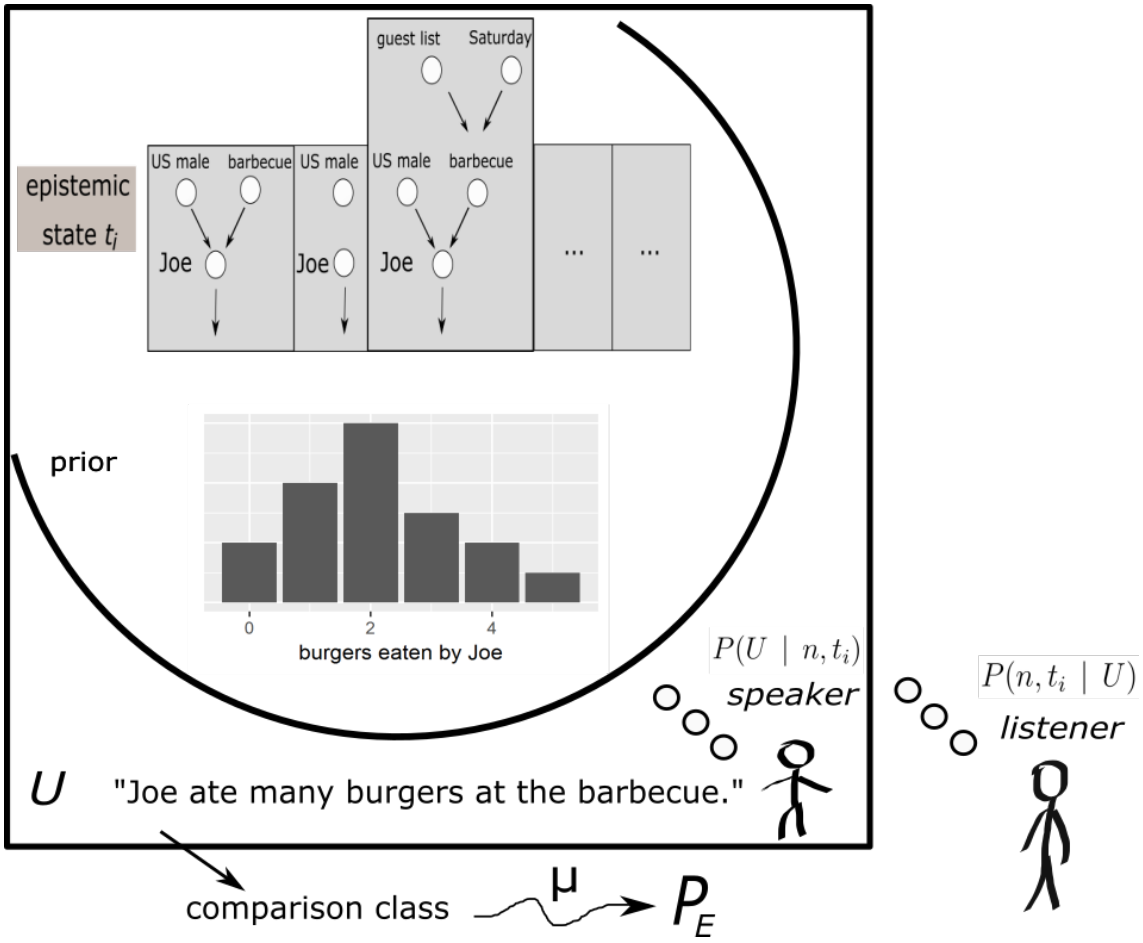


Figure 2.3: This figure depicts the reasoning process of speaker and listener when an utterance U expressing a surprise reading is made. On the basis of her underlying epistemic state t_i, the speaker forms prior expectations P_E about the number of burgers Joe might eat. The speaker then reasons whether to produce U as a description of n, the actual number of burgers eaten by Joe, or to remain silent: $P(U \mid n, t_i)$. Upon hearing U, the listener needs to jointly infer n and t_i: $P(n, t_i \mid U)$. From U, a comparison class can be derived, which constrains the inference of P_E via a measure function μ and thus ultimately also the inference of t_i.

burgers at this event, and this probability was determined before learning about the actual number n of burgers consumed by Joe.

The shape of the prior expectations P_E, and ultimately the decision to make an utterance U, are influenced by the speaker's epistemic state t_i, as illustrated by Figure 2.3 and the conditional production probability $P(U \mid n, t_i)$. This formula expresses the speaker's probability of producing utterance U given a cardinality n and an epistemic state t_i, see also in Section 5.3. The same shape of the distribution P_E can be triggered by different epistemic states t_i. But different epistemic states can also lead to different P_Es. Furthermore, epistemic states may include more or less concrete world knowledge. For example, the speaker can be aware that Joe is a man from the US, that Joe is a man from the US who attended a barbecue, that Joe is a US male who attended a barbecue on Saturday together with certain other guests, and maybe other things. An open question is which bits and pieces of information could or should inform the estimate expressed in a single P_E for the speaker. The information that goes into an estimate could be almost trivial (Joe is a human being, so he cannot eat more than, say, 50 burgers at the most), but it could also be quite elaborate (Joe is a meatlover, likes burgers especially, was not particularly hungry that day...).

Whereas a speaker's P_E is only dependent on her individual epistemic state, the listener's task in a talk exchange is more elaborate. P_E as inferred by the listener is influenced by the utterance (see more below), and by the listener's world knowledge. However, successful communication is only possible if the listener's prior expectations are sufficiently similar to the speaker's. The challenge for the listener is that the comparison class often goes unsaid. Consequently, we see the listener's role as inferring not only a cardinality upon hearing an utterance U, but also the speaker's epistemic state based on which she formed her P_E when uttering U (cf. Tessler et al., 2017). The formula $P(n, t_i \mid U)$ expresses the listener's probability of inferring a cardinality n and the epistemic state t_i given utterance U. Find this illustrated in Figure 2.3, which depicts a listener reasoning both about an utterance and about the speaker and her epistemic state.

The speaker can provide guidance about his epistemic state by explicitly mentioning the sentence's comparison class. Comparison classes can be made overt in the sentence by a *for*-phrase (Kennedy, 2007; Schwarzschild, 2013; Bylinina, 2014) or a *compared to* construction, for example.

(78) a. For a skinny man Joe ate many burgers at the barbecue.

b. Compared to his brother, Joe ate many burgers at the barbecue.

In most cases though, the comparison class is underspecified by the sentence meaning. This leaves the listener with a gap between the information provided by

the lexical material and information necessary to infer the exact epistemic state based on which expectations are formed. Have a look at the example uttered by the speaker in Figure 2.3.

(79) Joe ate many burgers at the barbecue.

The lexical meaning of this sentence only expresses that the number of burgers consumed by Joe at the barbecue is large. It does, however, not restrict the comparison class any further and only gives few hints on the speaker's epistemic state. We do not know from the lexical material whether the number is large for Joe, the vegetarian, or Joe, the meat-lover, for example. Moreover, the sentence itself does not state whether Joe is compared to other guests at the party ($\rightsquigarrow$ he ate more than most other guests) or whether the recent barbecue is compared to previous events ($\rightsquigarrow$ he ate more than at most earlier times). What sentence (79) expresses, however, is that the speaker knows that Joe participated in a barbecue, excluding at least those epistemic states from Figure 2.3 in which Joe went to a burger restaurant, for example.

Furthermore, the listener's inference of the speaker's epistemic state can be influenced by the information structure of the sentence, as exemplified by the contrast between (44) and (50) (repeated from above, see Sections 2.1.3, 2.2.1 and 2.2.2), suggesting that the associate of *POS* can be focus/contrastive topic-marked.

(44) Joe$_{F/CT}$ ate few burgers.

(50) Joe ate many burgers$_{F/CT}$.

In these examples, the speaker gives a hint of what exactly she compares the number of burgers eaten by Joe with, by focus/contrastive topic marking a constituent. Did she compare Joe to other guests, or burgers to other types of food? Note, however, that prosodic information is only a weak cue because it is easy to misperceive (if spoken). The same intonational contour can mark different focus/contrastive topic structures and focus/contrastive topic marking can have reasons other than signaling P_E. For example,

(80) a. Few of the faculty children had a good time.

b. No! Few of the faculty children had a bad$_F$ time.

Another option, which we want to pursue for now, is that the speaker compares the number of burgers actually eaten by Joe to the number of burgers Joe is or could have been expected to eat. Such an intensional comparison class suggests itself to explain surprise readings, which compare the degrees described by a gradable predicate with expectations about the degree. A comparison between the

number of burgers eaten by Joe in (79) and prior expectations about his burger consumption can be formalized as comparing the probability of sets of possible worlds. Further evidence for an intensional comparison class comes from examples like the lawyers-women scenario in (75) (repeated from above), for which a comparison class containing individuals does not provide the correct truth conditions.

(75) a. Imagine a conference of lawyers and policemen where normally 60 lawyers and 40 policemen attend. Also, on average, only 10 attendants are women. This year, there are only 20 lawyers, but a staggering 80 policemen. Strikingly, all the lawyers happen to be women and all the policemen are men. (Nouwen, 2010, 238)

b. Many lawyers attend the meeting this year.

c. Many women attended the meeting this year.

All of these cues in the linguistic material can influence how the listener infers the speaker's epistemic state. In the following, we want to dive further into the semantics. Even though we have seen that the information present in the sentence is not sufficient to identify a single candidate for the contextually free variable P_E, it certainly restricts the set of candidates. This is why we consider a compositional derivation of intensional comparison classes an important next step in bridging the gap between a compositional semantic analysis of the sentence and prior expectations, which are essential input for Fernando and Kamp's (1996) derivation of threshold values.

2.4.2 Compositional Derivation of Comparison Classes and Formalization of P_E

Even though P_E might not be fully specified by the sentence, the set of candidates is considerably restricted by the linguistic material. For example, the topic of sentence (79) is the consumption of burgers and it is far off to compare the number of burgers eaten by Joe with the number of books in the library. For this reason, the compositional analysis of a sentence with a context-dependent expression is a good starting point when its comparison class is to be determined. The biggest challenge will be to formally derive prior expectations while taking the sentence's comparison classes into account. To be clear about the terminology, we will use *comparison class* to refer to the set of focus/contrastive topic alternatives as proposed by Romero (2015, 2017). This set as we understand it here does not immediately fix threshold values, in contrast to e.g. Solt's (2011a) use of the term.

As demonstrated in Sections 2.2.1 and 2.2.2, focus-marking on Joe can be interpreted to derive an extensional comparison class over alternative individuals.

(44) Joe$_{F/CT}$ ate many burgers.

(81) $[\![\text{C}]\!] \subseteq \{\lambda d. \,|\, \{x : *\text{burgers}(x)\} \cap \{x : \text{Joe ate } x\} \,|\, \geq d,$
$\lambda d. \,|\, \{x : *\text{burgers}(x)\} \cap \{x : \text{Max ate } x\} \,|\, \geq d,$
$\lambda d. \,|\, \{x : *\text{burgers}(x)\} \cap \{x : \text{Sue ate } x\} \,|\, \geq d, ...\}$

This comparison class triggers intuitive truth conditions requiring that Joe ate more burgers than most other relevant individuals. Note, however, that input into the CFK-semantics is a probability distribution over cardinalities. A distribution which can be calculated directly from an extensional comparison class is a frequency distribution, resulting from summing up those individuals whose set of degrees in $[\![\text{C}]\!]$ has the same maximal degree. This distribution is then normalized to sum up to 1 to become a probability distribution. In the burger example, the probability assigned to a degree (i.e. cardinality) is just the normalized count (i.e. the proportion) of guests who have eaten this cardinality of burgers. Thus obtaining a probability distribution from a comparison class in (81) is mathematically possible and could provide the correct truth conditions for a case in which the speaker knows the number of burgers eaten by every guest at the party. This comparison class, cannot, however, account for subjective beliefs and expectations. Moreover, if extensional focus alternatives were all that matters for the inference of P_E, we would run into problems especially in the case where there are very few alternatives. If there were only two alternatives to Joe (i.e. two other people at the barbecue), we only ever get three cardinalities. The construction of P_E as a frequency distribution would give us equal probability for these three cardinalities and zero probability for all others. For three guests at the barbecue and $\theta_{\text{many}} > 0.67$ the sentence "Joe$_F$ ate many burgers" would be true if Joe ate most of all, independent of what this number may be. This does not always have to be the case.

If (79) (for now without overt focus) is interpreted as comparing Joe's burger consumption to beliefs about his individual burger eating habits or to beliefs about burger consumption in the overall population, the comparison class in (81) is not sufficient; an intensional one is necessary, as already pointed out by Romero (2017). We assume (and will show experimentally in the following chapters) that the most salient reading of *few* and *many* in sentences like (79) is the surprise reading, which interacts closely with subjective prior beliefs about the context. For this reason, we suggest to employ alternative possible worlds in the formalization of every surprise reading of *few* and *many*, independently of overt focus marking as in (44).

To achieve our goal of formalizing beliefs with an intensional semantic account of the sentence and to formalize the inference of P_E, some more effort has to be put into modifying the positive operator *POS*, which will be given more semantic and pragmatic power. Traditionally, *POS* serves to infer a value assignment for the

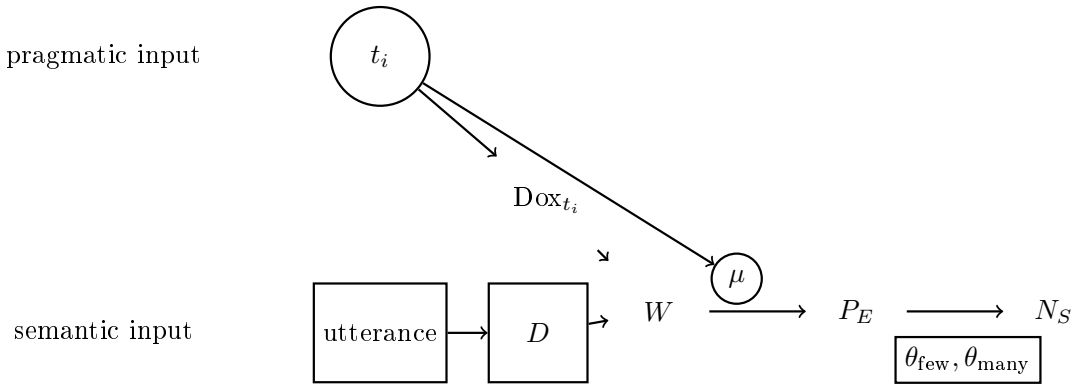


Figure 2.4: Visualization of the semantic (rectangles) and pragmatic (circles) components of the intensional degree operator POS^{surp} in (87)

free degree variable introduced by a gradable adjective or, in our case, a quantity word (Schwarz, 2010; Hohaus, 2015). We build on this role assigned to *POS* and expand its influence. So far, the version in (29) does not specify how exactly P_E or the neutral interval interact with the sentence meaning. *POS* takes as its first argument a comparison class C, which is input to the function L. L returns the so-called neutral interval $N_s = L([\![\text{C}]\!])$ of the comparison class. But in which way C is influenced by prior expectations and how exactly the neutral interval is determined is not spelled out in (29).

(29) $[\![POS]\!] = \lambda \text{C}_{\langle dt,t\rangle}.\lambda D_{\langle dt,t\rangle}.L_{\langle\langle dt,t\rangle,\langle d,t\rangle\rangle}(\text{C}) \subseteq D$

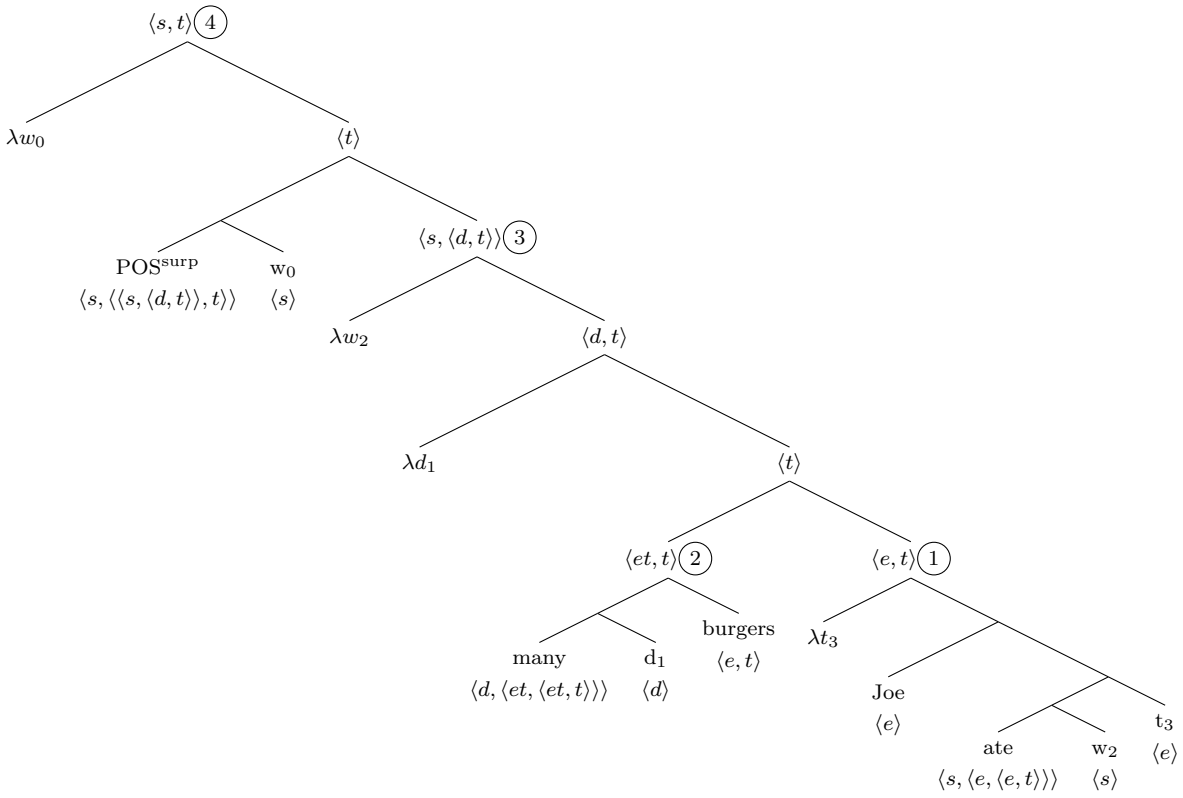
In the following, we propose an intensional degree operator POS^{surp} with a doxastic modal base. In this respect, POS^{surp} has interesting parallels with Meier's (2003) semantic proposal for *too* and *enough*. In order to connect our proposal for a "surprise-reading version" of *POS* with the discussion in the previous section and with Figure 2.3, let us start at the pragmatic level. The epistemic state t_i held by the speaker and inferred by the listener is a rich representation of beliefs about the world which incorporates all sorts of causal and epistemic dependencies. In a conversation, however, we are only concerned with those aspects of the world which are relevant for the evaluation of the recent utterance. These relevant aspects are focused by the question under discussion (QUD) (Roberts, 1996). We take it that the QUD is represented by an intensional comparison class W brought forward by the compositional analysis of the sentence. POS^{surp} introduces alternative possible worlds to the world variable w_0 in the intensionalized set of degrees which it takes as its argument, parallel to the Romero-style extensional comparison classes. This intensional comparison class W, is a set of properties of degrees which are linked to worlds compatible with $\text{Dox}_{t_i}(\text{w}_0)$, the beliefs held in the epistemic state t_i. W is the minimal semantic contribution necessary to derive P_E. In order to now systematically derive P_E from t_i while taking W into account, we introduce a free pragmatic

variable into POS^{surp}, the measure function μ. μ measures our beliefs given t_i, by assigning the worlds in $\text{Dox}_{t_i}(\text{w}_0)$ a probability, resulting in P_E. While being a representation of t_i, μ is more coarse-grained since it only takes into account those aspects of t_i which are specified by W. The listener will infer μ and possibly t_i from the utterance, but once she has done this, P_E is fully determined by μ and W. From P_E, POS^{surp} calculates $x_{\max}$ and $x_{\min}$ via θ_{few} and θ_{many}. Find the components of POS^{surp} visualized in Figure 2.4.

In the following, our proposal will be explained in detail by deriving the surprise reading of the sentence

(82) Joe ate many burgers.

The logical form for this sentence looks as follows[6]:



The first steps of the compositional analysis are given below:

(83) a. $(1) = \lambda x.$ Joe ate x in w_2

b. $[\![\text{many}_{card}]\!] = \lambda d_d.\lambda P_{\langle e,\,t\rangle}.\lambda Q_{\langle e,\,t\rangle}.\,|\,P \cap Q\,| \geq d$

c. $(2) = \lambda Q_{\langle e,\,t\rangle}.\,|\,\{x : {*}\text{burgers}(x)\} \cap Q\,| \geq \text{d}_1$

d. $(3) = \lambda w.\lambda d.\,|\,\{x : x \text{ are burgers \& Joe ate } x \text{ in } w\}\,| \geq d\,(=: D)$

On the semantic side, the role of POS will be extended to derive the sentence's intensional comparison class W. W can be thought of as a QUD and determines

[6]We continue with POS^{surp} of a different type from (29) ($\langle s, \langle\langle s, \langle d, t\rangle\rangle, t\rangle\rangle$), see (87).

what goes into P_E. To derive W, POS^{surp} takes as its argument an intensionalized set of degrees D and a world variable w_0. In the present example, D contains the intensionalized set of degrees corresponding to cardinalities of burgers eaten by Joe in some world, as given in (3) in the LF and spelled out in (83d). The version of POS we assume for the surprise reading then creates a set of sets of degrees by applying D to the doxastic alternatives of the evaluation world w_0. The result is the sentence's intensional comparison class W.

(84) $W := \{D(w) : w \in \mathrm{Dox}_{t_i}(\mathrm{w}_0)\}$
$= \{\lambda d'. \,|\, \{x : x \text{ are burgers \& Joe ate } x \text{ in } w\} \,|\, \geq d' : w \in \mathrm{Dox}_{t_i}(\mathrm{w}_0)\}$

The introduction of intensions via POS^{surp} proceeds in a similar way to the introduction of extensional alternatives proposed by Romero (2015, 2017), and triggers a set of alternative worlds. This set of alternative worlds is restricted by a conversational background $\mathrm{Dox}_{t_i}(\mathrm{w}_0)$, which is compatible with the speaker's or listener's beliefs in the actual world w_0 given their epistemic state t_i. For several reasons, we assign the introduction of alternative worlds to POS^{surp} and do not assume focus-marking on the world variable, which would be even more parallel to the extensional case. First, we believe that every surprise reading requires an intensional component to enable the inference of prior beliefs. Second, a surprise reading can also be available when overt focus marking as in (44) is present. Overt focus marking will be analyzed "traditionally" in the composition. Below, we will elaborate on how the extensional and the intensional alternatives interact to influence the derivation of P_E. Third, the association with prior expectations is not only driven by semantic, but rather by pragmatic mechanisms. These are contributed by POS^{surp} as will be demonstrated presently.

Once POS^{surp} has determined the intensional comparison class W, the next step towards a probability distribution over cardinalities assumed to be performed by POS^{surp} is to employ a probabilistic measure function μ to the worlds in $\mathrm{Dox}_{t_i}(\mathrm{w}_0)$ which are also linked to the properties made relevant by the QUD, as specified by W. The proposal in (85) is a formalization of the reasoning process from an epistemic state t_i and an utterance triggering W to a probability distribution P_E.

Note that it is crucially the intensional properties of degrees in W, see (84) for an example, which determines what goes into P_E. Among other things, W greatly influences the probability distribution's domain. As pointed out before, if W contains sets of degrees related to burger eating, P_E cannot express expectations about the number of girlfriends Joe had before he got married[7].

[7]For this reason, it is essential that W contains *properties* of degrees and not abstract degrees (i.e. numbers).

(85) $\mu : \mathrm{Dox}_{t_i}(\mathrm{w}_0) \rightarrow [0,1]$
and $P_E(m) = \sum_{w \in S_m} \mu(w)$
where $S_m = \{w : w \in \mathrm{Dox}_{t_i}(\mathrm{w}_0)\ \&\ \max(D(w)) = m\}$ and $m \in \mathbb{N}$

The measure function μ assigns each world in $\mathrm{Dox}_{t_i}(\mathrm{w}_0)$ a probability between 0 and 1, resulting in the probability distribution P_E. μ is constrained by the utterance, however, in that its domain is restricted to those worlds which are linked to a property of degrees as specified by W. The prior probability of a cardinality m under P_E is then the sum of μ applied to all doxastic alternatives to w_0 whose maximal degree in W is m. In our example, Joe would have eaten five burgers in every world contained in S_5. The prior probability of Joe eating five burgers, $P_E(5)$, would be calculated by applying μ to all $w \in \mathrm{Dox}_{t_i}(\mathrm{w}_0)$ which fulfill

(86) $\max(\lambda d. \,|\, \{x : x \text{ are burgers } \&\ \text{Joe ate } x \text{ in } w\} \,|\, \geq d) = 5$

and summing up their values. For simplicity we assume that the set $\mathrm{Dox}_{t_i}(\mathrm{w}_0)$ is finite[8].

The final step in the derivation of the truth conditions is to determine the neutral interval $N_S = [x_{\max}, x_{\min}]$ from the resulting P_E following the CFK semantics in (77). The context-independent threshold values θ_{few} and θ_{many} are applied to the cumulative density mass of P_E to determine $x_{\max}$ and $x_{\min}$. If the sister of POS^{surp}, the intensional set of degrees D as given in (3), fully contains N_S, the sentence is predicted to be true. For the hypothetical prior distribution in P_E in Figure 2.3 and a hypothetical value of $\theta_{\text{many}} = 0.7$, the sentence "Joe ate many burgers at the barbecue" would be true if he ate four burgers or more.

In the following, we want to integrate all of these individual components into one covert operator POS^{surp}, which can derive the surprise reading of *few* and *many*. After that, we present the entire compositional analysis of our example sentence and move on to apply the proposed version of POS^{surp} to a sentence with overt focus marking.

(87) $[\![POS^{\text{surp}}_{\mu,t_i}]\!] = \lambda w_0.\, \lambda D_{\langle s, \langle d, t\rangle\rangle} :\ N_S = [x_{\max}, x_{\min}]$ and
$x_{\max} = \max\{n : \sum_{m=0}^{n} P_E(m) \leq \theta_{\text{few}}\}$ and
$x_{\min} = \min\{n : \sum_{m=0}^{n} P_E(m) \geq \theta_{\text{many}}\}$
for $P_E(m) = \sum_{w \in S_m} \mu(w)$ and $m \in \mathbb{N}$
and $S_m = \{w : w \in \mathrm{Dox}_{t_i}(\mathrm{w}_0)\ \&\ \max(D(w)) = m\}$
$N_S \subseteq D(w_0)$

POS^{surp} predicts a sentence to be true iff the neutral interval N_S is fully contained in the set of degrees D denoted by the sentence. In the case of quantity

[8]Nothing hinges on this. For the infinite case, take a (Lebesgue-)integral instead of a sum and require that μ satisfies the necessary properties for (Lebesgue-) integration.

words, the boundaries of N_S, $x_{\max}$ and $x_{\min}$, are defined as the highest or lowest number for which the cumulative density mass of the speaker's prior expectations is lower or higher than the threshold value θ_{few} or θ_{many}. Prior expectations, in turn, are inferred via applying a measure function μ to the doxastic alternatives of w_0 compatible with D, relative to t_i[9]. The result of applying μ to $\text{Dox}_{t_i}(\text{w}_0)$ is the probability distribution P_E. μ and P_E are constrained by the sentence's intensional comparison class W because μ's domain is restricted to those worlds which are linked to a property of degrees specified by W. The probability of a single cardinality m is the sum of μ applied to all doxastic alternatives to w_0 in S_m, those worlds whose maximal degree in $W = \{D(w) : w \in \text{Dox}_{t_i}(\text{w}_0)\}$ is m.

Taking all this together, we modified the positive operator to account for surprise readings, to systematically calculate the neutral interval and to formally derive prior expectations which are compatible with the sentence's semantic contribution.

The compositional analysis of our example sentence "Joe ate many burgers" is given below:

(88) a. $\textcircled{1} = \lambda x.$ Joe ate x in w_0

b. $[\![\text{many}_{card}]\!] = \lambda d_d.\lambda P_{\langle e,t\rangle}.\lambda Q_{\langle e,t\rangle}.\,|\,P \cap Q\,| \geq d$

c. $\textcircled{2} = \lambda Q_{\langle e,t\rangle}.\,|\,\{x : *\text{burgers}(x)\} \cap Q\,| \geq \text{d}_1$

d. $\textcircled{3} = \lambda w.\lambda d.\,|\,\{x : x \text{ are burgers \& Joe ate } x \text{ in w}\}\,| \geq d =: D$

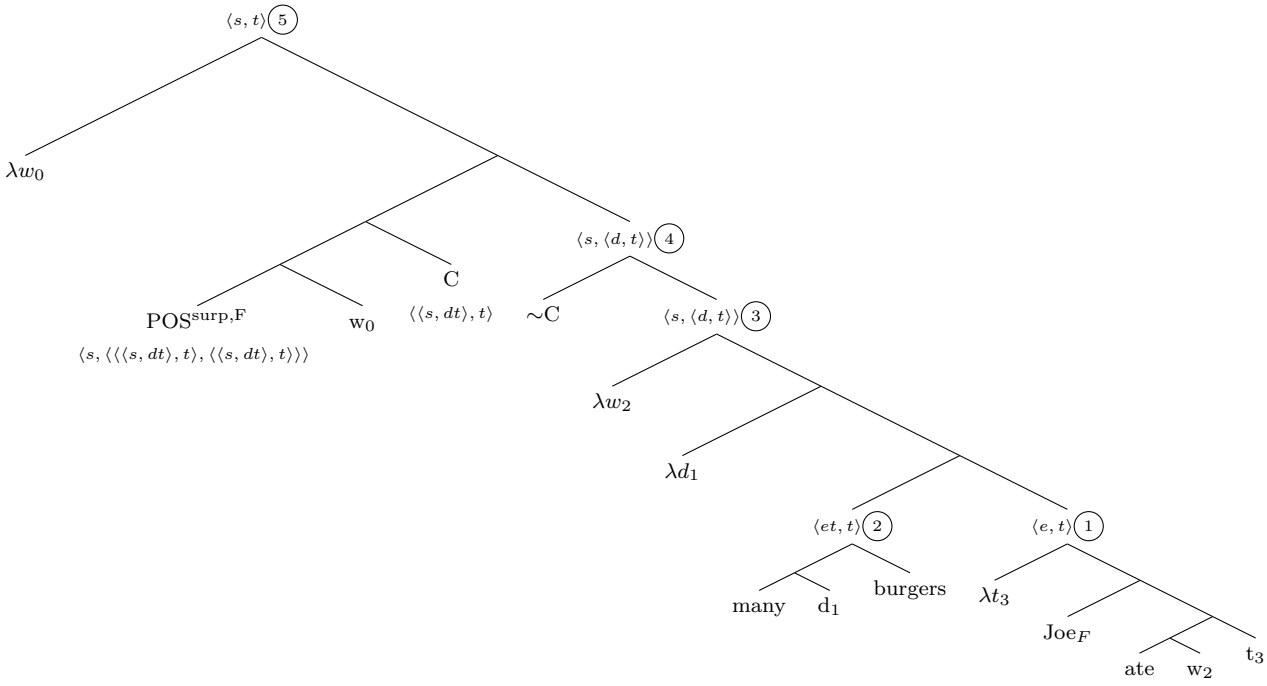
e. $W := \{D(w) :\ w \in \text{Dox}_{t_i}(\text{w}_0)\}$
$= \{\lambda d.\,|\,\{x : x \text{ are burgers \& Joe ate } x \text{ in } w\}\,| \geq d :\ w \in \text{Dox}_{t_i}(\text{w}_0)\}$

f. $\textcircled{4} = \lambda w_0.\, N_S = [x_{\max}, x_{\min}] \subseteq$
$\lambda d.\,|\,\{x : x \text{ are burgers\&Joe ate } x \text{ in } w_0\}\,| \geq d$
$\textcircled{4}$ is only defined iff N_S can be calculated from P_E for the contextually given, inferred or assumed μ

Next, we want to apply the developed account for surprise readings to sentences with overt focus. In contrast to Romero's (2015) purely extensional analysis, we demonstrate how prior expectations can be systematically derived from both an intensional and an extensional comparison class. Remember that one reason for outsourcing the introduction of alternative worlds to POS^{surp} was to keep open the possibility of having overt focus in the sentence, which can then be analyzed "conventionally" with a semantics of focus interpretation, as we have seen in Section 2.2.1. We assume the following LF for the sentence

(44) $\text{Joe}_{F/CT}$ ate many burgers.

[9] We add the subscripts μ and t_i to POS^{surp} to indicate their status as free variables.



To be able to account for the surprise reading of (44), to systematically derive P_E and N_S and to include an extensional comparison class C, we employ a version of surprise-*POS* which has an argument slot for C: the focus-sensitive, intensional degree modifier $POS^{surp,F}$.

$$
\begin{aligned}
(89)\quad [\![POS^{\text{surp,F}}_{\mu,t_i}]\!] = \lambda w_0.\, \lambda \text{C}_{\langle\langle s,dt\rangle,t\rangle}.\ \lambda D_{\langle s,\langle d,t\rangle\rangle} :\ & N_S = [x_{\max}, x_{\min}] \text{ and} \\
& x_{\max} = \max\{n : \textstyle\sum_{m=0}^{n} P_E(m) \leq \theta_{\text{few}}\} \text{ and} \\
& x_{\min} = \min\{n : \textstyle\sum_{m=0}^{n} P_E(m) \geq \theta_{\text{many}}\} \\
& \text{for } P_E(m) = \textstyle\sum_{w \in S_m} \mu(w) \text{ and } \text{C}(w_0) \sim P_E \text{ and } m \in \mathbb{N} \\
& \text{and } S_m = \{w : w \in \text{Dox}_{t_i}(\text{w}_0)\ \&\ \max(D(w)) = m\} \\
N_S \subseteq D(w_0)
\end{aligned}
$$

The lexical entry for surprise-*POS* in a sentence carrying focus, $POS^{\text{surp,F}}$, is nearly identical to (87)[10] with the exception of taking as its first argument the covert variable C (cf. Schwarz, 2010; Hohaus, 2015). $\text{C}(\text{w}_0)$ represents the sentence's extensional comparison class and contains the sets of degrees corresponding to the alternatives of the focus/topic-marked constituent in the sentence. The constraint on P_E introduced by C is that $\text{C}(\text{w}_0)$ has to be a likely sample of P_E. $\text{C}(\text{w}_0) \sim P_E$ is a mild pragmatic constraint for the listener's inference of μ/t_i. This means that the underlying epistemic state t_i requires P_E to be compatible with the sentence's extensional comparison class triggered by focus/topic-marking. Focus-marking on

[10]The subscripts μ and t_i are also added to $POS^{\text{surp,F}}$ to indicate their status as free variables.

Joe suggests a comparison between Joe and other guests at the barbecue and indicates that *a priori* Joe was not considered to be different from the other guests in terms of his burger eating habits. Consequently, P_E is once more the link between the hints in the linguistic material and the underlying epistemic state t_i.

Find a compositional analysis of sentence (44) below:

(90) a. $\textcircled{1} = \lambda x.$ Joe ate x in w_0

b. $[\![\mathrm{many}_{card}]\!] = \lambda d_d.\lambda P_{\langle e,t\rangle}.\lambda Q_{\langle e,t\rangle}.\,|\,P \cap Q\,| \geq d$

c. $\textcircled{2} = \lambda Q_{\langle e,t\rangle}.\,|\,\{x : {}^*\mathrm{burgers}(x)\} \cap Q\,| \geq \mathrm{d}_1$

d. $\textcircled{3} = \lambda w.\lambda d.\,|\,\{x : x \text{ are burgers \& Joe ate } x \text{ in } w\}\,| \geq d =: D$

e. $W := \{D(w) :\ w \in \mathrm{Dox}_{t_i}(\mathrm{w}_0)\}$
$= \{\lambda d.\,|\,\{x : x \text{ are burgers \& Joe ate } x \text{ in } w\}\,| \geq d :\ w \in \mathrm{Dox}_{t_i}(\mathrm{w}_0)\}$

f. $\textcircled{4}$ is defined iff
$[\![\mathrm{C}]\!] \subseteq \{\lambda w.\lambda d.\,|\,\{x : x \text{ are burgers \& Joe ate } x \text{ in } w\}\,| \geq d,$
$\lambda w.\lambda d.\,|\,\{x : x \text{ are burgers \& Max ate } x \text{ in } w\}\,| \geq d,$
$\lambda w.\lambda d.\,|\,\{x : x \text{ are burgers \& Sue ate } x \text{ in } w\}\,| \geq d, ...\}$

g. $\textcircled{5} = \lambda w_0.\, N_S = [x_{\max}, x_{\min}] \subseteq$
$\lambda d.\,|\,\{x : x \text{ are burgers \& Joe ate } x \text{ in } w_0\}\,| \geq d$
$\textcircled{5}$ is only defined iff N_S can be calculated from P_E for the contextually given, inferred or assumed μ

This formal semantic account of the concept of prior expectations takes the sentence's extensional comparison class C as input, but allows for a "moderately radical" influence of the pragmatics by including subjective beliefs via an independent, intensional comparison class W, which functions like a QUD.

2.4.3 Discussion

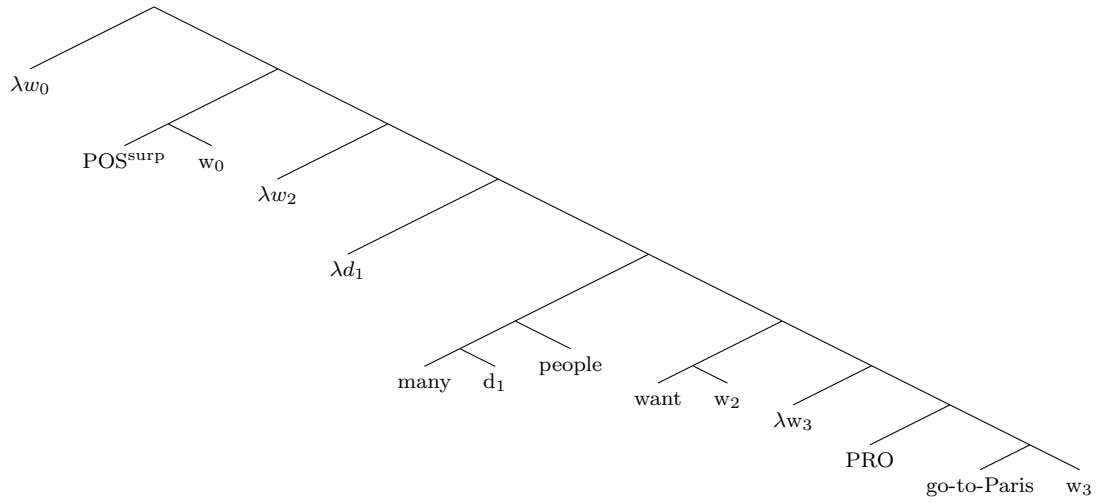
In this section, we proposed a modified version of the positive operator *POS*. With POS^{surp} we are able to account for surprise readings of *few* and *many*. This operator derives prior expectations which are compatible with the sentence's semantic contribution in a systematic way and calculates the neutral interval from them. However, there are several issues with the presented modification of POS^{surp} that still need to be addressed. Not all of them can be answered in detail in the scope of this dissertation though.

The first is how exactly the reasoning process of speaker and listener really takes place when producing or interpreting utterances based on prior expectations P_E. Which information is taken into account and how do we get from an epistemic state to the prior distribution P_E? And how much of this is part of the semantics? This is directly related to the question of which aspects influence the inference of the

measure function, the free variable μ. Moreover, what happens if several epistemic states are equally likely given the listener's inference and world knowledge? We have sketched a rough idea here, but future work with contributions from cognitive science, philosophy of mind and psycholinguistics is required to test the presented idea and factors influencing expectations in general.

Second, the surprise version of *POS* in (87) assumes that POS^{surp} takes as its argument its sister D of type $\langle s, \langle d, t\rangle\rangle$. We intensionalized the property of degrees to be also able to account for sentences which contain more than one world variable. It would be prudent to investigate more closely in future work whether we run into compositional problems when dealing with sentences containing other intensional operators like modals or *too*. An example of a sentence containing a modal would be

(91) Many people want to go to Paris.



As far as we can tell now, constructing the intensional comparison class via argument insertion into D should prevent problems with different scope configurations. This should be confirmed with more data though.

Third, the version of *POS* in (87) is inspired by the idea of the CFK semantics that *few* and *many* can express surprise readings, comparing the actual degree to expected degrees. A question that suggests itself is whether this assumption transfers to gradable adjectives as well. At first thought, it is not implausible that gradable adjectives draw on prior expectations. Nevertheless, to what extent gradable adjectives express surprise should be investigated more carefully. This also brings up the open questions of why there are different versions of *POS* and how they are related. Similarly, it is not clear whether *all* readings of *few* and *many* are surprise readings. An area of future research should be to investigate if there are cases which are not dependent on prior expectations and how to account for them. Barker (2002) claims

that gradable expressions can have more than descriptive uses. He suggests that a sentence like "Feynman is tall" can also have a metalinguistic use giving "guidance concerning what the prevailing relevant standard for tallness happens to be in our community" (Barker, 2002, 2). Whether this use of gradable adjectives really cannot be analyzed on the basis of prior expectations requires a more thorough reflection.

Another comment in relation to POS^{surp} is whether the mechanism of restricting the context relative to the speaker's or listener's epistemic state is not also applied elsewhere. It would only be economical if language transferred the strategy of dealing with context-dependence to other phenomena. An example would be nominal genericity. The domain of the universal quantifier *every* needs to be restricted to match the context of the utterance[11].

(92) Every student managed to explain the semantics of the definite determiner.

This sentence probably does not express that every student in our universe has some knowledge in formal semantics. Instead, the sentence would rather be interpreted as conveying that every student *in some relevant course* managed to explain the semantics of the definite determiner, even though this is not explicitly stated in the sentence. It would be interesting to further investigate possible parallels between POS^{surp} and the universal quantifier.

Last, Fernando and Kamp's (1996) theory is not the only one making predictions about the calculation of threshold values. Lassiter and Goodman (2015) defend a pragmatic approach whereas Qing and Franke (2014a) opt for an explanation based on evolutionary linguistics and optimal language use. The presented semantics of POS^{surp} in (87) only incorporates the CFK semantics and does not take into account competing theories. Which theory eventually makes the correct predictions of how speakers and listeners use context-dependent adjectives and quantity words will have to be tested experimentally. In the following chapters, we make a start on this undertaking by investigating the predictions of the CFK semantics for the surprise reading of cardinal and proportional *few* and *many*.

[11] Thanks to Vera Hohaus for pointing this out to me.

Chapter 3

Psychological Studies on *few* and *many*

In the previous chapter, an overview over the many readings of *few* and *many* was provided, accompanied by several semantic theories of how to account for them. After this rather theoretical introduction, we now want to make use of the concepts and terminology presented above and delve into experimental data about how *few* and *many* are used.

Section 2.3 introduced semantic theory by Fernando and Kamp (1996), which connects the meaning of *few* and *many* with expectations of the context. We called it the CFK semantics and a main part of this dissertation sets out to experimentally test its predictions. Before we start this venture, however, it is wise to be familiar with previous experiments on the context-dependence of *few* and *many* and, in particular, on their interaction with prior expectations. This chapter presents representative psychological work in this area.

Section 3.1 introduces a hypothesis by Clark (1991), which suggests to represent prior expectations as a probability distribution on which *few* and *many* impose a threshold. This idea was then picked up by Fernando and Kamp (1996), who develop it into a formal semantic account as described in the previous chapter. A large body of work produced by Moxey and Sanford is summarized in Section 3.2. They also identify prior expectations of the contexts as a factor which influences the use and interpretation of *few* and *many*. Newstead and Coventry (2000) and Coventry et al. (2005, 2010) point out expectations as a possible source of variation in the use of quantity words, too, but they explain subjects' behavior in terms of other factors such as visual cues. Their experiments on *few* and *many* are presented in Section 3.3. In Section 5.1 we replicate one of their experiments and show that the various visual factors with which they explain their findings can be summarized into the single factor *prior expectations*.

3.1 Early Work on the Context-Dependence of *few* and *many*

Early work on quantity words[1] at the interface of psychology and linguistics was conducted by Hörmann (1983) and Clark (1991). Both investigate the effects of the context on context-dependent expressions. Clark (1991) tentatively suggests an intuitive semantics for *few* and *many*. Clark, citing Hörmann (1983), argues that it is impossible to provide a dictionary theory for *few* and *many*. A *dictionary theory* would assume that the meaning of a word can be listed as a "a brief, partial description of some aspect of the world... every word has a lexical entry in memory that pairs a phonological shape, like 'dog', with a conventional meaning, like 'canine animal" (Clark, 1991, 264). For the meanings of *few* and *many*, Clark (1991) argues, it is impossible to come up with a short or even a finite list of denotations because conditions of use and interpretations vary highly between different situations. Variation is, for example, triggered by the physical size of the discussed objects which clearly influences how people judge possible amounts and estimate corresponding numbers (Hörmann, 1983). For example, "many people in front of the city hall" is interpreted as a larger cardinality than "many people in front of a hut" and "many bread crumbs" is interpreted to be more than "many mountains". This is the first time that the concept of prior expectations was mentioned in connection with *few* and *many*. Another problematic issue of a dictionary theory is that the numerical denotations of *few* and *many* are not "really fixed for each item on the list" (Clark, 1991, 270). As an alternative to a dictionary theory, Clark suggests that, e.g., *few* could rather be taken to denote "the 25th percentile (range: 10th to 40th percentile) on the distribution of items inferred possible in [the current] situation" (Clark, 1991, 271). The idea that a quantity word can be understood as denoting a simple function which takes a context-dependent value was taken up and formally spelled out by Fernando and Kamp (1996), see Section 2.3. The semantic account will be tested experimentally in the following chapters.

3.2 Influence of the Context and Expectations

A vast body of research on the use of *few* and *many* has been produced by Moxey and Sanford. They have not only investigated whether the meanings of quantifiers can be

[1] Chapter 2 introduces the linguistic background of *few* and *many* and discusses the controversy of how to classify them semantically. To remain non-committal, *few* and *many* are labeled "quantity words" in the remainder of this dissertation. In this chapter, though, the term "quantifier" will be used occasionally for consistency with the literature discussed. "Quantifier" is the common label of *few* and *many* in the psychological literature and they are often examined on par with other quantifiers like *a few, several* or *lots of*.

mapped onto a scale but also which extralinguistic knowledge they communicate. A word-to-scale mapping account of quantifier meaning assumes that "a mapping will hold between a distribution of numbers and a natural language quantifier" (Moxey and Sanford, 2000, 239-240). Even though this idea might seem appealing and intuitive at first sight, Moxey and Sanford (2000) argue against a rigid quantifier-to-number mapping and point out several problems. First, it is not possible to map the large number of quantity expressions onto a scale in a distinct way. In a study of ten quantifiers, Moxey and Sanford (1993) asked participants in a one-shot experiment to assign a number to a single quantifier chosen randomly from the data set. "Several quantifiers were simply not distinguishable from one another (*a few, only a few, not many, few,* and *very few*)". Second, there is convincing evidence against a stable, context-independent linking function between quantifier meaning and the cardinality scale since "values assigned by participants depend upon context" (Moxey and Sanford, 2000, 241). Plenty of examples of *few* and *many*'s context-dependence are given in the course of this dissertation.

Furthermore, a quantifier-to-number mapping account cannot capture that *few* and *many* express more than just a reference to a number or an interval on a scale. One core finding of Sanford et al. (1994) is that "quantifiers may be differentiated in terms of the patterns of focus which they produce"[2] (Sanford et al., 1994, 153). The example below shows that even when *a few* and *few* make reference to the same number, the difference in focus means that the sentence endings are not interchangeable (cf. Moxey, 2006, 423).

(93) Context: 5 out of 60 passengers were killed in an accident.
   a. A few of the passengers were killed in the accident, which is awful.
   b. Few of the passengers were killed in the accident, which is good news.

Many and *a few*, being "positive quantifier", typically make reference to the set whose cardinality is described and made reference to. This set is calles the *reference set*.

(94) Many of the football fans went to the match. They cheered loudly when the player scored.

Here, the pronoun *they* describes the reference set, the fans present at the match. The reference set is made salient by *many*.

Few, however, behaves differently from *many* being a "negative quantifier". *Few* expresses negation (cf. Heim, 2006) and thus is downward monotone and licenses negative polarity items (cf. Sanford et al., 1994, 157). Another interesting property

[2]The term "focus" is used here to mark the set that is made salient by the utterance, not in the sense of a prosodically marked constituent which triggers semantic alternatives (cf. Rooth, 1985).

that Sanford et al. (1994) investigate is that negative quantifiers can activate the complement set (compset) of the cardinality they describe.

(95) Few of the football fans went to the match. They watched the match at home instead.

They does not refer to the entities quantified over by *few*, the football fans present at the match. Instead, their complement set is activated: the fans who are *not* present. Moxey and Sanford (1987) confirm these observations experimentally. They asked participants in an experiment to continue sentences like

(96) Few of the football fans were at the match. They...

Negative quantifiers like *not all, not very many, not many, few, very few, hardly any* turned out to be complement set licensing in most cases. This means that they "can put focus on those As which are not Bs" (Sanford et al., 1994, 158). In the study, participants used to make reference to the set of fans who were not at the match and gave reasons for their absence, for example. Positive quantifiers like *nearly all, many, some, a few* rather make reference to the reference set, those As which are Bs. These interesting findings are further investigated by Moxey (2006), see below. We will pick up these results in Section 6.7 where we discuss the effect which *surprisingly* has on *many* but not on *few*.

Another study presented by Moxey and Sanford (1993) examined whether the choice of a positive or a negative quantifier expresses more than just the quantity and the reference set. Instead, *few* "might signal that the speaker's prior beliefs were to the effect that he expected more to be the case" (Sanford et al., 1994, 162). Subjects were presented with a quantified statement in a dialogue and were asked about the beliefs of the speaker and the listener. The experiment showed that negative quantifiers like *very few, few, not many* were associated with the listener believing that the speaker had expected more than turned out to be the case. In contrast, for the positive quantifier *a few* this did not hold. Moxey and Sanford take away from their studies that "quantifiers may convey information about the speaker's beliefs as well as about the current situation" (Sanford et al., 1994, 164). These experimental results constitute evidence for making the semantics of *few* and *many* dependent on prior expectations, as suggested in Section 2.3.

Moxey (2006) follows up on Moxey and Sanford (1993) and Sanford et al. (1994) and ascribes an even bigger role to prior expectations. Even though *few* and *many* are lexically different in terms of whether they express negation and are thus downward monotone, this is not the reason why *few* is complement set licensing and *many* is not. The claim is that complement set licensing is not a general lexical feature of negative quantifiers but triggered by unfulfilled expectations. Moxey (2006) proposes that negative quantity words like *few* "indicate an amount while at the

same time denying that this amount is as large as a supposed amount" (Moxey, 2006, 424). The difference between what was expected and what is fact is called the *shortfall*. Complement set focus occurs when there is a shortfall between the cardinality that was expected and the actual cardinality. "Focus on the shortfall leads to compset reference, and in fact this defines the complement set" (Moxey, 2006, 424). In a continuation task similar to the one presented in (96), participants' expectations were manipulated by describing that a character expected that a property holds for *none* or *all* objects or people in the context. The actual amount was described by several quantity expressions, for example *few* and *a few*. Then it was analyzed what the sentence continuation made reference to.

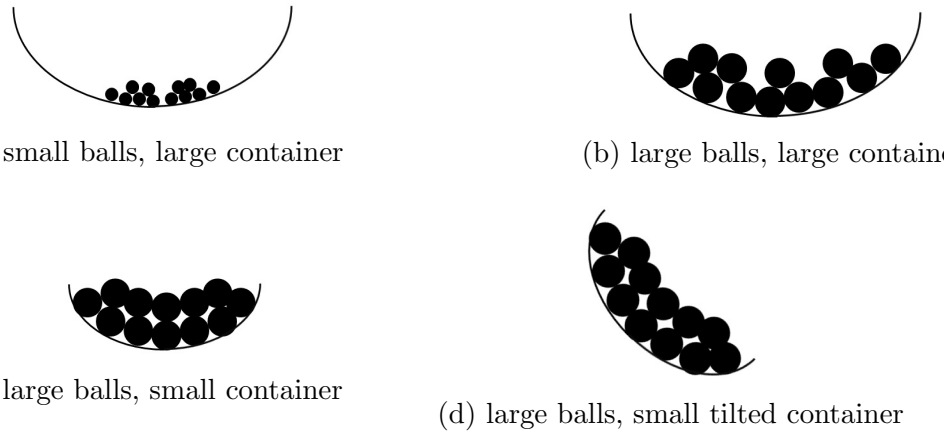
(97) Jill expected [none | all] of the glasses to be washed.
[Few | a few] of them were clean.
They...

The experiment confirms the hypothesis that complement set reference occurs when there is a shortfall. *Few* does not always refer to the complement set, only when the expressed cardinality is lower than expected. Moreover, positive quantifiers like *a few* can refer to the complement set when it is salient enough in the context.

We interpret these experiments to show that complement set focus occurs when the described cardinality is lower than expected whereas reference set focus is facilitated when the described quantity is higher than expected. However, it is not easy to draw further conclusions about *few* or *many* and their interaction with prior expectations from these results. First of all, *many* was not tested experimentally. Next, the expectations triggered by contexts like (97) affect the referents of other quantity expressions like *less than three* just as much, even though these quantifiers are not attributed to express surprise readings. Furthermore, although *few* does not make reference to the complement set in every case, this result does not speak against the CFK semantics. We argue against concluding that reference to the reference set for *few* suggests that it might express "more than expected" or that it does not relate to expectations at all. In contrast, we rather think, that *few* in Moxey's (2006) test sentences sounds at least marked which triggered the reference set and that participants used repair strategy. An example sentence from Moxey's (2006) data set is given below:

(98) Mrs. Smith expected none of the children to finish the essay.
Few of them completed the work.
They...

We think that *a few* would be a much more natural choice than *few* in such a context and that participants might have accommodated the sentence by replacing *few* with a better alternative. This suspicion would have to be confirmed experimentally.



(a) small balls, large container

(b) large balls, large container

(c) large balls, small container

(d) large balls, small tilted container

Figure 3.1: Stimuli from the replication of Newstead and Coventry's (2000) Superbowl study

To sum up, while investigating interesting properties of *few* and *many*, Moxey and Sanford identify prior expectations as a crucial factor for the meaning and use of *few* and *many*. That expectations play a role in linguistic experiments also supports the main objective of this dissertation: that it is worthwhile to experimentally test a theory like the CFK semantics, which is based on elusive concepts like prior expectations. Recent experimental methodology and a computational model as tools to approach this challenging task are presented in Chapter 5. Moreover, Moxey and Sanford's experiments also show that expectations can be manipulated experimentally. This methodology will be employed in Section 5.1 to challenge the claims made by Newstead and Coventry (2000) about the influence of visual clues (see next section) and in Chapter 7 in order to test the context-dependence of proportional *few* and *many*.

3.3 Subtle Effects of Visual Presentation

A series of studies on the influence of contextual and visual factors on the acceptability of vague quantifiers was conducted by Newstead and Coventry (2000) and Coventry et al. (2005, 2010). They found that the size and number of the objects described, the size of their container, position of the container, grouping and spacing of the objects as well as the number and properties of distracting objects influence how context-dependent quantifiers are rated in a judgment tast. They suggest that not only the actual number of objects matters but also their expected frequency.

The goal of Newstead and Coventry (2000) is to investigate the role of the physical properties size and functionality in the interpretation of the quantifiers *a few, few, several, many* and *lots of*. *Few* and *a few* are labeled as low magnitude quantifiers, *many* and *lots of* as high magnitude quantifiers. Participants saw images

of balls of varying size and number in a bowl of varying size and position as in Figure 3.1. The participants were then presented with sentences of the form "There are [quantifier] balls in the bowl" and asked to rate on a 7-point scale whether the statement is a good description of the image. The rating scale ranged from 1 (totally inappropriate) to 7 (totally appropriate). Newstead and Coventry (2000) found that, as expected, the number of balls had a significant effect on the ratings but that also their size had an effect: "Identical numbers of balls were given different ratings depending on ball size" (Newstead and Coventry, 2000, 243). Furthermore, the container made a difference. Small balls appear relative large in a small container and consequently the low magnitude quantifiers *a few* and *few* were rated lower. The authors suggest that this is due to the relative size of the balls. "What matters is how much space they take up of the container in which they are held" (Newstead and Coventry, 2000, 255).

Interestingly, even the position of the bowl has an effect. When a bowl which is filled with a high number of large balls, so many that they reach above the bowl's edge, was tilted, high magnitude quantifiers are rated higher than when the bowl is in its normal position. This is the case despite the fact that the balls look as though they should be falling out of the tilted bowl, see Figure 5.1d. The authors ascribe the higher ratings in this case to the fact that such a tilted bowl must be a "Superbowl" because the balls do not fall out. The authors conclude that this list of context effects indicates that quantifiers "carry little specific meaning in themselves but instead derive their meaning from the context in which they occur" (Newstead and Coventry, 2000, 243).

What we take to be the most striking explanation of all of these effects but what the authors do not pursue any further is whether "functionality reduces to the same thing as expected frequency" (Newstead and Coventry, 2000, 256). They suggest, just as we will do in Section 5.1, that the size of the bowl, its position and the size of the balls raises expectations about how many balls the bowl can hold. This is also how the surprise-based semantics by Fernando and Kamp (1996) (see Section 2.3) would explain these effects, but Newstead and Coventry (2000) dismiss the idea because it is "not entirely clear which way these expectations would work. For example, does the fact that a bowl is overflowing but the balls are still intact lead to a higher or lower expected frequency" (Newstead and Coventry, 2000, 256).

In Section 5.1 we will explicitly address the relationship between expectations and subjects' ratings in an experimental setup as in Figure 3.1 and replicate Newstead and Coventry's (2000) experiment.

Other effects of visually presented material are investigated by Coventry et al. (2005, 2010). Visual stimuli containing varying numbers of striped and white fish are presented by Coventry et al. (2005). The fish are either grouped or mixed and

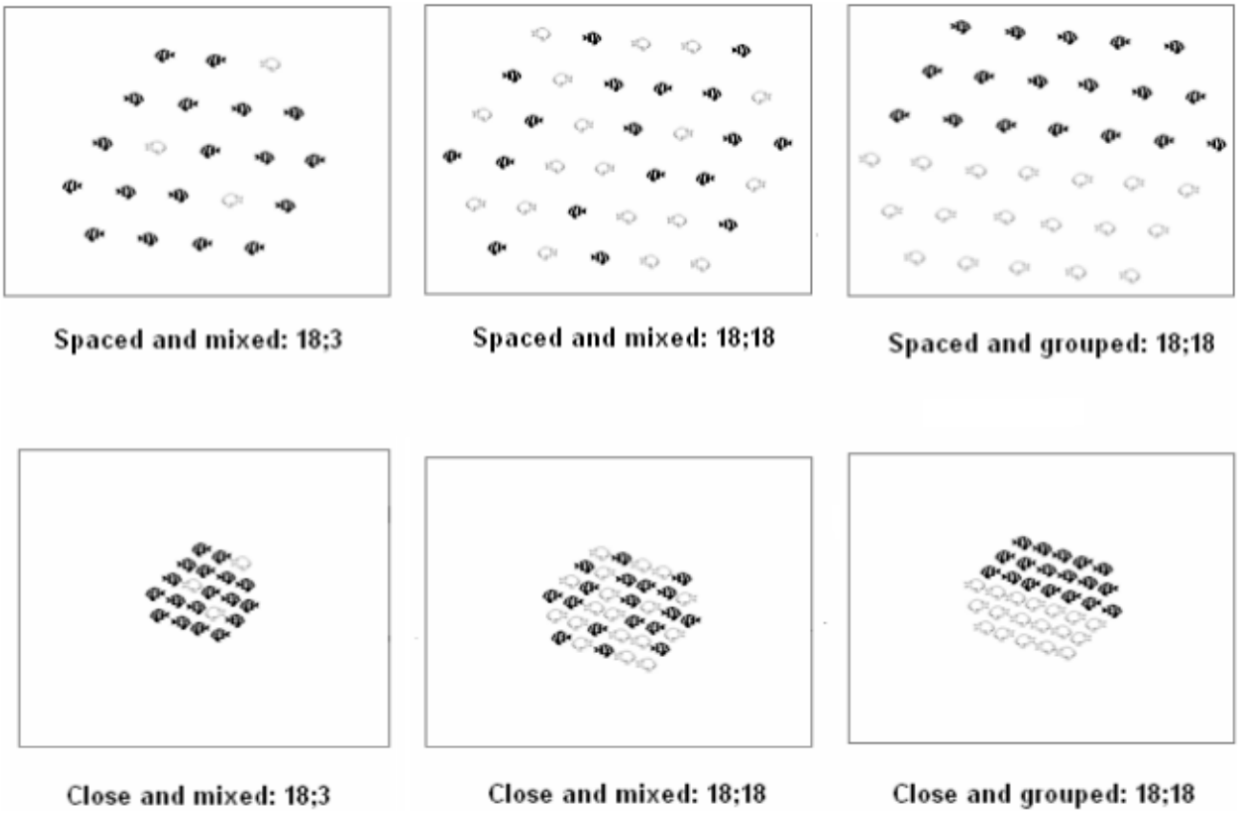


Figure 3.2: Example images from Coventry et al. (2005) with varying numbers of striped and white fish, varied spacing and grouping

the numbers of both striped and white fish differ. After having seen the image, participants are asked to rate sentences of the form

(99) There are [quantifier] striped/white fish.

with the quantifiers *a few, few, several, many* and *lots of* and a random choice of the color of the fish on a scale from 1 (totally inappropriate) to 7 (totally appropriate).

As expected, the number of fish on the display is a significant predictor of the appropriateness ratings. Additionally, the authors claim to have uncovered "three new effects on both quantifier rating and number judgements" (Coventry et al., 2005, 510). Both spacing and grouping of the objects in the scene affect quantifier ratings and number judgement, but "only when the number of focus objects rises above the subitizing region" (Coventry et al., 2005, 511). The subitizing region is the number of visually displayed objects that humans can immediately recognize without having to count them. Usually humans are able to subitize sets of the size of up to four (cf. Dehaene, 2011). In the mixed scenes (i.e. when striped and white fish are not grouped together), low magnitude quantifiers are rated higher and high magnitude quantifiers are rated lower. This suggests that participants estimated the reference set to be smaller when the grouping was mixed (see Figure 3.2), which was confirmed by the number estimation task. Also the factor spacing interacted

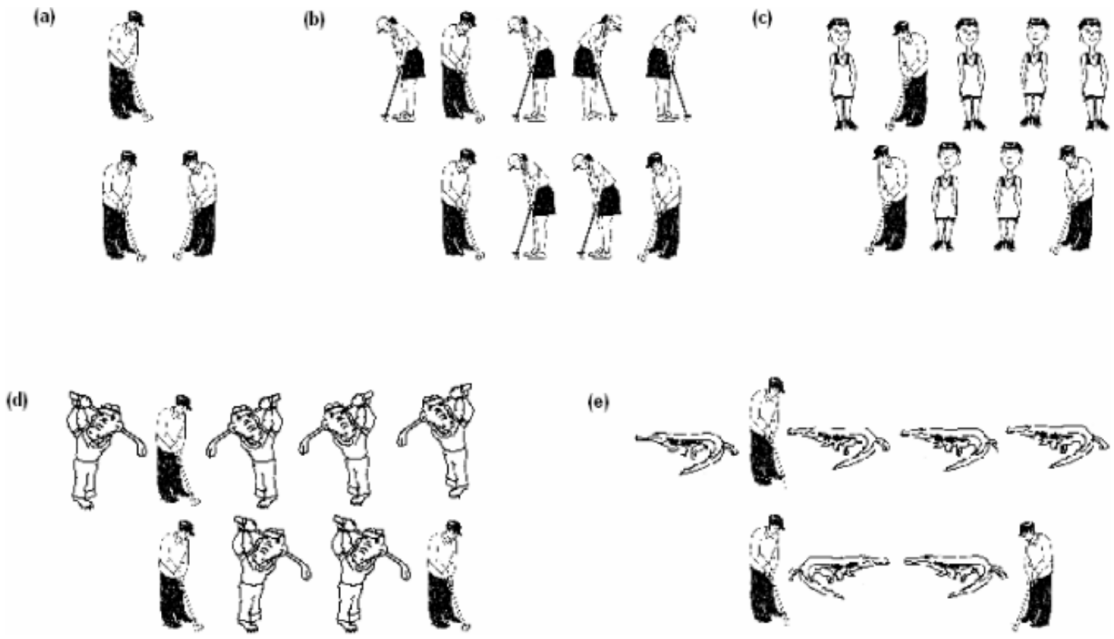


Figure 3.3: Example images from Coventry et al. (2010) with varying numbers of men and distractor objects. Distractors were manipulated in terms of form and function.

with the ratings of the quantifiers, but mainly for high magnitude quantifiers and at least nine objects in the display.

The authors interpret these results as a "correspondence between estimates of numbers of objects in a scene and context effects for quantifiers" (Coventry et al., 2005, 511). We are in line with the authors here that "knowing how many objects are in a set affects the likelihood that a certain number of objects from that set is present". What the authors do not discuss, however, is that the number of distractor objects (white fish if the quantified statement describes striped fish) is a main factor when the listener interprets the quantifiers proportionally. Even though the statements in (99) use the quantifiers cardinally, a proportional reading is plausible when the set size is explicitly fixed by the number of focused and distracting objects. Consequently, the effect of the total number of fish should be obvious.

Coventry et al. (2010) further investigate the influence of distractor objects in a visual display on the ratings of the quantifiers *a few, few, several, many* and *lots of*. In their experiment images of a varying number of men playing golf are presented. Coventry et al. (2010) manipulate whether the men are presented in isolation or whether a number of other objects is shown in the same display. The numbers of both focus and distractor objects are chosen from the set {3,6,9,12,15,18}. The distractor objects are of the same or a different species (women vs. crocodiles) and of the same or a different function (playing gold vs. not playing golf). Sample

stimuli are displayed in Figure (3.3). The images are described by sentences of the form

(100) There are [a few | few | several | many | lots of] men playing golf.

Participants were asked to rate how appropriate a description of each of the five sentences is (each with one of the five quantifiers) on a scale from 1 (totally inappropriate) to 7 (totally appropriate). The authors claim that "the number of objects in a scene impacts upon quantifiers judgments even when those objects are in a different category to the focus objects" (Coventry et al., 2010, 221). They link their findings to the mapping between approximate number system and language. However, the authors do not mention the possibility of a proportional interpretation of the quantifiers. We claim that presenting the objects as a group makes it natural to focus on their entirety. This, in turn, makes a proportional reading salient, although the quantifiers are presented in *there*-existentials. In this case, it is not surprising that the total number of objects in the display has an effect.

In their experiment, Coventry et al. found a main effect of quantifier and a significant interaction between quantifier and number, just as expected. Consistent with results in Coventry et al. (2005), the quantifier also interacted with the number of distractor objects. "Low magnitude quantifiers are rated as being more appropriate in the presence of larger numbers of other objects, and vice versa for high magnitude quantifiers" (Coventry et al., 2010, 231). Interaction with species and function did not reach significance.

Again, we argue against ascribing too much importance to visual cues since their influence can be subsumed under a smaller number of more general factors. Furthermore, we once more point out that Coventry et al.'s (2005) results indicate that the subjects interpreted the quantifiers proportionally. Even though the quantifiers are presented in *there*-existentials, the display makes the total amount of objects salient. This set size can be used as a standard of comparison for the cardinal reading but it certainly also facilitates a proportional interpretation. Speakers make use of this salient information and are likely to have interpreted the quantifiers proportionally. Coventry et al. (2010) once more do not draw this conclusion.

The possibility of a proportional reading brings up plans for follow-up studies. The test sentences could be phrased to make this reading salient by using a partitive construction.

(101) [A few | few | several | many | lots] of the men are playing golf.

Note that the quantifier *lots of* was used with this construction already in the original study, further strengthening our suspicions. When the factor number (the total number of objects as well as the number of distractor objects) influences the ratings

to the same extent, this would constitute evidence that also the *there*-existentials were interpreted proportionally. This result would be particularly interesting because it has been claimed that *there*-existentials force cardinal readings (Partee, 1989). Future studies on how the context can overcome this requirement and enable proportional reading anyway are therefore recommended. Apart from visual cues, we suggest that the focus structure of the sentence can help facilitate the shift in readings. Focusing a constituent that points out a difference between the described objects (men vs. crocodiles, playing golf vs. not playing) marks the comparison class and relevant alternatives. How the focus structure affects comparison classes was demonstrated in the last chapter when presenting two analyses by Romero (2015, 2017).

Chapter 4

Computational Modeling and Bayesian Inference

In Section 2.3 the surprise-based CFK semantics was introduced. The CFK semantics suggests that the lexical meaning of *few* and *many* comprises a function which takes as input prior expectations of the context and cuts off their cumulative density mass at a fixed percentage θ_{few} or θ_{many}. We have already pointed out the appeal of this account: it makes concrete predictions of how to derive the threshold values $x_{\max}$ and $x_{\min}$ in context and further proposes in which form contextual information is integrated.

Nevertheless, it poses empirical problems. How should such a fixed threshold theory and the predictions it makes be tested and verified or falsified, given that *few* and *many* are inherently vague and context-dependent? A first complication is that the theory is based on measures of surprise and prior expectations of contexts which are not fully specified by the sentence. Consequently, there is uncertainty about the exact comparison class and, depending on the speaker's knowledge, also uncertainty about its statistical properties. Prior expectations constitute *subjective beliefs* and are not frequency distributions of objective facts. For this reason, we opt to measure them experimentally and do not use objective statistics. Further detail of the experimental procedure are given in Sections 5.4, 7.2 and 7.3.

Another issue that complicates testing the CFK semantics is how to investigate the context-independent threshold parameters θ_{few} and θ_{many}. θ_{few} and θ_{many} are parameters which operate on representations of subjective prior expectations and thus determine threshold values $x_{\max}$ and $x_{\min}$ on a cardinality scale. It is exceedingly hard if not impossible to estimate the values of θ_{few} and θ_{many} based on solitary introspection. And even if we did, how should these estimates be verified in turn? Their values cannot be directly observed or measured in an experiment. We suggest to treat θ_{few} and θ_{many} as latent parameters in a computational model instead. Drawing conclusions from empirical data about values of latent variables

in a computational model is relatively straightforward for probabilistic models in concert with a Bayesian analysis. This is the path we trod here, as well. The basics of Bayesian inference and computational modeling will be explained in this chapter, mainly based on Kruschke (2011, 2014) and Lee and Wagenmakers (2013).

4.1 Terminology and Methods

Before we explain how the CFK semantics will be turned into a probabilistic model of language use which predicts the production and interpretation of *few* and *many* in Section 5.3, we take a step back and start by introducing relevant terminology and concepts with simpler examples. In general, we are interested in empirical data and the underlying processes which created and influenced it. Consider the following example by Lee and Wagenmakers (2013): Assume that a student, Anna, is sitting a test and has to answer 10 questions of equal difficulty. We want to estimate Anna's ability, which we define as the rate θ with which she answers questions correctly. θ is not directly observable. We can only observe Anna's score on the test (Lee and Wagenmakers, 2013, 3). To establish a relationship between observation (the score on the test) and its cause (the ability θ) we spell out a model. The term "model" is used here as a mathematical description, typically involving probabilities. The model formalizes the assumptions about exactly how the score on the test relates to the unobservable ability. For example, we can assume that we are prepared equally well for each question because all of them are about the same topic. $P(\mathrm{D} \mid \theta)$ is thus the probability that the observed data could be generated by the model with parameter values θ.

So far we have not thought much about θ itself. Even though we do not know Anna's ability, we have some beliefs about it. First of all, θ can range from 0 to 1. We could believe that Anna is very smart and assign higher values of θ a high probability. If we believe that she does not know much about the test's topic, we would find lower values more credible. But we could also not know anything about Anna's familiarity with the topic or about the difficulty level of the questions. In this case "a reasonable 'prior distribution', denoted by $P(\theta)$, is one that assigns equal probability to every value of θ between 0 and 1" (Lee and Wagenmakers, 2013, 3). This uniform distribution which represents $P(\theta)$ is shown by the dotted horizontal line in Figure 4.1.

Once the likelihood of the observed data given the model and parameter values is specified as well as prior expectations of the parameter values, we can return to our real objective: learn about which values of θ are credible given our observations. After observing Anna's score on the test we have modified beliefs. These so-called *posterior* beliefs are computed after taking into account a particular set of observa-

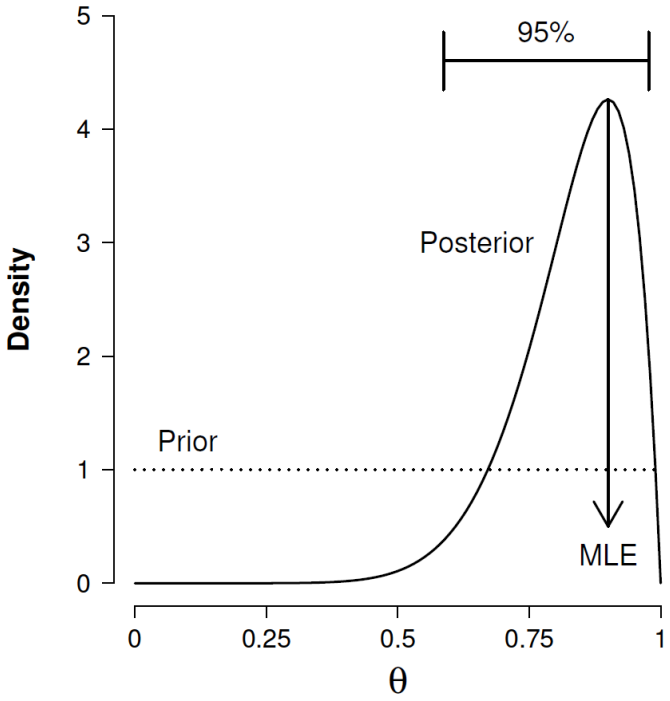


Figure 4.1: Example of a Bayesian parameter estimation from Lee and Wagenmakers (2013): The curve shows the posterior belief in Anna's ability θ, after observing 9 out of 10 correct responses. "The mode of the posterior distribution for θ is 0.9, equal to the maximum likelihood estimate (MLE), and the 95% credible interval extends from 0.59 to 0.98" (Lee and Wagenmakers, 2013, 3).

tions. The prior is simply the belief we hold by *excluding* a particular set of data, whereas the posterior is the belief we hold by *including* the dataset (Kruschke, 2011, 13).

Bayesian inference is what gets us from prior to posterior beliefs. A mathematical law called Bayes' rule specifies how to combine the information from the data - the likelihood $P(\mathrm{D} \mid \theta)$ - with the information from the prior distribution $P(\theta)$ to arrive at the updated, posterior distribution $P(\theta \mid \mathrm{D})$ (Lee and Wagenmakers, 2013, 3).

$$P(\theta \mid \mathrm{D}) = \frac{P(\mathrm{D} \mid \theta) \cdot P(\theta)}{P(\mathrm{D})} \tag{4.1}$$

The equation is often verbalized as

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}} \tag{4.2}$$

Bayes' rule is a quite simple mathematical law that helps us to 'reason backwards'. Since we cannot measure the values of latent parameters, we infer them from the data they trigger. Via Bayesian inference, three goals can be obtained:

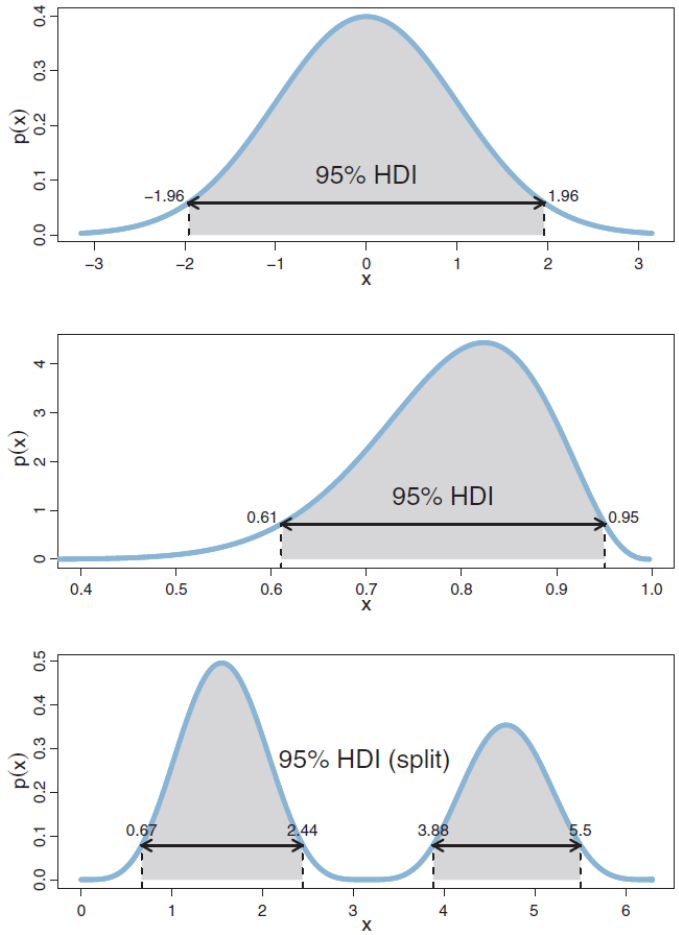


Figure 4.2: Examples of 95% highest density intervals (HDI) from Kruschke (2014): For each example, all the x values inside the interval have higher density than any x value outside the interval, and the total mass of the points inside the interval is 95%. The 95% area includes the zone below the horizontal arrow and is shaded in grey. The horizontal arrow indicates the width of the 95% HDI. The horizontal arrow's height marks the minimal density exceeded by all x values inside HDI (Kruschke, 2014, 88).

first, we can estimate parameter values to learn about Anna's ability or threshold values in the lexical semantics of *few* and *many*. Second, we may want to predict the probability of future data values, or, third, compare models which make distinct predictions about the same data generating process. Just as for parameter estimation, we will employ Bayesian inference to compare several model variants in Chapters 5 and 7.

Crucially, the posterior inferred via Bayes' rule does not deliver one 'true' value but another probability distribution over the parameter space. Since the prior distribution has been informed by the data, it is more peaked over the interval of parameter values which increase the data's likelihood. This is exemplified in Figure 4.1. To see more clearly which parameter values are most likely to have created the data, the posterior distributions' highest density interval (HDI) is easily calcu-

lated. The HDI is a way of summarizing a belief distribution. It indicates which points of a distribution we believe in most strongly. The HDI specifies an interval which covers most of the distribution's probability mass, for example 95% of it, such that every point inside has a higher probability than any point outside the interval (Kruschke, 2011). For an illustration look at Figure 4.2. Note that the HDI does not necessarily have to be one connected interval, it can also consist of several disjunct intervals, as can be seen in the bottom row of Figure 4.2. As a last remark, uncertainty of beliefs can be measured by the width of the HDI. "If the HDI is wide, then beliefs are uncertain. If the HDI is narrow, then beliefs are fairly certain" (Kruschke, 2011). The HDI's width and, correspondingly, the certainty of our posterior beliefs $P(\theta \mid \mathrm{D})$ in a parameter value θ can also be dependent on the size of the data set D with which we update the prior distribution $P(\theta)$. This point will be taken up in another example below. In the present example, a 95% credible interval of $[0.75; 1.0]$ for the mean rating of Anna's ability given that she answered 9 questions correctly would tell us that the model with its most likely parameters predicts Anna's test score to fall into this interval. The posterior's 95% credible interval is also marked in Figure 4.1.

Even though Bayes' rule is quite simple and easy to prove, it poses practical challenges. First of all, we need to specify our prior beliefs of the parameter values, a reasonable prior probability distribution $P(\theta)$. "Bayesian analysis tells us how much we should change our beliefs relative to our prior beliefs. Bayesian analysis does not tell us what our prior beliefs should be" (Kruschke, 2011, 224). Sceptics might argue that subjective beliefs manipulate the outcome of the data analysis to a too large extent. But we do not see this as a problem if the prior beliefs are made overt, are explicitly debated and consensual. The analysis will only convince its audience, if it uses priors that the audience finds palatable (cf. Kruschke, 2011). Often uniform prior distributions as in the exam example are employed.

Another practical problem is the exact mathematical calculation of the posterior. This does not only involve spelling out a model which predicts the data's likelihood and specifying prior beliefs in the parameter values, but also determining the denominator of Bayes' formula. The evidence, $P(\mathrm{D})$ is the "probability of the data according to the model, determined by summing across all possible parameter values weighted by the strength of the belief in those parameters" (Kruschke, 2011, 48).

$$P(\mathrm{D}) = \int P(\mathrm{D} \mid \theta) \cdot P(\theta)\, d\theta \tag{4.3}$$

The evidence $P(\mathrm{D})$ is independent of θ and is a single number which normalizes the posterior distribution to ensure that the area under its curve equals 1. Calculating the value of the denominator in Equation 4.1 usually means computing a difficult integral. This undertaking requires pure, analytical mathematics and can be difficult to achieve even with the help of modern computers and algorithms which numerically approximate the integral.

The analytical intractabilities have limited the scope of Bayesian parameter observation considerably, but they have now been overcome. The practicability of Bayesian statistics has changed dramatically with the evolution and refinement of computer-driven sampling methodology generally known as Markov Chain Monte Carlo (MCMC). MCMC techniques estimate the posterior distribution by randomly generating a high number of values from it. The approach is called a Monte Carlo method by analogy to the random events occurring when gambling in a casino. All that is required by this method is that for a specific value for a parameter θ, the value of $P(\theta)$ is easily calculated, especially by a computer. The same must hold for the likelihood $P(\mathrm{D} \mid \theta)$ for any value of D and θ and the product of prior and likelihood. "What the method produces for us is an approximation of the posterior distribution $P(\theta \mid \mathrm{D})$, in the form of a large number of θ values sampled from that distribution" (Kruschke, 2011, 98). This means that the posterior distribution can be approximated without having to calculate the difficult integral which constitutes the evidence $P(\mathrm{D})$. Based on this large sample of parameter values, useful characteristics of the posterior distribution, like its mean or credible region, can be estimated. Samples of the posterior distribution are generated by taking a random walk through the parameter space. A proposal distribution suggests a value for the next sample of θ, which is accepted with a probability α which is in turn dependent of whether the product of prior and likelihood is higher than for the previous sample (Kruschke, 2011; Lee and Wagenmakers, 2013).

"Each individual sample depends only on the one that immediately preceded it, and this is why the entire sequence of samples is called a *chain*. In more complex models, it may take some time before a chain converges from its starting value to what is called its stationary distribution" (Lee and Wagenmakers, 2013, 8f.). To speed up the process and to not have a too fuzzy beginning of the chain, it is common practice to run multiple chains, to discard the first samples (the *burn-in* samples) from each chain and to not record every sample, but only every second or third, for example. This is known as *thinning* (Lee and Wagenmakers, 2013).

With the processing power of computers constantly increasing, the "current adage is that *Bayesian models are limited only by the user's imaginations*" (Lee and Wagenmakers, 2013, 7). This quote shows the community's enthusiasm for MCMC sampling, which is implemented user-friendly in JAGS (Plummer, 2003),

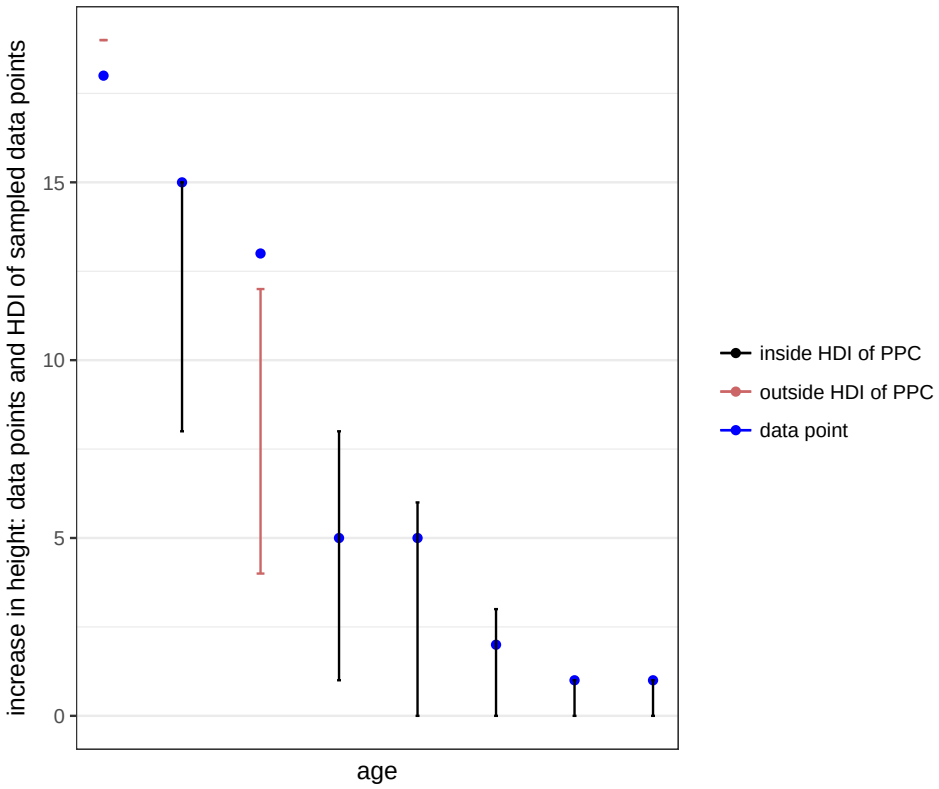


Figure 4.3: Comparing the HDIs of PPC samples with actual data

for example, allowing Bayesian statistics to gain practical use. JAGS is a general and versatile programing language for the specification of probabilistic models. It delivers a sequence of MCMC samples from the posterior $P(\theta \mid \mathrm{D})$.

After having approximated a posterior distribution, it is wise to not simply accept the model's predictions, but to check that the model, with its most credible parameter values, actually mimics the data reasonably well. This sanity check is called a "posterior predictive check" (PPC) (Kruschke, 2014). The idea of this test is to compare data predicted by the model with the actual data. Credible values of the parameters θ are plugged into the model to randomly generate sample data D^{sample}. This can be done particularly easy when using MCMC methods because in every link in the chain, the just sampled value for θ can be used to predict what data would look like according to the model. The results of this chain of simulations D^{sample} can then be summarized by once more calculating the 95% most credible predicted data values. Next, the predicted intervals are compared against the actual data.

As a simple demonstration of how to evaluate a PPC, imagine that we want to predict from a person's age her proportional increase in body height per year. After having specified a model (leaving aside the details for now) and having performed a PPC, we compare the HDIs of the PPC samples D^{sample} with the actual data (pair of age and proportional increase in body height). The hypothesized results of the

simulations are presented in Figure 4.3. The dots in the figure are the (hypothesized) actual data points. The vertical bars show the HDIs of the sampled data. Note that even within the same data set, the length of these intervals can vary quite a lot (for example, first vs. second vertical bar). By visual inspection of the graph, we can see that the decreasing tendency in the actual data seems to be well described by the predicted data and consequently by the model. However, two data points lie outside the PPC's HDIs, indicated by the red color of the vertical bars. Overall, the model manages to predict 75% of the data correctly. This suggests that it would be wise to contemplate alternative descriptive models. For example, the actual data might have a nonlinear trend or be better predicted by a different family of distributions (Kruschke, 2014).

This point leads us to the concept of Bayesian model comparison. It is often the case that several competing theories make predictions about how a certain data set was generated. And as we have seen in Figure 4.3, not every theory fits the data well. To discriminate between two or more theories, we can turn them into probabilistic models which predict the data generation process. By systematically comparing how well each model fits the data set at hand, we can draw conclusions about the validity of the theories. As a first measure we introduced the posterior predictive check. Another measure of the model's fit to the data which is easy to compute based on the output of our MCMC sampling results is the so-called *deviance information criterion* (DIC) (Spiegelhalter et al., 2002; Plummer, 2008). The DIC may be conceived of as a Bayesian cousin of classical model-choice criteria, in particular Akaike's information criterion (AIC). Like the AIC, the DIC weighs goodness of fit against the model's complexity. Where the AIC looks at a maximum likelihood fit for the model's free parameters, the DIC considers the full posterior distribution over these, given the data. A high value of the DIC indicates a lot of deviance of the model's predictions from the data it is applied to. This is undesirable, of course. At the same time, the model should stay as concise as possible and not include unnecessary parameters. This is measured by the pD, the number of effective free parameters, a measure of model complexity. Higher values of pD suggest higher model complexity.

4.2 Example: Estimating the Bias in a Coin

In the following, we present a simple example of Bayesian inference by Kruschke (2014, 108ff.) to see the just introduced concepts and methods at work. The example demonstrates how to infer latent parameter values from a sample of observed data. From the number of heads we observe when flipping a coin several times, we set out to estimate its underlying bias θ, i.e. the underlying probability of the coin coming

up heads. Note that we know that θ's value is between 0 and 1, but all we can observe is the proportion of heads coming up in a sequence of coin flips, not the bias θ itself.

As a first step in Bayesian data analysis, we identify the data at hand. In the present example, the data consist of heads and tails. The number of heads coming up will be referred to as z, the total number of coin flips as N, and, consequently, the number of tails will be N - z in a dataset called D. In a next step, Kruschke (2014) describes the likelihood of observing z heads in N coin flips with a simple, descriptive model containing a bias parameter θ. When the outcome of the ith flip of a coin with bias θ is denoted as y_i and the set of outcomes as $\{y_i\}$, the probability of observing a set of outcomes given bias θ is the multiplicative production of the probabilities of the individual outcomes:

$$p(\{y_i\} \mid \theta) = \theta^z(1-\theta)^{N-z} \tag{4.4}$$

After having defined a likelihood function in Equation 4.4, we specify a prior distribution over the values of the parameter θ. The prior formalizes what we believe about the factory's production of coins. Kruschke (2014) decides to use an unrealistic but illustrative prior distribution, and assume that "there are only a few discrete values of the parameter θ", namely the values $\theta = 0.0$, $\theta = 0.1$, $\theta = 0.2$ and so forth up to $\theta = 1.0$. "You can think of this as meaning that there is a factory that manufactures coins, and the factory generates coins of only those 11 types" (Kruschke, 2014, 110). Furthermore, Kruschke (2014) supposes that the factory tends to produce fair coins, with θ near 0.5, and assigns lower prior credibility for biases far above or below $\theta = 0.5$. This prior distribution is shown in the top panel of Figure 4.4.

Collecting data and applying Bayes' rule to update our beliefs in the possible parameter values is the next step. For a simple example, Kruschke (2014) assumes that we flip the coin only once and observe heads (i.e. $z = 1, N = 1$). For these data, the likelihood function becomes $p(D \mid \theta) = \theta$, as illustrated by the linear function in the middle panel of Figure 4.4. The lower panel shows the posterior distribution, which is computed by multiplying prior and likelihood for each possible value of θ, divided by $P(D)$. We can observe that the posterior distribution is different from the prior distribution. Because the coin showed a head, our belief in higher values of θ increases. However, the prior's effect shows because even though we observed 100% heads, the posterior probability of high θ values is still low. "This illustrates a general phenomenon of Bayesian inference: The posterior is a compromise between

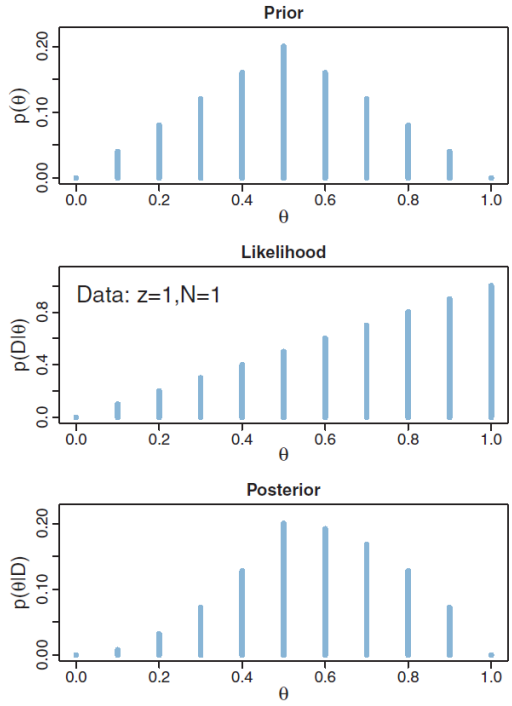


Figure 4.4: Bayes rule is applied to estimating the bias θ of a coin when flipping the coin only once and observing one head. There are discrete candidate values of θ and the posterior is computed by multiplying prior and likelihood for each θ, normalized (Kruschke, 2014, 111).

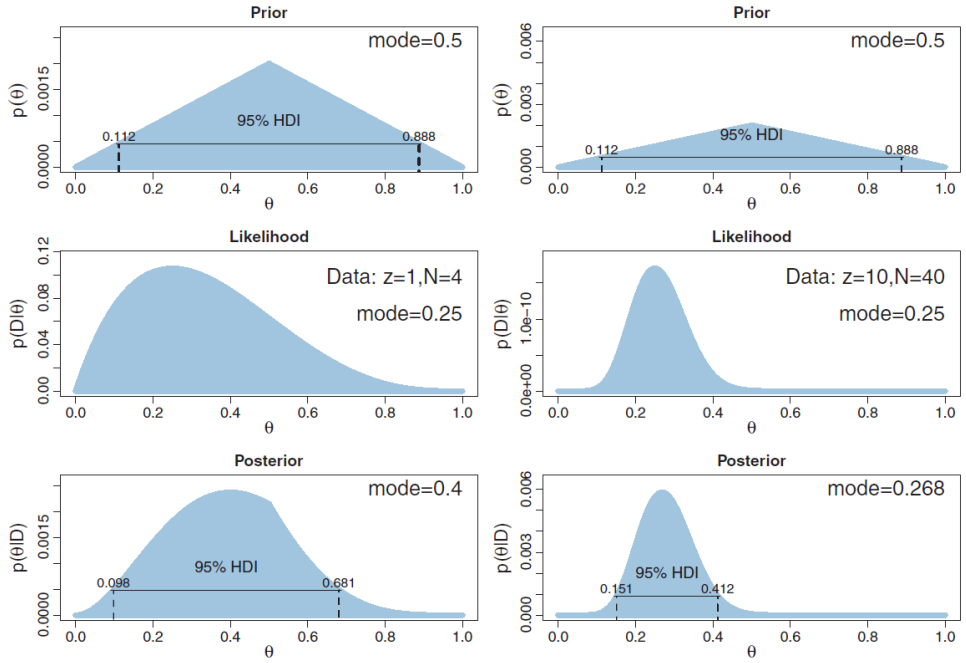


Figure 4.5: The two columns show the influence of different sample sizes, while keeping the proportion of heads constant. The prior is the same in both columns but plotted on a different vertical scale. The prior's influence is overwhelmed by larger samples (right column), resulting in the posterior's peak being closer to the peak of the likelihood function. Moreover, the posterior HDI is narrower for the larger sample (Kruschke, 2014, 113).

the prior distribution and the likelihood function", or, in other words, between the prior and the data (Kruschke, 2014, 112).

The previous example has shown that the influence of the prior can be large especially when the sample size is small. In the following we want to expand on the influence of the sample size on the posterior. To do so, Kruschke (2014) fills in candidate values of θ with 1,001 options, from 0.000, 0.001, 0.002, up to 1.000, resulting in a still discrete but more dense distribution in the top row of Figure 4.5. In the left column of the figure, the data is a small sample of $N = 4$ coin flips with 25% heads. In the right column, the proportion of heads is still 25%, but for a larger sample of $N = 40$ coin flips. Whereas both likelihood functions (as given in the middle row of Figure 4.5) have the same mode (peak), the posterior distributions differ in this respect, among other things. For the small sample size, the posterior's mode is at $\theta = 0.40$, which is closer to the prior's mode than to the likelihood's mode. For the large sample size, we observe the opposite. With the mode of the posterior being at $\theta = 0.268$, it is closer to the mode of the likelihood. From this example, we can learn that the influence of the posterior rises with a larger sample size. This shows also in the highest density interval. Even though both samples contained the same proportion of heads, the HDI is smaller for the larger sample size. In general, the more data is available, the more precise is the model's estimate of the parameters (Kruschke, 2014).

To finalize the Bayesian data analysis, we would next conduct a PPC to test whether the model manages to predict the data well enough. If there was another model available which explained the data generating process differently, we would compute each model's *DIC* and *pD* to compare their fit to the data relative to the model's complexity.

With this introduction into Bayesian inference and computational modeling in mind, we can now move on to our ultimate goal, testing the CFK semantics. In the next chapter, the predictions of the CFK semantics are transformed into a computational model of production and interpretation behavior. We will use experimental data and MCMC sampling to infer a posterior distribution about θ_{few} and θ_{many} and to compare several models which make predictions about the same set of data.

Chapter 5

Cardinal *few* and *many*

In this chapter we explore the cardinal reading of *few* and *many*. The general properties of this reading have been introduced in Section 2.1. This chapter's main goal is to test the surprise-based semantics by Fernando and Kamp (1996). They stipulate that the lexical meaning of cardinal *few* and *many* comprises fixed threshold values θ_{few} and θ_{many} which operate on prior expectations of the context.

Section 5.1 presents a pre-study which tests whether prior expectations really comprise sufficient contextual information to predict the acceptability of quantifiers across contexts. We do so by replicating the "Superbowl" study by Newstead and Coventry (2000) and investigate whether the factor *prior expectations* has as much explanatory power as the various visual cues they propose. Section 5.2 once more presents Fernando and Kamp's (1996) fixed threshold semantics which is translated into testable predictions and embedded in a computational model in Section 5.3. Section 5.4 introduces the behavioral experiments to elicit representations of a priori expectations, as well as production and comprehension behavior of cardinal *few* and *many*. Section 5.5 describes how we employ Bayesian inference to learn about latent parameters and the use of Bayesian model comparison to assess the plausibility of the hypothesis that a context-independent threshold parameter governs the production and comprehension of *few* and *many*. The results of the model evaluation and their implications are discussed in Section 5.6.

5.1 Pre-study: The Superbowl

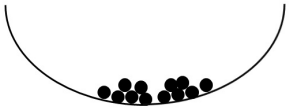
As a starting point into our investigations of the cardinal reading of *few* and *many*, this section presents a replication of the study conducted by Newstead and Coventry (2000). We investigate the influence of the context, in this case of visual cues, on the use of the vague and context-dependent quantifiers *a few, few, several, many* and *lots of*. Newstead and Coventry (2000) present participants with images of bowls of varying size and position which contain balls of varying size and number. Depend-

ing on the number of balls in the bowl, the bowl turned into what Newstead and Coventry (2000) labeled a "Superbowl" because even if the balls reached over the edge of the bowl, they did not fall out. A sample of the images we used in our replication is provided in Figure 5.1. The number of balls is described by sentences which include one of the quantifiers and participants are asked to rate the acceptability of the statements. Our goal is to show that the variance in acceptability ratings which Newstead and Coventry (2000) explain by the factors BALL SIZE, BOWL SIZE and BOWL POSITION can be summarized by one other factor: the number of balls that are expected to additionally fit into the bowl, on top of the ones which are already in the bowl (CAPACITY). This would constitute first experimental evidence supporting Fernando and Kamp's (1996) semantics which assumes that the truth conditions of cardinal *few* and *many* are dependent on expected cardinalities in the respective context.

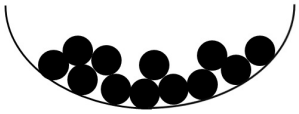
Newstead and Coventry (2000) themselves point out that expected frequency is probably a key factor, but they do not follow up on their suspicion. "It is possible that all these effects reduce to expected frequency but we believe the picture is more complicated than this. The concept of expected frequency is currently too vague to make specific predictions, and hence more detailed studies of the factors involved are necessary before any overarching theories can be adopted." (Newstead and Coventry, 2000, 258). We agree that expectations come with a certain vagueness but we do not necessarily share the worry that expectations can only be elicited in elaborate studies. For this reason, we additionally asked participants for their best guess of the number of balls which still fit into the bowl. If this measure turned out to not produce informative data, it is still possible to turn to more elaborate elicitation methods. Recently, progress has been made in the methodology of measuring expectations and their validation (see Kao et al. (2014) and Franke et al. (2016) and the experiments presented in this and the following chapters). In the following, we spell out a number of hypotheses and present a replication of Newstead and Coventry's (2000) experiment to test them.

5.1.1 Hypotheses

Uncontroversially, we predict the ratings of the quantified statements to differ by QUANTIFIER and NUMBER of the balls in the bowl. *Few* and *a few* are expected to be rated higher for small numbers whereas *many* and *lots of* are rated higher for higher numbers of balls in the bowl. Furthermore, we predict that *few* and *a few* are a good description of scenarios in which a high number of balls is expected to additionally fit into the bowl, when the CAPACITY of the bowl is large. CAPACITY is defined as the difference between the maximal EXPECTED NUMBER of balls that can fit in the bowl



(a) small balls, large container



(b) large balls, large container

(c) large balls, small container

(d) large balls, small tilted container

Figure 5.1: Stimuli of the superbowl replication

	scenarios of which quantifiers are a good description	
QUANTIFIER	NUMBER	CAPACITY = EXPECTED NUMBER - NUMBER
few, a few	small	small
many, lots of	large	large

Table 5.1: Hypotheses of Superbowl replication

in total and the NUMBER of balls that is already in the bowl. For the sample of images in Figure 5.1, the bowl in Figure 5.1a would have the largest capacity, Figure 5.1b a medium capacity and Figure 5.1c and Figure 5.1d a small capacity. We take it that the empirically measured CAPACITY is an approximation of prior expectations, for example in the form of a flat distribution over the interval [0, EXPECTED NUMBER]. A higher expected number would imply a lower probability for each $i <$EXPECTED NUMBER. *Many* and *lots of*, on the other hand, are applicable in scenarios in which the CAPACITY is small and the bowl is nearly full, when the difference between the EXPECTED NUMBER and the actual NUMBER is small. We predict that it does not make a difference which visual cues are available to estimate the maximal number of balls the bowl can hold.The hypotheses tested in this experiment are summarized in Table 5.1.

Newstead and Coventry (2000) utter concerns about including expectations as a factor because of their inherent vagueness. Even though they suggest that "functionality reduces to the same thing as expected frequency", they dismiss the idea because it is "not entirely clear which way these expectations would work. For example, does the fact that a bowl is overflowing but the balls are still intact lead to a higher or lower expected frequency" (Newstead and Coventry, 2000, 256). We do not see this uncertainty as a problem for our predictions though. First, whether people are really uncertain about overflowing balls could be tested experimentally if necessary. Second, we do not see a problem in whether balls can reach above the

bowl's edge or not, since this is directly related to the way the capacity is determined. If a subject expects that the balls can flow out of the bowl, this results in a higher estimate of the balls the bowl can hold in total and consequently in a higher capacity than for the case in which the balls are expected to not be able to reach above the bowl's edge. In this case, the estimated maximum and also the bowl's capacity decrease accordingly.

5.1.2 Experiment

This judgment task is a replication of Newstead and Coventry's (2000) "Superbowl" experiment. It investigates the influence of visual cues and related prior expectations on the acceptability of context-dependent quantifiers.

Design. Participants saw one picture of a bowl varying in SIZE and POSITION filled with balls of varying SIZE and NUMBER. The balls were either small or large, just as the bowl was either small or large. The position of the bowl was either normal or tilted. The number of balls in the bowl was 6, 12, 18 or 24. At this point we diverge from Newstead and Coventry's (2000) design who included more number conditions. When the bowl was tilted and contained a large number of balls, the balls reached over the edge of the bowl. But instead of falling out of the bowl, the balls sticked somewhat unnaturally together and remained in the bowl, turning the bowl into a "Superbowl". A sample of the stimuli depicting 12 balls is presented in Figure 5.1. The picture participants saw was chosen randomly. Participants were asked to look at the picture and then read five sentences as in (102). The sentences each contained one of the quantifiers *a few, few, several, many* and *lots of* and were presented in a random order underneath each other. Participants were then asked to rate on a horizontal 7-point scale whether the sentence is a good description of the picture. The value 1 was labeled "very bad", the value 7 was labeled "very good". Each sentence contained a different quantifier and all quantified statements were presented at once. The *there*-existential construction made a cardinal reading of the sentences salient, as discussed in Section 2.1.3.

(102) There are [few | a few | several | many | lots of] balls in the bowl.

Additionally, participants estimated the maximal number of balls that fit in the bowl. Note that there is uncertainty of whether the balls should be able to fit into the bowl or whether they can reach above its edge, as was indicated in some comments. We intensionally left the decision to the participants since it is accounted for in the statistical analysis. A smaller maximal EXPECTED NUMBER (when the balls cannot reach above the bowl's edge) is directly coarrelated with a smaller CAPACITY.

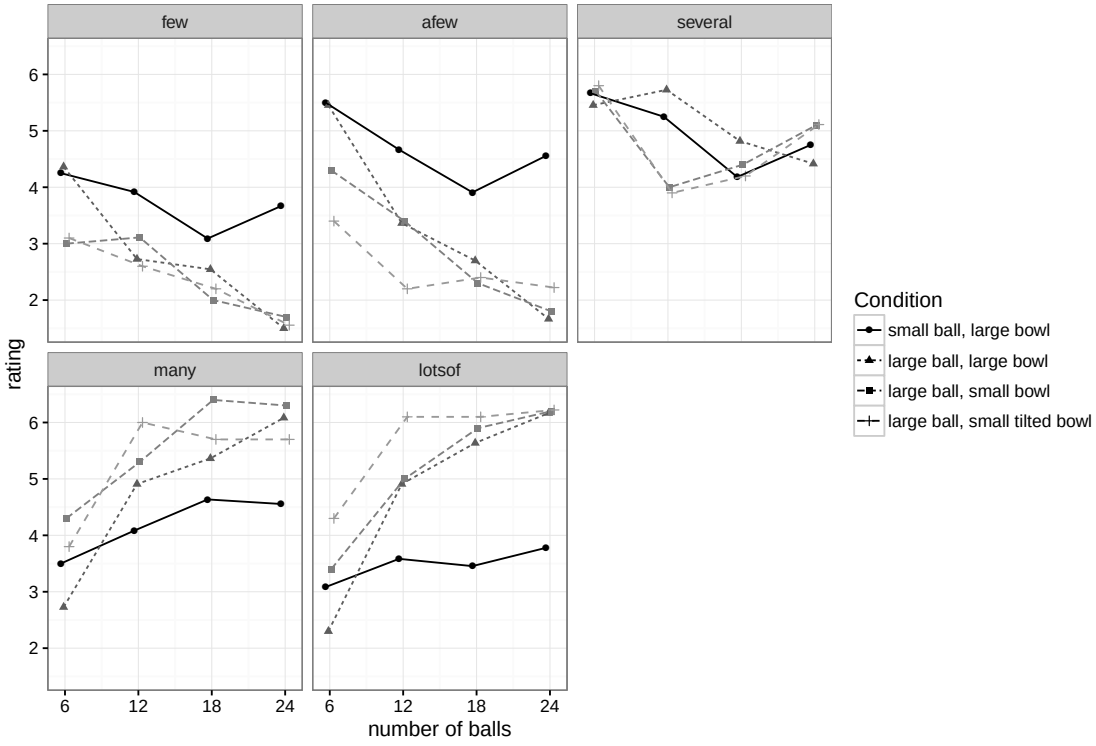


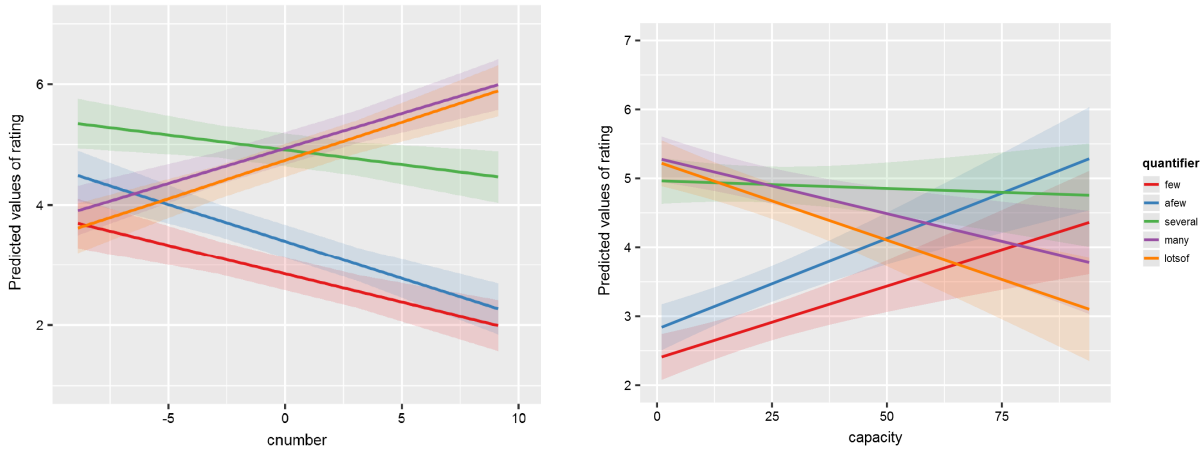
Figure 5.2: Mean ratings of the Superbowl replication

Participants. 126 subjects were recruited via Amazon's Mechanical Turk with US-IP addresses. Each subject participated only once.

Materials & Procedure. After initial instructions that explained the task, each subject saw one image and five quantified statements presented in a random order. In addition to the rating task and the estimation of the bowl's capacity, we asked participants for the color of the balls. Participants could choose between the answers green and black. This tested whether participants paid minimal attention to the task. The correct answer was black because only black balls were presented.

Results. Since every subject answered the question for the balls' color correctly, no data had to be excluded. The mean rating of each quantifier per condition is visualized in Figure 5.2. A high mean rating corresponds to a high acceptability of the sentence.

To analyze the data, we specified a linear mixed effects regression model predicting acceptability ratings. During a guided search through the model space, we started out with a model containing only the random effect PARTICIPANT and added fixed effects if this significantly increased the model's fit to the data (measured by AIC). The final model includes the fixed effect QUANTIFIER (levels *few, a few, several, many, lots of*) and its interaction with the NUMBER of balls presented and the expected CAPACITY of balls that the bowl can additionally hold. The CAPACITY



(a) Interaction between QUANTIFIER and NUMBER

(b) Interaction between QUANTIFIER and CAPACITY

Figure 5.3: Interactions with QUANTIFIER in the Superbowl replication

was calculated as the difference between the maximal EXPECTED NUMBER of balls the bowl can hold and the NUMBER of balls in the bowl.

Participants gave the lowest ratings for the QUANTIFIER *few* ($\beta = 2.39, SE = 1.59, p < 0.001$). The ratings for *a few* are not significantly different ($\beta = 0.42, SE = 0.281, p = 0.066$), but the quantifiers *several* ($\beta = 2.60, SE = 0.23, p < 0.001$), *many* ($\beta = 2.90, SE = 0.23, p < 0.001$) and *lots of* ($\beta = 2.90, SE = 0.23, p < 0.001$) were rated significantly higher than *few*. This effect is modulated by an interaction between QUANTIFIER and NUMBER of balls and QUANTIFIER and CAPACITY. The interaction of QUANTIFIER and NUMBER of balls reaches significance for the difference between *few* and the high magnitude quantifiers *many* ($\beta = 0.22, SE = 0.03, p < 0.001$) and *lots of* ($\beta = 0.26, SE = 0.03, p < 0.001$). For higher numbers of balls, *few* and *a few* were rated lower, whereas the ratings for *lots of* and *many* increased. The interaction is visualized in Figure 5.3a. In comparison to *few*, the quantifiers *several* ($\beta = -0.03, SE = 0.01, p < 0.001$), *many* ($\beta = -0.04, SE = 0.01, p < 0.001$) and *lots of* ($\beta = -0.05, SE = 0.01, p < 0.001$) were rated significantly lower for a larger CAPACITY of the bowl. For *a few* the difference is not significant, see Figure 5.3b. The fixed effects BALL SIZE, BOWL SIZE and BOWL POSITION do not significantly increase the model's fit to the data.

5.1.3 Discussion

The data support the hypotheses from Section 5.1.1, as summarized in Table 5.2. *Few* and *a few* are applicable to small cardinalities whereas *many* and *lots* of describe large cardinalities. *Several* seems to sojourn in the middle. This is confirmed by the interaction between QUANTIFIER and NUMBER. Further support is provided for the

	scenarios of which quantifiers are a good description			
QUANTIFIER	NUMBER	results	CAPACITY	results
few, a few	small	✓	small	✓
many, lots of	large	✓	large	✓

Table 5.2: Results of Superbowl replication

hypothesis that vague quantifiers express expectations towards a cardinality: low magnitude quantifiers can be interpreted to express that a number is lower than expected, whereas high magnitude quantifiers state that a number is higher than expected. This is predicted by the factor CAPACITY, the difference between the estimated maximum of balls the bowl can hold and the actual number of balls in the bowl. *Many* and *lots* were rated higher when the bowl was expected to not be able to hold more balls. *Few* and *a few*'s acceptability increased when the bowl's capacity was high.

These two factors, NUMBER and CAPACITY manage to account for the variance in the data. The predictors BALL SIZE, BOWL SIZE and BOWL POSITION did not reach significance when the factors NUMBER and CAPACITY were included in the model. This supports the hypothesis that these visual cues can be subsumed in the predictor expectations. The size and position of the presented objects naturally influence prior expectations, but it is not necessary to include each of them as a separate predictor or to ascribe the tilted bowl the magical properties of a "super-container" (Fernando and Kamp, 1996, 254). This role is better filled by the more general factor *prior expectations*.

Prior expectations are also a major factor in the surprise based semantics proposed by Clark (1991) and Fernando and Kamp (1996). In the following sections we investigate prior expectations in real-world contexts and learn how context-dependent expressions are assigned meaning based on them. The ultimate goal is to quantitatively predict the production and interpretation of *few* and *many* based on experimentally measured prior expectations of the respective context.

5.2 The CFK Semantics and How To Test It

In Chapter 2, we noted an omission in the semantic literature, which does not specify how the threshold values which determine the use of *few* and *many* are calculated. One theory that makes concrete predictions for sentences exhibiting a cardinal surprise reading, however, is the surprise-based semantics by Clark (1991) and Fernando and Kamp (1996). The CFK semantics was introduced in Section 2.3 and will be briefly summarized in the following. This theory stipulates that *few* and

many comprise fixed threshold values θ_{few} and θ_{many} operating on prior expectations of the context. According to this approach, a sentence of the form "Many As are B" is true if the actual cardinality $n = |\, A \cap B \,|$ exceeds a fixed threshold θ_{many} on a measure of surprise, which is derived from a *a priori* expectations P_E about likely values of n provided by the context. In simpler terms, "Many As are B" is true if the actual number of $n = |\, A \cap B \,|$ is surprisingly high, higher than $x_{\min}$. The truth conditions are repeated from above:

(77) **CFK Semantics**

a. $[\![\text{Few As are B}]\!] = 1$ iff $|\,\text{A} \cap \text{B}\,| \leq x_{\max}$

$$\text{where } x_{\max} = \max\{n \in \mathbb{N} \mid P_E(|\,\text{A} \cap \text{B}\,| \leq n) < \theta_{\text{few}}\}$$

b. $[\![\text{Many As are B}]\!] = 1$ iff $|\,\text{A} \cap \text{B}\,| \geq x_{\min}$

$$\text{where } x_{\min} = \min\{n \in \mathbb{N} \mid P_E(|\,\text{A} \cap \text{B}\,| \leq n) > \theta_{\text{many}}\}$$

Even with a fixed and contextually-stable threshold for what counts as sufficiently surprising, whether a certain n counts as surprisingly high can still vary dramatically with the context. For example, for numbers of children a family has and points scored in a basketball match, we may have dramatically different prior expectations P_E. For this reason context-dependence and vagueness can be possible despite a systematic, calculable and learnable stable core meaning.

To assess whether the CFK semantics in (77) is on the right track is a challenge to classical methods from theoretical linguistics because they require *intuitions* about truth, entailment and the like as input. This is because, in almost all cases, a precise conception of what we consider to be "likely" is hard to get hold of. Still, it could be the case that (77) captures speakers' non-introspective use of *many* and *few* well enough. What can we do? Certainly, we can probe intuitions (be it our own, or those of informants in a controlled experiment) about applicability and interpretation of relevant sentences in laboratory conditions that provide perfect or near-perfect information. This approach poses practical problems that may or may not be solvable by clever design.

But there is also an alternative that is worth exploring: data-oriented computational modeling, introduced in the previous chapter. Focusing for now on *few* and *many* and the CFK semantics for their cardinal surprise uses, our main goal here is to give one constructive example of how data-oriented computational modeling could be useful for formal semantic theory. For one, we demonstrate how recent experimental methodology (e.g., Kao et al., 2014; Franke et al., 2016) can help obtain approximate empirical measures of introspectively inaccessible "prior expectations." For another, we show how the core semantics in (77) can be turned into probabilistic

models of speaker production and listener interpretation behavior. Finally, feeding empirically measured prior expectations into production and interpretation models, we show that production and interpretation data from suitable experimental tasks can be used to infer plausible values of θ_{many} and θ_{few}.

We will propose a relatively simple computational model in the next section. For instance, we will not consider genuine pragmatic competition between alternative expressions. Other models are conceivable and may or may not give rise to similar conclusions about the tenability of a CFK semantics. We believe that this is normal: testing an abstract hypothesis (like the CFK semantics) alongside empirical data will require auxiliary assumptions about how the hypothesis relates to data observations (e.g., Quine, 1951). Yet, given data and a model about how latent variables generate possible observations, we can then draw inferences about the unobservable latent variables of interest.

In the following pages we want to test the CFK semantics by contrasting two competing hypotheses. Hypothesis 1 assumes one fixed, context-independent pair of threshold values θ_{few} and θ_{many}, which apply to probability distributions representing prior expectations about the respective context. This is what the CFK semantics predicts. Hypothesis 2 assumes that the thresholds $\theta_{\text{few,i}}$ and $\theta_{\text{many,i}}$ vary for each context i. There is no deeper theoretical motivation for this hypothesis except that it is the negation of the fixed-threshold hypothesis. Since thresholds on prior expectations cannot be directly observed, we use computational modeling to infer their most credible values. We will spell out one model for each hypothesis and compare their fit to an experimentally-gathered data set. But not only the fit to the data is a crucial factor for discriminating between two models, their complexity matters, too. A model with less free parameters (in our case less threshold values which need to be inferred) is less complex and therefore *ceteris paribus* preferable.

5.3 Computational Model: The Fixed Thresholds Model

Evaluating the CFK semantics in (77) is a challenge for standard methods from theoretical linguistics insofar as they rely on intuitions about truth, entailment and the like. This is because, in almost all real-world cases, a precise enough determination of prior expectations P_E seems to elude solitary introspection. To test a CFK semantics, we therefore turn to data-driven computational modeling.

This approach effectively considers the contextually stable thresholds θ_{many} and θ_{few} as *latent parameters*: their values cannot be directly observed but must instead be reconstructed from observable behavior. Bayesian inference is one way to do

so. Given values for latent parameters, a probabilistic model makes predictions about how likely certain observable choices in production and comprehension of relevant sentences are. In technical terms, the model specifies a likelihood function $P(\text{observation} \mid \theta_{\text{many}}, \theta_{\text{few}})$ mapping values of latent parameters onto a probability of seeing a particular choice in a suitable experiment. We will use data from a production and a comprehension task to infer, via Bayes rule, which values of the latent parameters are credible, given the likelihood function and some prior over latent parameters:[1]

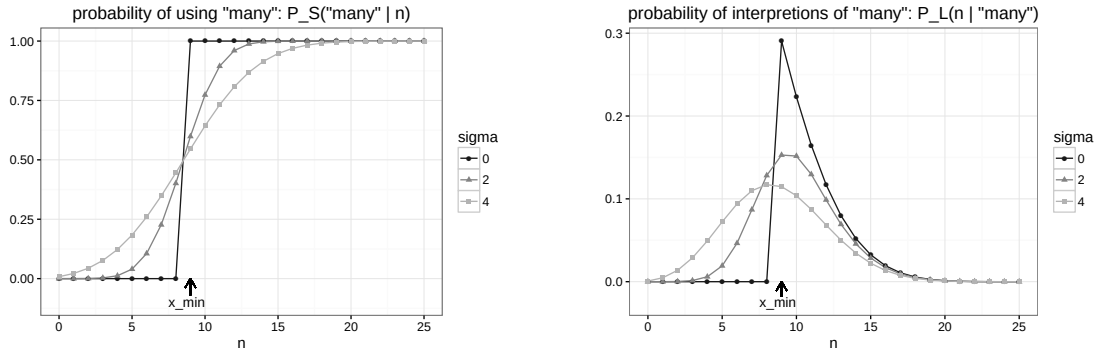
$$P(\theta_{\text{many}}, \theta_{\text{few}} \mid \text{observation}) \propto P(\theta_{\text{many}}, \theta_{\text{few}})\ P(\text{observation} \mid \theta_{\text{many}}, \theta_{\text{few}}) \tag{5.1}$$

Our goal, then, is to see whether a single pair of threshold values θ_{many} and θ_{few} explains our empirical data well enough. We focus on *many* in the exposition, but the case for *few* is parallel.

Our computational model consists of a production and a comprehension rule, both probabilistic. A probabilistic production rule is a function that assigns a probability distribution over expressions or utterances to any given meaning, while a probabilistic comprehension rule is the same in reverse, assigning a probability distribution over meanings or interpretations for each possible utterance that needs to be interpreted (e.g., Franke and Jäger, 2016; Goodman and Frank, 2016). Here, a production rule should give us the probability $P_S(\text{“many”} \mid n, P_E)$ with which a speaker, or speakers in general, would find the sentence “Many As are B” applicable to $n = |\, A \cap B \,|$ under prior expectation P_E. A comprehension rule should give us the probability $P_L(n \mid \text{“many”}, P_E)$ with which a listener, or listeners in general, would believe in interpretation n when they hear the relevant statement with *many* in a context where P_E captures the relevant statistical properties of the assumed comparison class.

A production rule that implements the CFK semantics in (77) is straightforward: $P_S(\text{“many”} \mid n, P_E\ ;\ \theta_{\text{many}}) = 1$ if $n \geq x_{\text{min}}$ and otherwise 0, where x_{min} is derived from P_E, as in (77), based on θ_{many}, which is a free parameter for this rule (indicated by writing it after a semicolon). This probabilistic production rule is only a degenerate probabilistic rule: it only assigns the extreme values 0 and 1; it does not allow for slack, mistakes or other trembles. As such, it would not apply well

[1] The notation “$\propto$” for “proportional to” says that the expression on the right must yet be normalized. So, $P(x) \propto f(x)$ for some function f is short for $P(x) = \frac{f(x)}{\sum_{x'} f(x')}$.



(a) Production rule examples (b) Comprehension rule examples

Figure 5.4: Illustration of production and comprehension rules for the example from Figure 2.2

to noisy empirical data. So, instead of a step-function we look at a parameterized, smoothed-out version.

$$P_S(\text{“many”} \mid n, P_E \ ; \ \theta_{\text{many}}, \sigma) = \sum_{k=0}^{n} \int_{k-0.5}^{k+0.5} \mathcal{N}(y; x_{\min}, \sigma) \mathrm{d}y \tag{5.2}$$

P_S: the probability of producing “many” to describe a cardinality n given $P_E, \theta_{\text{many}}, \sigma$ is a smoothed-out version of a step function with a step at $x_{\min}$

Here, σ is another free model parameter that regulates the steepness of the curve, and $\mathcal{N}(y; x_{\min}, \sigma)$ is the probability density of y under a normal distribution with mean $x_{\min}$ and standard deviation σ. Essentially, this gives us a noisy implementation of speaker behavior under a CFK semantics where the amount of noise is controlled by σ. Illustrations of this probabilistic production rule are shown in Figure 5.4a for the example started in Figure 2.2. The degenerate, non-noisy production rule is the case of $\sigma = 0$.

The idea behind Equation (5.2) is this. Assume that a hypothetically true value of θ_{many} exists. Then, given a prior expectation P_E over the contextually relevant domain, the CFK semantics in (77) gives a clear cutoff for the minimum number $x_{\min}$ of, say, cups of coffee that some particular Andy must minimally drink per week to license applicability of *many* in a sentence like (76). We should assume that speakers do not know for sure the actual $x_{\min}$ that is entailed by θ_{many} and P_E, most likely because they do not know P_E for certain, but that speakers nonetheless approximate it[2]. More concretely, we assume that when a speaker decides whether some n licenses *many*, she “samples”, so to speak, a noise-perturbed “subjective threshold” $x'_{\min}$

[2]For most contexts, speakers do not know the exact statistical properties, but nevertheless they are able to approximate them (Griffiths and Tenenbaum, 2006). How much knowledge we have and how certain we are of it is very domain-specific. Estimating can be facilitated by asking for measurements which are often stated explicitly, relevant in daily life and in restricted domains with little variation. For example, it is easier to estimate the length of a radio song than the size in acres

from a Gaussian distribution whose mean is $x_{\min}$ and whose standard deviation σ is a free model parameter that captures speaker uncertainty (about θ_{many}, P_E, and perhaps other things). If the sampled value is below n, the speaker finds *many* applicable to cardinality n; otherwise, she does not. This gives us a probabilistic prediction of how likely a speaker would, on occasion, find *many* applicable to n as a probabilistic function of θ_{many}, P_E and noise parameter σ.

A derivation of a reasonable probabilistic comprehension rule follows suit:

$$P_L(n \mid \text{“many”}, P_E \,;\, \theta_{\text{many}}, \sigma) \propto P_E(n) \cdot P_S(\text{“many”} \mid n, P_E \,;\, \theta_{\text{many}}, \sigma)\,. \qquad (5.3)$$

P_L: the probability of choosing a cardinality n as the interpretation of “many” given $P_E, \theta_{\text{many}}$ and σ is the prior probability $P_E(n)$ of n weighted by P_S (probability of producing "many" to describe n)

This rule, which is illustrated in Figure 5.4b, can be motivated in two conceptually distinct ways that yield the same mathematical result. For one, we can think of Equation (5.3) as an application of Bayes' rule. Under this interpretation, the listener tries to infer likely world states based on a model of reverse production by taking into account how likely each world state is and how likely the speaker would use the observed *many*-statement in these states. But since the production rule in Equation (5.2) is just encoding “noisy truth-conditions” (rather than a genuine pragmatic choice of which out of several alternatives to use), the formulation in (5.3) also follows from the same considerations that motivated the production rule in (5.2): the formula in (5.3) captures interpretation based on the CFK semantics, given (Gaussian) uncertainty about threshold $x_{\min}$.

5.4 Experiments

To test the CFK semantics through the lens of the computational model from the previous section we need two types of empirical data. First, we need estimates of subjects' prior expectations P_E. Second, we need data on how sentences with *few* and *many* are used and interpreted. This section presents three experiments aimed to give us such data. All three experiments use the same 14 contexts about everyday events, objects or people which all involve a quantity of some sort (see Appendix 5.A for the full list of test items). Test items were designed so as to make a cardinal reading salient by choosing sentences with stage-level or existential predicates (Partee, 1989; Solt, 2009) and by not providing contexts in which an upper bound could be inferred. Test items are aimed to tap into general expectations of

of a park in the US. Uncertainty about a distribution could, for example, also reflect uncertainty about potential mechanisms which generate the distribution beyond mere noise.

common, every-day situations, not specific expectations about some possibly abnormal or non-stereotypical individual. We did not include fillers and no subject participated in more than one experiment.

5.4.1 Elicitation of Prior Expectations

Design. To get an empirical estimate of participants' prior expectations, we used the *binned histogram task* of Kao et al. (2014). Participants saw descriptions of a context as in (103a) and a question as in (103b). Subjects were presented with 15 intervals per item and rated the likelihood that the true value lies in each interval, by adjusting a slider labeled from "extremely unlikely" to "extremely likely". The intervals' ranges were determined by a pre-test. For each context, the pre-test asked 20 participants for the most likely, the lowest and the highest possible cardinality. Based on their answers, we determined a range of plausible values which we divided into 15 equally spaced intervals. For example, they would adjust a slider each for the probability that Andy drank 0–1, 2–3, . . . , 26–27 or more than 28 cups of coffee last week.

(103) **Prior elicitation example**

a. BACKGROUND: Andy is a man from the US.

b. QUESTION: How many cups of coffee do you think Andy drank last week?

Participants. 80 subjects were recruited via Amazon's Mechanical Turk with US-IP addresses.

Materials & Procedure. After initial instructions that explained the task, each subject saw all of the 14 contexts from Appendix 5.A one after another. For each context, the 15 intervals were presented horizontally on the screen in ascending order from left to right. On top of each interval was a vertical slider. Participants had to adjust or at least click on each slider before being able to proceed.

Results. We excluded one participant for not being a self-reported native speaker of English. Another participant was excluded for blatantly uncooperative behavior because she had not adjusted any slider. To convert likelihood judgments into probability distributions, participants' ratings for each item were normalized and these normalized ratings were then averaged across participants and once more normalized. The outcome is visualized in Figure 5.5. These probability distributions can be conceived of as approximations to the central tendencies of the beliefs held

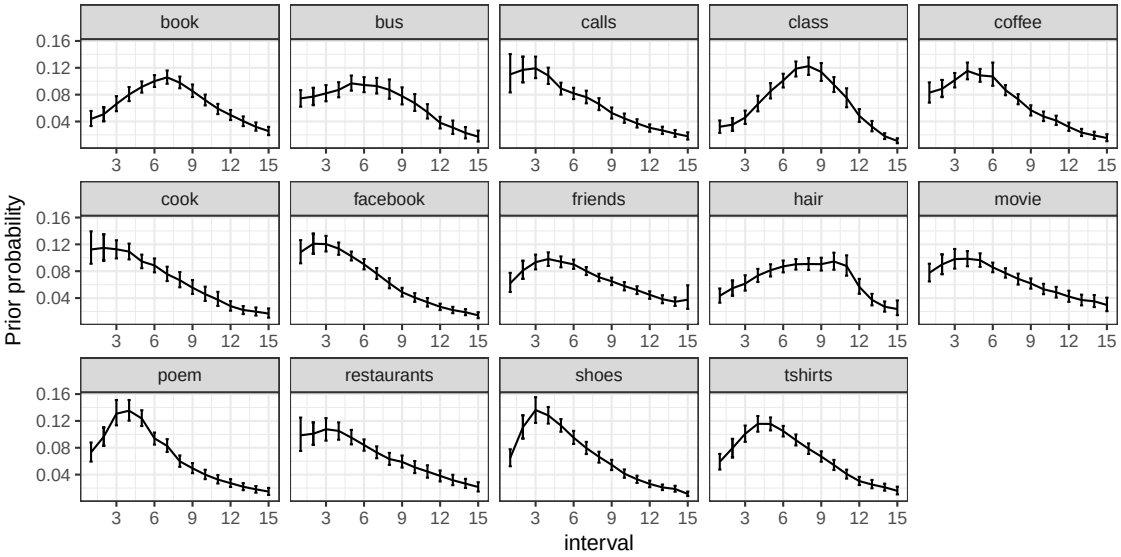


Figure 5.5: Empirically measured prior expectations. Error bars are estimated 95% confidence intervals.

within the population of participants (Franke et al., 2016). This average measure of P_E from Figure 5.5 will be input to the model.

5.4.2 Production Study: Judgment Task

Design. In a binary judgment task we measured participants' production behavior of *few* and *many*. Participants were presented with a context which introduced a situation and an interval as in (104a). The interval was randomly chosen from 8 of the 15 intervals from the prior elicitation task, for example 10-12; see Appendix 5.A. We presented only every other interval to avoid too large a number of combinations. The context was described by a statement as in (104b) which contained either *few* or *many*. Participants were asked to rate whether the statement is a good description of the context by clicking on TRUE or FALSE.

(104) **Production study example**

a. CONTEXT: Andy is a man from the US who drank [2–3 | 6–7 | . . . | 26–27] cups of coffee last week.

b. STATEMENT: Compared to other men from the US, Andy drank [few | many] cups of coffee.

c. QUESTION: Is this statement a good description of the context?

Participants. We recruited 301 participants with US-IP addresses via Amazon's Mechanical Turk.

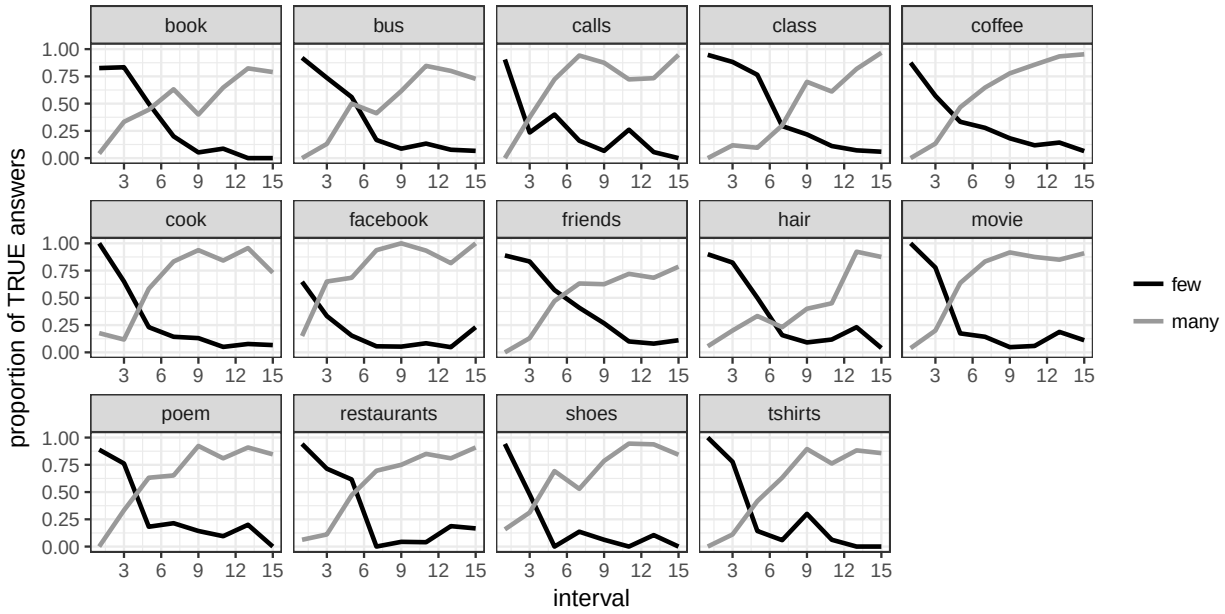


Figure 5.6: Proportion of TRUE answers from Experiment 2

Materials & Procedure. After reading a short explanation of the task, each subject saw all of the 14 contexts from Appendix 5.A one after another. For each context, one of 8 intervals and *few* or *many* were assigned randomly. Participants had to click on one of two radio buttons labeled with TRUE or FALSE before being able to proceed to the next item.

Results. Data was excluded of nine participants who reported not to be native speakers of English. Figure 5.6 shows the proportion of TRUE answers. We want the production rule P_S in Equation (5.2) to predict the data from this experiment. The decision to produce *few* or *many* to describe a certain number in the respective context is binary. Our production rule from Equation (5.2) tries to capture exactly this: the probability of whether *few* or *many* fit a given context.

5.4.3 Interpretation Study

Design. To measure how participants interpret *few* and *many* in different contexts, we used a forced-choice task. Participants saw descriptions of a context containing one of the quantifiers as in (105a) and a question as in (105b). They were presented with all 15 intervals for the given context and were asked to choose the interval that they thought is most likely given the background information.

(105) **Comprehension task example**

a. BACKGROUND: Andy is a man from the US who drank [few | many] cups of coffee last week.

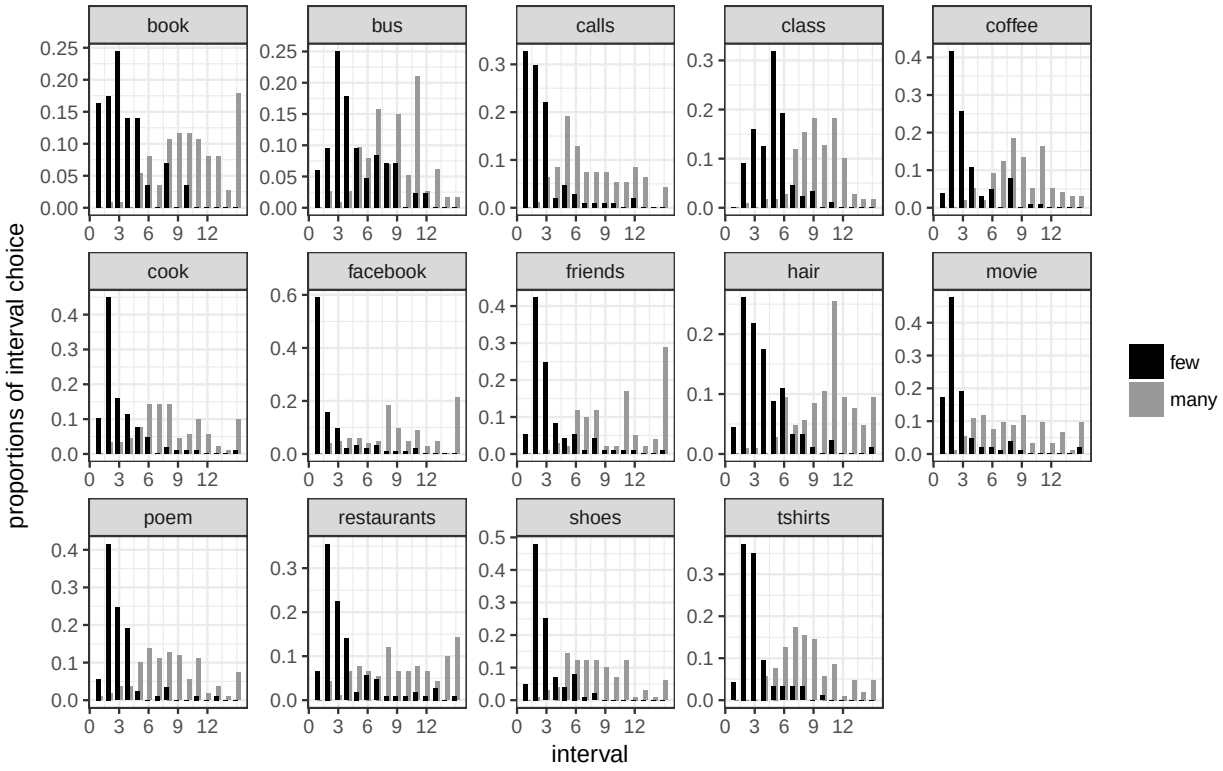


Figure 5.7: Proportions of interval choices from Experiment 3

b. QUESTION: How many cups of coffee do you think Andy drank last week?

c. INTERVALS: 0-1, 2-3, 4-5, 6-7, 8-9, 10-11, 12-13, 14-15, 16-17, 18-19, 20-21, 22-23, 24-25, 26-27, 28 or more

Participants. 200 subjects were recruited via Amazon's Mechanical Turk with US-IP addresses.

Materials & Procedure. First participants read a short introduction that explained the task. Then each subject saw all of the 14 contexts in a random order. For each context, the quantifier was selected randomly and the 15 intervals were presented horizontally on the screen in ascending order from left to right. Participants had to select one interval before being able to proceed.

Results. Data from two subjects who did not identify themselves as native speakers of English was excluded. Figure 5.7 shows the proportions of interval choices. The comprehension rule P_L in Equation (5.3) is to predict the data from this experiment.

5.5 Model Evaluation

As explained in Section 5.3, our goal is to learn about θ_{many} and θ_{few} from the observed experimental data. To this end, we feed the empirically measured prior expectations P_{E_i} for each item i (see Figure 5.5) into the production and comprehension rules in (5.2) and (5.3). This gives us likelihood functions for the production and comprehension data, which are visualized in the graphical model in Figure 5.8 and described presently. We only explicitly cover the case of *many* wherever that for *few* is analogous.

Let $O_{ij}^{p_m}$ be the number of *true* answers for item i and interval j in production experiments for *many* and let $O_{ij}^{c_m}$ be the number of times interval j has been selected as the interpretation for the relevant *many*-statement about item i in comprehension experiments. Let $N_{ij}^{p_m}$ be the number of participants that saw a production trial for *many*, item i and interval j. Likewise, $N_i^{c_m}$ is the number of participants that saw a comprehension trial for *many* and item i. $O_{ij}^{p_f}$, $O_{ij}^{c_f}$, $N_{ij}^{p_f}$ and $N_i^{c_f}$ hold the same information for conditions involving *few*. Finally, let I_{ij} be the j^{th} interval of numeric values for item i. Let $|\, I_{ij} \,|$ be the length of interval I_{ij}. The probabilistic rules from Section 5.3 then give us (parameterized) likelihood functions for observable data.

$$P(O_{ij}^{p_m} \mid \theta_{\text{many}_i}, \sigma_i) = \text{Binomial}\left(O_{ij}^{p_m},\, N_{ij}^{p_m},\, \sum_{n \in I_{ij}} \frac{P_S(\text{“many”} \mid n, P_{E_i}\ ;\ \theta_{\text{many}_i}, \sigma_i)}{|\, I_{ij} \,|} \right)$$

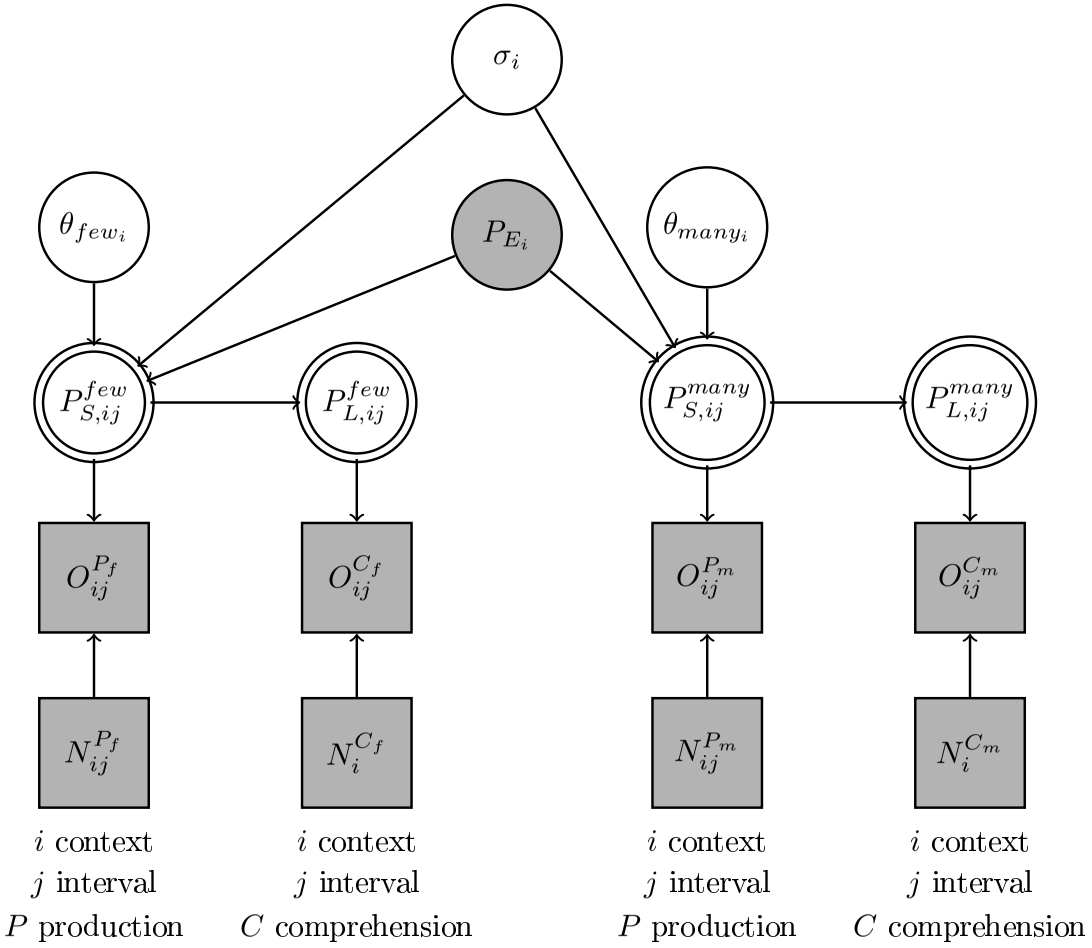
$$P(O_{ij}^{c_m} \mid \theta_{\text{many}_i}, \sigma_i) = \text{Binomial}\left(O_{ij}^{c_m},\, N_i^{c_m},\, \sum_{n \in I_{ij}} P_L(n \mid \text{“many”}, P_{E_i}\ ;\ \theta_{\text{many}_i}, \sigma_i) \right)$$

Here, $\text{Binomial}(k, n, p)$ is the probability of observing k instances of a coin coming up heads out of n coin tosses when each toss has an (independent) chance p of coming up heads.

Using Bayes rule, we can therefore make inferences about credible parameter values given the data that we observed.

$$P(\theta_{\text{many}_i}, \theta_{\text{few}_i}, \sigma_i \mid O^{p_m}, O^{c_m}, O^{p_f}, O^{c_f}) \propto \quad P(\theta_{\text{many}_i}, \theta_{\text{few}_i}, \sigma_i) \cdot \prod_j P(O_{ij}^{p_m} \mid \theta_{\text{many}_i}, \sigma_i) \cdot P(O_{ji}^{c_m} \mid \theta_{\text{many}_i}, \sigma_i) \ \cdot P(O_{ij}^{p_f} \mid \theta_{\text{few}_i}, \sigma_i) \cdot P(O_{ji}^{c_f} \mid \theta_{\text{few}_i}, \sigma_i) \tag{5.4}$$

Two remarks. Firstly, we assume here that each item has its own σ_i, but that σ_i is the same for production and comprehension, as well as for *many* and *few*. This is because we think of σ_i (and the vagueness it brings) as mainly affected by uncertainty about the contextual distribution P_{E_i}. Secondly, the formula above contains as a factor the joint prior probability $P(\theta_{\text{many}_i}, \theta_{\text{few}_i}, \sigma_i)$ of parameter values



$$\sigma_i \sim \text{Uniform}(0, 10)$$

$$\theta_{few_i}, \theta_{many_i} \sim \text{Uniform}(0, 1)$$

$$P^{few/many}_{L,ij} \propto P_{E_i} \cdot P^{few/many}_{S,ij}$$

$$P^{few/many}_{S,ij} = \sum_{k=0}^{j} \int_{k-0.5}^{k+0.5} \mathcal{N}(y; x_{\text{max/min,i}}, \sigma_i)\text{d}y$$

$$x_{\text{max,i}} = \max\{n \in \mathbb{N} \mid P_{E_i}(|\, A \cap B \,| \leq n) < \theta_{few_i}\}$$

$$x_{\text{min,i}} = \min\{n \in \mathbb{N} \mid P_{E_i}(|\, A \cap B \,| \leq n) > \theta_{many_i}\}$$

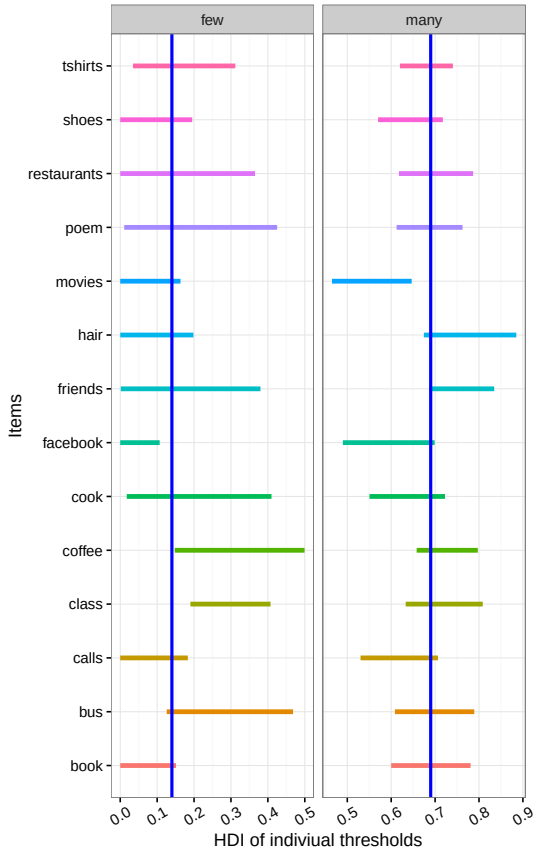
$$O^{P_f}_{ij} \sim \text{Binomial}(P^{few}_{S,ij}, N^{P_f}_{ij})$$

$$O^{C_f}_{ij} \sim \text{Binomial}(P^{few}_{L,ij}, N^{C_f}_i)$$

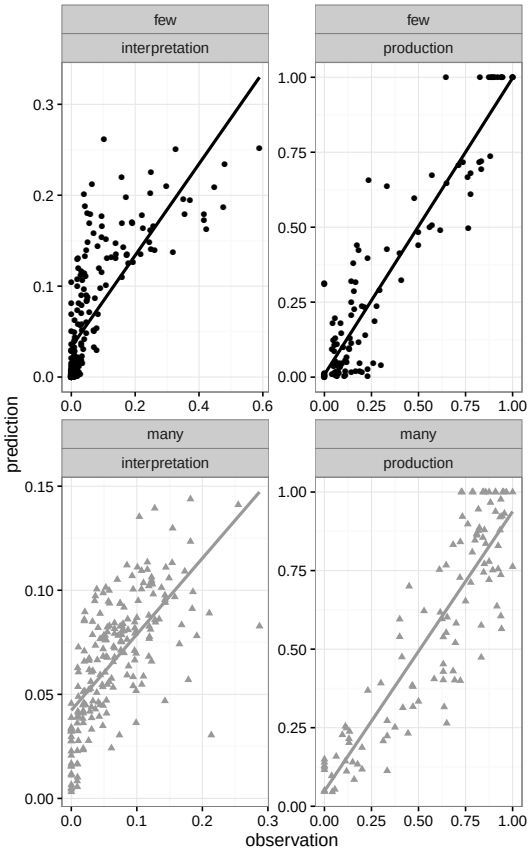
$$O^{P_m}_{ij} \sim \text{Binomial}(P^{many}_{S,ij}, N^{P_m}_{ij})$$

$$O^{C_m}_{ij} \sim \text{Binomial}(P^{many}_{L,ij}, N^{C_m}_i)$$

Figure 5.8: Graphical model of the CFK semantics



(a) 95% HDIs of the estimated posteriors for thresholds for different contexts i. The vertical lines give the biggest interval in which most contexts' HDIs overlap.



(b) Correlation of GTM's predictions and observations. For production we plot the proportion of TRUE ratings, for interpretation the proportion of interval choices.

Figure 5.9: Model Predictions

θ_{many_i}, θ_{few_i} and σ_i for each item i. Here, we simply assume that θ_{many_i}, θ_{few_i} and σ_i are independent of each other and that they have uniform priors over a large-enough interval of a priori plausible values.

$$P(\theta_{\text{many}_i}, \theta_{\text{few}_i}, \sigma_i) = \text{Uniform}_{[0;1]}(\theta_{\text{many}_i}) \cdot \text{Uniform}_{[0;1]}(\theta_{\text{few}_i}) \cdot \text{Uniform}_{[0;10]}(\sigma_i)$$

To approximate the joint posterior distribution defined in (5.4), we used MCMC sampling, as implemented in JAGS (Plummer, 2003). We collected 10,000 samples from 2 MCMC chains after a burn-in of 10,000. This ensured convergence, as measured by $\hat{R}$ (Gelman and Rubin, 1992). Figure 5.9a shows the estimated 95% credible intervals for the marginalized posteriors over θ_{many_i} and θ_{few_i} for all items.[3]

[3] A 95% credible interval is, intuitively put, an interval of values that are sufficiently plausible to warrant belief in (see Kruschke, 2014), see Section 4.1. For example, a 95% credible interval for θ_{many_i} of $[0.6; 0.8]$ for some item i would tell us that, given the data used to condition the inference, we should be reasonably certain that the true value of θ_{many_i} is in $[0.6; 0.8]$.

If for all i the credible intervals for θ_{many_i} in Figure 5.9a overlapped, and likewise for θ_{few_i}, then this would very clearly speak in favor of a CFK semantics. Such clear evidence is not forthcoming. For *many*, 13 of the 14 items' credible intervals overlap in [0.687 , 0.699]. For *few*, 12 of the 14 items' credible intervals overlap in [0.148 , 0.151]. This is close to uniformity, but there are exceptions: "movies watched per year" for *many* as well as "students in class" and "facebook friends" for *few*. In effect, we do not see clear evidence in favor of a uniform CFK semantics, but we also do not see clear evidence against it.

Another possibility of assessing the idea of a uniform CFK semantics is to compare different models. The approach in (5.4) assumes that each item i has its own semantic threshold values θ_{many_i} and θ_{few_i}. Let us call it the Individual Threshold Model (ITM). It incorporates what we called hypothesis 2 at the end of Section 5.2. We can compare the ITM with the outcome of a model that allows for only one θ_{many} and one θ_{few}, call this the General Threshold Model (GTM). GTM represents hypothesis 1. Its posterior is defined as follows:

$$P(\theta_{\text{many}}, \theta_{\text{few}}, \sigma_i \mid O^{p_m}, O^{c_m}, O^{p_f}, O^{c_f}) \propto \quad P(\theta_{\text{many}}, \theta_{\text{few}}, \sigma_i) \cdot \\ \prod_j P(O_{ij}^{p_m} \mid \theta_{\text{many}}, \sigma_i) \cdot P(O_{ji}^{c_m} \mid \theta_{\text{many}}, \sigma_i) \ \cdot P(O_{ij}^{p_f} \mid \theta_{\text{few}}, \sigma_i) \cdot P(O_{ji}^{c_f} \mid \theta_{\text{few}}, \sigma_i)\,.$$

It is also possible to use information from either only the production or the comprehension data to make inferences about latent thresholds. We will make use of that possibility too in order to see whether a uniform CFK semantics might work well for production or comprehension only. For example, an inference about likely item-specific thresholds based on production data only would use the posterior distribution given by:

$$P(\theta_{\text{many}_i}, \theta_{\text{few}_i}, \sigma_i \mid O^{p_m}, O^{p_f}) \propto \ P(\theta_{\text{many}}, \theta_{\text{few}}, \sigma_i) \cdot \\ \prod_j P(O_{ij}^{p_m} \mid \theta_{\text{many}}, \sigma_i) \cdot P(O_{ij}^{p_f} \mid \theta_{\text{few}}, \sigma_i)\,. \tag{5.5}$$

The question we are interested in is then: which model is better suited to explain the data? This question can be addressed by statistical model comparison. There are different measures for model comparison, all based on different purposes and reasons for preferring one model over another (Vehtari and Ojanen, 2012). Given our modest theoretical purposes here, we use an approach that is easy to compute based on the output of our MCMC sampling results, the so-called *deviance information criterion* (DIC) (see Chapter 4). The DIC weighs goodness of fit (here: the likelihood of the data given the model "trained" on the data) against the model's complexity (here: the number of its effective free parameters). A high value of the DIC indicates

	data used		
model	production	interpretation	both
GTM	DIC = 4191.6, $pD = 16$	DIC = 2239.6, $pD = 17$	DIC = 6546.7, $pD = 17$
ITM	DIC = 4196.0, $pD = 40$	DIC = 2182.4, $pD = 46$	DIC = 6529.5, $pD = 40$

Table 5.3: Estimated DIC values and effective free parameters

a lot of deviance of the model's predictions from the data it is applied to. This is undesirable, of course. At the same time, the model should stay as concise as possible and not include unnecessary parameters. This is measured by the pD, the number of effective free parameters, a measure of model complexity. Higher values of pD suggest higher model complexity.

Table 5.3 gives estimated DICs for the GTM and the ITM, based only on production data, based only on comprehension data and based on both data sets at once. We see that the GTM is roughly equal to, if not better than the ITM based on the production data only. It is a bit worse based on interpretation data and both data sets combined. Still, both models are clearly in the same ballpark. What the GTM misses in terms of goodness of fit, it makes up in terms of reduced model complexity. Based on our data alone, there is no clear reason to prefer either model in terms of DICs. That means that there is no reason, provided by our data, to reject the "null assumption" that a single θ_{many} and a single θ_{few} governs the use of *many* and *few*. The alternative model ITM did not do any better.

What is more, the ITM allows no possibility to generalize beyond the 14 items used here. Put differently, the ITM would assume that θ_{many} would be anywhere between 0 and 1 (its prior) for a context which was not part of the data used to condition it on. The GTM would be able to use its posterior distribution for θ_{many}. The utter lack of generalizability in ITM speaks, at least conceptually, in favor of GTM. Whether this is an empirical advantage would have to be tested. Given the data at hand and the fact that the ITM is not obviously better for this data set, there is no good reason to dismiss the hypothesis that a single pair of fixed thresholds θ_{many} and θ_{few} may have generated the production and interpretation data that we have seen.

Figure 5.9b shows the correlation between the GTM's predictions and the observed data. For production task, the correlation between predicted and observed data is 0.89 for *many* and 0.92 for *few*. For the interpretation task, we find a correlation of 0.54 for *many* and 0.73 for *few*. The GTM's predictions are less accurate for the interpretation data because in this task participants had much more freedom. They could choose between 15 intervals, whereas in the production task, only two options (TRUE and FALSE) were available. The posterior values of the noise parameters σ_i for each item i are given in Table 5.4. They express the steepness of

the production probability's curve formalized as the standard deviation of a normal distribution. σ_i ranges over the 15 intervals, whose length is dependent on the item (see Section 5.4.1).

5.6 Discussion

This chapter tried to make a methodological contribution, exemplifying a potential use of data-driven computational modeling in formal semantics/pragmatics. By measuring subjects' prior expectations about real-world events experimentally, we set out to test a proposal for a semantics of *few* and *many* that is hard to assess introspectively. We showed how to couch the CFK semantics for *few* and *many* in a probabilistic model for production and comprehension. With the help of this model, we inferred *a posteriori* credible values for latent threshold parameters θ_{many} and θ_{few} from experimental data that aimed to measure production and comprehension behavior. Posterior credible values of individual threshold parameters θ_{many_i} and θ_{few_i} for different experimental items i are very similar, with overlap in the 95% HDIs of almost all items. Moreover, statistical model comparison in terms of DICs does not favor a model with individual thresholds for each item over a more parsimonious model that assumes only one fixed threshold for *many* and one for *few*. The model comparison based on fit to the data and model complexity supports Hypothesis 1. Consequently, the question whether a fixed threshold CFK semantics is plausible can be answered positively, at least for the data set at hand. This finding is especially credible in the light of language acquisition because it is not a plausible assumption that the meaning of *few* and *many* would have to be learned anew for each context. Factors that might be responsible for the observed finding that not all credible intervals do fully overlap could be a methodological issue in the elicitation of priors, too much uncertainty about the threshold or that not all speakers shared the same comparison classes. For now, we are not able to solve all of these factors. Instead, we want to point out some of them as questions for further investigation.

The benefits of theoretically informed statistical modeling of this kind are many. The computational model makes explicit all modeling assumptions including any linking hypotheses regarding how theoretical notions relate to each other in producing the observable data (e.g. Chemla and Singh, 2014; Franke, 2016). The model

	book	bus	calls	class	coffee	cook	facebook	friends	hair	movie	poem	restaurants	shoes	tshirts
σ	4.1	5.5	4.8	4.2	4.7	4.6	4.5	5.6	4.7	4.9	4.6	5.7	3.6	3.4

Table 5.4: Mean of GTM's posterior distribution of σ per item

considered here, for instance, assumes that the production and comprehension data are only driven by considerations of truth. In other words, this quite simple semantic model assumes that participants in, say, Experiment 3 would not reason about what other expressions a speaker may have used other than *many* or other than *few*. However, alternative utterances containing *a few, lots of* or *surprisingly few* are very likely also taken into account during the speaker's precision process.

Furthermore, some experimental items from Schöller and Franke (2015) revealed other relevant factors which can influence the use of *few* and *many* and might want to be included in a more elaborate model. The first is the grammatical number feature which requires that the quantity words combine with a plural noun. This constraint turned out to be a stronger factor than participants' prior expectations.

(106) John is a man from the US who has few children.
How many children do you think John has?

Even though the prior elicitation task confirmed that participants consider it very likely that an American man has 0 or 1 children, *few* was nevertheless not interpreted as describing a singular noun. This issue is also related to a pragmatic competition of *few* with the quantifiers *none* and *one*.

Another observation from Schöller and Franke (2015) is that participants do not only employ their expectations of the statistical properties in some contexts, but also their moral standards of which cardinalities are considered too low or too high. For the context of a smoker's cigarette consumption, the model's inferred threshold values differed to a large extent from those of the remaining items

(107) Margaret is a woman from the US who smokes few/many cigarettes a day.
How many cigarettes do you think Margaret smokes a day?

Participants' answers were very low, compared to the prior expectations measured. Most people judged a sentence with *few* true only for the lowest presented interval and true for a sentence with *many* for all of the other intervals. Maybe participants did not use the prior expectations as they did for the other context. Since smoking has fallen in disrepute in the US, people might not only use their plain "statistical" prior expectations when they form a judgment about "few or many cigarettes." They might factor in their "moral expectations" as well (cf. Égré and Cova, 2014). In principle, a CFK semantics is compatible with this idea. The prior expectations P_E would not only have to be sensitive to statistical beliefs about, in this case, actual number of cigarettes smoked, but also to a deontic dimension about how many cigarettes should be smoked.

A last issue directly related to the model is the role of the noise parameters σ_i. We introduced them as capturing uncertainty about P_E "and perhaps other things". What these other things might be is not answered by the model and should also

receive more attention since the values of σ given in Table 5.4 are quite large. The posterior distribution predict a standard deviation of roughly five intervals, which makes up one third of the quantity word's scale in the experiment. If σ really turned out to capture uncertainty about P_E, this uncertainty could be reduced by backing away from population-level expectations. The prior expectations which are input into the model were obtained by measuring the subjective beliefs of 80 subjects and then averaging across the (normalized) observed slider ratings. These representations might not be adequate for the individual subject carrying out a production or interpretation task resulting in a worse fit to the data and maybe even wrong model predictions. One way to address this problem is to replicate the experiments from the Section 5.4 as a within-subjects design and have the same individual complete all three tasks: give her expectations and subsequently produce and interpret *few* and *many*. This way, the CFK semantics could be tested based on individual expectations and not on average beliefs held within a sample of the population. If σ, however, accounts for uncertainty about other factors as well, say the threshold values or alternative utterances, the noise parameter could be split up and made explicit in the model to help us learn about the vagueness and context-dependence of *few* and *many*

Taken together, we see that the model presented here is a stark simplification. Nevertheless, it fits the data surprisingly well given its simplicity and contributes to our understanding of *few* and *many*. The benefit of probabilistic modeling is not only in bringing these assumptions and simplification to the fore, but in providing direct means of testing whether they are correct or, by means of model comparison, which linking hypotheses may actually be better suited to explain the data. An interesting next step is therefore to develop further model variants which include pragmatic, grammatical or moral considerations and compare their predictions.

Besides considerations of how to further develop and improve the present model, the methodological approach introduced here opens even more interesting venues for future research. Firstly, inference of latent thresholds could naturally be applied beyond our example case of *few* and *many*. Context-dependent threshold values are also assumed to form part of the semantics of gradable adjectives (Kennedy and McNally, 2005; Kennedy, 2007) and of other vague quantifiers like *most* (Hackl, 2009). Computational models in combination with experimental data put themselves forward as a promising method to investigate these phenomena within a uniform framework.

Secondly, we can use probabilistic modeling to compare the CFK semantics against alternatives. For example, a different account for the meaning of *few* and *many* was proposed by Solt (2011b). Here, the threshold is derived as a positive or negative deviation from the median of the comparison class. This theory can just as

well be couched in a probabilistic model and its predictions can then be compared against the CFK semantics, using statistical model comparison.

Thirdly, it is an open issue whether a CFK semantics, as formulated here, can also account for other readings of *many* and *few*. Fernando and Kamp (1996) apply a similar idea to proportional readings. The key feature of the proportional reading is the existence of an upper bound on the scale *few* and *many* operate on. Consequently, also the prior expectations P_E are a distribution on a bounded interval. In contrast to the cardinal reading, this opens the possibility of employing an uninformed, uniformly distributed prior expectation which captures whether the described proportion is small or large. This idea is linked to the question whether the proportional reading expresses a fixed proportion. We set out to address this question in Chapter 7. But there may be even more potential readings of *few* and *many*, such as the *inverse proportional reading*, as already discussed in detail in Chapter 2. This reading makes sentence (108) true if the proportion of Scandinavians among Nobel prize winners was bigger than the proportion of people from other contextually salient alternative world regions who won a Nobel prize (c.f. Westerståhl, 1985; Eckardt, 1999; Cohen, 2001; Romero, 2015).

(108) Many SCANDINAVIANS won the Nobel prize.

(109) Inverse proportional reading of "Few/Many As are B"

a. *Few*: $\frac{|A \cap B|}{|A|} \leq \frac{|\bigcup Alt(A) \cap B|}{|\bigcup Alt(A)|}$ b. *Many*: $\frac{|A \cap B|}{|A|} \geq \frac{|\bigcup Alt(A) \cap B|}{|\bigcup Alt(A)|}$

It could be hypothesized that it is just a matter of specifying the right P_E to account for these cases as well within a CFK-approach. For the inverse proportional reading of (108) in (109) we would need to consult the cumulative probability of the actual number of Scandinavians with a Nobel prize to an expectation P_E that takes, presumably, the average number of Nobel laureates in the set of all relevant world regions. It would need to be seen how far the CFK-approach can be pushed in this direction (c.f. Fernando and Kamp, 1996). Still, data-driven computational modeling seems like just the right tool to help in this investigation.

Finally, it would be interesting to not only infer plausible threshold values but to try to *explain why* we see the threshold values that we apparently see. Focusing on the case of gradable adjectives, Lassiter and Goodman (2015) give a model that suggests that threshold values are the result of pragmatic inferences; another approach tries to explain why particular threshold values are evolutionarily optimal for successful communication (Franke, 2012; Qing and Franke, 2014a). Testing these theoretical accounts with data-driven inferences of credible thresholds and applying statistical model comparison would be a natural next step.

5.A Experimental material

1. **book** — A friend's favorite book has been published only recently (and has few/many pages). — How many pages do you think the book has? — intervals: 0-40, 41-80, 81-120, 121-160, 161-200, 201-240, 241-280, 281-320, 321-360, 361-400, 401-440, 441-480, 481-520, 521-560, 560 or more
2. **bus** — Vehicle No. 102 is a school bus (which has seats for few/many passengers). — How many passengers do you think can sit in Vehicle No. 102? — intervals: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70 or more
3. **calls** — Lisa is a woman from the US (who made few/many phone calls last week). — How many phone calls do you think Lisa made last week? — intervals: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70 or more
4. **class** — Erin is a first grade student in primary school. (There are few/many children in Erins class.) — How many children do you think are in Erin's class? — intervals: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-38, 39-41, 42 or more
5. **coffee** — Andy is man from the US (who drank few/many cups of coffee last week). — How many cups of coffee do you think Andy drank last week? — intervals: 0-1, 2-3, 4-5, 6-7, 8-9, 10-11, 12-13, 14-15, 16-17, 18-19, 20-21, 22-23, 24-25, 26-27, 28 or more
6. **cook** — Tony is a man from the US (who cooked himself few/many meals at home last month). — How many meals do you think Tony cooked himself at home last month? — intervals: 0-3, 4-7, 8-11, 12-15, 16-19, 20-23, 24-27, 28-31, 32-35, 36-39, 40-43, 44-47, 48-51, 52-55, 56 or more
7. **facebook** — Judith is a woman from the US (who has few/many Facebook friends). — How many Facebook friends do you think Judith has? — intervals: 0-69, 70-139, 140-209, 210-279, 280-349, 350-419, 420-489, 490-559, 560-629, 630-699, 700-769, 770-839, 840-909, 910-979, 980 or more
8. **friends** — Lelia is a woman from the US (who has few/many friends). — How many friends do you think Lelia has? — intervals: 0-1, 2-3, 4-5, 6-7, 8-9, 10-11, 12-13, 14-15, 16-17, 18-19, 20-21, 22-23, 24-25, 26-27, 28 or more
9. **hair** — Betty is a woman from the US (who washed her hair few/many times last month). — How many times do you think Betty washed her hair last month? — intervals: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-38, 39-41, 42 or more

10. **movie** — Nick is a man from the US (who saw few/many movies last year). — How many movies do you think Nick saw last year? — intervals: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-38, 39-41, 42 or more

11. **poem** — A friend wants to read you her favorite poem (which has few/many lines). — How many lines do you think the poem has? — intervals: 0-3, 4-7, 8-11, 12-15, 16-19, 20-23, 24-27, 28-31, 32-35, 36-39, 40-43, 44-47, 48-51, 52-55, 56 or more

12. **restaurants** — Sarah is a woman from the US (who went to few/many restaurants last year). — To how many restaurants do you think Sarah went last year? — intervals: 0-3, 4-7, 8-11, 12-15, 16-19, 20-23, 24-27, 28-31, 32-35, 36-39, 40-43, 44-47, 48-51, 52-55, 56 or more

13. **shoes** — Melanie is a woman from the US (who owns few/many pairs of shoes). — How many pairs of shoes do you think Melanie owns? — intervals: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-38, 39-41, 42 or more

14. **tshirts** — Liam is a man from the US (who has few/many T-shirts). — How many T-shirts do you think Liam has? — intervals: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-38, 39-41, 42 or more

Chapter 6

Surprise Readings

The previous chapter demonstrated how the CFK semantics can be tested, a theory which makes concrete predictions about the calculation of the threshold values for the applicability of *few* and *many* in context. The computational modeling approach supports the idea that speaker and listener behavior can be explained by a fixed pair of thresholds θ_{many} and θ_{few} which apply to a probability distribution representing prior expectations of the context.

We have claimed that such a semantics is able to account for the so-called "cardinal surprise reading". This reading of cardinal *few* and *many* in a sentence like (110) compares the cardinality described by the quantity word with quantitative expectations about cardinalities in the respective context, as exemplified below, and thus makes *few* and *many* dependent on prior expectations. This assumption is in line with a long tradition in psychology which has acknowledged the role of prior expectations in the use of vague and context-dependent expressions like *tall, heavy, few* and *many* (e.g. Clark, 1991; Sanford et al., 1994; Lassiter and Goodman, 2013; Qing and Franke, 2014b).

(110) For a man from the US, Chris saw few/many movies last year.
$\rightsquigarrow$ Chris saw less/more movies than expected for a US male.

Apart from the question of how to test the semantic account of surprise readings, the theory brings up further interesting questions: So far the assumption that *few* and *many* receive a surprise reading has not been challenged. In Section 6.1, we discuss how the surprise reading can be made salient by a *compared to*-phrase (111).

(111) Compared to what you would expect for a man from the US, Chris saw few / many movies last year.

One question arises: if sentences with *few* and *many* express that a cardinality is surprising anyway, are they different from sentences in which the surprise element is overtly marked?

In Section 6.2, *few* and *many* are modified by *surprisingly* as in (112).

(112) For a man from the US, Chris saw surprisingly few / many movies last year.

Whether this adverb functions as a marker of surprise or as an intensifier and whether the surprise reading is the most salient reading of *few* and *many* is tested with an experiment and a computational model in Sections 6.3 and 6.4. The computational model is essentially the same as in the previous chapter. It incorporates the CFK semantics and infers plausible threshold values on a cumulative probability distribution of contextual expectations. After an interim summary in Section 6.5, we present a follow up experiment in Section 6.6 and discuss the findings and in particular the semantics of *few* in Section 6.7.

6.1 Making Surprise Readings Overt

As briefly mentioned above, language has several means of making the surprise element overt that we assume to be part of the reading of *few* and *many*. In this chapter, some options are introduced, in particular the *compared to* construction and the adverb *surprisingly*. The ultimate goal of this chapter is to learn about the meaning contribution of surprise and to test whether the surprise reading is really the salient reading of *few* and *many*. We will do so by contrasting unmodified *few* and *many* with constructions in which surprise is explicitly made salient.

The first construction which can mark surprise is the frame setter *compared to*. The *compared to*-phrase "serves to indirectly (contextually) fix the intended value for the comparison standard" (Beck, 2009) and its function is to "set the context for the following sentence" (Beck et al., 2004).

(111) Compared to what you would expect for a man from the US, Chris saw few / many movies last year.

Loosely following Hohaus (2015), (111) (repeated from above) is true iff the cardinality of movies seen by Chris is lower/higher than some contextually provided standard and only defined for situations in which comparison is with expectations about men in the US. I refer the reader to Hohaus (2015) for an in-depth introduction into the semantics of *compared to* constructions.

Under the assumption of a surprise reading of *few* and *many*, the meaning of sentence (111) should be very similar if not identical to sentence (110) in which the comparison class, namely US men, is marked by a *for*-phrase but the surprise component is contributed by the quantity word. *For*-phrases denote comparison classes which affect the standard involved in the semantics of positive forms of gradable adjectives and they presuppose that the subject of the gradable predicate is included in the comparison class set (Klein, 1980; Kennedy, 2007).

	hypothesis		
	intensifier	**marker of surprise**	**salient surprise reading**
predictions	*many* $\leq$ *surprisingly many* *few* $\geq$ *surprisingly few* *surprisingly* = *incredibly*	*many* = *surprisingly many* *few* = *surprisingly few* *surprisingly* = *compared to*	*many* = *compared to... many* *few* = *compared to... few*

Table 6.1: Hypotheses for sentences expressing surprise readings

(110) For a man from the US, Chris saw few / many movies last year.

This sentence is defined iff Chris is a man from the US and true iff the number of movies seen by Chris is lower/higher than the standard number of movies an American man is expected to watch.

In a nutshell, when assuming that *few* and *many* express a surprise reading and make reference to expectations, sentences like (110) are predicted to have very similar truth conditions to sentences in which expectations are overtly marked by a *compared to*-phrase as in (111). For this reason, we expect no difference in a judgment task when either (110) or (111) is used to describe the same cardinality. In the following we spell out this claim in terms of its predictions about the threshold values θ_{few} and θ_{many} as assumed by Fernando and Kamp (1996). A first hypothesis about the influence of a *compared to* phrase marking surprise in sentences with *few* and *many* is given below (**salient surprise reading**). In the next section, hypotheses of the interaction between *surprisingly* and the quantity words are developed (**marker of surprise, intensifier**). The hypotheses are tested with a computational model which infers these threshold values on the basis of experimental data in the remainder of this chapter and are summarized in Table 6.1.

Salient surprise reading. We cannot exclude that *few* and *many* may also denote a small or large cardinality, independent of prior expectations. Nevertheless, we assume that the most salient readings of our experimental test sentences (see Appendix 6.A) are cardinal surprise readings given the comparison class for which we measure information about subjects' prior expectations (see below). To test this assumption, we contrast sentences with bare *few* and *many* with sentences modified by the *compared to* phrase in (111) which makes the relevant expectations overt. It is necessary to test this because if *few* and *many* did not have the intended surprise reading, differences between *few/many* and *surprisingly few/many* could be due to different readings and possibly different threshold values associated with them. Alongside *few* and *many*'s intrinsic surprise reading, we test another related assumption: the *for-* phrase used to mark the comparison class triggers the same prior expectations P_E as the *compared to* phrase which openly addresses expectations, see (111) and (112).

6.2 *Surprisingly*: Marker of Surprise Readings or Intensifier

Another way of marking surprise in sentences with *few* and *many* is by adding the adverb *surprisingly*. *Surprisingly* can appear in two positions in the sentence.

(113) a. Surprisingly, Chris saw few / many movies last year.
⇝ It is surprising that Chris saw few/many movies last year.

b. Chris saw surprisingly few / many movies last year.
⇝ The number of movies which Chris saw last year is surprisingly low/high.

In a sentence initial position, *surprisingly* takes over the role of a sentence adverbial, as we can see in example (113a). It marks the entire proposition as being surprising. When uttering (113a), the speaker expresses his surprise about the fact that Chris watched few/many movies. Which number of movies counts as *few* or *many* is determined independently of the adverb. For this reason, the second occurence is more interesting in the scope of this dissertation. When *surprisingly* precedes the quantity word, it functions as a degree modifier as in (113b). In its most salient reading (113b) expresses that the number of movies watched by Chris is surprisingly low or high.

With respect to our assumption that *few* and *many* express a surprise reading, two views are prima facie plausible for the meaning contribution of the adverb *surprisingly*. Note that our hypotheses for *surprisingly* apply to sentences with a salient cardinal surprise reading and a restricted comparison class. On the one hand, *surprisingly* can be taken to intensify the meaning of *few* and *many* just like other intensifiers like *incredibly* or *very* do. As a result, *surprisingly many* might be associated with a threshold $\theta_{\text{surpr. many}}$ higher than θ_{many}. The contrasting view is to classify *surprisingly* as a marker of the surprise reading, which overtly marks that truth-conditions must draw on a threshold on a measure of surprise. In this view, the threshold of *surprisingly few/many* should not be different from unmodified *few/many* under a surprise reading. The later view is supported by the semantic literature.

Katz (2005) and Nouwen (2011) discuss the relation between a gradable adjective modified by *surprisingly* and its unmodified positive form.

(114) a. Jasper is surprisingly tall.

b. Jasper is tall.

For the relation between the sentences (114a) and (114b), Nouwen (2011) suggests that "being surprisingly tall comes to mean taller than expected". Crucial to his

proposal is the role of inferences and the assumption that gradable predicates are monotone. If Jasper is surprisingly tall, this means that there exists a degree to which Jasper is tall that is surprising (for someone like Jasper). Had Jasper been taller, he would also have been tall to a surprising degree (by monotonicity). So we infer that had Jasper been taller, he would also be called *surprisingly tall*. "This is why we can only use [*surprisingly tall*] to refer to someone who is *taller than (what is considered) [expected]*" (Nouwen, 2011, 154). Note that statements about degrees actually license downward directed inferences, but adverbs like *surprisingly* reverse such inferences and license upward directed inferences. If Andy is tall to degree d, then he is also tall do any degree lower than d. For *surprisingly tall*, the opposite holds. If Andy is surprisingly tall, his height would still be surprising if he were taller, but not necessarily if he were shorter.

Furthermore, Katz (2005) and Nouwen (2011) agree that sentences with *surprisingly* do not entail the positive form of the same gradable predicate. Sentence (114a) does not entail (114b), since even though Jasper may be surprisingly tall (given that his parents are very short, for example) he is not necessarily tall for general standards. Note that this argument can be misunderstood as evidence against the intensifier hypothesis. However, the entailment relation between (114a) and (114b) only fails, when the comparison class is changed, for example from people with short parents to people in general. For this reason, Katz's (2005) and Nouwen's (2011) observation does not constitute evidence for or against the intensifier hypothesis.

Semantically, the degree modifier *surprisingly* is analyzed as a propositional modifier of type $\langle st, st\rangle$ or $\langle st, t\rangle$ (Nouwen, 2011; De Vries, 2012), which expresses the speaker's surprise about the information she is conveying. A sample denotation by Nouwen (2011) is given below:

(115) Nouwen (2011)

a. $[\![\text{surprisingly}]\!] = \lambda p.\lambda w.p(w) \& \text{surprising}_w(p)$
$\rightsquigarrow$ via type-shift and existential closure

b. $[\![\text{surprisingly tall}]\!] = \lambda x.\lambda w.\exists d[\text{tall}_w(x, d) \& \text{surprising}_w(\lambda w'.\text{tall}_{w'}(x, d))]$

c. $[\![\text{Ann is surprisingly tall}]\!](\text{w}) = 1$
iff $\exists d[\text{tall}_w(\text{A}, d) \& \text{surprising}_w(\lambda w'.\text{tall}_{w'}(\text{A}, d))]$

Nouwen (2011) predicts (115c) to be true if there exists a degree d such that Ann is d-tall and it is surprising that Ann is d-tall. In other words, the sentece is true if Ann is taller than expected. *Surprisingly* in (115b) functions like *POS* by being type-shifted to be able to apply the modifier semantics in (115a) to gradable predicates. The adjective's degree variable is existentially bound before combining with the subject. Since the standard of comparison is only inferred on the basis of the monotonicity assumption described above, there is no explicit prediction of how the

expected degrees are determined in the respective context. I take it that *POS* and (115b) function in a parallel way by defining which degrees count as significantly higher than the contextual standard.[1] Sentences like (113a) in which *surprisingly* functions as a sentence adverbial can be accounted for by applying (115a) to the entire sentence. De Vries's (2012) and Piñón's (2005) proposals are similar in spirit: *surprisingly* is a modifier of propositions and expresses that the proposition (that the gradable predicate holds for a certain degree) is surprising. All in all, the three accounts (Piñón, 2005; Nouwen, 2011; De Vries, 2012) derive essentially the same result. *Suprisingly* is semantically a modifier of propositions and expresses that the proposition is surprising. We do not find predictions about its influence on the threshold of a gradable predicate; it is not explicitly classified as an intensifier which raises the threshold of applicability. The semantic literature can be interpreted to predict that *surprisingly* only marks the surprise reading, just like the *compared to* phrase in (111) is expected to do.

At the same time, the suspicion that *surprisingly* functions as an intensifier cannot be ruled out. It behaves parallel to other intensifiers like *incredibly, extremely* or *very* which also modify gradable predicates and with which *surprisingly* is in complimentary distribution. Note that the following linguistic test does not only apply to *few*, but also to *many* and other gradable predicates like *tall*.

(116) a. Liam has surprisingly few T-shirts.
b. Liam has incredibly few T-shirts.
c. ?? Liam has surprisingly incredibly few T-shirts.
d. ?? Liam has surprisingly very few T-shirts.
e. ?? Liam has incredibly very few T-shirts.

The following two paragraphs develop an intensifier semantics for *surprisingly* which can only account for its degree modifier variant. The sentence adverbial cannot be explained with this lexical entry.
Even though the semantics does not make explicit predictions about the influence of *surprisingly* on the threshold of the gradable predicate it modifies, there is an interesting parallel to the semantics of intensifiers. Heim (2006) and von Stechow (2006) both assume that *very* also takes over the role of *POS* with which it stands in complimentary distribution. *Very* raises the boundaries of the neutral interval. "The very interval must be a superinterval of the neutral interval of N(S) that sym-

[1] Nouwen (2011) does not further comment on the consequences of assuming that *surprisingly* takes over the role of *POS*. He does not spell out whether the same scope interaction that we see for *surprisingly* in (113a) and (113b) would also be predicted for *POS*. It is not elaborated on what exactly triggers the type shift and the binding of the degree argument when *surprisingly* combines with a gradable adjective.

metrically includes both bounds or N(S)" (von Stechow, 2006, 7) and it is possible to iterate the very-operation (for example, in *very very many*).

(117) von Stechow (2006)
$[\![\text{very}_{N,S}]\!]^c = \lambda D_{\langle d,t\rangle}$: c specifies an Interval I that symmetrically includes N(S) and is considerably bigger than N(S). $(\forall d \in I)D(d)$

I understand a symmetrical inclusion to mean that the distance from the lower bound of I to the lower bound of N(S) must be the same as from the upper bound of N(S) to the upper bound of I.

The semantics for *very* can be extended to *surprisingly* to form an intensifier which functions like *POS* and is compatible with surprise readings and the CFK semantics, see our proposal for POS^{surp} in (87) in Section 2.4.

(118) $[\![\text{surprisingly}]\!] = \lambda D_{\langle s,\langle d,t\rangle\rangle}.\lambda w_0 :$ I symmetrically includes $N_S = [x_{\max}, x_{\min}]$,
$x_{\max} = \max\{n : \sum_{m=0}^{n} P_E(m) \leq \theta_{\text{few}}\}$ and
$x_{\min} = \min\{n : \sum_{m=0}^{n} P_E(m) \geq \theta_{\text{many}}\}$
for $P_E(m) = \sum_{w \in S_m} \mu(w)$ and $m \in \mathbb{N}$
and $S_m = \{w : w \in \text{Dox}_{t_i}(\text{w}_0)\ \&\ \max(D(w)) = m\}$
$I \subseteq D(w_0)$

The neutral interval N(S) is the result of determining the cut-off points $x_{\max}$ and $x_{\min}$ based on prior expectations P_E, θ_{few} and θ_{many}. *Surprisingly* would then impose a superinterval I on N(S), just like von Stechow's (2006) *very* does. This way we don't contradict the stable core meaning hypothesis since θ_{few} and θ_{many} as the lexical meaning of *few* and *many* can remain unchanged. *Surprisingly* only modifies the already determined N(S) to intensify the meaning of *few* or *many*. What remains to be tested is whether *surprisingly* keeps up the symmetrical inclusion requirement from (117). This is why for now we only tentatively suggest the semantics in (118).[2]

From a pragmatic point of view, an intensifying effect of *surprisingly* is plausible since a speaker makes the effort of uttering a longer and thus more costly sentence when she could also only have used unmodified *few* or *many*. Consequently, a speaker who adds *surprisingly* can be taken to assume that this utterance is more informative. For example, *surprisingly* would rise the threshold of *many* and make it thus applicable to a smaller range of cardinalities, which results in a stronger statement than the alternative with bare *many*. This is in line with work about the pragmatic effects of intensifiers.

The intensifier hypothesis is further supported by a pragmatic theory by Bennett and Goodman (2015). They explain the strength of an intensifying degree adverb

[2]Note that this preliminary lexical entry only captures occurrences of *surprisingly* as a degree modifier.

as "pragmatic inference based on differing cost [(their length and frequency)] rather than differing semantics" (p. 1). Bennett and Goodman (2015) test 40 intensifiers, like *amazingly*, *terribly* or *seriously*, which have a high frequency in the Google Web 1 T 5grams corpus and do not signal affect (like *depressingly* would). Each intensifier was paired with the adjective *expensive* to describe three categories of objects (laptop, watch and coffee maker). In a free production task, 30 participants on Amazon's Mechanical Turk were asked to give their estimate of the prize of the objects as described by "[intensifier] expensive". Bennett and Goodman (2015) find a linear relationship between the meaning of intensifiers and their length and frequency. The adverb *surprisingly* is not part of their set of intensifiers though. From the adverbs tested by Bennett and Goodman (2015), *incredibly* comes closest to *surprisingly*, as they have the same number of syllables and the most similar frequency in an updated version of the corpus Bennett and Goodman (2015) used, the Google Web 1 T 5grams corpus (4,987,059 occurrences of *incredibly* as compared to 4,373,670 occurrences of *surprisingly*).

To discriminate between the two views on *surprisingly*, we deduce two experimentally testable hypotheses. Another auxiliary hypothesis of *incredibly* is tested alongside to complement our understanding of modified *few* and *many*, see Table 6.1. In what follows, we again spell out these general hypotheses in terms of their predictions about the threshold values θ_{few} and θ_{many}. We run a judgment task to gather experimental data, which will be input to a theory-driven, computational model.

Marker of surprise. If the function of *surprisingly* is to mark a cardinal surprise reading, thresholds are the same as for unmodified *few/many*, where these cardinal surprise readings are most salient anyway (see above). Furthermore, sentences with *surprisingly* should not be different from sentences with *compared to*, as in (111).

Intensifier. Modification by *surprisingly* raises the threshold of *many* and makes it applicable to a smaller range of cardinalities, resulting in a stronger statement than the alternative with bare *many*. *Few*'s threshold decreases.

Bennett & Goodman. The intensifier hypothesis is in line with work by Bennett and Goodman (2015) who explain the strength of an intensifying degree adverb as "pragmatic inference based on differing cost [(their length and frequency)] rather than differing semantics" (p. 1). Following Bennett and Goodman (2015), we hypothesize that the thresholds of *surprisingly few/many* are roughly the same as for *incredibly few/many*.

6.3 Experiments

To test the hypotheses in Table 6.1, two experiments were conducted to gather acceptability ratings of sentences with (modified) *few* and *many* and to measure representations of participants' prior expectations. The prior expectations task and its results are the same as in Section 5.4.1. It is summarized briefly. Prior expectations will be input to the computational model from 5.3, which is presented again briefly in the next section.

6.3.1 Elicitation of Prior Expectations

Design. To get an empirical estimate of participants' prior expectations, we used a *binned histogram task*. A sentence as in (119a) introduced a comparison class and a question as in (119b) asked about typical cardinalities of every-day situations. Subjects rated the likelihood that the true value lies in the 15 intervals, by adjusting a slider each labeled from "extremely unlikely" to "extremely likely."

(119) **Prior elicitation example**

a. BACKGROUND: Chris is a man from the US.

b. QUESTION: How many movies do you think he saw last year?

Participants. 80 subjects were recruited via Amazon's Mechanical Turk with US-IP addresses.

Materials & Procedure. Materials and procedures were the same as in Section 5.4.1, see Appendix 5.A.

Results. For each item, each participant's ratings were normalized and these normalized ratings were then averaged across participants. The results displayed in Figure 6.1 were already reported in Section 5.4.1.

6.3.2 Production Study: Judgment Task

Design. In a binary judgment task we measured acceptance of sentences with *few* and *many* with and without modifiers (*surprisingly, incredibly* or *compared to*). Participants were presented with a context which introduced a situation and an interval as in (120a). The interval was randomly chosen from 8 of the 15 intervals from the prior elicitation task (see Appendix 6.A). We presented only four low intervals for *few* and four high intervals for *many* to avoid a large number of combinations. The context was described by a statement as in (120b) which contained either *few* or

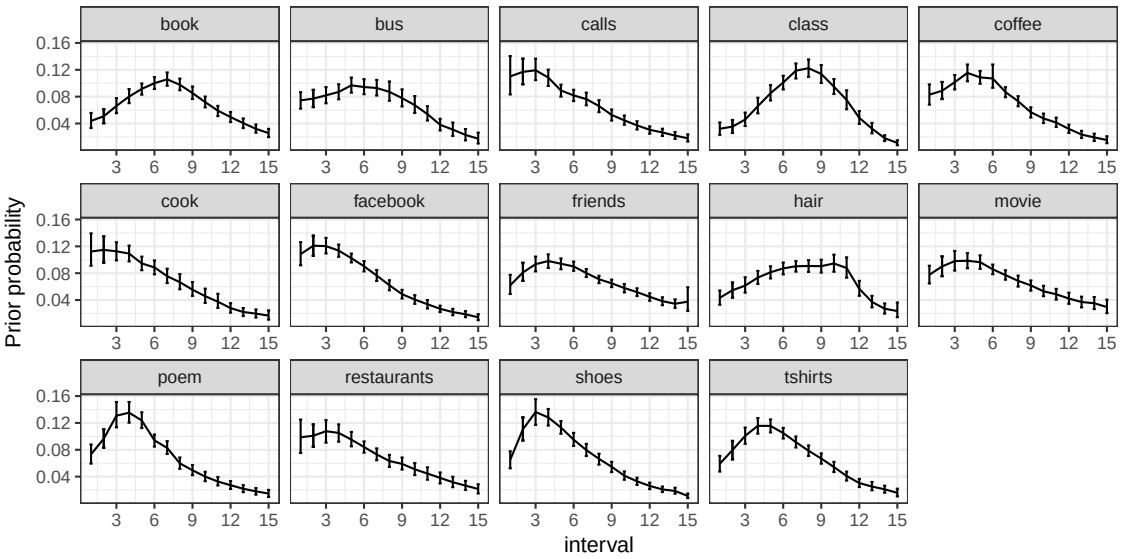


Figure 6.1: Empirically measured prior expectations from Section 5.4.1. Error bars are estimated 95% confidence intervals.

many. We elicited data of four groups of participants which each saw a different modifier.

(120) **Production study example**

a. CONTEXT: Chris is a man from the US who saw [0–2 | 6–8 | . . . | 42 or more] movies last year.

b. STATEMENT: [For | Compared to what you would expect for] a man from the US, Chris saw [- | surprisingly | incredibly] [few | many] movies last year.

Materials & Procedure. Each participant was randomly assigned to one modifier condition (unmodified, *compared to* construction, *surprisingly*, *incredibly*). After reading a short explanation of the task, each subject saw all of the 14 contexts from Appendix 6.A one after another in random order. Sentences with unmodified *few* and *many* or *incredibly* or *surprisingly* were introduced by a *for*-phrase which made the intended comparison class overt. The fourth group saw a *compared to* phrase which additionally made expectations salient. For each context, a quantity word and one of its four associated intervals were assigned randomly. Participants had to click on one of two radio buttons labeled with TRUE or FALSE before being able to proceed to the next item.

Participants. We recruited 787 participants with US-IP addresses via Amazon's Mechanical Turk, among them 301 participants in the unmodified condition and 162 participants each in the other three conditions. The unmodified condition had more

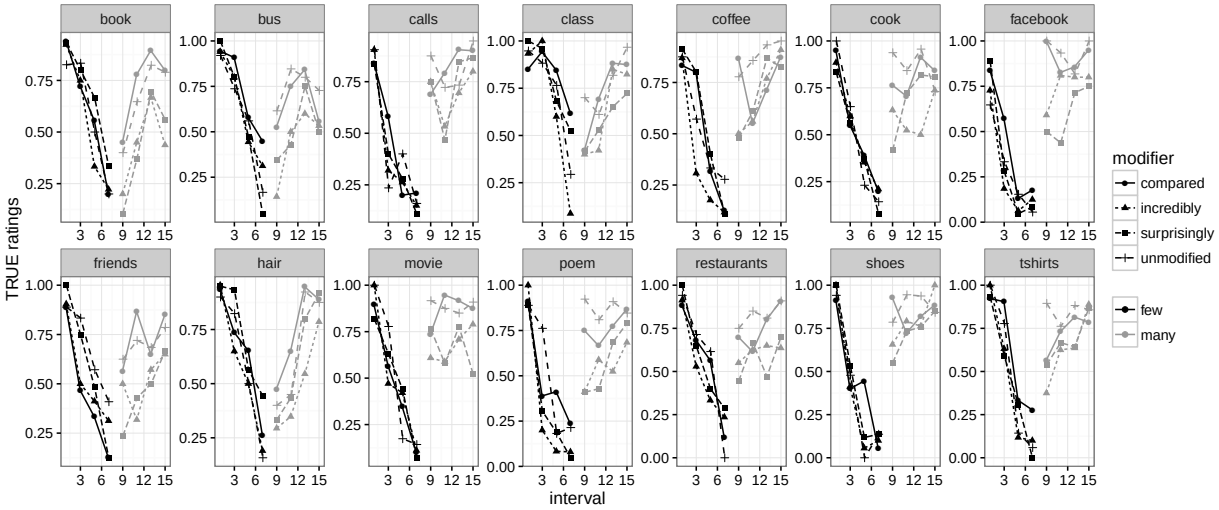


Figure 6.2: Proportion of TRUE answers per modifier condition

participants because it was part of the production experiment from Section 5.4.2 in which we presented 8 of 15 intervals for both *few* and *many*. For the analysis only data from those intervals presented in the other three conditions was used.

Results. Data was excluded of 25 participants who reported not to be native speakers of English or to not have understood the task. Figure 6.2 shows the proportion of TRUE answers.

For each of the quantity words *few* and *many* we specified a linear mixed effects regression model predicting the proportional acceptance of statements as in (104b). During a guided search through the model space, we started out with a model containing only the random effect ITEM and added fixed effects if this significantly increased the model's fit to the data (measured by AIC).

For *many*, the final model includes the fixed effects INTERVAL and MODIFIER and their interaction. Significantly more participants accepted the statements for higher intervals ($\beta = 0.02, SE = 0.007, p < 0.01$). The modification of *many* by *surprisingly* leads to a lower acceptance ($\beta = -0.59, SE = 0.12, p < 0.001$) than of sentences with unmodified *many*. This suggests that *surprisingly* intensifies the meaning of *many*. The same is the case for sentences with *incredibly*, which were also rated lower than unmodified *many* ($\beta = -0.53, SE = 0.12, p < 0.001$). There is no difference between sentences with a *compared to* phrase and unmodified *many* ($\beta = -0.17, SE = 0.12, p < 0.15$), which suggests that *many* receives a surprise reading in both cases. *Surprisingly* and *compared to* are rated significantly different ($\beta = -0.42, SE = 0.12, p < 0.001$), but there is no difference between *surprisingly* and *incredibly*. Furthermore, there is a significant interaction between INTERVAL and MODIFIER for *surprisingly* ($\beta = 0.03, SE = 0.01, p < 0.001$) and *incredibly* ($\beta = 0.02, SE = 0.01, p < 0.01$).

For *few*, the final model, obtained by the same procedure, includes the fixed effects INTERVAL and MODIFIER. The proportion of participants accepting the statement is significantly lower for higher numbers ($\beta = -0.12, SE = 0.004, p < 0.001$). Among the modifiers only *incredibly* is significantly different from bare *few* ($\beta = -0.05, SE = 0.02, p < 0.05$); for *surprisingly* and *compared to* this is not the case. No significant difference between *surprisingly* and *compared to* is found, but *incredibly* is rated significantly lower than *surprisingly* ($\beta = -0.05, SE = 0.02, p < 0.05$).

These results are expected under the "salient surprise reading" hypothesis. While *surprisingly* seems to behave like an intensifier for *many*, it seems to redundantly mark surprise for *few*.

6.4 Computational Model and Model Evaluation

The regression models reported above include a random effect for items, but do not constrain these to reflect prior expectations. Moreover, regression models do not predict judgments as a function of thresholds on expectations. It is therefore insightful to complement regression modeling with an explicit theory-driven model of a possible data-generating process. We use the computational model of Section 5.3 for this purpose. The model takes empirically measured prior expectations as input and treats $\theta_{\text{[i]few}}$ and $\theta_{\text{[i]many}}$ for each modifier condition i (unmodified, *surprisingly, incredibly, compared to*) as latent parameters, whose values will be estimated from experimental data. The model specifies a likelihood function $P(\text{Observation} \mid \theta_{\text{[i]many}}, \theta_{\text{[i]few}})$ which assigns to values of latent parameters a probability of seeing a particular experimental observation. Bayesian inference is one way to infer plausible threshold values, given the likelihood function and a prior distribution on parameter values:

$$P(\theta_{\text{[i]many}}, \theta_{\text{[i]few}} \mid \text{O}) \propto P(\theta_{\text{[i]many}}, \theta_{\text{[i]few}}) \cdot P(\text{O} \mid \theta_{\text{[i]many}}, \theta_{\text{[i]few}}) \tag{6.1}$$

Our goal, then, is to see for each modifier which pairs of threshold values $\theta_{\text{[i]many}}$ and $\theta_{\text{[i]few}}$ are likely given the data. We estimate the a posteriori credible threshold values and compare how similar they are across conditions. We focus on *many* in the exposition, but the case for *few* is parallel.

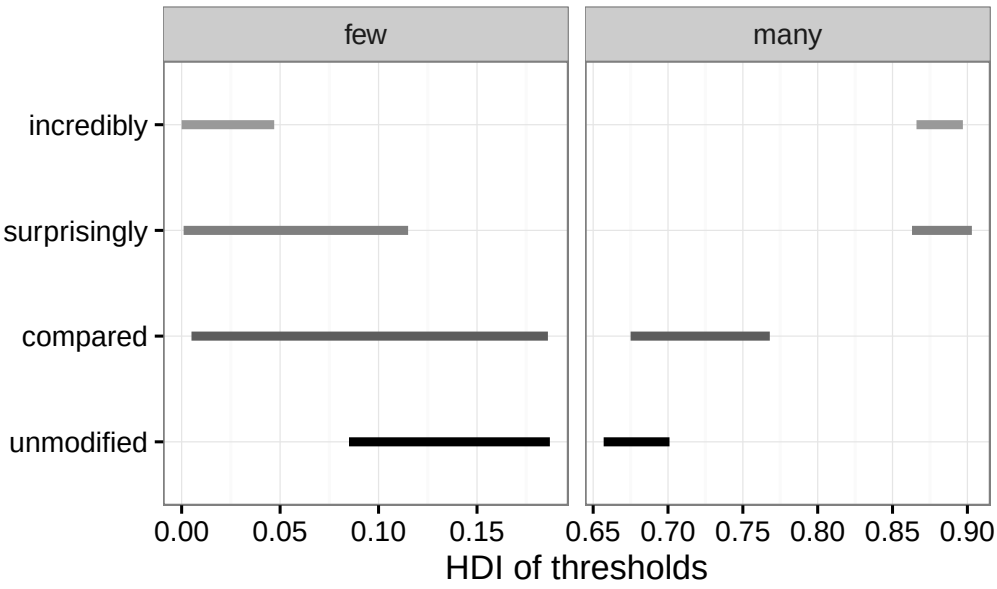
(77) **CFK Semantics**

a. $[\![\text{Few As are B}]\!] = 1$ iff $|\,\text{A} \cap \text{B}\,| \leq x_{\max}$

$$\text{where } x_{\max} = \max\{n \in \mathbb{N} \mid P_E(|\,\text{A} \cap \text{B}\,| \leq n) < \theta_{\text{few}}\}$$

b. $[\![\text{Many As are B}]\!] = 1$ iff $|\,\text{A} \cap \text{B}\,| \geq x_{\min}$

$$\text{where } x_{\min} = \min\{n \in \mathbb{N} \mid P_E(|\,\text{A} \cap \text{B}\,| \leq n) > \theta_{\text{many}}\}$$



	few	*many*
incredibly	[0.000 , 0.047]	[0.866 , 0.897]
surprisingly	[0.001 , 0.115]	[0.863 , 0.903]
compared to	[0.005 , 0.186]	[0.675 , 0.768]
unmodified	[0.085 , 0.187]	[0.657 , 0.701]

Figure 6.3: Estimated 95% credible intervals for $\theta_{\text{few,i}}$ & $\theta_{\text{many,i}}$

Straightforwardly, the CFK semantics repeated from (77) translates into a probabilistic rule $P(\text{“[modifier } i\text{] many”} \mid n, P_E \; ; \; \theta_{\text{[i]many}}) = \delta_{n \geq x_{\text{min,i}}}$, where $x_{\text{min,i}}$ is derived from P_E, as in (77), based on $\theta_{\text{[i]many}}$. This is a degenerate probabilistic rule because it maps the applicability of “many” to 0 and 1 only. To allow for noise, we look at a parameterized, smoothed-out version.

$$P(\text{“}[i] \text{ many”} \mid n, P_E; \theta_{\text{[i]many}}, \sigma_j) = \sum_{k=0}^{n} \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} \mathcal{N}(y; x_{\text{min,i}}, \sigma_j) \mathrm{d}y \tag{6.2}$$

The steepness of the curve is regulated by another free model parameter σ_j. $\mathcal{N}(y; x_{\text{min,i}}, \sigma_j)$ is then the probability density of y under a normal distribution with mean $x_{\text{min,i}}$ and standard deviation σ_j. This rule predicts noisy acceptability ratings under a surprise-based semantics where the amount of noise is controlled by σ_j, see Figure 5.4a from the previous chapter. Noise can be caused by uncertainty about the exact shape of P_E and the amount of uncertainty differs across contexts. This is why we allow an individual value of σ_j for each context j. Furthermore, we assume that the parameter values $\theta_{\text{[i]many}}$, $\theta_{\text{[i]few}}$ and σ_j are independent of each other and that they have uniform priors over an interval that is large enough to accommodate a range of plausible values without weighting them.

	hypothesis		
	intensifier	**marker of surprise**	**salient surprise reading**
predictions	*many* $\leq$ *surprisingly many* *few* $\geq$ *surprisingly few* *surprisingly* = *incredibly*	*many* = *surprisingly many* *few* = *surprisingly few* *surprisingly* = *compared to*	*many* = *compared to... many* *few* = *compared to... few*
results	*few*: × *many*: ✓	*few*: ✓ *many*: ×	*few*: ✓ *many*: ✓

Table 6.2: Results for sentences expressing surprise readings

$$P(\theta_{\text{[i]many}}, \theta_{\text{[i]few}}, \sigma_j) = \text{Uniform}_{[0;1]}(\theta_{\text{[i]many}}) \cdot \text{Uniform}_{[0;1]}(\theta_{\text{[i]few}}) \cdot \text{Uniform}_{[0;10]}(\sigma_j) \tag{6.3}$$

To approximate the joint posterior distribution, we used MCMC sampling, as implemented in JAGS (Plummer, 2003) and briefly introduced in Chapter 4. We collected 10,000 samples from 2 MCMC chains after a burn-in of 10,000. This ensured convergence, as measured by $\hat{R}$ (Gelman and Rubin, 1992). Figure 6.3 shows the estimated 95% credible intervals for the marginalized posteriors over thresholds per modifier. Where intervals (clearly) do not overlap, there is reason to believe that thresholds differ. For example, $\theta_{\text{surpr.many}} \in [0.863, 0.903]$ tells us that *surprisingly many* describes cardinalities which are higher than at least 86% of the cumulative density mass of P_E. This threshold is higher than bare *many*'s, $\theta_{\text{many}} \in [0.657, 0.701]$. Taken together, the model predicts that *surprisingly many* is restricted to describe higher cardinalities than unmodified *many*.

6.5 Interim Summary

Table 6.2 summarizes the results from regression and theory-driven modeling. The data supports the "salient surprise reading" hypothesis assumed by Fernando and Kamp (1996) and suggests that an expectation-based reading is the canonical interpretation of cardinal *few* and *many* in our test sentences. There is no difference between unmodified sentences and sentences in which expectations are made salient by a *compared to*-phrase.

For *surprisingly*, the picture is less clear. Sentences with *many* provide support for the intensifier hypothesis. Speakers prefer it for higher cardinalities than those which render unmodified *many* or sentences with a *compared to* construction true. Furthermore, we do not find a difference to *incredibly*. When combined with *few*, however, *surprisingly* does not appear to be an intensifier. Sentences with *few*, *surprisingly few* and *compared to* are rated equally, speaking in favor of a "marker of surprise" hypothesis. For the comparison between *surprisingly* and *incredibly*, we get conflicting results from the regression and the theory-driven model. The regression

analysis finds that *incredibly few* is rated lower than *surprisingly few*, but the computational model identifies an overlap in the estimated credible intervals. However, we want to once more stress that we are here comparing conclusions based on models which are decidedly different. Whereas the computational model is theory-driven and includes experimentally measured prior expectations, the regression model only looks at numerical differences in the average ratings. If we were forced to make a decision, we would believe in the computational model.

Keeping in mind that *few* only applies to small cardinalities, the lack of a difference could also be due to a floor effect. In the judgment task, participants were presented with intervals instead of single numbers and moreover, we only presented four out of 15 intervals per quantity word. This setup might not be adequate to reveal a potential difference between *surprisingly few* and *few*. Due to the lower bounded scale, the difference for *few* is probably more subtle than the difference between *surprisingly many* and *many*. This is where future research should tie in. *Few* should be presented in contexts like **book** or **facebook** (see 6.A), in which large cardinalities are plausible and *few* can operate away from 0. Additionally, the presented intervals should be more fine-grained. We opt for a third option and follow up on the presented judgment task with an interpretation experiment. To investigate a possible floor effect which might conceal an intensifying effect of *surprisingly* on *few*, we present the same items in a free choice interpretation task. Such a task gives participants much more freedom in their choice. If the lack of a difference between *surprisingly few* and *few* was due to a floor effect, we hope to be able to reveal it with this task type. The follow-up experiment is presented in the next section.

6.6 Follow-up study: Refining Modified *few*

The production task discussed in the previous sections produced puzzling results. The adverb *surprisingly* seems to function like an intensifier in combination with *many*, but not with *few*. This is very surprising under the assumption that *surprisingly* contributes the same meaning in both cases. To learn whether the lack of a difference between *few* and *surprisingly few* was really due to a floor effect, we run an interpretation task as a follow-up experiment. We hope to gain more insight into the interaction between *surprisingly* and *few* by allowing participants to choose a single number, instead of an interval as we did in the judgment task presented in Section 6.3.2. We expect to see more fine-grained results in this free choice interpretation task which could identify a difference between *surprisingly few* and *few* that could previously not be revealed due to a floor effect.

6.6.1 Interpretation Task

Design. In a free choice task we measured participants' interpretation of modified and unmodified *few* and *many*. Participants saw the same 14 contexts as in the production study, see Appendix 6.A. Each item was paired with either *few* or *many* and one of the four modifier conditions (unmodified, *compared to, surprisingly, incredibly*). Participants were asked to give their interpretation of the quantifier term by adjusting a slider on a scale. The sliders were presented on a horizontal scale on the screen. The slider's label on the lower end was 0, the label on the upper end was the highest of the 15 intervals, which were already used in the prior elicitation task (for example '28 or more'). The interim interval boundaries were not marked on the scale, however and participants could select any number, not just intervals.

(121) **Interpretation study example**

a. CONTEXT (unmodified, *surprisingly, incredibly*):
Andy is a man from the US who drank [- | surprisingly | incredibly] [few | many] cups of coffee last week.
CONTEXT (*compared to*):
Andy is a man from the US. Compared to what you would expect for a man from the US, Andy drank [few | many] cups of coffee last week.

b. QUESTION: How many cups of coffee do you think Andy drank last week? (0 - 28 or more)

Participants. We recruited 170 participants with US-IP addresses via Amazon's Mechanical Turk.

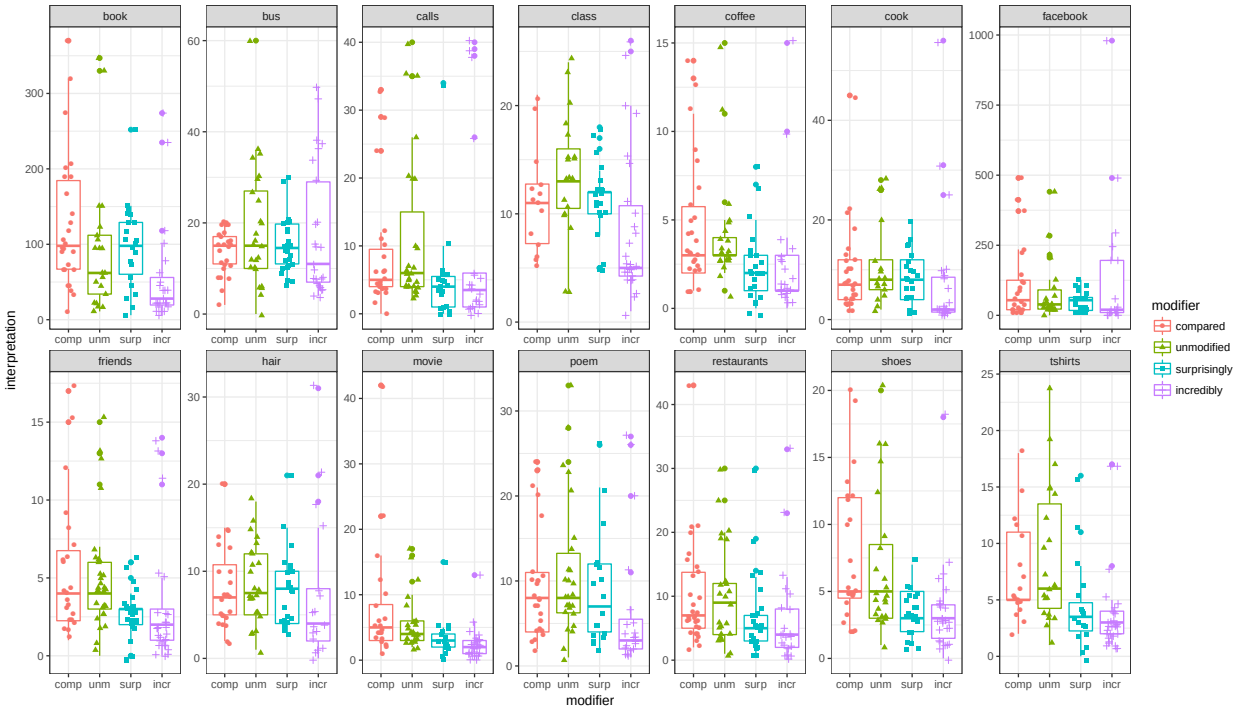
Materials & Procedure. After reading a short introduction, each subject saw all of the 14 contexts from Appendix 6.A one after another in a random order. Participants saw each context in the same modifier condition. For each context, *few* or *many* were assigned randomly. Participants had to adjust the slider or at least click on it before being able to proceed to the next item.

Results. The data of four participants were excluded because they reported not to be native speakers of English. The interpretations per condition as well as the median answer is plotted in Figure 6.4.

For each of the quantity words *few* and *many* we specified a linear mixed effects regression model predicting interpretations. The model contained the random effect ITEM and the fixed effect MODIFIER. Note that there were no other factors.

The modification of *many* by *surprisingly* leads to significantly higher interpretations ($\beta = 28.37, SE = 6.66, p < 0.001$) than of sentences with unmodified *many*.

(a) Interpretion data of *few*



(b) Interpretation data of *many*

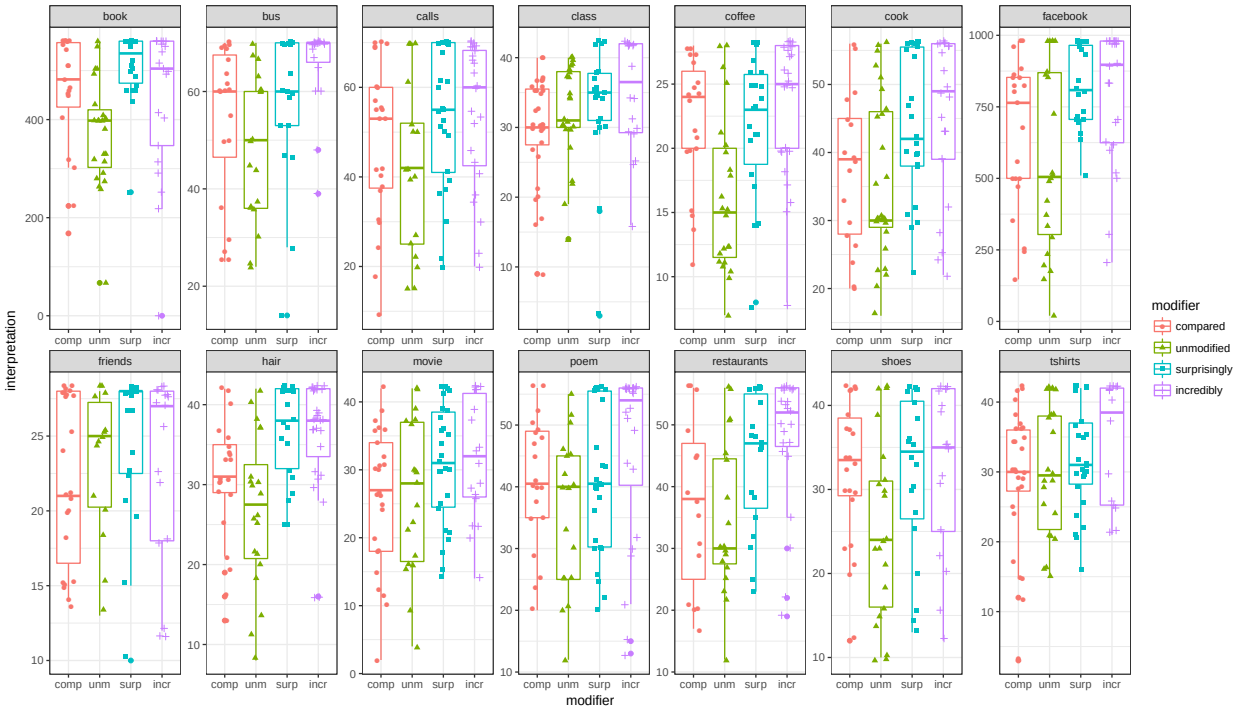


Figure 6.4: Interpretation data of modified *few* and *many*

This supports the production study's finding that *surprisingly* intensifies the meaning of *many*. The same is the case for sentences with *incredibly*, which also received higher interpretations than unmodified *many* ($\beta = 29.66, SE = 6.93, p < 0.001$). There is no significant difference between sentences with a *compared to* phrase and unmodified *many* ($\beta = 13.12, SE = 6.79, p = 0.06$). We interpret this result as suggesting that the most salient reading of *many* is the surprise reading. *Surprisingly* and *compared to* are rated significantly different ($\beta = -15.25, SE = 6.42, p < 0.05$), but there is no difference between *surprisingly* and *incredibly*.

For *few*, no modifier triggered a significantly different interpretation from bare *few* ($\beta = -0.05, SE = 0.02, p < 0.05$). Furthermore, no significant difference between *surprisingly* and *compared to* nor between *surprisingly* and *incredibly* is found.

These results are expected under the "salient surprise reading" hypothesis. While *surprisingly* seems to behave like an intensifier for *many*, for *few* it seems to redundantly mark surprise.

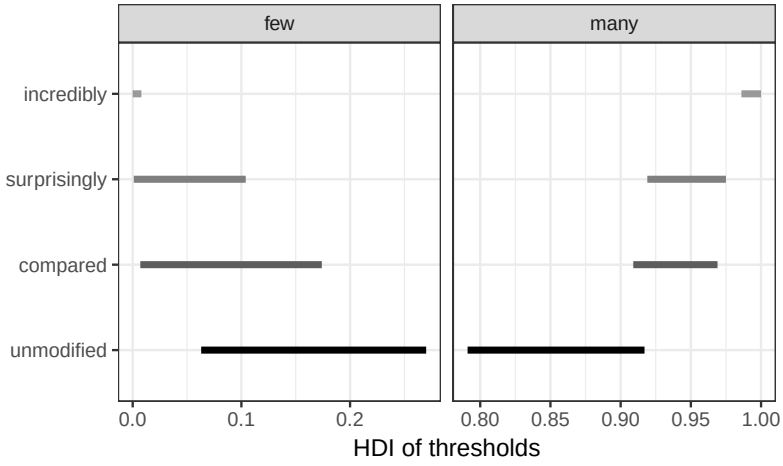
6.6.2 Computational Model

The interpretation data was also analyzed with a theory-driven model of listener behavior. The interpretation rule developed in Section 5.3 specifies a likelihood function which assigns to each interval the probability of being chosen as the interpretation of "[modifier] *few/many*".

$$P_L(n \mid \text{“}[i] \text{ many”}, P_E \; ; \; \theta_{\text{[i]}many}, \sigma_j) \propto P_E(n) \cdot P_S(\text{“}[i]\text{many”} \mid n, P_E \; ; \; \theta_{\text{[i]many}}, \sigma_j) \tag{6.4}$$

Via Bayesian inference we again infer a pair of threshold values $\theta_{\text{[i]many}}$ and $\theta_{\text{[i]few}}$ for each modifier i, which are most likely to have generated the observed data. Note that for this purpose participants' answers were fused into the respective intervals to be able to relate them to the prior data from Section 6.3. The 95% highest density intervals of the posterior distribution of $\theta_{\text{[i]many}}$ and $\theta_{\text{[i]few}}$ are displayed in Figure 6.5. Unfortunately, these results do not allow us to see the picture more clearly. For *many*, the findings of the production experiment could be replicated. We do not find an overlap between $\theta_{\text{surpr.many}}$ and θ_{many} (even though the two HDIs come very close), suggesting that *surprisingly* intensifies the meaning of *many*, like *incredibly* does. For *few*, however, the HDIs of *surprisingly few* and *few* overlap, suggesting that *surprisingly* does not raise *few*'s threshold.

To get a more nuanced quantitative measure for the likelihood of θ_{surp} being more extreme than $\theta_{\text{unmodified}}$, we used another version of the model which jointly infers the threshold values on both production and interpretation data in these two modifier conditions.



	few	*many*
incredibly	[0.000 , 0.008]	[0.986 , 1.000]
surprisingly	[0.001 , 0.104]	[0.919 , 0.975]
compared to	[0.007 , 0.174]	[0.909 , 0.969]
unmodified	[0.063 , 0.270]	[0.791 , 0.917]

Figure 6.5: Estimated 95% credible intervals for $\theta_{\text{few,i}}$ & $\theta_{\text{many,i}}$ for interpretation data

$$P(\theta_{\text{many}}, \theta_{\text{few}}, \theta_{\text{surp.many}}, \theta_{\text{surp.few}} \mid \text{O}) \propto \tag{6.5}$$
$$P(\theta_{\text{many}}, \theta_{\text{few}}, \theta_{\text{surp.many}}, \theta_{\text{surp.few}}) \cdot P(\text{O} \mid \theta_{\text{many}}, \theta_{\text{few}}, \theta_{\text{surp.many}}, \theta_{\text{surp.few}})$$

For each sample of θ_{many} and $\theta_{\text{surp.many}}$ and of θ_{few} and $\theta_{\text{surp.few}}$ we calculate their difference. From the posterior distribution of $\theta_{\text{surp.many}} - \theta_{\text{many}}$ and $\theta_{\text{surp.few}} - \theta_{\text{few}}$ we then calculate their highest density interval. If the HDI of the differences does not contain 0, the model supports the hypothesis that the threshold values are really different from each other, resulting in an intensifier interpretation of *surprisingly*.

For *many*, the model once more supports the intensifier hypothesis for *surprisingly*. The HDI of the difference between the thresholds of *surprisingly many* and *many*, $\theta_{\text{surp.many}} - \theta_{\text{many}}$, does not contain 0. It is estimated to be [0.101, 0.145]. This suggests that the value of $\theta_{\text{surp.many}}$ is credibly higher than the value of θ_{many}. For *few*, the HDI of the difference between the thresholds of *surprisingly few* and *few*, $\theta_{\text{surp.few}} - \theta_{\text{few}}$, is and [-0.074, 0.022] and does contain 0. Strictly speaking, the model does not constitute clear evidence for an intensifying effect of *surprisingly* in combination with *few*. Nevertheless, we find that 80% of the difference's values which carry the highest density mass are below 0. In other words, the posterior probability is 80% that $\theta_{\text{surp.few}}$ is lower than θ_{few} after all. Even though this result cannot be taken as strong evidence for the intensifier hypothesis, it encourages us collect more data on *surprisingly few*, hoping to arrive at a definite verdict.

6.7 Discussion

If the lack of an intensifying effect of *surprisingly* on *few* was due to a floor effect, the free choice task could not resolve it even though it allowed for more freedom by not restricting the choices to intervals. In contrast, even the significant difference between *incredibly few* and *few* vanishes. Apart from this, the interpretation task's results replicate the production task's findings. In both production and interpretation, there is no significant difference between sentences with bare *few* and *many* and sentences in which expectations are made overt with a *compared to*-phrase as in (121a). We conclude that the surprise reading is the most salient reading of *few* and *many* in our test sentences. This result constitutes experimental support for an expectation-based semantics of *few* and *many*.

For *surprisingly*, the interpretation task does not provide new insights. There is no difference between *surprisingly few* and *few*, but in combination with *many* the adverb still has an intensifying effect. As mentioned above, these results are unexpected under the assumption of an unambiguous *surprisingly* that contributes the same meaning in combination with both *few* and *many*.

For now, we can only speculate about possible reasons and will have to back off from further promoting the lexical entry of an intensifier version of *surprisingly*, suggested in (118). In contrast to von Stechow's (2006) semantics for *very*, *surprisingly* does not seem to maintain the symmetrical inclusion presupposition from (117). So far the experimental data suggests that *surprisingly* raises the boundary for *many*, but for *few* the evidence is not clear enough to draw the same conclusion.

A first explanation of the missing intensifying effect on *few* could be a decompositional analysis of *few*, as introduced in Section 2.2.4. If *few* decomposes into *many* and a scopally mobile, negative operator *little*, the negation could prevent *surprisingly*'s intensifying effect. What exactly this might look like in a compositional semantic analysis is not clear though. The only meaning component of *surprisingly* that would have to be negated is its intensifying effect, not the entire sentence. For example, the sentence "Chris saw surprisingly few movies last year" still expresses that the number is small, even though the number is maybe not *very small*. It is not obvious how *little* could only negate the intensifying effect of *surprisingly* while not affecting anything else. This technical problem may be solvable, but for now it remains a puzzle what a compositional analysis would have to look like.

Nevertheless, the idea that the semantics of *few* blocks an intensifying effect of *surprisingly* is supported by several observations in which *few* behaves differently from *many*. A series of experiments by Moxey and Sanford (1987) and Sanford et al. (1994) shows that *few* licenses a complement set reference whereas *many* makes reference to the set whose cardinality it describes. For example, in a continuation

task repeated from Section 3.2, participants associate the pronoun with a different set for *few* than for *many*.

(96) Few of the football fans were at the match. They...

Participants would associate *they* not with the entities quantified over by *few*, the football fans present at the match. Instead, the complement set is activated: the fans who are *not* present. For example they would continue the sentence with "watched the match at home instead". For *many* and other positive quantifiers, like *a few*, this is not the case (cf. Moxey and Sanford, 1987). Transferring these results to the present case in a sentence like (122), *surprisingly few* tends to shift the attention to, say,the students who had not passed whereas *surprisingly many* does not, as exemplified by a continuation similar to Sanford et al.'s (1994) items.

(122) a. Surprisingly many students passed the test. They... had a big party to celebrate their success.

b. Surprisingly few students passed the test. They... had underestimated the test's difficulty.

Surprisingly many's reference set is not only part of the utterance's asserted meaning, it is additionally highlighted by the cardinality word. For this reason, *surprisingly many* might be perceived as a stronger description than *surprisingly few*. For now, these speculations can only be put on the agenda for future research, however.

Furthermore, many uses of bare *few* are perceived as sounding marked. For example, when presented with alternatives, many speakers prefer *not many, only few* or *a few* to bare *few*. See the example below:

(123) a. Andy drank few cups of coffee.

b. Andy drank only few cups of coffee.

c. Andy drank not many cups of coffee.

d. Andy drank a few cups of coffee.

This was the feedback that we often got from participants, especially for the items in Schöller and Franke (2015).

All of these features of *few* confirm the difficult undertaking of providing a semantics for *few*. As already pointed out in Section 2.2.4, the semantics of *few* is quite elusive and will have to be further investigated. The open issues in the semantics of *few* will have to be answered on par with the more general question of whether antonyms are really decomposed in the syntax or not, see Section 2.2.4.

When looking at the computational model's inferred threshold values, we see that they are in general more "extreme" in interpretation tasks than in production tasks. For *few*, $\theta_{\text{[i]few}}$ are slightly lower in the interpretation task, and for *many*, $\theta_{\text{[i]many}}$ are

higher in the interpretation task. The same could be observed in the experiments on cardinal *few* and *many* in the previous chapter. We believe that the reason for the difference between the tasks is the range of choices participants have. The judgment task, on the one hand is a binary rating task in which participants only have two options; either the quantity word is a felicitous description of the presented cardinality or it is not. On the other hand, an interpretation task leaves participants with a much broader range of choices. Even though the quantity word might also have been used to describe lower cardinalities, participants seem to play it safe and choose a higher number as *many*'s interpretation (or a lower number for *few*). For this reason we do not yet want to reject the floor effect hypothesis for *surprisingly few*. We cannot exclude that we just have not yet found the right method to test it. This suspicion is also fueled by the model version which estimates the difference between $\theta_{\text{surp.few}}$ and θ_{few} and assigns a probability of 80% to a potential difference of the threshold values.

Apart from the puzzle of whether *surprisingly* is an intensifier, another interesting area of future research is to investigate which kind of knowledge is necessary to be able to form the expectations which are required to license *surprisingly*. A surprising observation is that *surprisingly* is felicitous in (124), but not in (125).

(124) Grandma walked into a bar. Even though she didn't have many drinks, she had surprisingly many.

(125) ?? A random Joe walked into a bar. Even though he didn't have many drinks, he had surprisingly many.

In (124), we both communicate expectations about people in general ($\rightsquigarrow$ *not many*) and about the well-known individual Grandma ($\rightsquigarrow$ *surprisingly many*). But in (125) there is not sufficient information to form special expectations about a "random Joe". Why a stereotypical Joe renders *surprisingly* infelicitous remains to be elucidated.

6.A Experimental Material

1. **book** — A friend's favorite book has been published only recently and has [0-40, 81-120, 161-200, 241-280, 321-360, 401-440,481-520, 560 or more] pages. — [For | Compared to what you would expect for] a recently published book, the book has [- | surprisingly | incredibly] [few | many] pages.

2. **bus** — Vehicle No. 102 is a school bus which has seats for [0-4, 10-14, 20-24, 30-34, 40-44, 50-54, 60-64, 70 or more] passengers. — [For | Compared to what you would expect for] a school bus, Vehicle No. 102 has seats for [- | surprisingly | incredibly] [few | many] passengers.

3. **calls** — Lisa is a woman from the US who made [0-4, 10-14, 20-24, 30-34, 40-44, 50-54, 60-64, 70 or more] phone calls last week. — [For | Compared to what you would expect for] a woman from the US, Lisa made [- | surprisingly | incredibly] [few | many] phone calls last week.
4. **class** — Erin is a first grade student in primary school. There are [0-2, 6-8, 12-14, 18-20, 24-26, 30-32, 36-38, 42 or more] children in Erin's class. — [For | Compared to what you would expect for] a primary school class, there are [- | surprisingly | incredibly] [few | many] children in Erin's class.
5. **coffee** — Andy is man from the US who drank [0-1, 4-5, 8-9, 12-13, 16-17, 20-21, 24-25, 28 or more] cups of coffee last week. — [For | Compared to what you would expect for] a man from the US, Andy drank [- | surprisingly | incredibly] [few | many] cups of coffee last week.
6. **cook** — Tony is a man from the US who cooked himself [0-3, 8-11, 16-19, 24-27, 32-35, 40-43, 48-51, 56 or more] meals at home last month. — [For | Compared to what you would expect for] a man from the US, Tony cooked himself [- | surprisingly | incredibly] [few | many] meals at home last month.
7. **facebook** — Judith is a woman from the US who has [0-69, 140-209, 280-349, 420-489, 560-629, 700-769, 840-909, 980 or more] Facebook friends. — [For | Compared to what you would expect for] a woman from the US, Judith has [- | surprisingly | incredibly] [few | many] Facebook friends.
8. **friends** — Lelia is a woman from the US who has [0-1, 4-5, 8-9, 12-13, 16-17, 20-21, 24-25, 28 or more] friends. — [For | Compared to what you would expect for] a woman from the US, Lelia has [- | surprisingly | incredibly] [few | many] friends.
9. **hair** — Betty is a woman from the US who washed her hair [0-2, 6-8, 12-14, 18-20, 24-26, 30-32, 36-38, 42 or more] times last month. — [For | Compared to what you would expect for] a woman from the US, Betty washed her hair [- | surprisingly | incredibly] [few | many] times last month.
10. **movie** — Chris is a man from the US who saw [0-2, 6-8, 12-14, 18-20, 24-26, 30-32, 36-38, 42 or more] movies last year. — [For | Compared to what you would expect for] a man from the US, Chris saw [- | surprisingly | incredibly] [few | many] movies last year.
11. **poem** — A friend wants to read you her favorite poem which has [0-3, 8-11, 16-19, 24-27, 32-35, 40-43, 48-51, 56 or more] lines. — [For | Compared to what you would expect for] a poem, the poem has [- | surprisingly | incredibly] [few | many] lines.

12. **restaurants** — Sarah is a woman from the US who went to [0-3, 8-11, 16-19, 24-27, 32-35, 40-43, 48-51, 56 or more] restaurants last year. — [For | Compared to what you would expect for] a woman from the US, Sarah went to [- | surprisingly | incredibly] [few | many] restaurants last year.

13. **shoes** — Melanie is a woman from the US who owns [0-2, 6-8, 12-14, 18-20, 24-26, 30-32, 36-38, 42 or more] pairs of shoes. — [For | Compared to what you would expect for] a woman from the US, Melanie owns [- | surprisingly | incredibly] [few | many] pairs of shoes.

14. **tshirts** — Liam is a man from the US who has [0-2, 6-8, 12-14, 18-20, 24-26, 30-32, 36-38, 42 or more] T-shirts. — [For | Compared to what you would expect for] a man from the US, Liam has [- | surprisingly | incredibly] [few | many] T-shirts.

Chapter 7

The Proportional Reading of *few* and *many*

The previous chapters focused on the experimental investigation of the cardinal surprise reading of *few* and *many*. We found evidence for Fernando and Kamp's (1996) theory that the quantity words comprise stable core meanings θ_{few} and θ_{many}, which operate on prior expectations of the context. To determine which cardinalities count as *few* or *many* in the respective context, the cumulative density mass of said prior expectations is cut off at a fixed percentage θ_{few} or θ_{many}, deriving thresholds on the cardinality scale. In this chapter we turn to another very prominent reading of *few* and *many*, the proportional reading. We investigate whether the stable core meaning hypothesis can be transferred to this reading. To do so, we briefly summarize the characteristics of the proportional reading and test whether proportional *few* and *many* are equally context-dependent, by manipulating prior expectations in an interpretation task in Section 7.1. Minimal pairs of sentences are compared which introduce contrasting properties. For example, the number of muffins eaten by hungry person vs. the number of muffins eaten by a person feeling full. Complementing experiments in real-world contexts are presented in Section 7.2, eliciting the production of proportional *few* and *many* and prior expectations of the presented contexts. By real-world contexts we refer to contexts which deal with every-day situations and proportions, like the proportion of all muffins on the table a person ate or the ratio of tennis matches a player lost. Whether world knowledge is all that matters or whether the sheer size of the described proportion has to be taken into account as well is investigated by testing *few* and *many* in an abstract urn scenario in Section 7.3.

Experimental findings suggest that the proportional reading can both express that a proportion is (i) numerically high and (ii) surprisingly high. Consequently, we assume that the contextual contribution is two-fold: the first is an uninformed, uniform belief about proportions and the second is an informed prior expectation

about likely proportions based on world knowledge. For this reason, the computational model from the previous Chapter 5.2 does not manage to predict the data in a satisfying way. Nevertheless, we assume that *few* and *many* have a stable core meaning. We propose a linear combination model in Section 7.5 which incorporates that the amount of world knowledge employed depends on its saliency in the context. We are interested in whether the estimated threshold values θ_{many} and θ_{few} are the same as those inferred for cardinal *few* and *many*. This could provide further evidence for a potential lexical ambiguity between the proportional and the cardinal reading, as discussed in Section 2.1.1. The model is evaluated in Section 7.6 before concluding with a discussion in Section 7.7.

7.1 The Proportional Reading in Context

In Section 2.1.1, the two most prominent readings of *few* and *many* were introduced: the cardinal and the proportional reading. Whereas the cardinal reading describes the *cardinality* of a set of objects, the proportional reading describes the *proportion* of a set relative to its superset. This section briefly calls to mind the key semantic properties and presents an interpretation study from Schöller and Franke (2016). The study investigates whether the size of the proportions which count as *few* or *many* is fixed or whether it varies with the context. If proportional *few* and *many* turned out to express that a proportion is lower or higher than expected, a natural next step would be to transfer the CFK semantics from the previous chapters to this reading.

7.1.1 Proportional *few* and *many*

Since proportional *few* and *many* describe a proportion, the existence of an upper bound on the quantity word's scale is required for this reading to arise. This upper bound can be implicit or it can be spelled out overtly with a partitive construction, as exemplified below:

(126) a. Many of the 1,000 women testing the new contraceptive became pregnant.

b. Many Germans love bread.

(127) a. Few of Cornwall's residents speak more than four languages.

b. Few of the 28 students passed the exam.

According to Partee (1989), sentence (126b) is true if a large proportion of the German citizens like bread; at least k, where "[w]e may think of k either as a

fraction between 0 and 1 or as a percentage" (Partee, 1989, 2). Truth-conditions of "Few/Many A are B" under a proportional reading are repeated from above.

(11) **Proportional reading**

a. *Few*: $|A \cap B| : |A| \leq k_{\max}$

b. *Many*: $|A \cap B| : |A| \geq k_{\min}$

This simple semantics is intuitively appealing (see Chapter 2 for an extensive discussion of the semantics of proportional *few* and *many*), but it leaves several questions unanswered. How to define the size of the fraction $k_{\min/\max}$ which determines the use of *few* and *many* is left unspecified. Furthermore, (11) does not tell us what the influence of the context on thresholds $k_{\min/\max}$ is, or whether it is assumed to be a fixed proportion. In contrast to the cardinal reading whose scale is not bounded, it would theoretically be possible to hardwire a value for a fixed proportion $k_{\max}$ and $k_{\min}$ in the semantics, but already the few examples in (126) suggest to dismiss this idea as being implausible. The proportion of women needed to make (126a) true is much lower than the proportion of Germans that are needed to make (126b) true. Similarly, the proportion of Cornwall residents for (127a) to be true is much lower than the proportion of students in (127b). With an interpretation experiment, we confirm the suspicion that proportional *many* is dependent on expectations of the context and that its use cannot be captured by a fixed threshold on proportions.

7.1.2 Interpretation Experiment: the Context-Dependence of Proportional *many*

The objective of this interpretation experiment of sentences with proportional *many* is to verify the hypothesis that also the proportional reading is influenced by expectations of the context. Furthermore, we aim to find out whether it makes a difference to use *many* in the plain form ("many") or in the partitive construction ("many of the") and whether the number of objects in the context influences the interpretation.

Design A sentence introduced the context and the amount of the objects under discussion, see (128) and (129) for sample items. Each item was randomly paired with one of two numbers of the form [3/4N | N] (labeled as low or high NUMBER condition below). The number in the context sentence was described by a sentence containing either "many" or "many of the". The sentence was randomly chosen from two PROBABILITY conditions [HP | LP], high probability or low probability. The two conditions differed in the comparison class set in the relative clause. We set

the comparison classes in a way that we expect higher answers in high probability contexts. We made sure that the two relative clauses per item are a minimal pair. Most of them differed only in contrasting adjectives. Participants were asked to guess the number that they think "many" or "many of the" refers to.

(128) CONTEXT: There were [9|12] muffins on the kitchen table in Eds flat.

HP: Ed, who arrived feeling hungry, ate [many|many of the] muffins.

LP: Ed, who arrived feeling full, ate [many|many of the] muffins.

How [many|many of the] muffins do you think Ed ate?

(129) CONTEXT: When moving flat, Martha packed [15|20] big boxes.

HP: Martha, who is a strong woman, carried [many|many of the] boxes herself.

LP: Martha, who is a weak woman, carried [many|many of the] boxes herself.

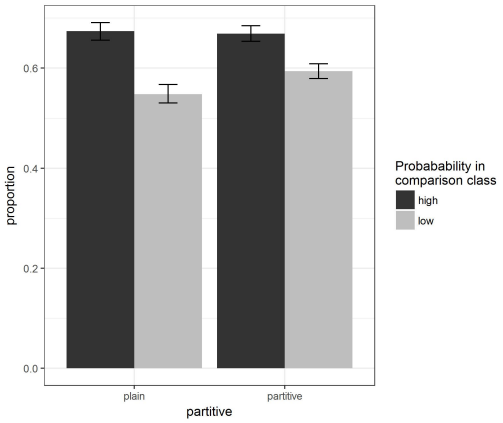
QUESTION: How [many|many of the] boxes do you think Martha carried?

Participants The experiment was conducted via Amazon's Mechanical Turk and elicited data from 160 participants. Participants who are not self-reported native speakers of English were excluded. No participant participated more than once.

Methods & Material At the beginning of the experiment, each participant was randomly assigned to the PARTITIVE condition [-|+ partitive]. [-partitive] means, that every sentence was presented with plain "many", whereas in the [+partitive] condition "many of the" was used. Every participant saw all 16 items from Appendix 7.A.1 in a random order. The PROBABILITY and the *number* condition were assigned randomly. Participants could only proceed to the next item after having entered a number into a text box. In this experiment, only *many* is presented. The following experiments in this chapter, however, test *few* as well.

Results Figure 7.1 displays the mean proportions of N that were given as the interpretation of *many*. A first visual inspection of the data suggests a difference between LP and HP condition, which supports the hypothesis that prior expectations influence interpretation of proportional *many*. Furthermore, the difference between low and high probability seems to be greater in the plain condition than when the partitive is used. Whether these differences are statistically significant will be analyzed in the following.

At first we specified a mixed linear effects regression model predicting proportional interpretations for *many* which included the factors PROBABILITY (high or low probability sentence), NUMBER (number in context), PARTITIVE (plain or partitive "many") and a 3-way interaction as well as three 2-way interactions of these



condition	mean rating
plain & low	0.55
plain & high	0.68
partitive & low	0.58
partitive & high	0.67

Figure 7.1: Mean ratings for the interpretation of *many* in proportions of the total number of objects N

three predictors. In terms of random effects, the initial model had the maximal random effects structure as justified by the design (Bates et al., 2013). We removed redundant random effects by running a principle component analysis and arrived at a parsimonious model (Bates et al., 2015).

The final model included both varying intercepts for PARTICIPANT and ITEM, as well as a random PARTICIPANT slope for PROBABILITY. In terms of the fixed effects, only PROBABILITY was included as a main effect. We found that participants gave significantly lower ratings in the low-level condition ($\beta = -0.128, SE = 0.013, p < 0.001$). The data suggests that participants interpret *many* as a lower proportion of N when it is presented in a low probability context than when *many* occurs in a high probability context. We can interpret the fact that the factor PROBABILITY was identified as a main effect as evidence that the context influences the interpretation of proportional *many*. This effect was modulated by an interaction of PROBABILITY with PARTITIVE ($\beta = -0.052, SE = 0.018, p < 0.005$).

The linear mixed effects regression model suggests that the comparison class has a significant effect on the interpretation of *many*. This contradicts a theory which assumes one fixed value for the proportion $k_{\max}$ and $k_{\min}$ respectively. Rather, the semantics should comprise *many*'s interaction with the context. Interestingly, neither the factor NUMBER nor the factor PARTITIVE were significant main effects. This result leaves open the possibility of a unified semantics. Cardinal *many* cannot be combined with the partitive and our results show that this construction does not lead to a significant difference in interpretations when combined with the proportional use of *many*. Furthermore, cardinal *many*'s range is not restricted by an upper bound. An upper bound is available for the proportional reading, but its exact numeric value is not decisive. It is the size of the proportion that matters. Overall, we see that even though an upper bound as well as the partitive construction are

only available with the proportional reading, these two factors do not disqualify a unified semantics.

The CFK semantics postulates a stable core meaning for cardinal *few* and *many*. For the proportional reading, this stable core meaning cannot be a fixed proportion, as the experimental data shows. However, we do not yet want to rule out the idea of a fixed threshold on expectations, as proposed by Fernando and Kamp (1996). As a next step, we examine more closely how the interpretation of proportional *many* is affected by the context. To do this, we measure data from people's prior expectations of typical cardinalities in the contexts we used and also apply a computational model to the data. Then, we set out to test whether the CFK semantics can be transferred to the proportional reading. This undertaking might appear straightforward for real-world contexts as tested in the recent experiment and in another study presented in Section 7.2. As we have discussed in the context of the Superbowl experiment presented in Sections 3.3 and 5.1, the proportional reading can also appear in very abstract contexts though. To take a look at the big picture, we also present an experiment in which draws of colored balls from an urn are described by proportional *few* and *many*. In this abstract context, we predict that expectations can be manipulated in a more controlled way because they are not influenced by participants' probably differing world knowledge. We want to learn whether also in such a case, in which world knowledge does not play a role, prior expectations are sufficient to capture the use of proportional *few* and *many* or whether the actual size of the proportion needs to be taken into account.

It turns out that the computational model from the previous chapters does not make correct predictions and conclude that the size of the described proportion cannot be neglected. For this reason, an extension of the CFK model is proposed in Section 7.5, which preserves the hypothesis of a stable core meaning of *few* and *many*, but allows for two different kinds of prior expectations: an uninformed, uniform distribution over proportions and a distribution informed by world knowledge. By applying this model to the data, we test whether the fixed threshold hypothesis from Chapter 5 can be confirmed also for the proportional reading. Furthermore, it will be interesting to see whether the same values for the threshold values θ_{many} and θ_{few} apply to both kinds of prior expectations. This would provide further evidence for the CFK semantics' fixed threshold theory. Moreover, if the cardinal and the proportional threshold turn out to be the same, this would speak against the much debated lexical ambiguity hypothesis of proportional and cardinal *few* and *many* (see Partee (1989), Krasikova (2011) and references therein).

7.2 Experiments in Real-World Contexts

These experiments on proportional *few* and *many* in real-world contexts follow up on the interpretation experiment reported in the previous section. The interpretation task confirmed that prior expectations matter when proportional *few* or *many* are produced or interpreted. In a next step, we test whether the CFK semantics can be transferred to this reading and measure P_E for both *probability* conditions per item. We decided to continue with a smaller set of items and chose those items from the previous experiment whose ratings were most sensitive to the manipulation of the PROBABILITY condition. We elicit prior expectations of the contexts and set a judgment task to measure the production of the proportional reading. Note that in these two follow-up experiments only one total amount per item is presented and *few* and *many* are used in a partitive construction.

7.2.1 Elicitation of Prior Expectations

This experiment gathers data about people's prior expectations concerning the contexts used in the interpretation task reported above. The obtained probability distributions will be input to a computational model being an empirical measure of P_E, see below.

Design. We used a slider-rating task to collect data about the participants' prior expectations about likely world states relevant for each experimental context. Participants saw a description of a context as in (130a), which introduced the total quantity of objects and the PROBABILITY condition, and a question as in (130b). We again manipulated the PROBABILITY condition of the context by presenting one of two statements which influenced expectations of the context. For a *high probability* condition we expect higher proportional answers than in a *low probability* condition. For example, we expect that a strong woman carries more boxes than a weak woman. Depending on the quantity, we presented participants with 10, 13 or 16 slider-interval pairs and asked them to rate the likelihood of each interval by adjusting a slider labeled from "extremely unlikely" to "extremely likely". For example, participants would adjust a slider for the probability that Martha carried 0, 1, 2, ...15 boxes. The task was the same as before (Kao et al., 2014; Franke et al., 2016), see Section 5.4.1.

(130) **Prior elicitation example**

a. BACKGROUND: When moving flat, Martha packed 15 big boxes. Martha is a [strong | weak] woman.

b. QUESTION: How many of the boxes do you think Martha carried?

Participants. We elicited data from 160 participants with US-IP addresses via Amazon's Mechanical Turk.

Materials & Procedure. After having read an explanation of the task, each participant saw all of the 10 contexts from Appendix 7.A.2 in a random order, one after another. For each of the contexts, the PROBABILITY condition was assigned randomly. For each context, the 10, 13 or 16 intervals were presented horizontally on the screen in ascending order from left to right. On top of each interval was a vertical slider. Participants had to adjust or at least click on each slider before being able to proceed (Kao et al., 2014; Franke et al., 2016).

Results. We excluded data from one participant who reported to be a native speaker of Russian. We normalized each subject's rating by item and condition and then averaged each item-condition pair over all subjects, as before in Section 5.4.1. Figure 7.2 displays the probability distributions for each item in both conditions.

7.2.2 Production Study: Judgment Task

Design. To assess production behavior of proportional *few* and *many* in real-world contexts we presented participants with a binary judgment task. They read a context as in (131a) which introduced the total quantity of objects, the PROBABILITY of the context condition, and a PROPORTION of objects. The contexts differed in the PROBABILITY condition specified in the relative clause. The relative clauses were minimal pairs and differed in most cases only in the adjective. To make the prior expectation salient, a statement was paired with a *for*-phrase, see (131b). The quantity words were included in a partitive construction ("many/few of the") to hint at a proportional use. Participants rated whether the statement is a good description of the sentence by clicking on TRUE or FALSE.

(131) **Production study example**

a. CONTEXT: When moving to a new flat, Martha packed 15 boxes. Martha is a [strong | weak] woman. She carried [1 | 3 | 5 | 8 | 10 | 12 | 14] of the boxes herself.

b. STATEMENT: For a [strong | weak] woman, Martha carried [few | many] of the boxes herself.

c. QUESTION: Is this statement a good description of the context? TRUE / FALSE

Participants. On Amazon's Mechanical Turk, we elicited data from 456 participants with US-IP addresses.

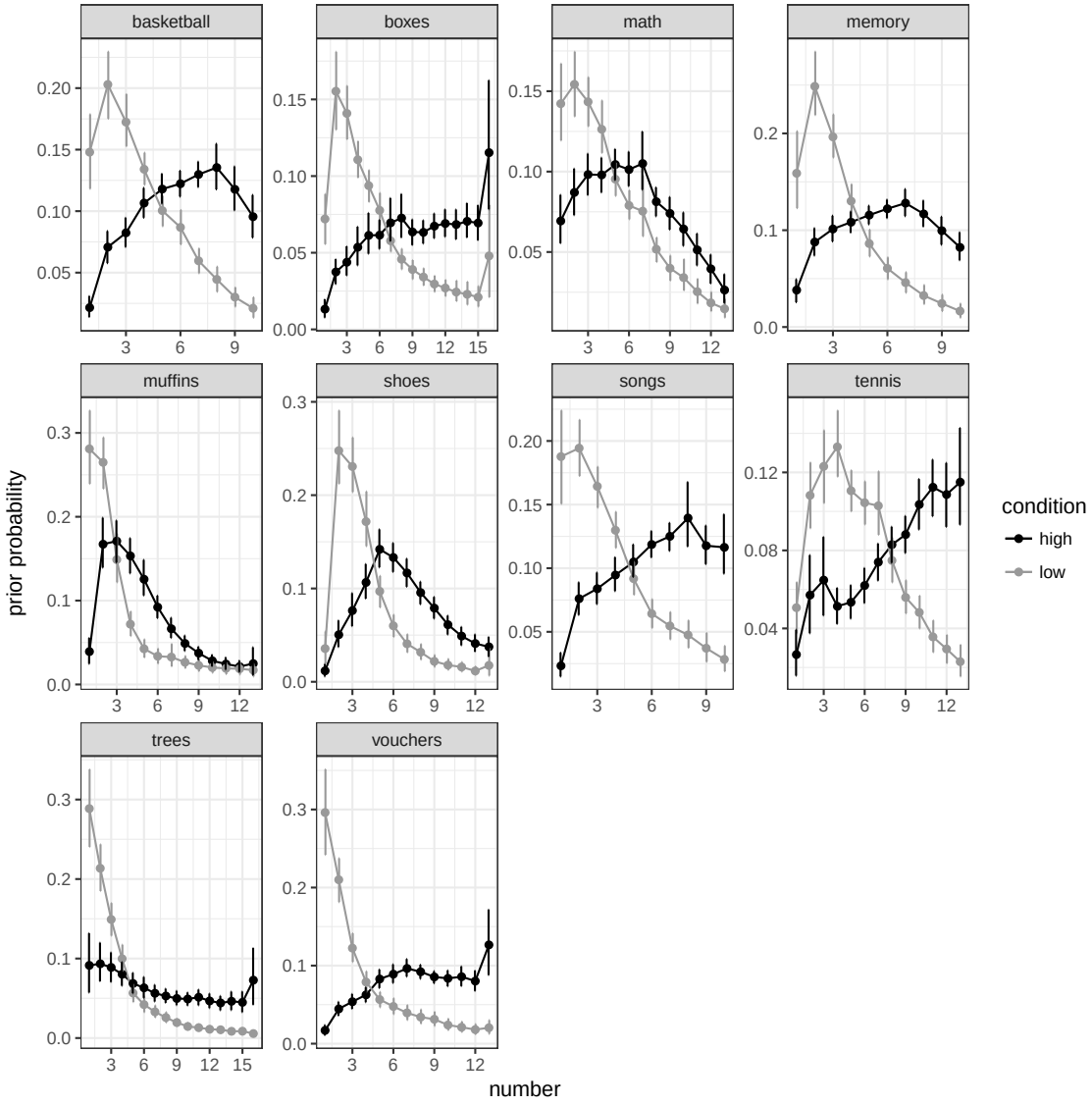


Figure 7.2: Proportional *many*, prior expectations for both context conditions. Error bars are estimated 95% confidence intervals.

Materials & Procedure. After reading instructions that explained the task, participants saw each of the 10 contexts one by one. *Few* or *many* as well as the PROBABILITY condition and the PROPORTION were assigned randomly for each context. We presented only 7 proportions per context and hence not every number in the interval from 0 to the total QUANTITY (10, 13 or 16) to avoid too many combinations. Participants had to rate each statement before being able to proceed.

Results. We excluded 9 participants for not being self-reported native speakers of English. We calculated the proportion of TRUE answers for each context-quantifier-number-prior combination. The proportion of TRUE answers per combination is presented in Figure 7.3. The computational model's production rule will have to predict these data.

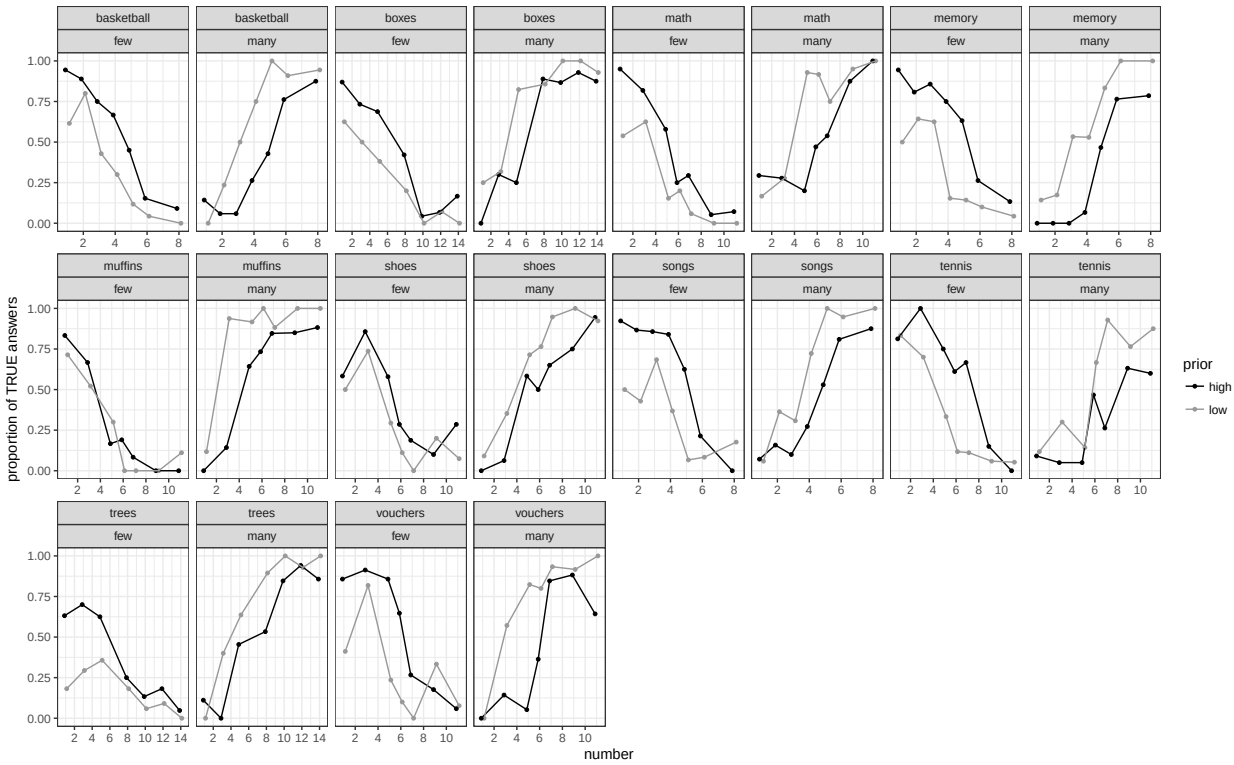


Figure 7.3: Proportion of TRUE responses in real-world contexts. Black lines show the high, gray lines the low prior condition.

For each of the quantifiers *few* and *many* we again specified a mixed linear effects regression model predicting the percentage of TRUE answers. During a guided search through the model space, we started out with a model containing only the random effect ITEM and added fixed effects if this significantly increased the model's fit to the data (measured by AIC).

few. For *few*, the final model included the fixed effect PROPORTION and PROBABILITY. We found that participants gave significantly lower ratings for a higher proportion ($\beta = -1.07, SE = 0.07, p < 0.001$). The sentences were rated significantly lower in the low probability condition ($\beta = -0.21, SE = 0.03, p < 0.001$). The factor QUANTITY did not turn out to be significant main effect. The regression suggests that *few* is used to describe numerically and surprisingly low proportions.

many. For *many*, we found the same pattern in reverse. Participants gave significantly higher ratings in the low probability condition ($\beta = 0.23, SE = 0.03, p < 0.001$) and for higher proportions ($\beta = 1.10, SE = 0.07, p < 0.001$). *Many* seems to convey that a proportion is numerically and surprisingly high.

7.3 Experiments in Abstract Contexts

As pointed out above, the proportional reading of *few* and *many* does not only occur in real-world contexts, in which the world knowledge is very salient, thereby making Fernando and Kamp's (1996) theory plausible. We have seen in Sections 3.3 and 5.1 that the proportional reading suggests itself in abstract contexts, too, for example when proportions of balls in a bowl are described. We propose that in such abstract scenarios, not only statistical information about real-world events is available, but also abstract probabilistic beliefs about pure chance processes. The latter kind of beliefs can be better manipulated, but subjects may have trouble to get clear about these abstract expectations.

Data of proportional *few* and *many* is elicited in an abstract scenario in which balls are drawn from an urn. We test whether the computational model from Section 5.2 is able to account for the proportional reading in an abstract setting in which the sheer size of the proportion might play a greater role than world knowledge. It turns out that prior expectations only based on world knowledge are not sufficient to explain the proportional reading in abstract contexts. We propose an extension of the model in Section 7.5. In the following, a judgment task using various images and the underlying prior expectations we assume are presented.[1]

7.3.1 Production Study: Judgment Task

Design. In a rating task we measured participants' production behavior of proportional *few* and *many* in an abstract context. We assume that contextual information is represented as prior expectations about which quantities are considered typical or normal. In order to restrict the influence of real-world knowledge and corresponding opinions and personal experiences, which are hard to control, we present a very abstract setting. A sample of the material can be found in Figure 7.4. Participants were presented with a situation in which a character draws balls from an urn of varying content. The character describes the draw with a statement about the proportion of blue balls. The statement contained either *few* or *many*, as exemplified in (132b). Participants were then asked to rate on a 7-point scale whether the sentence is a good description of the situation, see (132d). The value 1 was labeled "disagree", the value 7 was labeled "agree" (see Figure 7.4).

(132) **Production study example**

a. CONTEXT: visual display as in Figure 7.4.

[1]Note that no interpretation experiment has been conducted because this task would require an extensive display of a large number of images. We leave this experiment as a follow-up study for future research.

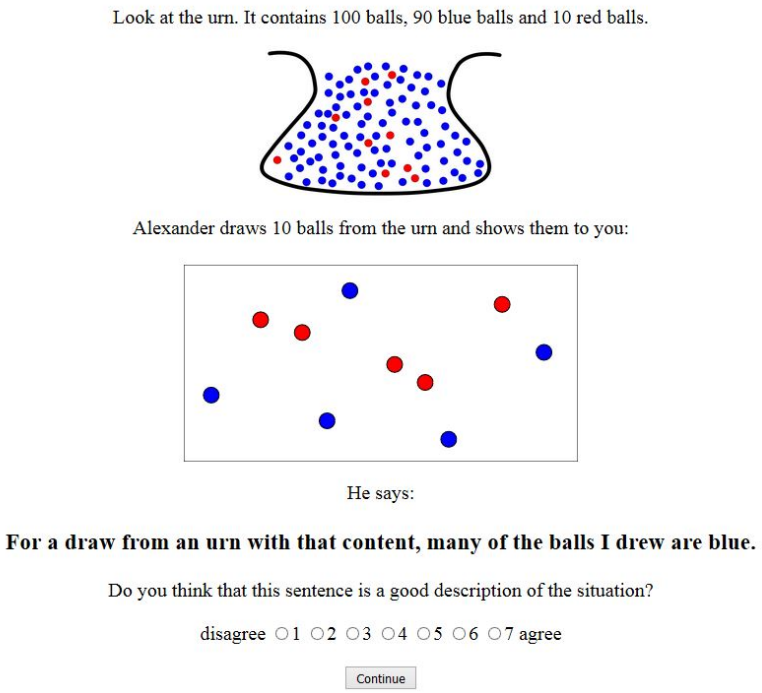


Figure 7.4: Sample item in the rating task

b. STATEMENT: For a draw from an urn with that content, [few|many] of the balls I drew are blue.

c. FILLER: For a draw from an urn with that content, this number of [blue|red] balls is [impossible |unexpected|surprising|expected].

d. QUESTION: Is this statement a good description of the situation?

We manipulated the factors PRIOR EXPECTATIONS, PROPORTION, and DRAW. To manipulate PRIOR EXPECTATIONS, picture of the urn was presented from which the balls were drawn. The urn's content varied. From a total of 100 balls either [25, 50, 75 or 90] balls were blue, the rest red. Depending on the proportion of blue balls in the urn, we hypothesize that participants expect a similar proportion of blue balls in their draw. These prior expectations can be formalized as a draw without replacement. This is discussed in detail in Section 7.3.2. The content of the urn was explicitly mentioned in a *for*-phrase to make the prior salient.

A character draws balls from the urn. We varied the size of the DRAW. The character either draws [10] or [20] balls from the urn. We investigate if the quantity of the superset has an influence on subjects' behavior. The outcome of the draw is presented visually and the balls are randomly placed on the screen. The character describes the PROPORTION of blue balls drawn [10% to 90%] with a statement including the QUANTITY WORD *few* or *many* as in (132b).

quantity word	prior expectation	draw	proportion of blue balls in draw
few, *many*	[25, 50, 75, 90] blue balls out of 100 balls in urn	10, 20	*few*: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7] *many*: [0.4, 0.5, 0.6, 0.7, 0.8, 0.9]

Table 7.1: Experimental conditions in production study

Participants. 300 subjects were recruited via Amazon's Mechanical Turk with US-IP addresses.

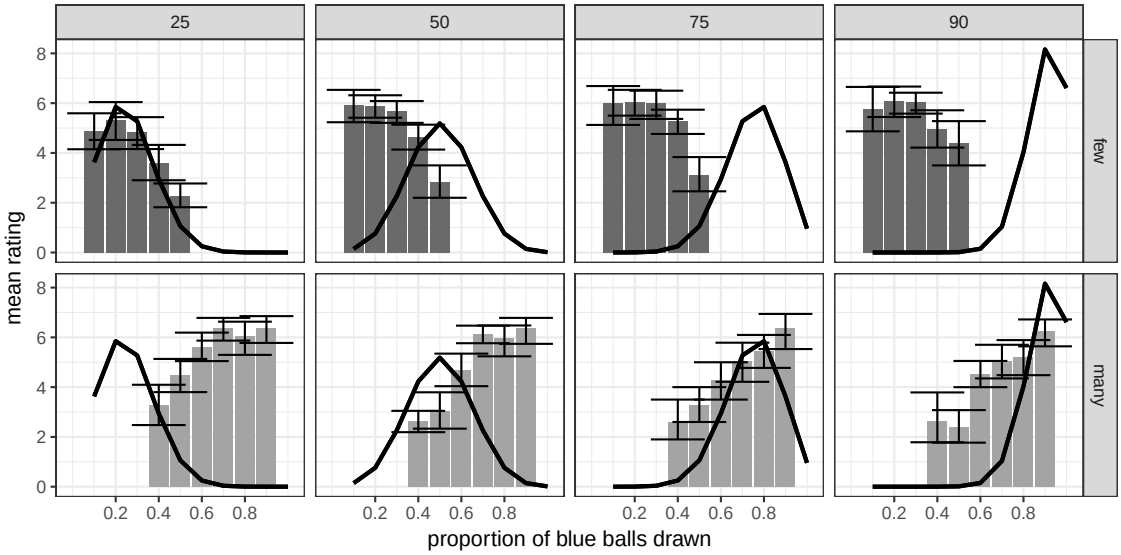
Materials & Procedure. After reading a short explanation of the task, each subject saw 16 items, one after another in a random order. Eight of them are fillers which draw attention to expectations, the remaining eight are target items with *few* or *many*. The arrangement of the balls in the urn and the gender and name of the characters were chosen randomly. All items were presented with either 10 or 20 drawn balls. For each item, one of four prior expectation conditions and a statement were assigned randomly. The statements included either *few, many* or one of the fillers *impossible, unexpected, surprising* and *expected*. The quantity words *few* and *many* were presented four times each just as the prior expectation conditions, see (132b) and (132c). For *few* we presented the proportions [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7] of blue balls among the drawn balls, for *many* [0.4, 0.5, 0.6, 0.7, 0.8, 0.9]. Participants had to click on one of the radio buttons before being able to proceed to the next item. An overview of all experimental conditions is provided in Table 7.1.

Results. Data was excluded of 9 participants who reported not to be native speakers of English. Figure 7.5 shows the mean rating of the prior-proportion pairs. We also plot the hypothesized underlying prior distribution, as described in the next section. Note that the scale of this graph is stretched but its shape remains unchanged. A computational model is to account for the data from this experiment.

A first visual inspection of the mean ratings of each number-prior pair suggests that the manipulation of the urn content makes a difference. In general, the manipulation seemed to have worked out for *few* and *many*, see Figures 7.5a and 7.5b.

For each of the quantifiers *few* and *many* we specified a mixed linear effects regression model predicting ratings for the quantified statements. During a guided search through the model space, we started out with a model containing only the random effect PARTICIPANT and added fixed effects if this significantly increased the model's fit to the data (measured by AIC). Possible factors which could turn out to be a main effect are PROPORTION of blue balls (10%-70% for *few*, 40%-90% for *many*), PRIOR EXPECTATIONS (25, 50, 75 or 90 blue balls out of 100 in the urn) and DRAW (10 or 20 balls drawn from the urn).

(a) 10 balls drawn



(b) 20 balls drawn

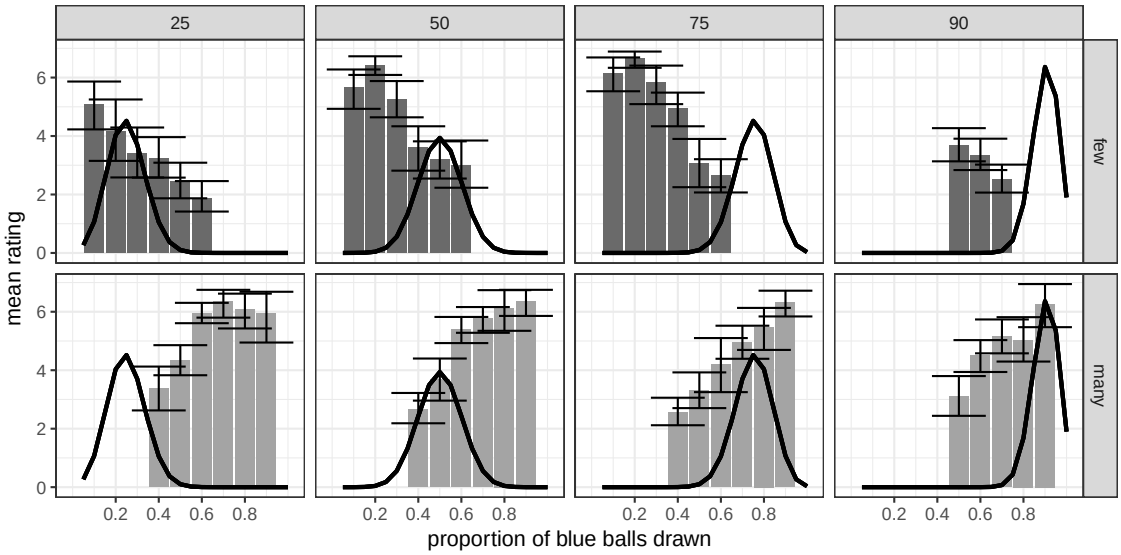


Figure 7.5: Mean rating of sentences with proportional *few* or *many* in the urn scenario. Lines are the hypothesized underlying prior distributions (stretched for presentation), bars are mean ratings of each prior-proportion pair, error bars are estimated 95% confidence intervals.

For *few*, the final model included the fixed effects PROPORTION, PRIOR and DRAW. We found that participants gave significantly lower ratings for a rising proportion of balls ($\beta = -6.80, SE = 0.34, p < 0.001$). The factor DRAW was significant as well ($\beta = 0.03, SE = 0.01, p < 0.01$). This suggests that proportional *few* is not necessarily applicable to low numbers (see mismatch in number feature discussed in Section 7.7) but rather to low proportions. PRIOR EXPECTATIONS turned out to be

significant, too. We found that a higher prior (say 75 or 90 blue balls in the urn) led to higher ratings ($\beta = 0.02, SE = 0.003, p < 0.001$). This once more confirms the idea of the surprise semantics that *few* applies to surprisingly low cardinalities or proportions.

For *many*, the final model included the fixed effects PROPORTION, PRIOR, and DRAW. We found that participants gave significantly higher ratings for an increasing number of balls ($\beta = 7.10, SE = 0.29, p < 0.001$). DRAW did not turn out to be significant, which suggests that in contrast to *few*, *many* applies to numbers which count as both large quantities and large proportions. We found a significant effect of the factor PRIOR EXPECTATIONS as well. We found that a higher prior condition led to lower ratings ($\beta = -0.01, SE = 0.002, p < 0.001$). As expected, *many* can express that a cardinality is surprisingly high.

7.3.2 Prior Expectations

In order to use Bayesian inference to estimate likely threshold values, the underlying prior expectation of the outcome of the draws is necessary. As mentioned in the previous subsection, we chose an abstract scenario for the experiment with the goal that the participants' prior expectations are not influenced by real-world knowledge or personal experiences. Consequently, we expect that there is very little variance in the prior expectations based on which participants form their judgments. These expectations may be formalized as a hypergeometrical probability distribution, see below:

$$P(X = k) = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}} \tag{7.1}$$

$$\text{where} \quad \binom{n}{k} = \frac{n!}{(n-k)! \cdot k!} \tag{7.2}$$

The hypergeometrical distribution is introduced in stochastics using the example of a draw from an urn without replacement, exactly what we find in the experiment from the previous section. n balls are drawn from an urn containing N balls, K of them blue. The distribution states the probability of having drawn k blue balls. Let us illustrate this by calculating the probability of Alexander's draw from Figure 7.4. Alexander drew k=5 blue balls in a draw of n=10 balls from an urn containing K=90 blue balls and N=100 balls in total:

$$P(X = 5) = \frac{\binom{90}{5} \cdot \binom{100-90}{10-5}}{\binom{100}{10}} = 0.0006 \tag{7.3}$$

The probability of this outcome is 0.06%, so it is extremely unlikely to draw only five blue balls in a draw of 10 balls when there are 90 blue balls in the urn. If the participants make use of these prior expectations and if *many* expresses that a quantity is higher than expected, then *many* is not a good description of Alexander's draw. We expect that this item receives a low rating.

Once more, the prior expectations for each possible prior-proportion combination, which are input into the computational model, are assumed to be hypergeometrically distributed. Whether this is really a valid assumption is scrutinized in Section 7.7.3.

7.3.3 Data Evaluation with Computational Model

Before we set out to analyze both data sets (real-world and abstract context) with a computational model, we first test whether the computational model from Section 6.4 is applicable at all. Does this model, which assumes that the production of *few* and *many* is dependent on prior expectations informed by world knowledge, also make good predictions for the proportional reading in abstract contexts? Note that this model variant is a special case of the model presented in Section 5.2 because it only predicts production data and not also interpretation data. Data from the urn scenario is tested in isolation first because we expect that this context is a particularly hard case for the model. The size of the proportion is not taken into account by the model, and if it manages to capture the data anyway, this can be taken as evidence that the proportional reading is also only dependent on world knowledge. If the model's predictions do not match the experimental data though, we learn that the size of the proportion might have to be included as another factor in the computational model.

The computational production model from Section 6.4 is applied to the urn data from Section 7.3.1 while taking the hypothesized priors from the previous section as input. Since the rating were given on an ordinal ratings scale and not on a binary scale as before, we used a link function to be able to predict ordinal data from the model's binary predictions. This link function for ordinal data is adopted from Kruschke (2014, Chapter 23) and Franke (2016) and explained in detail in Section 7.6. Via Bayesian inference we test which threshold values θ_{many} and θ_{few} are most likely after having seen the production data of proportional *few* and *many*. We assumed the same priors for θ_{many} and θ_{few} and assumed one noise parameter σ_i per urn condition (25, 50, 75 or 90 blue balls in urn). The prior over σ_i is uniformly distributed over the interval [0 ; 1] because it ranges over proportions this time.

$$P(\theta_{\text{many}}, \theta_{\text{few}}, \sigma_i) = \text{Uniform}_{[0;1]}(\theta_{\text{many}}) \cdot \text{Uniform}_{[0;1]}(\theta_{\text{few}}) \cdot \text{Uniform}_{[0;1]}(\sigma_i) \quad (7.4)$$

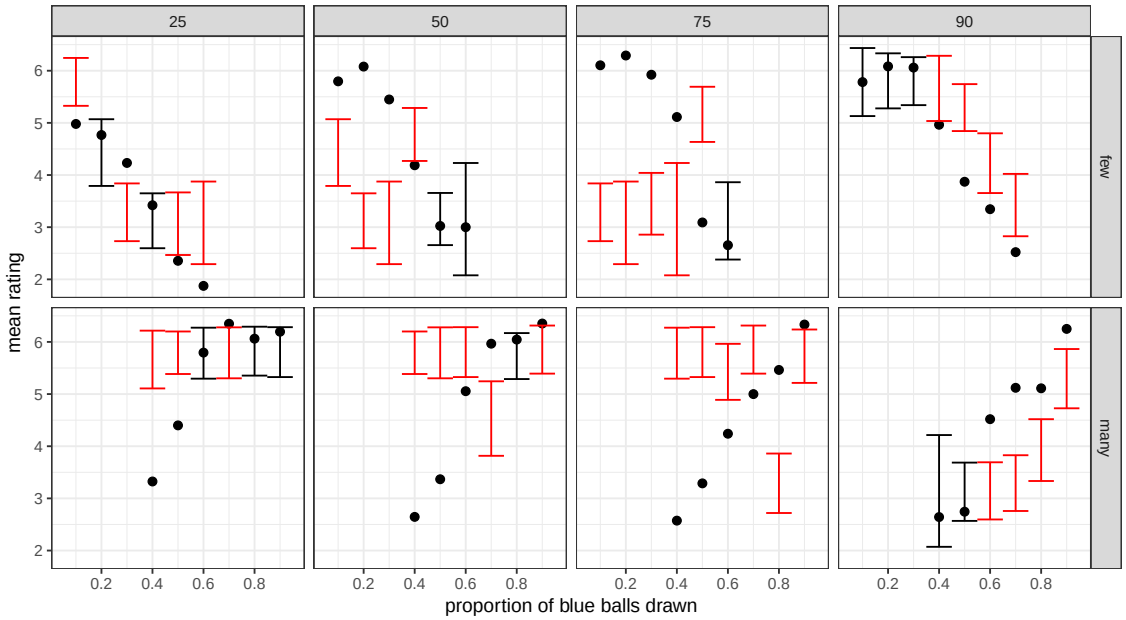


Figure 7.6: PPC of the production model from Section 6.4 applied to the urn data. Bars are the 95% HDIs of the model's predicted mean ratings, points are mean ratings as measured with the judgment task in Section 7.3.1. Bars are printed in red if the experimentally measured mean ratings do not fall into the PPC's HDIs.

To approximate the joint posterior distribution defined in Equation (5.4), we used MCMC sampling, as implemented in JAGS (Plummer, 2003). We collected 10,000 samples from 2 MCMC chains after a burn-in of 10,000. This ensured convergence, as measured by $\hat{R}$ (Gelman and Rubin, 1992).

The highest density intervals predicted by the model are [0.002 , 0.003] for θ_{few} and [0.659 , 0.831] for θ_{many}. For *few*, this model makes very implausible predictions. If the threshold on the use of *few* were really this low, *few* could only be applied to the lowest proportions, at most 10%. This is not what we see in the data. The results of the judgment task visualized in Figure 7.5 show clearly that *few* is also used to make reference to higher proportions. The mean correlation between the observed data and the model's predictions is 0.72 for *few* and 0.66 for *many* which is lower than for the cardinal data from Section 5 (0.92 for *few* and 0.89 for *many*, see Section 5.5). As a further sanity check of the model's predictions, we conducted a posterior predictive check (PPC), as introduced in Chapter 4. In each step of the chain, we created a sample set of ratings, as predicted by the model and the sampled parameter values. From these sampled ratings, we calculated the mean ratings and, next, their 95% HDI. When comparing the sample mean ratings with the mean ratings from the judgment task in Section 7.3.1, we find that the model manages to predict 32% of the data for *few* correctly and 28% for *many*. These numbers show clearly, that the model does not describe the experimental data well. Find the PPC's results visualized in Figure 7.6. The bars show the 95% HDIs of the predicted mean

ratings, the points experimentally measured mean ratings. When the bar's color is red, we see that prediction and actual data do not match. Last also the model's fit to the data as measured by DIC = 1857.2 and pD = 62.1 is not convincing (in contrast to a much better fit of DIC = 1347.5 with a mean posterior deviance of 28.1 by the linear combination model to the same data set, see Section 7.5).

7.4 Adapting the CFK Semantics to the Proportional Reading

From the urn data reported in the previous sections and the predictions the computational model from Section 5.3 makes, we conclude that a version of the CFK semantics which makes its predictions only based on prior expectations informed by world knowledge is not apt to account for the use of the proportional use in such an abstract context. We propose that also the size of the described proportion needs to be taken into account. As mentioned above, we therefore assume that the contextual contribution for the proportional reading is two-fold: the first is an uninformed, uniform distribution about proportions and the second is an informed prior belief about likely proportions based on world knowledge. Consequently the proportional reading can express that a proportion is numerically high or surprisingly high.

Prior expectations informed by world-knowledge have been employed in the previous chapters as the contextual input into the CFK semantics. Fernando and Kamp (1996) formulate their expectation-based semantics also for the proportional reading by adding the assumption of a bounded scale. The proportional surprise reading of *few* and *many* expresses that a proportion is surprisingly high or low, see the sentence below taken from the experimental material in Appendix 7.A.1.

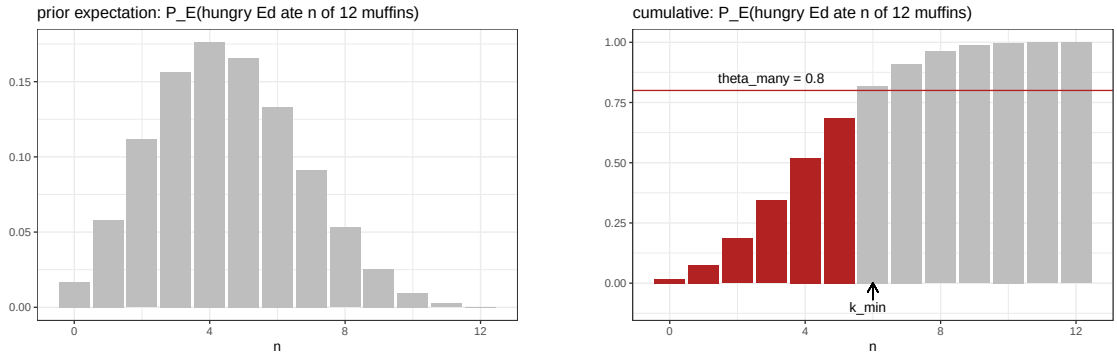
(133) There were 12 muffins on the kitchen table in Ed's flat. Ed, who arrived feeling hungry, ate few / many of the muffins.
$\rightsquigarrow$ Ed ate fewer/more muffins than expected.

To briefly summarize once more, the idea behind the CFK semantics is that, e.g., *few* could be taken to denote "the 25th percentile (range: 10th to 40th percentile) on the distribution of items inferred possible in [the current] situation" (Clark, 1991, 271). The CFK semantics are repeated from above:

(77) **CFK Semantics**

a. $[\![\text{Few As are B}]\!] = 1$ iff $|\, A \cap B \,| \leq x_{\max}$

where $x_{\max} = \max\{n \in \mathbb{N} \mid P_E(|\, A \cap B \,| \leq n) < \theta_{\text{few}}\}$



(a) Prior (b) Cumulative

Figure 7.7: Illustration of the CFK-semantics

b. $[\![\text{Many As are B}]\!] = 1$ iff $|\, A \cap B \,| \geq x_{\min}$

where $x_{\min} = \min \{n \in \mathbb{N} \mid P_E(|\, A \cap B \,| \leq n) > \theta_{\text{many}}\}$

This approach explains the surprise reading of *few* and *many* in sentences like above as a comparison between the actual proportion of muffins that the hungry Ed ate and a probabilistic belief P_E about the expected proportion of consumed muffins in some contextually provided *comparison class* (say, hungry American males relevantly similar to Ed). This prior belief P_E is very individual and clearly dependent on the speaker's world knowledge, see Figure 7.7a for a sample distribution. A probability is assigned to any natural number n in the interval [0,12], indicating how likely the speaker thinks it is that hungry Ed ate n of the 12 muffins on the table.

In addition to this context-dependent input, the CFK semantics proposed a context-independent lexical meaning of *few* and *many*. A pair of fixed thresholds $\theta_{\text{few,E}}$ and $\theta_{\text{many,E}}$ applies to the cumulative distribution of P_E, see Figure 7.7. Figure 7.7b shows the cumulative distribution of the distribution in Figure 7.7a. If $\theta_{\text{many,E}}$ was fixed to, say, 0.8, then the CFK-semantics would identify $k_{\text{min,E}}$ to be 6. Accordingly, for this P_E, the *many*-sentence in (133) would be false for any $n < 6$ and true for any $n \geq 6$.

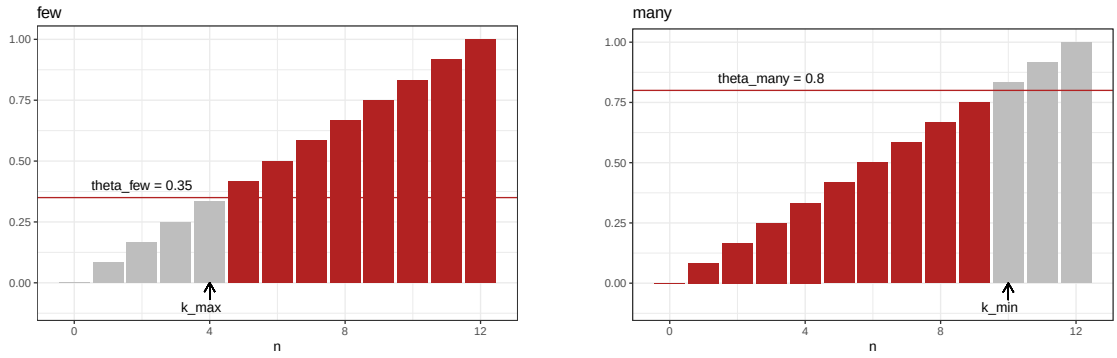
Figure 7.7 shows an example of a prior expectation which a speaker might have in mind when talking about a real-world context in which she can make use of her knowledge and experience. That proportional *few* and *many* are context-dependent has been shown by the interpretation experiment in Section 7.1.2. We now set up a way of making the underlying idea of the CFK semantics work for the proportional use of the quantity words. We will propose that two kinds of prior expectations, P_E and P_U are required as input. Note that this is just one possibility of adapting the CFK semantics to the proportional reading. We explore this variant in the following.

In contrast to the cardinal reading, the reasoning process about the use of proportional *few* and *many* is not necessarily only dependent on surprise though; in theory it can be completed without employing world knowledge. An uninformed, flat prior expectation P_U about the proportions of the total amount $|\,A\,|$ could be used as input[2]. This is not (easily) possible and usually quite implausible for the cardinal reading where no upper bound on the interval exists. For mathematical reasons, a uniform distribution on an unbounded interval is peculiar. Consequently, the speaker cannot remain ignorant and *must* employ world knowledge when deciding whether cardinal *few* or *many* are a true description of a certain cardinality.

Coming back to the proportional reading, we predict that in the abstract urn scenario (Section 7.3), the sheer size of the proportion has a greater effect on the production of *few* and *many* than prior expectations informed by world knowledge. In these abstract contexts it seems natural that the contextual input is a uniform distribution about proportions of $|\,A\,|$. Just like the informed prior, this distribution can be input into the CFK semantics, which determines the proportional threshold $k_{\max}$ and $k_{\min}$ in exactly the same fashion. The context-independent threshold $\theta_{\text{few,U}}$ and $\theta_{\text{many,U}}$ apply to the cumulative density mass of this distribution. For a flat prior, it does mathematically not make a difference whether the threshold is applied to distribution P_U or directly to $|\,A\,|$, resulting in some proportion $\frac{m}{|\,A\,|}$. For a total number of 12 muffins as in example (133) and an uninformed distribution as in Figure 7.8b, the CFK semantics would predict that *few* can be felicitously used to describe cardinalities up to 4, for hypothetical threshold $\theta_{\text{few,U}}$ of 0.35 on the cumulative density mass. Similarly, *many* would apply to cardinalities 10 and higher, if $\theta_{\text{many,U}}$ were 0.8 (note that this cut-off point is higher for P_U than for P_E even though the threshold value is 0.8 in both cases). See Figure 7.8 for an illustration. Such a view on the proportional reading seems to be especially appropriate in abstract contexts in which we cannot rely on world knowledge. But also for real-world examples like (133) there is reason to believe that the size of the described proportion matters. Just think of example (12) from Section 2.1.1 again, for which Partee (1989) argued that *few* can never mean *all*.

Finally, an open issue is how the proportional and the surprise-based threshold are combined in the proportional reading. Once more, we hypothesize that when the proportional reading is formed, two kinds of prior expectations are available: (i) P_U expresses a context-independent, flat distribution about proportions of $|\,A\,|$ and (ii) P_E is informed by *world knowledge* about likely proportions in the situation.

[2]Note that in contrast to P_E, the prior expectations P_U over proportions are uniformly distributed because mathematically there is no reason to prefer one proportion over the other; the bare proportions are independent of expectations of the context. Contextual information comes in in the form of N, however, because the total quantity obviously determines the cardinality which the proportion corresponds to.



(a) *Few* (b) *Many*

Figure 7.8: Illustration of a fixed threshold theory

The speaker is probably uncertain about which contextual information she should draw on. Depending on which prior the speaker makes use of, the cut-off points $k_{\max}$ and $k_{\min}$ can differ. This is why we suggest that the proportional reading of *few* and *many* is best described by a linear combination of both readings. The linear combination contains a weight parameter, a contextually free variable α. α expresses the salience of world knowledge in the respective context. When α is high, the speaker cannot rely on his world knowledge and forms his judgment based on the numerical size of the proportion. When α is low, world knowledge is very salient and the proportional reading expresses that a proportion is surprisingly high or low. This extension is not as dramatic as it might look at first sight. The computational model from Section 5.2 is just a special case of the linear combination model proposed in the following with $\alpha = 0$. The next section explains how the idea of integrating two kinds of contextual contribution can be translated into a computational model.

7.5 Linear Combination Model

This section shows how the idea that the proportional reading is influenced by two kinds of prior expectations can be turned into a probabilistic, linear combination model of speaker production behavior.

In technical terms, the model specifies a likelihood function $P(\text{observation} \mid \text{parameters})$ mapping values of latent parameters onto a probability of seeing a certain behavior in a suitable experiment, parallel to Section 5.3. The latent parameters contained in this model are the weight parameter α and contextually stable thresholds $\theta_{\text{many,E}}$, $\theta_{\text{many,U}}$, $\theta_{\text{few,E}}$ and $\theta_{\text{few,U}}$. $\theta_{\text{many,E}}$ and $\theta_{\text{few,E}}$ apply to the prior distribution P_E which is based on world knowledge. $\theta_{\text{many,U}}$ and $\theta_{\text{few,U}}$ express the cut-off point of the uninformed, flat prior distribution P_U. By applying Bayes rule, credible values of the latent parameters will be inferred from the data from the

judgment tasks reported above, given the likelihood function and some prior over latent parameters.

$$P(\alpha, \theta_{\text{many,U}}, \theta_{\text{many,E}}, \theta_{\text{few,U}}, \theta_{\text{few,E}} \mid \text{observation}) \propto \tag{7.5}$$
$$P(\alpha, \theta_{\text{many,U}}, \theta_{\text{many,E}}, \theta_{\text{few,U}}, \theta_{\text{few,E}}) \cdot P(\text{observation} \mid \alpha, \theta_{\text{many,U}}, \theta_{\text{many,E}}, \theta_{\text{few,U}}, \theta_{\text{few,E}})$$

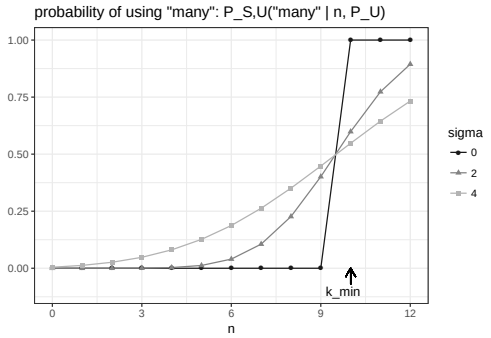
In order to test the predictions of a fixed threshold semantics which takes two distinct prior distributions as input, we once more draw on Bayesian inference. Our goal is to see whether a quintuplet consisting of a linear combination weight and threshold parameters for different quantity words and readings explains the empirical data well enough. If this is the case, it would constitute experimental evidence for the assumption that the proportional reading is a weighted combination of thresholds based on an uninformed and an informed prior distribution and also that proportional *few* and *many* have a stable core meaning. Moreover, we are interested in $P(\theta_{\text{many,E}} = \theta_{\text{many,U}} \mid \text{observation})$. If the two thresholds $\theta_{\text{many,E}}$ and $\theta_{\text{many,U}}$ are the same, this would speak against a lexical ambiguity theory.

The computational model consists of two probabilistic production rules, one for *few* and one for *many*. We focus on *many* in the exposition, but the case for *few* is parallel. Just as for the previous model, a production rule should give the probability $P_S(\text{“many”} \mid n, N, P_U, P_E)$ with which a speaker would produce sentence “Many of the As are B” to describe $n = |A \cap B|$ as a proportion of $N = |A|$ under prior expectations P_U and P_E, where P_U is a flat distribution over numbers $n \in [0, N]$ and P_E captures the relevant statistical properties of the assumed comparison class.

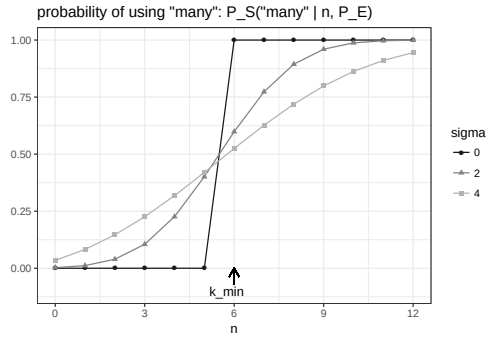
$$P_{\text{speaker}}(\text{”many”} \mid n, N, P_U, P_E) = \alpha \cdot P_{\text{speaker,U}} + (1 - \alpha) \cdot P_{\text{speaker,E}} \tag{7.6}$$

The idea behind (7.6) is this: a speaker reasons about using *many* to describe n as a proportion of N given his prior expectations P_U and P_E of the situation. The production probability P_{speaker} of proportional *many* is then the weighted sum of the production probability $P_{\text{speaker,U}}$ of a speaker who reasons in terms of a flat prior P_U about proportions of the total quantity N and the production probability $P_{\text{speaker,E}}$ of a speaker whose input are his informed prior expectations P_E of likely proportions in the respective context. The weight parameter α regulates the amount of world knowledge expressed in the respective context.

Let us now have a more detailed look at the two summands of P_{speaker}. The CFK semantics in (77) is once more translated into a production rule. The production rule $P_{\text{speaker,E}}$ implements the surprise-based CFK semantics. $P_{\text{speaker,E}}(\text{”many”} \mid n, N, P_E;\ \theta_{\text{many,E}}) = 1$ if $n \geq k_{\text{min,E}}$ and otherwise 0, where $k_{\text{min,E}}$ is derived from P_E, as in (77b). $P_{\text{speaker,U}}(\text{”many”} \mid n, N, P_U;\ \theta_{\text{many,U}})$ is derived in a parallel fashion



(a) Production rule examples for $P_{\text{speaker,U}}$



(b) Production rule rule examples for $P_{\text{speaker,E}}$

Figure 7.9: Illustration of production rule for the example from Figures 7.7 and 7.8

but uses $\theta_{\text{many,U}}$, derived from a uniform prior distribution P_U. Once more, $P_{\text{speaker,U}}$ and $P_{\text{speaker,E}}$ are smoothed-out versions of a binary production rule. Both contain a free model parameter σ that regulates the steepness of the curve.

$$P_{\text{speaker,U}}(\text{"many"} \mid n, N, P_U;\ \theta_{\text{many,U}}, \sigma) = \sum_{k=0}^{n} \int_{k-0.5}^{k+0.5} \mathcal{N}(y; k_{\text{min,U}}, \sigma)\text{d}y \qquad (7.7)$$

Illustrations of the probabilistic production rule $P_{\text{speaker,U}}$ can be found in Figure 7.9a for the example started in Figure 7.8b.

The production rule $P_{\text{speaker,E}}$ implements the surprise-based CFK semantics as well and applies to an informed prior distribution P_E.
$P_{\text{speaker,E}}(\text{"many"} \mid n, P_E;\ \theta_{\text{many,E}}) = 1$ if $n \geq k_{\text{min,E}}$ and otherwise 0, where $k_{\text{min,E}}$ is derived from P_E, based on $\theta_{\text{many,E}}$. $\theta_{\text{many,E}}$ is a free parameter just like the noise parameter σ, which again controls the steepness of the smoothed-out curve. Examples of various values for σ are presented in Figure 7.9b.

$$P_{\text{speaker,E}}(\text{"many"} \mid n, N, P_E;\ \theta_{\text{many,E}}, \sigma) = \sum_{k=0}^{n} \int_{k-0.5}^{k+0.5} \mathcal{N}(y; k_{\text{min,E}}, \sigma)\text{d}y \qquad (7.8)$$

Equations (7.7) and (7.8) express the following idea. For $Y \in \{U, E\}$, assume that a hypothetically true value of $\theta_{\text{many,Y}}$ exists. Then, given a total number N and prior expectations P_Y over the contextually relevant domain, the CFK semantics in (77) gives a clear cutoff for the minimum number $k_{\text{min,Y}}$ of, say, muffins Ed must have eaten to license applicability of *many* in a sentence like (133). We should assume that speakers do not know for sure the actual $k_{\text{min,Y}}$ that is entailed by $\theta_{\text{many,Y}}$ and P_Y, most likely because they do not know P_E for certain, but that speakers nonetheless approximate it. The same holds for N, when it is not provided by the context. More concretely, we assume that when a speaker decides whether some n

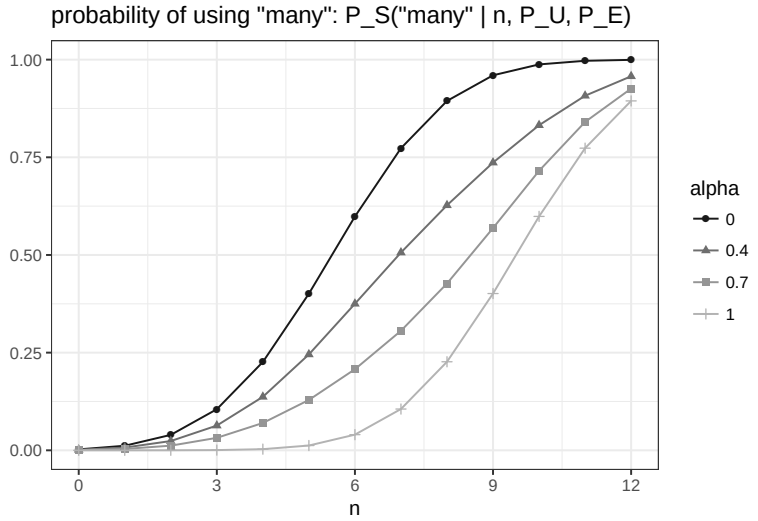


Figure 7.10: Linear Combination of $P_{\text{speaker,U}}$ and $P_{\text{speaker,E}}$ with $\sigma = 2$

licenses *many*, she "samples", so to speak, a noise-perturbed "subjective thresholds" from a Gaussian distribution whose mean is $k_{\text{min,Y}}$ and whose standard deviation σ is a free model parameter that captures speaker uncertainty (about θ_{many}, N, P_E, and perhaps other things). If $k_{\text{min,Y}}$ is below n, the speaker finds *many* applicable to proportion n under the respective reading; otherwise, she does not.

In a next step, reasoning based on proportion and world knowledge are fused, when the general production probability P_{speaker} is derived as a linear combination of $P_{\text{speaker,U}}$ and $P_{\text{speaker,E}}$, as in Equation (7.6) with weight parameter α capturing the amount of world knowledge expressed. This gives us a probabilistic prediction of how likely a speaker would, on occasion, find *many* applicable to n as a probabilistic function of $\theta_{\text{many,U}}$, $\theta_{\text{many,E}}$, P_U, P_E and noise parameter σ. See Figure 7.10 for an example of a linear combination of $P_{\text{speaker,U}}$ and $P_{\text{speaker,E}}$ with several α values. When $\alpha = 1$, only the production probability of *many* is just $P_{\text{speaker,U}}$ and world knowledge is not employed. For $\alpha = 0$, only $P_{\text{speaker,E}}$ determines the use of *many* and we see that *many* is also applicable to lower cardinalities. For values of α between 0 and 1, $P_{\text{speaker,U}}$ and $P_{\text{speaker,E}}$ are combined resulting in the curves in the middle.

7.6 Model Evaluation

After having defined a computational model which makes predictions about the production of proportional *few* and *many*, it will be used to find out to what extent which prior expectations, flat or informed, influence the production of proportional *few* and *many*. Furthermore, we test whether proportional *few* and *many* comprise a fixed threshold on prior expectations as well. We do this by testing whether a

model which incorporates the production rule in (7.6) can explain data from two very different contextual settings, real-world situations and abstract urn scenarios. Is one pair of threshold values, θ_{many} and θ_{few}, applied to P_U and P_E enough to explain the two data sets? Or is it necessary to differentiate between $\theta_{\text{many,U}}$ and $\theta_{\text{many,E}}$ and between $\theta_{\text{few,U}}$ and $\theta_{\text{few,E}}$? In terms of the two data sets, do the model's predictions only differ in the salience of world knowledge represented by α? These questions will be addressed in the course of this section.

We want to learn about $\alpha, \theta_{\text{many}}$ and θ_{few} from the observed experimental data. To do so, we feed the total quantity N, P_U and the hypothesized prior expectations $P_{E_i}^{urn}$ for each number of blue balls i in the urn (see Figure 7.5) or respectively the experimentally measured $P_{E_i}^{RW}$ about real-world contexts (see Figure 7.2) into the partitive production rules in (7.7) and (7.8) which are then combined to form the general production rule in (7.6). This gives us likelihood functions for the production data as described presently. We only explicitly cover the case of *many* wherever that for *few* is analogous. Note that from this point on we will differentiate between σ_U and σ_{E_i}. These noise parameters apply to the production rules in (7.7) and (7.8) respectively. We do so because the parameters are assumed to be independent of each other. Moreover, since σ_U expresses the standard deviation in proportions between 0 and 1 and σ_{E_i} in intervals 0-14 of the respective item, their prior distributions are different. See more below.

For the real-world data, let $O_{ij}^{p_m,RW}$ be the number of TRUE answers for item i and proportion j in production experiments for *many*. Let $N_{ij}^{p_m,RW}$ be the number of participants that saw a production trial for *many*, item i, interval j and total number N (see Section 7.2). $O_{ij}^{p_f,RW}$ and $N_{ij}^{p_f,RW}$ hold the same information for conditions involving *few*. The probabilistic rules from the previous section then give a (parameterized) likelihood function for observable data. Binomial(k, n, p) gives the likelihood of observing k TRUE ratings among n with probability p.

$$P(O_{ij}^{p_m,RW} \mid \theta_{\text{many}}, \sigma_U, \sigma_{E_i}) = \text{Binomial}(O_{ij}^{p_m,RW},\ N_{ij}^{p_m,RW},\ P_S(\text{“many”} \mid j, P_U, P_{E_i}^{RW}\ ;\ \alpha_{many}, \theta_{\text{many}}, \sigma_U, \sigma_{E_i}))$$

For the urn data from Section 7.3, let $\vec{O_{ij}^{p_m,U}}$ be the vector of length 7 which contains the number of times a rating $d \in 1, ..., 7$ has been selected for a proportion j and a prior i in production experiments for *many*. Let $N_{ij}^{p_m,U}$ be the number of participants that saw a production trial for *many*, proportion j and prior i. The same information is contained in $\vec{O_{ij}^{p_f,U}}$ and $N_{ij}^{p_f,U}$ in the case of *few*. Multinomial$(\vec{k}, n, \vec{p})$ gives the likelihood of observing a vector of counts k (here: $\vec{O_{ij}^{p_m,U}}$), where k_d are the number of choices of rating $d \in 1, ..., 7$, for n (here: $N_i^{p_m}$) observations and $\vec{p}$ is a probability vector of length 7. Remember, however, that the observations $\vec{O_{ij}^{p_m,U}}$

are ordinal data, whereas our production rules predict binary data, namely whether *many* is true or not for a certain k. Consequently, a link function needs to transform the binary predictions of (7.6) to the multinomial distribution which predicts ratings on a 7-point scale.

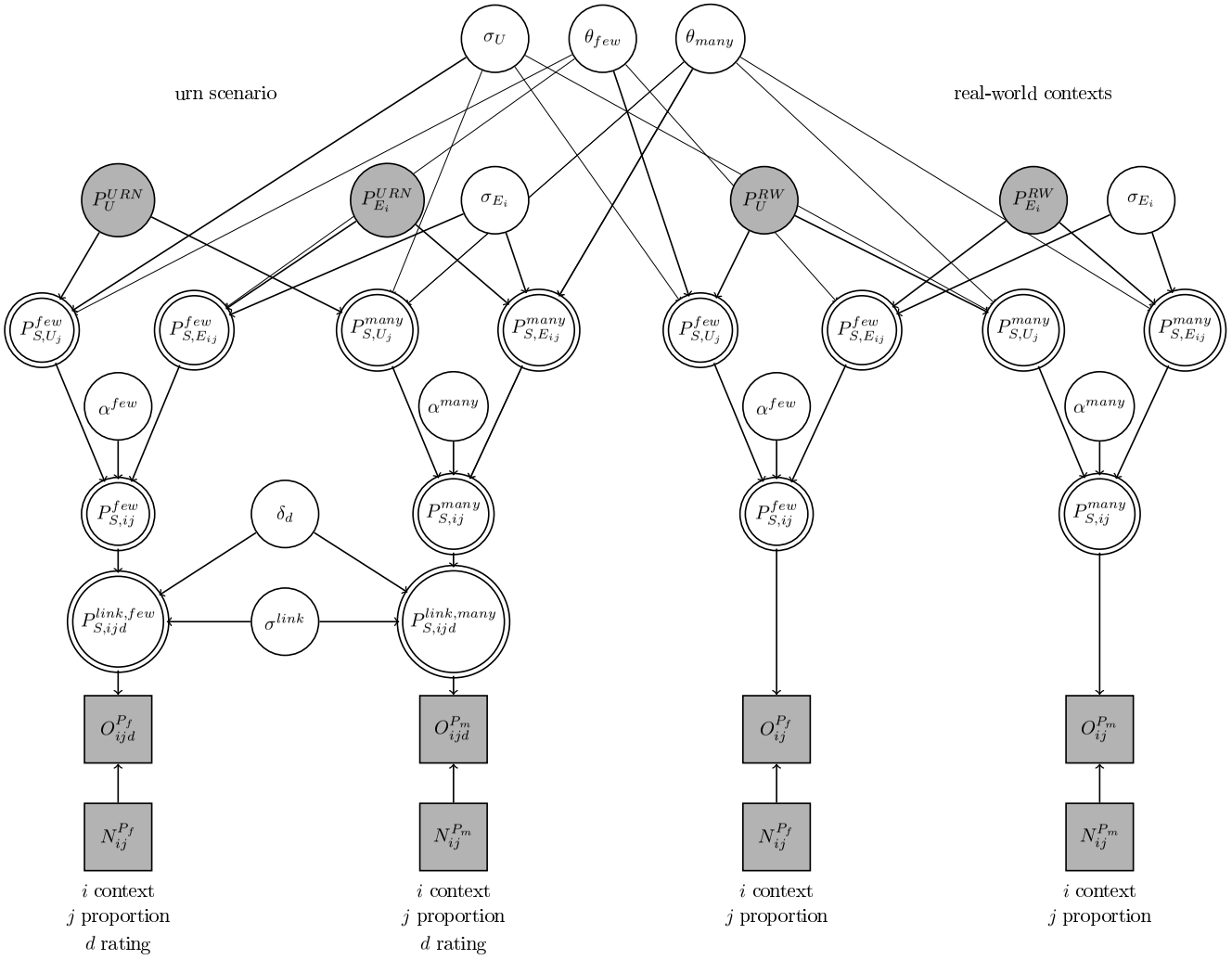
The basic concept of this link function for ordinal data is borrowed from Kruschke (2014, Chapter 23). Each rating d is associated with an interval I_d. The boundaries of (some of) these intervals are free model parameters. Intervals for all ratings form a partition of the range of the scale. For the choice of a certain interval I_d (i.e. rating) for a prior-proportion pair, the binary predictor $P_{\text{speaker}}(\text{"many"} \mid n, N, P_U, P_E)$ from the previous section is perturbed by Gaussian noise with standard deviation σ. Finally, the rating is chosen which corresponds to the interval into which the perturbed value falls. This means that the probability p_d that rating d on the rating scale is chosen in the experiment is the probability that the Gaussian perturbation of $P_{\text{speaker}}(\text{"many"} \mid n, N, P_U, P_E)$ lies in I_d (Franke, 2016). Via this link function we arrive at (parameterized) likelihood functions for data from the urn experiment.

$$P(\vec{O_{ij}^{p_m,U}} \mid \alpha_{many}, \theta_{\text{many}}, \sigma_U, \sigma_{E_i}) =$$
$$\text{Multinomial}\left(\vec{O_{ij}^{p_m,U}},\ N_{ij}^{p_m,U},\ \vec{P}_{speaker_i}(\text{“many”} \mid j, P_U, P_{E_i}^{urn}\ ;\ \alpha_{many}, \theta_{\text{many}}, \sigma_U, \sigma_{E_i})\right)$$

We can make inferences about credible parameter values given the data that we observed by applying Bayes rule, see Equation 7.9. A graphical model version can be found in Figure 7.11.

$$\begin{aligned} P(\alpha_{few}, \alpha_{many}, \theta_{\text{many}}, \theta_{\text{few}}, \sigma_U, \sigma_{E_i} \mid O_{ij}^{p_m,RW}, O_{ij}^{p_f,RW}, \vec{O_{ij}^{p_m,U}}, \vec{O_{ij}^{p_f,U}}) \propto \\ P(\alpha_{few}, \alpha_{many}, \theta_{\text{many}}, \theta_{\text{few}}, \sigma_U, \sigma_{E_i}) \cdot \\ P(O_{ij}^{p_m,RW} \mid \alpha_{many}, \theta_{\text{many}}, \sigma_U, \sigma_{E_i}) \cdot P(O_{ij}^{p_f,RW} \mid \alpha_{few}, \theta_{\text{few}}, \sigma_U, \sigma_{E_i}) \cdot \\ P(\vec{O_{ij}^{p_m,U}} \mid \alpha_{many}, \theta_{\text{many}}, \sigma_U, \sigma_{E_i}) \cdot P(\vec{O_{ij}^{p_f,U}} \mid \alpha_{few}, \theta_{\text{few}}, \sigma_U, \sigma_{E_i}) \end{aligned} \tag{7.9}$$

Several remarks about Equation 7.9. Firstly, we assume here that each summand in Equation 7.6 has its own standard deviation, σ_U or σ_{E_i}, and that the surprise-based one differs again for each prior condition i, but that the parameters are the same for *many* and *few*. This is because we think that uncertainty about world knowledge connected to P_E is distinct from uncertainty about which proportion counts as high or low independent of the context, as captured by P_U. Secondly, the formula above contains as a factor the joint prior probability $P(\alpha_{few}, \alpha_{many}, \theta_{\text{many,prop}}, \theta_{\text{many}}\ \theta_{\text{few}}, \sigma_U, \sigma_{E_i})$ of parameter values $\alpha_{few}, \alpha_{many}, \theta_{\text{many}}, \theta_{\text{few}}, \sigma_U$ and σ_{E_i}. Here, we



$$\theta_{few}, \theta_{many} \sim \text{Uniform}(0,1)$$

$$\sigma_{E_i} \sim \text{Uniform}(0,10)$$

$$\sigma_U, \sigma^{link} \sim \text{Uniform}(0,1)$$

$$P_{S,E_{ij}}^{few/many} \text{ see main text}$$

$$P_{S,U_j}^{few/many} \text{ see main text}$$

$$\alpha^{few}, \alpha^{many} \sim \text{Uniform}(0,1)$$

$$P_{S,ij}^{few} = \alpha^{few} \cdot P_{S,U_j}^{few} + (1-\alpha^{few}) \cdot P_{S,E_{ij}}^{few}$$

$$P_{S,ij}^{many} = \alpha^{many} \cdot P_{S,U_j}^{many} + (1-\alpha^{many}) \cdot P_{S,E_{ij}}^{many}$$

$$\delta_d \sim \text{Normal}(d/7, (2\cdot 7)^{-1})$$
$$\delta_0 = -\infty, \delta_7 = \infty$$

$$P_{S,ijd}^{link,few/many} = \int_{\delta_{d-1}}^{\delta_d} \text{Normal}(x, P_{S,ij}^{few/many}, \sigma^{link}) \mathrm{d}x$$

$$O_{ij}^{RW,P_f} \sim \text{Binomial}(P_{ij}^{few}, N_{ij}^{P_f})$$

$$O_{ij}^{RW,P_m} \sim \text{Binomial}(P_{ij}^{many}, N_{ij}^{P_m})$$

$$O_{ijd}^{URN,P_f} \sim \text{Multinomial}(P_{ijd}^{link,few}, N_{ij}^{P_f})$$

$$O_{ijd}^{URN,P_m} \sim \text{Multinomial}(P_{ijd}^{link,many}, N_{ij}^{P_m})$$

Figure 7.11: Graphical model of the Universal Thresholds Model (UTM), assumes θ_{many} and θ_{few}

simply assume that these parameters are independent of each other and that they have uniform priors over a large-enough interval of a priori plausible values.

$$\begin{aligned} &P(\alpha_{few}, \alpha_{many}, \theta_{\text{many}}, \theta_{\text{few}}, \sigma_U, \sigma_{E_i}) = \\ &\text{Uniform}_{[0;1]}(\alpha_{few}) \cdot \text{Uniform}_{[0;1]}(\alpha_{many}) \cdot \text{Uniform}_{[0;1]}(\theta_{\text{many}}) \cdot \\ &\text{Uniform}_{[0;1]}(\theta_{\text{few}}) \cdot \text{Uniform}_{[0;1]}(\sigma_U) \cdot \text{Uniform}_{[0;10]}(\sigma_{E_i}) \end{aligned} \tag{7.10}$$

Thirdly, we allow for two distinct weight parameters α_{few} and α_{many} to discover whether the influence of the two prior expectations is different for each of the quantity words. Last, we want to learn about θ_{many} and θ_{few} and whether the fixed threshold hypothesis holds also for the proportional reading. Can one pair account for the data from the two very distinct experiments? Another interesting aspect is a potential ambiguity between thresholds applying to P_U or P_E.

To address these questions, different model variants are compared regarding their fit to the experimental data. The formula in (7.9) assumes that one pair of threshold values θ_{many} and θ_{few} is sufficient. Let us call it the Uniform Threshold Model (UTM). The UTM assumes the CFK semantics' fixed threshold hypothesis of one pair of context-independent threshold values, which apply equally to P_U and P_E. The UTM can be compared with another model's fit to the data, which allows for one pair of thresholds per prior ($\theta_{\text{many,U}}$, $\theta_{\text{many,E}}$, $\theta_{\text{few,U}}$ and $\theta_{\text{few,E}}$), the Ambiguous Threshold Model (ATM). This model still assumes context-independent threshold parameters but leaves open the possibility of an ambiguity. This ambiguity, however, would not differentiate the cardinal from the proportional reading, but a judgment based on world knowledge from an uninformed judgment based on the numerical size of the proportion. If the thresholds on P_U and P_E converge to the same value, we take this as evidence for fixed threshold semantics à la Fernando and Kamp (1996). The ATM's posterior is defined as follows (see Figure 7.18 in Appendix 7.B for a graphical model version):

$$\begin{aligned} &P(\alpha_{few}, \alpha_{many}, \theta_{\text{many,U}}, \theta_{\text{many,E}}, \theta_{\text{few,U}}, \theta_{\text{few,E}}, \sigma_U, \sigma_{E_i} \mid O_{ij}^{p_m,RW}, O_{ij}^{p_f,RW}, O_{ij}^{\vec{p_m},U}, O_{ij}^{\vec{p_f},U}) \\ &\propto \; P(\alpha_{few}, \alpha_{many}, \theta_{\text{many,U}}, \theta_{\text{many,E}}, \theta_{\text{few,U}}, \theta_{\text{few,E}}, \sigma_U, \sigma_{E_i}) \cdot \\ &\; P(O_{ij}^{p_m,RW} \mid \alpha_{many}, \theta_{\text{many,U}}, \theta_{\text{many,E}}, \sigma_U, \sigma_{E_i}) \cdot P(O_{ij}^{p_f,RW} \mid \alpha_{few}, \theta_{\text{few,U}}, \theta_{\text{few,E}}, \sigma_U, \sigma_{E_i}) \cdot \\ &\; P(O_{ij}^{\vec{p_m},U} \mid \alpha_{many}, \theta_{\text{many,U}}, \theta_{\text{many,E}}, \sigma_U, \sigma_{E_i}) \cdot P(O_{ij}^{\vec{p_f},U} \mid \alpha_{few}, \theta_{\text{few,U}}, \theta_{\text{few,E}}, \sigma_U, \sigma_{E_i}) \end{aligned} \tag{7.11}$$

Finally, it is also imaginable that the proportional reading cannot be captured by a stable core meaning, contradicting the CFK semantics. In such a case the data would be explained best by a model variant which does not only allow for a

threshold per prior distribution but also further differentiates between the data sets. We will call this model the Individual Threshold Model (ITM) (see Figure 7.19 in Appendix 7.B for a graphical model version).

(7.12)

$$
\begin{aligned}
&P(\alpha_{few}, \alpha_{many}, \theta^{U}_{\text{many,U}}, \theta^{U}_{\text{few,U}}, \theta^{RW}_{\text{many,U}}, \theta^{RW}_{\text{few,U}}, \theta^{U}_{\text{many,E}}, \theta^{U}_{\text{few,E}}, \theta^{RW}_{\text{many,E}}, \theta^{RW}_{\text{few,E}}, \sigma_U, \sigma_{E_i} \\
&\quad \mid O_{ij}^{p_m,RW}, O_{ij}^{p_f,RW}, O_{ij}^{\vec{p_m},U}, O_{ij}^{\vec{p_f},U}) \propto \\
&\quad P(\alpha_{few}, \alpha_{many}, \theta^{U}_{\text{many,U}}, \theta^{U}_{\text{few,U}}, \theta^{RW}_{\text{many,U}}, \theta^{RW}_{\text{few,U}}, \theta^{U}_{\text{many,E}}, \theta^{U}_{\text{few,E}}, \theta^{RW}_{\text{many,E}}, \theta^{RW}_{\text{few,E}}, \sigma_U, \sigma_{E_i}) \cdot \\
&\quad P(O_{ij}^{p_m,RW} \mid \alpha_{many}, \theta^{RW}_{\text{many,U}}, \theta^{RW}_{\text{many,E}}, \sigma_U, \sigma_{E_i}) \cdot P(O_{ij}^{p_f,RW} \mid \alpha_{few}, \theta^{RW}_{\text{few,U}}, \theta^{RW}_{\text{few,E}}, \sigma_U, \sigma_{E_i}) \cdot \\
&\quad P(O_{ij}^{\vec{p_m},U} \mid \alpha_{many}, \theta^{U}_{\text{many,U}}, \theta^{U}_{\text{many,E}}, \sigma_U, \sigma_{E_i}) \cdot P(O_{ij}^{\vec{p_f},U} \mid \alpha_{few}, \theta^{U}_{\text{few,U}}, \theta^{U}_{\text{few,E}}, \sigma_U, \sigma_{E_i})
\end{aligned}
$$

We see that for each model variant a higher number of free parameters needs to be inferred. This is not theoretically motivated, rather it makes the model more complex and is thus not desirable. We see that we are facing the same problem as in Section 5.5, where we compared the GTM and the ITM to test the CFK semantics for the cardinal reading.

The question we are interested in is then: which model is best suited to explain the data? Statistical model comparison is the methodology of choice to address this question. Different arguments for preferring one model over another make use of different measures for model comparison (Vehtari and Ojanen, 2012). Given our modest theoretical purposes here, we use the same approach as in Chapter 5, the *deviance information criterion* (DIC) (Spiegelhalter et al., 2002; Plummer, 2008). This measure is easy to compute based on the output of our MCMC sampling results. The DIC weighs goodness of fit (here: the likelihood of the data given the model "trained" on the data) against the model's complexity (here: the number of its effective free parameters). A high value of the DIC indicates a lot of deviance of the model's predictions from the data it is applied to. This is undesirable, of course. At the same time, the model should stay as concise as possible and not include unnecessary parameters. This is measured by the pD, the number of effective free parameters, a measure of model complexity. Higher values of pD suggest higher model complexity.

To approximate the joint posterior distribution defined in (7.6) and compute the DIC, we used MCMC sampling, as implemented in JAGS (Plummer, 2003). Per model variant, we collected 10,000 samples from 2 MCMC chains after a burn-in of 10,000. This ensured convergence, as measured by $\hat{R}$ (Gelman and Rubin, 1992).

Table 7.2 gives estimated DICs for all three model variants. Given its high DIC value, the UTM has the worst fit to the data. The two candidates that are still in the running, ATM and ITM, are roughly equal in their DIC. The difference is less

UTM	ATM	ITM
DIC = 2879.7, $pD = 42.7$	DIC = 2761.7, $pD = 40.6$	DIC = 2751.5, $pD = 53.3$

Table 7.2: Estimated DIC values and effective free parameters for the three variants of the linear combination model

	α^{urns}	$\alpha^{\text{real-world}}$	θ_U	θ_E
few	0.687	0.383	0.374	0.046
many	0.713	0.388	0.596	0.623

Table 7.3: Estimated posteriors for weight and threshold parameters by the Ambiguous Thresholds Model (ATM)

than 1%. What the ATM misses in terms of goodness of fit, it makes up in terms of reduced model complexity. Consequently, there is no clear reason to prefer either model in terms of DICs. We follow the same line of reasoning as in Section 5.5 and interpret the result as there being no reason, provided by our data, to reject the "null assumption" that proportional *many* and *few* have a stable core meaning. The alternative model ITM did not do any better. Parallel to the cardinal reading, the ITM does not allow to generalize beyond the 24 contexts used here. Put differently, the ITM would assume that θ_{many} would be anywhere between 0 and 1 (its prior) for a context which was not part of the data used to condition it on. In contrast, the ATM would be able to use its posterior distribution for $\theta_{\text{many,U}}$ and $\theta_{\text{many,E}}$. The utter lack of generalizability in ITM speaks, at least conceptually, in favor of ATM. Whether this is an empirical advantage would have to be tested. Given the data at hand and the fact that the ITM is obviously not better for this data set, there is no good reason to dismiss the hypothesis that also the proportional reading has a stable core meaning. The data suggests that single quadruplet of fixed thresholds $\theta_{\text{many,U}}$, $\theta_{\text{many,E}}$, $\theta_{\text{few,E}}$ and $\theta_{\text{few,U}}$ may have generated the production data that we have seen. The posterior credible values inferred by the ATM are presented in Table 7.3.

We find that the influence of world knowledge is different for the two data sets, but its effect is the same for both *few* and *many*. The urn data is more influenced by the uninformed, flat prior P_U with a weight α of about 0.70. In the abstract scenario, the participants seemed to have used *few* and *many* to express that a proportion is numerically high or low, rather than being guided by their expectations of likely proportions (triggered by the ratio of balls in the urn). For the real-world contexts, we find that the influence of world knowledge rises, which corresponds to a lower α of about 0.38.

Even though we have concluded that the ATM is the model with the *best* data fit, this does not necessarily mean that its predictions describe the data well. A

sanity check of the model's fit to the data is a Posterior Predictive Check (PPC), see Section 4.1. A PPC tests whether the parameter values inferred in each step of the chain manage to predict the observed production behavior. For each set of parameter values inferred in one step of the MCMC chains, we have the likelihood function $\vec{P}_{speaker}$ predict a set of sampled observations. $\vec{O}^{p_f,U}_{\text{sample}}$ and $\vec{O}^{p_m,U}_{\text{sample}}$ are vectors containing sampled counts of ratings on a 7-point scale, for the same number of participants as in the production experiment. $\vec{O}^{p_f,RW}_{\text{sample}}$ and $\vec{O}^{p_m,RW}_{\text{sample}}$ contain sampled counts of TRUE ratings. For this chain of simulated data sets, we calculated the 95% highest density intervals (HDI) of their mean ratings for the urn data as well as the HDIs of the real-world data by running the function HDIofMCMC.R from Kruschke (2014). We then check which observed mean ratings fall into the predicted HDIs. If the model estimated the latent weight and threshold parameters well, the model should manage to predict all of the ratings we measured in the production task. For the urn data, 96% of the observed mean ratings fall into the predicted HDIs. For the real-world data, the model managed to correctly predict 93% of the data.

Figures 7.12 and 7.13 show the results of the PPC. The black circles are the participants' mean ratings in the judgment task. The error bars represent the 95% credible intervals of the sampled data's mean ratings per prior-proportion pair. The error bars are printed in black when the participants' mean ratings fall into the credible interval. It is printed in red when the model's predictions do not match the participants' ratings. The ATM model managed to predict all but two of the mean ratings for *few* and *many* in the urn scenario correctly. Of the real-world contexts, the model's predictions matched 243 out of 280 conditions. Another test of the model's fit to the data is to calculate the distribution of correlation coefficients between predicted data sets and observed data. For the urn data, the mean of this distribution was 0.84, and it was 0.92 for the real-world sentences. Next, we turn to the most credible threshold values inferred via Bayesian inference. That the ATM turned out to fit the data well enough supports the hypothesis that *few* and *many* have a stable core meaning. Table 7.3 presents the mean posterior distributions' mean values. The posterior values of $\theta_{\text{many,U}}$ and $\theta_{\text{many,E}}$ are very close and their HDIs overlap in [0.572 , 0.619]. When looking at these results for *many*, we might wonder why the UTM's fit to the data did not turn out better. For *many*, the same threshold seems to apply to P_U and P_E. However, the picture looks different for *few*. The posteriors' mean values as well as their HDIs are very different: $\theta_{\text{few,U}} \in [0.353, 0.397]$ and $\theta_{\text{few,E}} \in [0.017, 0.119]$. The proportional use of *few* cannot be explained with one threshold value; it is necessary to differentiate between contexts in which world knowledge is salient and contexts in which the sheer size of the proportion is described.

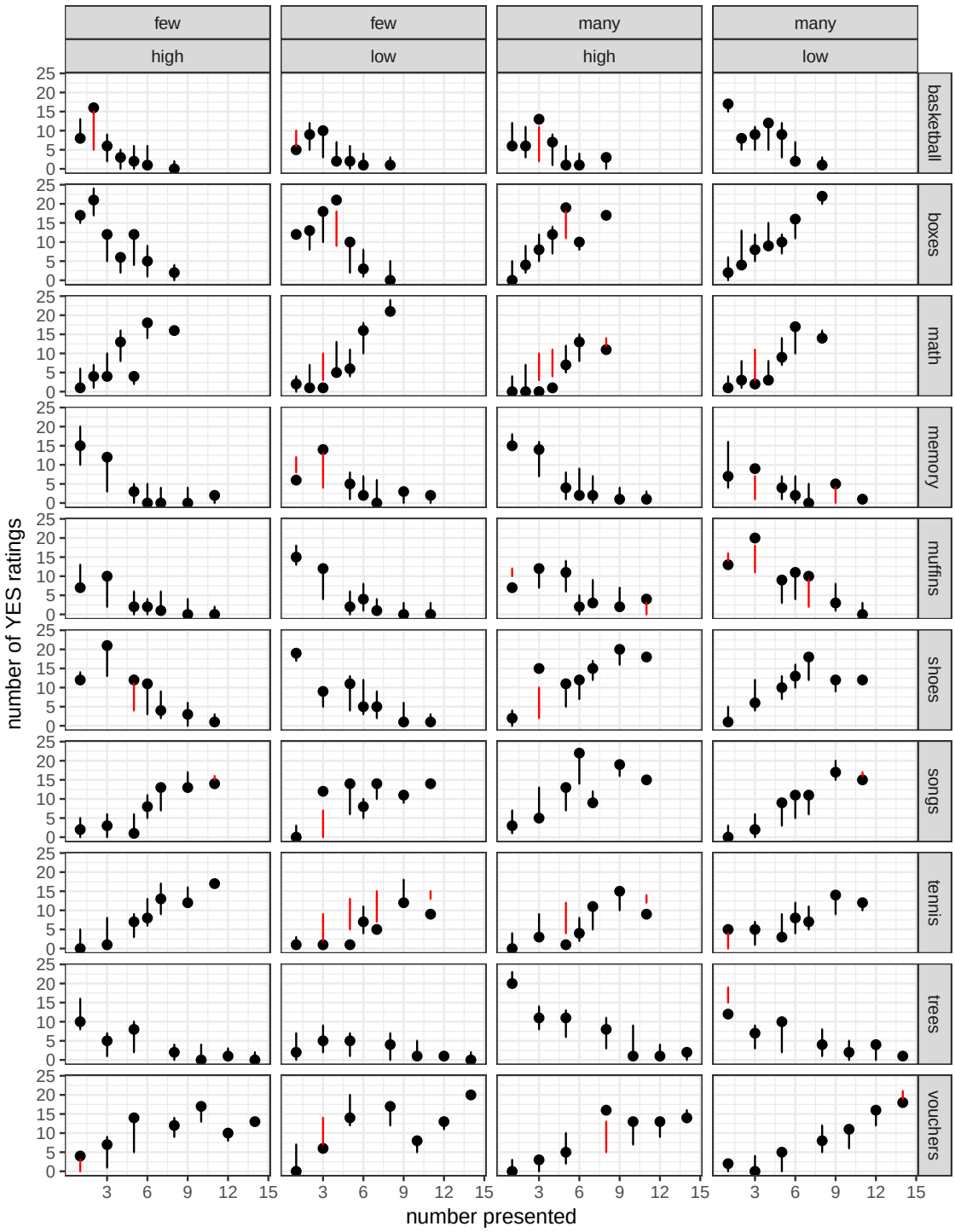


Figure 7.12: ATM's PPCs of real-world data. Bars are the 95% HDIs of the model's predictions, points are mean ratings measured experimentally (see Section 7.2.2). Bars are printed in red if the experimental data and the model's predictions do not coincide

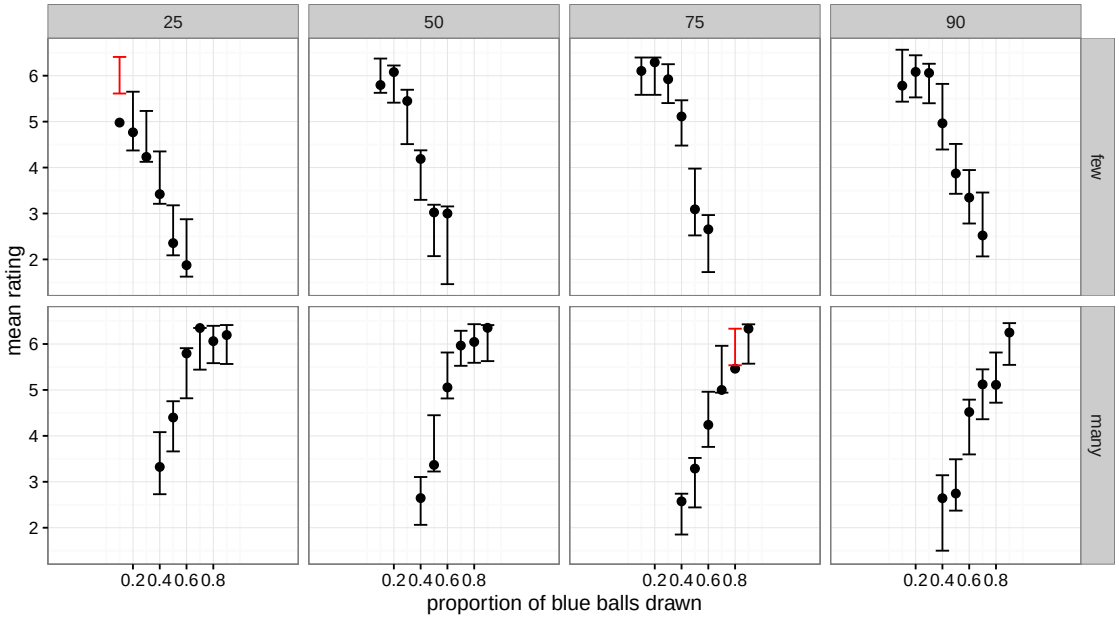


Figure 7.13: ATM's PPC of urn data. Bars are the 95% HDIs of the model's predicted mean ratings, points are mean ratings measured with in Section 7.3.1. Red bars indicate that the experimentally measured mean ratings do not fall into the PPC's HDIs.

7.7 Discussion

This chapter set out to identify two sources which influence the use of *few* and *many* when describing proportions. Furthermore, it investigated whether proportional *few* and *many* have a stable core meaning and whether this threshold is the same for expectations based on world knowledge and for an uninformed distribution over proportions.

7.7.1 Contextual Factors Influencing the Proportional Reading

The results of the model comparison support the ATM and lets us conclude that the findings for the cardinal reading in Chapter 5 transfer to the proportional reading. We identified a stable core meaning and showed that speaker behavior is correlated with prior expectations. For the proportional reading, the influence of world knowledge differs between contexts. The urn data is more influenced by the uninformed, flat prior with a weight α of about 0.7. In this abstract scenario, the participants seemed to have used *few* and *many* to mainly express that a proportion is numerically high or low, rather than being guided by their expectations of likely proportions (triggered by the ratio of balls in the urn). For the real-world contexts, world knowledge is more salient. We find that its influence rises, which corresponds to a lower

α of about 0.38. Interestingly, the salience of world knowledge has the same effect on *few* and *many*.

To sum up, we have seen that contextual information is employed both in the form of pure numerical information and also as world knowledge. As an additional sanity check of this assumption, future research might set out to test even more extreme prior distributions. So far, we have mainly presented participants with real-world contexts with rather left-skewed priors whereas the urn scenario used right-skewed prior distributions. Ideal candidates for scenarios for a follow-up experiments are contexts from Degen et al. (2015) in which expectations vary extremely depending on the noun in the minimal pair.

(134) a. CONTEXT: John threw 15 [ballons | cups | marbles] into a pool.

b. QUESTION: How many of the [ballons | cups | marbles] do you think sank?

With these extremely left- or right-skewed priors the context-dependence of α and the influence of a uniform prior (the numerical size of the proportion) on the proportional reading could be tested further. In our experiments, world knowledge was very influential in real-world contexts, but not in abstract scenarios. With Degen et al.'s (2015) contexts, we could check whether it is possible anyway to have the numerical size of the proportion described take over a greater role in real-world contexts (i.e. a high α) or whether world knowledge always holds the upper hand.

However, also other factors might play a role. In some contexts a speaker might choose a quantity word for reasons of politeness. He also might not know the exact cardinality of objects or, on the other side, only "express his value judgment at the number [of objects in question] more or less regardless of what that number is" (Keenan and Stavi, 1986).

7.7.2 A Possible (Lexical) Ambiguity of *few* and *many*

Next, we turn to the most credible threshold values inferred via Bayesian inference. Even though the hypothesis of a fixed threshold on prior expectations could be confirmed, the posterior distributions' mean values in Table 7.3 confront us once more with the by now familiar difference between *few* and *many*. The posterior values of $\theta_{\text{many,U}}$ and $\theta_{\text{many,E}}$ are very close and their HDIs overlap in [0.572 , 0.619]. This suggests a fixed threshold semantics for proportional *many* since the same threshold seems to apply to P_U and P_E. For *few*, the picture looks different again. The posteriors' mean values as well as their HDIs are very different: $\theta_{\text{few,U}} \in [0.353, 0.397]$ and $\theta_{\text{few,E}} \in [0.017, 0.119]$. The proportional use of *few* cannot be explained with one threshold value so that a fixed threshold hypothesis for proportional *few* needs to be questioned. A higher threshold seems to be applied when *few* is used to describe

the size of a proportion than when it compares a proportion with beliefs about its size. A reason why the lower threshold of $\theta_{\text{few,E}}$ is not compatible with P_U is that *few* likes to be applied to small proportions but not necessarily to small *numbers*. We conclude this from the fact that the same proportion is rated significantly higher if it corresponds to a larger number, see Section 7.3. For low numbers, *few* competes with alternatives like *no, a few, some* and number words, especially when a plural noun used to describe a single ball results in a number mismatch. This is why a threshold value of about 0.05 predicts too low numbers to be applicable to P_U. Since many of the prior expectations in the real-world contexts are left-skewed (see Figure 7.2), a low threshold on the cumulative density mass can still allow for higher numbers.

Nevertheless, the PPC and correlation scores show that the ATM makes very good predictions. Even though the values for $\theta_{\text{few,U}}$ and $\theta_{\text{few,E}}$ differ, we find that the strategy of applying a fixed threshold to a distribution representing prior expectations seems to be employed across contexts, for both *few* and *many* in cardinal and proportional readings. The ITM, which does not require this strategy, does not fit the data any better. The puzzling result of the incompatible threshold parameters for proportional *few* could also be due to methodological issues as discussed in the next section.

Another interesting observation is that even though there is evidence that one value θ_{many} from the interval [0.572 , 0.619] can explain the production of proportional *many*, this value is lower than the cardinal threshold which is predicted to fall into the interval [0.687 , 0.699], see Section 5.5. Several explanations are conceivable. The first would be to assume a lexical ambiguity between cardinal and proportional *many*, as proposed by Partee (1989) and Krasikova (2011) among others. It is also possible, however, that the different threshold values are not due to the two readings but due to the data sets on the basis of which these parameters were inferred. For the proportional reading, only production data was used whereas the model for the cardinal case made predictions for both production and interpretation. As already discussed in Section 6.7 on *surprisingly*, the threshold values seem to be more "extreme" when interpretation data is involved. In interpretation tasks, participants seem to play it safe and choose lower numbers as the interpretation of *few* and higher numbers for *many* than they rate to be true in a production task. This greater freedom of choice might result in different threshold values when also interpretation data is predicted by the model. We consider both options to be plausible explanations for the discrepancy between cardinal and proportional θ_{many}. For now, we do not want to commit to any of them since our observations are made on the basis of very limited data sets. We propose to further investigate the difference

between cardinal and proportional *many* and *few* by collecting more data to validate the models' predictions.

7.7.3 Measuring Priors in Abstract Contexts

The attentive reader might have noticed that in the urn experiment from Section 7.3 we departed from our usual procedure. Prior expectations were not measured experimentally, but we simply assumed that participants made use of the normative hypergeometrical distribution of blue balls in the draw, see Section 7.3.2. The reasoning behind this decision was that we assumed that in such an abstract context there is less variation between the individual participants' expectations since they are not influenced by their world knowledge. To test this we ran a follow-up study which elicited participants' prior expectations of the number of blue balls in a draw from an urn of varying content.

Design. To measure participants' prior expectations in the abstract urn scenario from Section 7.3 and to test whether they really employ the hypergeometrical distribution we have assumed, we again used a slider-rating task (e.g., Kao et al., 2014; Franke et al., 2016). Participants were presented with an urn which contained 100 balls. Prior expectations were manipulated by varying the content of the urn. [25|50|75|90] of the balls were blue, the rest red. A character then drew 10 balls from the urn. Subjects were presented with 11 slider-interval pairs, labeled from 0 to 10, and rated the likelihood that the draw contains the respective number of blue balls, by adjusting a slider labeled from "extremely unlikely" to "extremely likely". We formulated the task in a way to make the prior salient:

(135) TASK: For a draw from an urn with this content, please rate how likely it is that from the 10 balls the following numbers are blue.

Participants. 25 subjects were recruited via Amazon's Mechanical Turk with US-IP addresses.

Materials & Procedure. After initial instructions that explained the task, each subject saw the four prior conditions, [25|50|75|90] blue balls of 100 balls in the urn, rest red, one after another in a random order. We used the same images of the urns as in the production study, see Figure 7.4. For each prior condition, the 11 intervals were presented horizontally on the screen in ascending order from left to right. On top of each interval was a vertical slider. Participants had to adjust or at least click on each slider before being able to proceed.

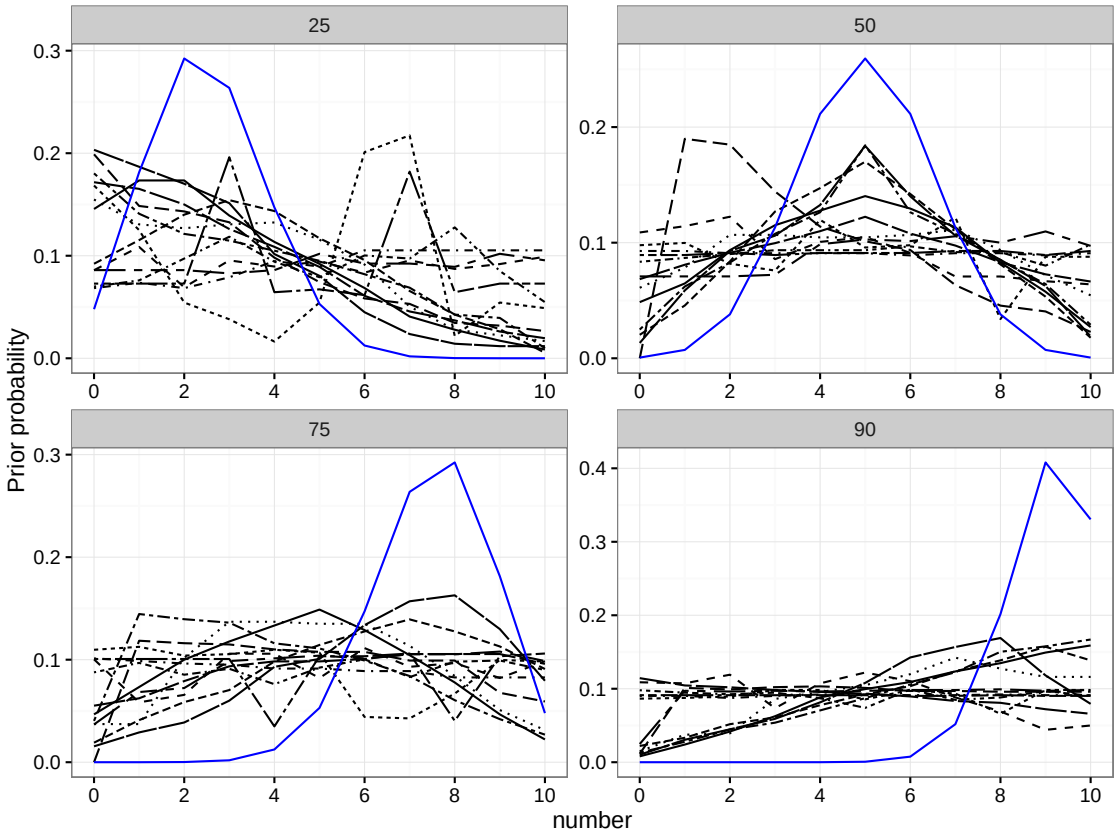


Figure 7.14: Measured (one black line per participant) and normative (blue line) prior expectations in the abstract urn context

Results. Participants' ratings per item were normalized by subject-prior-condition. The black lines in Figure 7.14 show the individual distributions. We can see that they show inter-subjective variance and that they appear to differ significantly from the normative hypergeometrical distribution printed in blue.

These results pose a serious problem to our computational modeling and Bayesian inference approach. Prior expectations are vital input for these methods and if this input cannot be trusted to represent participants' beliefs, the models' predictions and the inferred parameter values cannot be used to draw conclusions about the behavior of speakers and listeners. If we choose to use the experimental priors as input for the model, predictive success of the model decreases. The model with experimentally measured priors has a DIC of 1378.8 with pD = 16.0, which is higher than when the normative priors were used (DIC = 1349.0 and pD = 22.0). Furthermore, the inferred posteriors predicted for the urn data set become very implausible. The threshold values for *few* are predicted to be far too high to be realistic, $\theta_{\text{few,U}} \in [0.608, 0.991]$ and $\theta_{\text{few,E}} \in [0.328, 0.372]$, whereas the threshold values for *many* come out too low, $\theta_{\text{many,U}} \in [0.027, 0.495]$ and $\theta_{\text{many,E}} \in [0.674, 0.742]$. Additionally, also the noise parameters σ_E of the production rule in Equation 7.8 are extremely high with values close to 1. This means that the uncertainty about

the thresholds $k_{\max}$ and $k_{\min}$ could be so large that the standard deviation of the Gaussian noise spans the entire interval of proportion. The result would be a very flat production rule, which in turn makes very imprecise predictions. Given these dubious results, we doubt that the slider task was a reliable measure to elicit representations of participants' prior expectations in these abstract contexts. In the following, we want to discuss several aspects with respect to the discrepancy between the experimentally measured and the normative priors.

A first problem might have been caused by the experimental setup of the prior elicitation task. So far the slider task asking for likelihood judgments from Kao et al. (2014) has produced reliable results, however, and Franke et al. (2016) showed that using averaged normalized slider ratings for binned quantities can fairly well approximate inferred population-level beliefs. Furthermore, Herbstritt and Franke (2016) successfully used a slider-based task to assess the mode of beliefs in even more complex urn scenarios and the results reported do not diverge substantially from the hypothesized normative priors. It seems that participants are (on average) able to form judgments about the likelihood of such abstract events.

In the present case, the puzzling results might have been influenced by other factors. It is possible that we did not make the task clear enough, given the abstract scenario with which participants are probably not familiar. Moreover, a test round might have been useful to familiarize participants with the material and the task of giving likelihood judgments, especially since the experiment was very short and a training effect cannot have ensued. In order to test whether the unexpected results in the previous experiment were caused by a too demanding or unclear task, we conducted an additional study. We had subjects choose the outcome of the draw they consider most likely as well as the range of expected numbers of blue balls in the draw, instead of asking for likelihood judgments. We checked whether their answers are compatible with the hypergeometrical distribution we assumed.

Design. To test a possible task-effect on the discrepancy between the probability distributions measured with the slider-task from Kao et al. (2014) above and the normative hypergeometrical distribution, we opted for a different, probably conceptually easier dependent measure. Once more, participants were presented with an urn which contained 100 balls. We manipulated prior expectations by showing urns of varying content. [25|50|75|90] of the balls were blue, the rest red. A character then drew 10 balls from the urn. Subjects were asked to give three judgments: the number of blue balls they expect the character to draw, as well as the lowest and highest number of blue balls the character might probably draw. For each answer, subjects chose a number between 0 and 10 by adjusting a slider on a scale (see Figure 7.15).

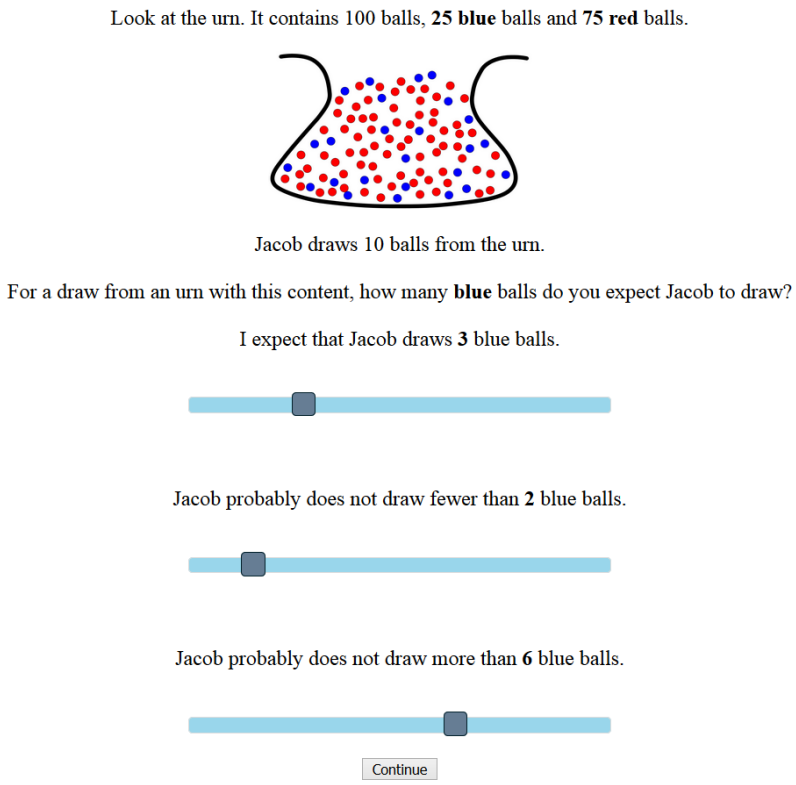


Figure 7.15: Sample item in the prior elicitation task

Participants. 50 subjects were recruited via Amazon's Mechanical Turk with US-IP addresses.

Materials & Procedure. Two subjects were excluded from the data analysis because they reported not to be native speakers of English. After reading an explanation, each subject was presented with the four prior conditions, [25|50|75|90] blue balls of 100 balls in the urn, rest red, in a random order. We used the same images of the urns as in the previous prior elicitation task. For each prior condition, participants provided their answers by adjusting three vertical sliders on the screen, one for the number of blue balls in the draw they consider to be most likely, and one each for the lowest and highest number they expect. The sliders ranged from 0 to 10. We accepted only those answers in which the lowest number, the most likely number and the highest number were identical or in ascending order. Only if an acceptable triplet of numbers was selected, participants could proceed to the next trial.

Results. The histogram in Figure 7.16 shows the frequency distribution of expected numbers of blue balls. The mode of the ratings coincides with the number which the hypergeometrical distribution assigns the highest probability in all urn conditions but one. For the urn containing 75 blue balls, the mode of the ratings is

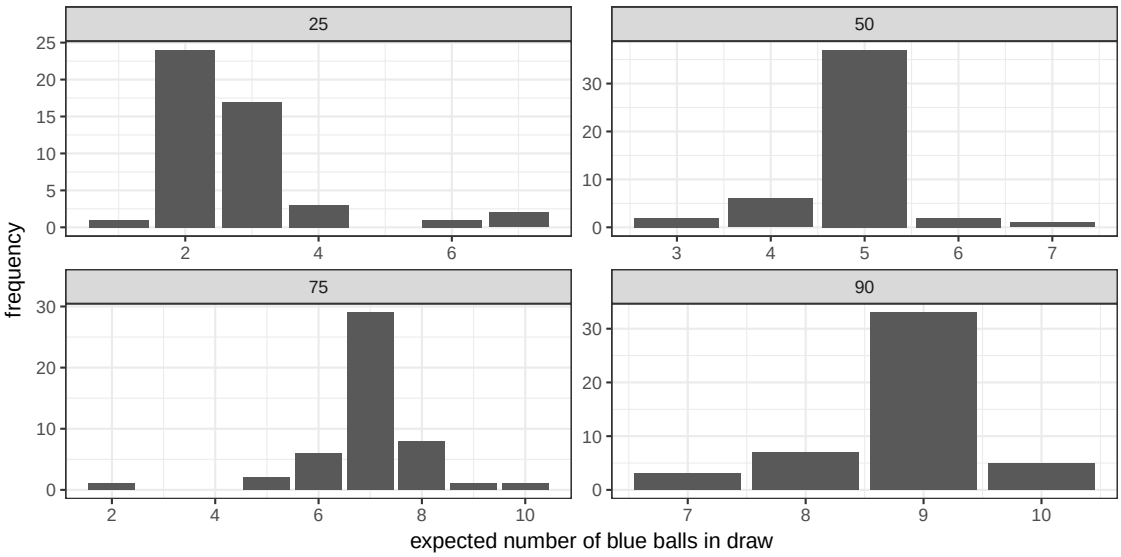


Figure 7.16: Frequency distribution of the expected number of blue balls in the draw

at seven whereas eight blue balls are most likely for a normative prior. Moreover, we analyzed the intervals that participants considered likely outcomes of the draw. We considered a flat distribution with equal, non-zero mass on all the numbers between the lowest and highest expected number, assuming that the remaining numbers are considered unlikely and therefore assigned probability 0. For example, when a participant answered three as the lowest and seven as the highest number, all five numbers in the interval [3, 7] were assigned probability 0.2. If the interval were [6, 9], the probability of each number in the interval would be 0.25. We then summed up all probabilities per condition and normalized again. The resulting distributions are presented in Figure 7.17. When comparing them to the normative hypergeometrical distributions plotted as the dashed lines, we see that the difference between the experimentally measured, average prior expectations and the normative prior is much smaller then in the previous experiment.

We conclude that the participants in the experiment, at least on the aggregate population-level, were able to form prior expectations about the outcome of the draw and that these expectations are (at least on average) compatible with the normative priors we assumed, hypergeometrical distributions. These findings are in line with Herbstritt and Franke (2016). Nevertheless, the elicitation of prior expectations is quite a young field of study. More work is necessary to understand the influence of the different task types on the elicitation of prior expectations as well as the relationship between their mental representations and experimental data.

To conclude, the Bayesian analysis of the experimental data confirmed the hypothesis that the use of proportional *few* and *many* is both influenced by prior beliefs

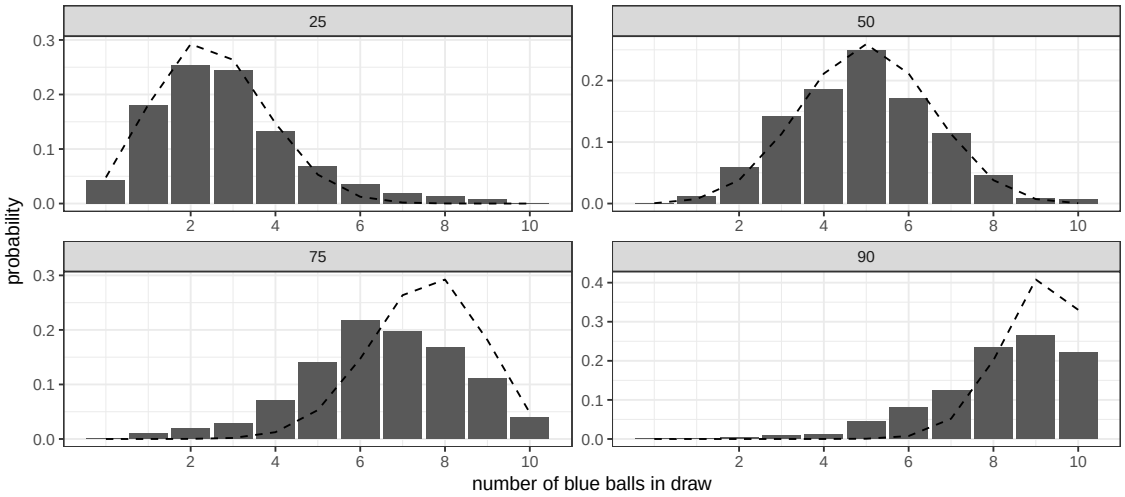


Figure 7.17: Probabilities derived from participants' judgments of plausible intervals (bars) and normative hypergeometrical distribution (dashed lines) in the abstract urn context

about likely proportions as well as the numerical size of the proportion. The salience of world knowledge can be modeled with a linear combination of speaker probabilities based on these two distributions. For *many*, a fixed threshold hypothesis could be confirmed for both the cardinal and the proportional reading even though more research needs to be conducted to draw conclusions about a potential lexical ambiguity. For *few*, the same strategy seems to be at work, applying a fixed threshold to prior expectations. As in several other cases, however, *few* seems to behave differently from *many* and the CFK semantics can neither be confirmed nor rejected for proportional *few*. As pointed out above, it is therefore worthwhile to follow up on the presented experiments and a Bayesian analysis of the data. To avoid the problem of normative priors or normalized, population-level priors, a follow-up experiment could elicit a participant's contextual expectations in a controlled lab experiment and then have the same participant carry out a production and interpretation task of *many* and *few* using the same contexts. This way we hope to gain further insight, especially about a lexical ambiguity and about proportional *few*'s fixed thresholds.

7.A Experimental Material: Real-World Contexts

7.A.1 Interpretation Study

1. **basketball** — Alex took part in a basketball competition and was allowed 9/12 shots from the three-point line. — HIGH: Alex, who is a professional player, made many (of the) shots. — LOW: Alex, who is an amateur player,

made many (of the) shots. — How many (of the) shots do you think Alex made?

2. **boxes** — When moving to a new flat, Martha packed 15/20 boxes. — HIGH: Martha, who is a strong woman, carried many (of the) boxes herself. — LOW: Martha, who is a weak woman, carried many (of the) boxes herself. — How many (of the)boxes do you think Martha carried?

3. **cinema** — 30/40 people attended the late-night performance in a small cinema. — HIGH: At the end of the movie, which was longer than people had expected it to be, many (of the) people had fallen asleep. — LOW: At the end of the movie, which was shorter than people had expected it to be, many (of the) people had fallen asleep. — How many (of the) people do you think fell asleep?

4. **math** — A math teacher presented a tricky problem to the 18/24 students in his course. — HIGH: Many (of the) students in his course, which focuses on problem-solving strategies, could solve the problem. — LOW: Many (of the) students in his course, which does not teach problem-solving strategies, could solve the problem. — How many (of the) students do you think could solve the problem?

5. **memory** — For a memory test 9/12 three-digit numbers were read out to Chris. — HIGH: Chris, who has a great memory, memorized many (of the) numbers. — LOW: Chris, who has a bad memory, memorized many (of the) numbers. — How many (of the) numbers do you think Chris memorized?

6. **muffins** — There were 9/12 muffins on the kitchen table in Ed's flat. — HIGH: Ed, who arrived feeling hungry, ate many (of the) muffins. — LOW: Ed, who arrived feeling full, ate many (of the) muffins. — How many (of the) muffins do you think Ed ate?

7. **raffle** — Deborah bought 9/12 tickets in a raffle. — HIGH: Many (of the) tickets bought by Deborah, who is always lucky, were winning tickets. — LOW: Many (of the) tickets bought by Deborah, who is never lucky, were winning tickets. — How many (of the) tickets that Deborah bought do you think were winning tickets?

8. **shoes** — Melanie had to choose which among 9/12 pairs of shoes to bring on holiday. — HIGH: Melanie, who loves fashion, packed many (of the) pairs of shoes. — LOW: Melanie, who doesn't care about fashion, packed many (of the) pairs of shoes. — How many (of the) pairs of shoes do you think Melanie packed?

9. **slats** — Jimmy jumped onto grandma's old slatted bed frame which only had 18/24 slats left. — HIGH: Jimmy, who is a fat boy, broke many (of the) slats. — LOW: Jimmy, who is a skinny boy, broke many (of the) slats. — How many (of the) slats do you think Jimmy broke?

10. **songs** — In a music quiz the beginnings of 9/12 pop songs were played. — HIGH: Heidi, who loves pop songs, recognized many (of the) songs. — LOW: Heidi, who hates pop songs, recognized many (of the) songs. — How many (of the) songs do you think Heidi recognized?

11. **tennis** — Bruno played 12/16 tennis matches last season. — HIGH: Bruno, who is an unathletic person, lost many (of the) matches. — LOW: Bruno, who is a fit person, lost many (of the) matches. — How many (of the) matches do you think Bruno lost?

12. **tents** — On a camping trip 15/20 tents had to be put up. — HIGH: Dave, who loves camping, pitched many (of the) tents. — LOW: Dave, who doesn't like camping, pitched many (of the) tents. — How many (of the) tents do you think Dave pitched?

13. **training** — A football coach named Max invited 12/16 boys to come to practice training. — HIGH: Max, who is an easy-going coach, allowed many (of the) boys to come back in the next week. — LOW: Max, who is a strict coach, allowed many (of the) boys to come back in the next week. — How many (of the) boys do you think were allowed to come back in the next week?

14. **trees** — Jim had 15/20 trees in his garden. — HIGH: Jim, who is a strong man, cut down many (of the) trees. — LOW: Jim, who is a weak man, cut down many (of the) trees. — How many (of the) trees do you think Jim cut down?

15. **vacuum cleaner** — Walter is a door-to-door salesman. Yesterday he presented a vacuum cleaner in 18/24 households. — HIGH: Walter, who offered his product at a low price, sold a vacuum cleaner to many (of the) households. — LOW: Walter, who offered his product at a high price, sold a vacuum cleaner to many (of the) households. — To how many (of the) households do you think Walter sold a vacuum cleaner?

16. **vouchers** — Carla won 9/12 vouchers for roller coaster rides on a fair. — HIGH: Carla, who is an adventurous person, used many (of the) vouchers. — LOW: Carla, who is a fearful person, used many (of the) vouchers. — How many (of the) vouchers do you think Carla used?

7.A.2 Judgment Task

1. **basketball** — Alex took part in a basketball competition and was allowed 9 shots from the three-point line. Alex is a [professional |amateur] player. He made [1 |2 |3 |4 |5 |6 |8] of the shots. — For a [professional |amateur] player, Alex made [few |many] of the shots.

2. **boxes** — When moving to a new flat, Martha packed 15 boxes. Martha is a [strong |weak] woman. She carried [1 |3 |5 |8 |10 |12 |14] of the boxes herself. — For [strong |weak] woman, Martha carried [few |many] of the boxes herself.

3. **math** — A math teacher presented a tricky problem to the 24 students in his course. The course [focuses on |does not teach] problem-solving strategies. [2-3 |6-7 |10-11 |12-13 |14-15 |18-19 |22-23] of the students could solve the problem. — For a course which [focuses on |does not teach] problem-solving strategies, [few |many] students could solve the problem.

4. **memory** — For a memory test 9 three-digit numbers were read out to Chris. Chris has a [great |bad] memory. He remembered [1 |2 |3 |4 |5 |6 |8] of the numbers. — For a man with a [great |bad] memory, Chris memorized [few |many] of the numbers.

5. **muffins** — There were 12 muffins on the kitchen table in Ed's flat. Ed arrived feeling [hungry |full]. He ate [1 |3 |5 |6 |7 |9 |11] of the muffins. — For a man feeling [hungry |full], Ed ate [few |many] of the muffins.

6. **shoes** — Melanie had to choose which among 12 pairs of shoes to bring on holiday. Melanie [loves |doesn't care about] fashion. She packed [1 |3 |5 |6 |7 |9 |11] of the pairs of shoes. — For a woman who [loves |doesn't care about] fashion, Melanie packed [few |many] of the shoes.

7. **songs** — In a music quiz the beginnings of 9 pop songs were played. Heidi [loves |hates] pop songs. She recognized [1 |2 |3 |4 |5 |6 |8] of the songs. — For a pop song [lover |hater], Heidi recognized [few |many] of the songs.

8. **tennis** — Bruno played 12 tennis matches last season. Bruno is an [unathletic |fit] person. He lost [1 |3 |5 |6 |7 |9 |11] of the matches. — For an [unathletic |fit] person, Bruno lost [few |many] of the matches.

9. **trees** — Jim had 15 trees in his garden. Jim is a [strong |weak] man. He cut down [1 |3 |5 |8 |10 |12 |14] of the trees. — For a [strong |weak] man, Jim cut down many of the trees.

10. **vouchers** — Carla won 12 vouchers for roller coaster rides on a fair. Carla is an [adventurous |fearful] person. She used [1 |3 |5 |6 |7 |9 |11] of the vouchers. — For a [adventurous |fearful] person, Carla used [few |many] of the vouchers.

7.A.3 Prior Elicitation Study

1. **basketball** — Alex took part in a basketball competition and was allowed 9 shots from the three-point line. — Alex is a [professional |amateur] player. — How many of the shots do you think Alex made?

2. **boxes** — When moving to a new flat, Martha packed 15 boxes. — Martha is a [strong |weak] woman. — How many of the boxes do you think Martha carried?

3. **math** — A math teacher presented a tricky problem to the 24 students in his course. — The course [focuses on |does not teach] problem-solving strategies. — How many of the students do you think could solve the problem?

4. **memory** — For a memory test 9 three-digit numbers were read out to Chris. — Chris has a [great |bad] memory. — How many of the numbers do you think Chris memorized?

5. **muffins** — There were 12 muffins on the kitchen table in Ed's flat. — Ed arrived feeling [hungry |full]. — How many of the muffins do you think Ed ate?

6. **shoes** — Melanie had to choose which among 12 pairs of shoes to bring on holiday. — Melanie [loves |doesn't care about] fashion. — How many of the pairs of shoes do you think Melanie packed?

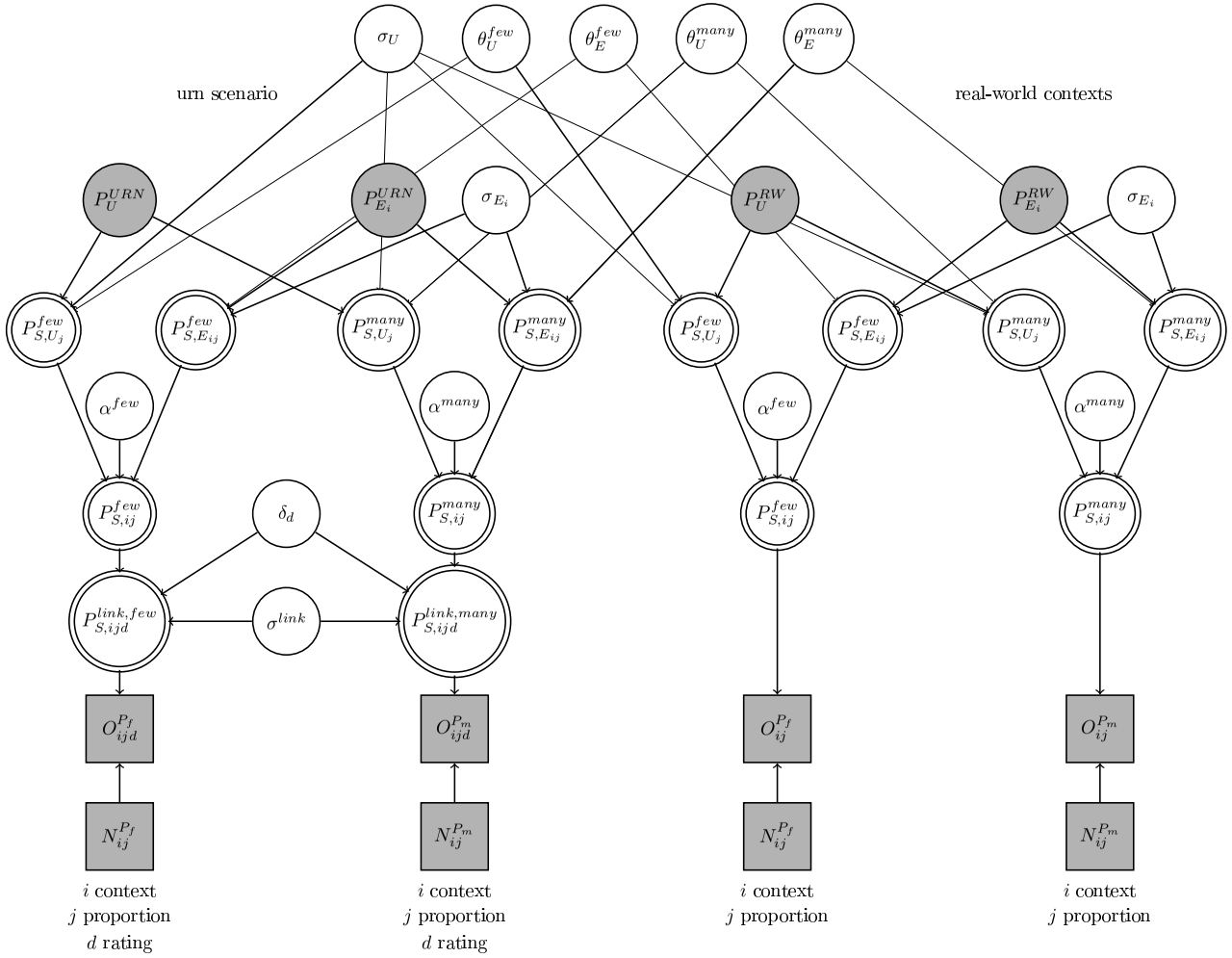
7. **songs** — In a music quiz the beginnings of 9 pop songs were played. — Heidi [loves |hates] pop songs. — How many of the songs do you think Heidi recognized?

8. **tennis** — Bruno played 12 tennis matches last season. — Bruno is an [unathletic |fit] person. — How many of the matches do you think Bruno lost?

9. **trees** — Jim had 15 trees in his garden. — Jim is a [strong |weak] man. — How many of the trees do you think Jim cut down?

10. **vouchers** — Carla won 12 vouchers for roller coaster rides on a fair. — Carla is an [adventurous |fearful] person. — How many of the vouchers do you think Carla used?

7.B Graphical models of the Ambiguous Thresholds Model (ATM) and of the Individual Thresholds Model (ITM)



$$\theta_E^{few/many}, \theta_U^{few/many} \sim \text{Uniform}(0,1)$$

$$\sigma_{E_i} \sim \text{Uniform}(0,10)$$

$$\sigma_U, \sigma^{link} \sim \text{Uniform}(0,1)$$

$$P_{S,E_{ij}}^{few/many} \text{ see main text}$$

$$P_{S,U_j}^{few/many} \text{ see main text}$$

$$\alpha^{few}, \alpha^{many} \sim \text{Uniform}(0,1)$$

$$P_{ij}^{few} = \alpha^{few} \cdot P_{S,U_j}^{few} + (1-\alpha^{few}) \cdot P_{S,E_{ij}}^{few}$$

$$P_{S,ij}^{many} = \alpha^{many} \cdot P_{S,U_j}^{many} + (1-\alpha^{many}) \cdot P_{S,E_{ij}}^{many}$$

$$\delta_d \sim \text{Normal}(d/7, (2\cdot 7)^{-1})$$
$$\delta_0 = -\infty, \delta_7 = \infty$$

$$P_{S,ijd}^{link,few/many} = \int_{\delta_{d-1}}^{\delta_d} \text{Normal}(x, P_{S,ij}^{few/many}, \sigma^{link}) \mathrm{d}x$$

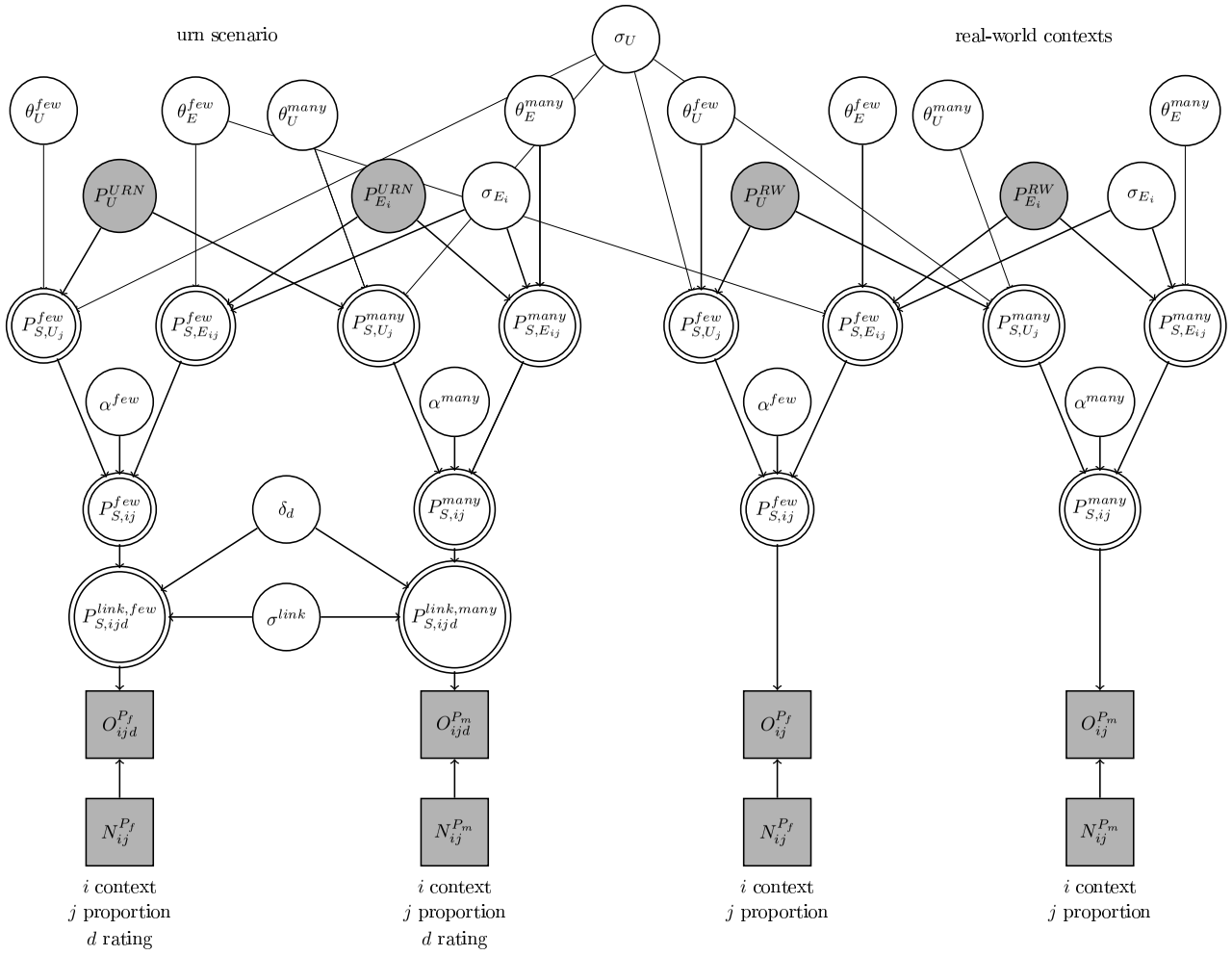
$$O_{ij}^{RW,P_f} \sim \text{Binomial}(P_{S,ij}^{few}, N_{ij}^{P_f})$$

$$O_{ij}^{RW,P_m} \sim \text{Binomial}(P_{S,ij}^{many}, N_{ij}^{P_m})$$

$$O_{ijd}^{URN,P_f} \sim \text{Multinomial}(P_{S,ijd}^{link,few}, N_{ij}^{P_f})$$

$$O_{ijd}^{URN,P_m} \sim \text{Multinomial}(P_{S,ijd}^{link,many}, N_{ij}^{P_m})$$

Figure 7.18: Graphical model of the Ambiguous Thresholds Model (ATM), assumes $\theta_{\text{many,U}}$, $\theta_{\text{many,E}}$, $\theta_{\text{few,U}}$ and $\theta_{\text{few,E}}$



$$\theta_E^{few/many}, \theta_U^{few/many} \sim \text{Uniform}(0,1)$$

$$\sigma_{E_i} \sim \text{Uniform}(0,10)$$

$$\sigma_U, \sigma^{link} \sim \text{Uniform}(0,1)$$

$$P_{S,E_{ij}}^{few/many} \text{ see main text}$$

$$P_{S,U_j}^{few/many} \text{ see main text}$$

$$\alpha^{few}, \alpha^{many} \sim \text{Uniform}(0,1)$$

$$P_{S,ij}^{few} = \alpha^{few} \cdot P_{S,U_j}^{few} + (1-\alpha^{few}) \cdot P_{S,E_{ij}}^{few}$$

$$P_{S,ij}^{many} = \alpha^{many} \cdot P_{S,U_j}^{many} + (1-\alpha^{many}) \cdot P_{S,E_{ij}}^{many}$$

$$\delta_d \sim \text{Normal}(d/7, (2\cdot 7)^{-1})$$
$$\delta_0 = -\infty, \delta_7 = \infty$$

$$P_{S,ijd}^{link,few/many} = \int_{\delta_{d-1}}^{\delta_d} \text{Normal}(x, P_{S,ij}^{few/many}, \sigma^{link}) \text{d}x$$

$$O_{ij}^{RW,P_f} \sim \text{Binomial}(P_{S,ij}^{few}, N_{ij}^{P_f})$$

$$O_{ij}^{RW,P_m} \sim \text{Binomial}(P_{S,ij}^{many}, N_{ij}^{P_m})$$

$$O_{ijd}^{URN,P_f} \sim \text{Multinomial}(P_{S,ijd}^{link,few}, N_{ij}^{P_f})$$

$$O_{ijd}^{URN,many} \sim \text{Multinomial}(P_{ijd}^{link,many}, N_{ij}^{P_m})$$

Figure 7.19: Graphical model of the Individual Thresholds Model (ITM), assumes that $\theta_{\text{many,U}}$, $\theta_{\text{many,E}}$, $\theta_{\text{few,U}}$ and $\theta_{\text{few,E}}$ further differ per data set

Chapter 8

Concluding Remarks

8.1 Summary and Conclusions

This dissertation set out to investigate how the context-dependent quantity words *few* and *many* receive their meaning in context. Concretely, we tested a particular formalization of one theory by Fernando and Kamp (1996) which makes precise predictions about how the contextual information might be integrated in the semantics. We called it CFK semantics because it goes back to ideas from Clark (1991) and Fernando and Kamp (1996). This theory assumes that the "surprise reading" of *few* and *many* expresses that a number or a proportion is lower or higher than expected. Prior expectations of the context are formalized as probability distributions P_E over cardinalities and the cardinalities which count as *few* or *many* are determined by applying fixed, context-independent thresholds θ_{few} and θ_{many} to the cumulative density mass of these distributions. In other words, *few* and *many* comprise a stable core meaning, which explains why speakers and listeners manage to successfully communicate with these context-dependent expressions and how children can acquire proficiency in their use. Even though the quantity words are assumed to have a fixed meaning, their denotation can vary to an extreme degree because the contextual input, prior expectations P_E, may be dramatically different depending on the context.

Fernando and Kamp's (1996) surprise-based semantics may seem intuitively appealing, but it is hard to test it with the standard methods of the field. The threshold values θ_{few} and θ_{many} cannot be directly measured nor can their existence or uniqueness be validated based on intuitions alone. For this reason, we treat them as latent parameters in a probabilistic model of language use whose values are estimated based on experimental data by applying Bayesian inference. We measure representations of prior expectations by applying recent experimental methodology

(Kao et al., 2014; Franke et al., 2016) and conduct production and interpretation experiments.

The cardinal surprise reading of *few* and *many* was investigated first. We showed how Fernando and Kamp's (1996) theory can be couched in a computational model. The model was used to infer those values for θ_{few} and θ_{many} which, by taking experimentally measured prior expectations as input, are most likely to have generated the production and interpretation data we measured. To test the CFK semantics, we compared two variants of the model. The first, the General Thresholds Model (GTM), assumed one pair of fixed threshold values, whereas the Individual Thresholds Model (ITM) allowed for an individual value per context. Given their nearly identical fit to the data, as measured by DIC, and the fact that the ITM is not theoretically motivated and more complex because of the higher number of free parameters, the GTM is preferred. Consequently, the data-driven computational modeling approach supports a cardinal surprise reading of *few* and *many* which expresses that a cardinality is higher than a fixed threshold on a measure of surprise. What is surprising is in turn dependent on the contextual contribution in the form of prior expectations.

The existence of a cardinal surprise reading brings up another interesting topic. Since *few* and *many* can express that a cardinality is surprisingly low or high anyway, does it make a difference if surprise is overtly marked? To answer this question, the quantity words were combined with the adverb *surprisingly*. We presented two possible views on the influence of *surprisingly*. On the one hand, it could simply function like a frame setter and mark a comparison class of expectations in respect to which the quantity word is evaluated. If this is the case, the production of sentences with *surprisingly few/many* should not be different from sentences in which expectations are explicitly made reference to by a *compared to* phrase (for example, "compared to what you would expect for a man from the US"). On the other hand, the presence of *surprisingly* could have an intensifying effect on *few* and *many*, just as Bennett and Goodman (2015) predict for the adverb *incredibly*. To discriminate between the two views, we collected production data of sentences with and without the three modifiers and applied a production model incorporating the CFK semantics for cardinal *few* and *many* to see if the modifiers have an effect on the threshold values.

The data from the production task was analyzed both with a linear mixed effects regression model and the computational model, delivering conflicting and surprising results. To start with the uncontroversial findings, there was no significant difference between ratings of sentences with unmodified quantity words and ratings of sentences in which a *compared to* phrase made reference to expectations. Moreover, their threshold values' HDIs overlapped. We take this as support for our assumption

that *few* and *many*'s most salient reading is the surprise reading. When it comes to *surprisingly*, though, the results are less clear. The adverb seems to intensify the meaning of *many* but not of *few*. For *many*, *surprisingly* patterns with *incredibly*, resulting in lower ratings and a higher threshold than for sentences with unmodified *few* and *many*. For *few*, on the other hand, we find no difference between sentences with unmodified *few* and *surprisingly few*. The inferred threshold's HDI for *surprisingly few* overlaps both with the HDIs of the thresholds for *incredibly few* and unmodified *few*, identifying it neither as an intensifier nor contradicting this option. That the adverb *surprisingly* might have a different effect on *many* than on *few* was not expected, neither was it predicted by the semantic literature. This is not the only time where we find a difference between *few* and *many*. An overview will be provided below after a review of our findings for the proportional reading of the quantity words.

The computational model used to test the CFK semantics for the cardinal reading of *few* and *many* was extended to test whether its predictions can be transferred to the proportional reading. We proposed this extension because an interpretation study showed that the proportional reading is both context-dependent, excluding a fixed threshold on proportions, and sensitive to the size of the proportion described. From these results, we conclude that the contextual contribution for the proportional reading is two-fold, resulting in two kinds of prior expectations. The first is an uninformed, uniform belief about proportions P_U and the second are informed prior expectations P_E about likely proportions based on world knowledge. We propose a linear combination model which incorporates the assumption that the amount of world knowledge employed depends on its salience in the context.

The linear combination model turned out to be a good predictor for the proportional reading in both real-world contexts and in an abstract scenario in which balls are drawn from an urn. In total, it manages to explain 95% of the data correctly. We take this as support for the assumption of two-fold prior expectations and an influence of the context on the saliency of world knowledge. In terms of the fixed threshold hypothesis for *many*, we find that a unique threshold value applies to both P_U and P_E. Nevertheless, this threshold value is lower than for the cardinal reading and their HDIs do not overlap. However, we do not yet want to jump to the conclusion of suggesting a lexical ambiguity between cardinal and proportional *many*. For the proportional reading, the threshold values were inferred only on the basis of production data whereas for the cardinal reading also interpretation data were available. We find that the inclusion of interpretation data generally leads to more extreme threshold values because of the greater range of choices in this task. Future research could shed new light on whether the difference in threshold values

is caused by greater uncertainty in interpretation or whether the present results are evidence for a lexical ambiguity of *many*.

For *few*, the results are once again less coherent. Even though the influence of world knowledge is the same as for *many*, the model predicts widely separated threshold values on P_U and P_E. Given the fact that a unique θ_{many} could be identified, this result for *few* is surprising and is in line with several other aspects in which *few* and *many* seem to differ.

8.2 Differences between *few* and *many*

In the following, the most substantial differences between *few* and *many* are summarized once more. Experimental work by Sanford et al. (1994) finds that the use of *few* and *many* differs in terms of the referents they highlight. *Many* tends to make reference to the objects in the set whose size it describes (the reference set or refset), whereas *few* makes reference to the objects which are *not* in the described set (but in the complement set or compset). This is exemplified in the example repeated from Section 3.2.

(136) a. Many of the football fans went to the match. They cheered loudly when the player scored.

b. Few of the football fans went to the match. They watched the match at home instead.

Sanford et al.'s (1994) observation might be linked with a speculation about the semantics of the negative member in an antonym pair. Heim (2006, 2008) and Büring (2007a,b) wonder whether "negative adjectives" like *short* or *cheap* have negation as part of their semantic meaning by decomposing them into their positive counterpart and a negative operator. When transferring this idea to quantity words, *few* would decompose into *NEG* + *many*, with the negative operator being scopally mobile. Even though Heim (2006, 2008) claims that this decomposition analysis can account for ambiguous sentences with *few*, the theory cannot solve the puzzle we are facing here. The difference in referents preferred by *few* but not by *many*, as identified by Sanford et al. (1994), goes beyond the semantics of the sentence since we are here describing an observation at the level of contextual enrichment. A lexicalized negative operator, however, could only have an impact on the sentence's asserted meaning, i.e. the description of the cardinality's size, and therefore cannot explain why the quantity words prefer to highlight different referents. Coming back to the modification by *surprisingly*, we again do not see how the surprising results of an intensifying effect on *many* but not on *few* could be explained by the presence or absence of a negative operator in the semantics. This negative operator in the

semantics of *few* would only have to cancel the intensifying effect of *surprisingly* while leaving everything else unaffected. At this point we do not see how this could be achieved compositionally.

A last puzzling observation concerning the difference between *many* and *few* was made when investigating their proportional reading. For *many*, a unique threshold θ_{many} applying to both P_U and P_E could be identified, which might also be compatible with *many*'s cardinal reading. We conclude that Fernando and Kamp's (1996) predictions for a surprise-based semantics of *many* could be confirmed. For *few*, this is not the case, however. In sum, a major contribution of this thesis is to provide further evidence that *many* and *few* differ in crucial respects. To complement our findings and to naturally progress our work, we once more suggest conducting a more extensive experiment and eliciting prior expectations as well as production data in a within participants design to validate our findings. We expect that the model's predictions might be even more reliable when using individual priors instead of normalized, population-level expectations.

Apart from testing the CFK semantics experimentally, this dissertation provided an overview over three semantic accounts of the quantity words. Depending on the semantic analysis, *few* and *many* are treated as quantifiers, adjectives or semantically empty degree modifiers. The key features of the three accounts and their advantages and disadvantages are summarized in Table 2.2. Building on the alternative semantic account of *few* and *many* by Romero (2015, 2017), we proposed a modification of the positive operator *POS* in order to formally integrate subjective beliefs into the compositional analysis of sentences containing the quantity words. POS^{surp} in (87) can account for surprise readings by introducing an intensional comparison class and inferring compatible prior expectations. Truth conditions are then determined by employing a particular formalization of Fernando and Kamp's (1996) fixed threshold theory. This analysis was extended to be able to also capture the surprise reading of *few* and *many* in sentences with overt focus. While developing a semantics for POS^{surp}, we realized that there seems to be a gap between semantic theory and the available empirical data, even more so when it comes to the findings brought forward by probabilistic modeling. We made an attempt to bridge this gap and to incorporate our empirical findings into a semantics of *few* and *many*. We are aware, however, that more work is necessary to shed light on all the details of this complicated undertaking and to make sure that our proposal for POS^{surp} makes the right predictions also in more complex constructions and contexts.

So what have we learned about the context-dependence of *few* and *many*? We can identify three core results: first, *few* and *many* are dependent on prior expectations of the context. Second, a compositional analysis of *few* and *many*'s surprise reading

can both integrate the CFK semantics and make use of the sentence's comparison class to infer prior expectations. Third, a data-driven computational modeling approach in concert with Bayesian inference could support the CFK semantics for both the cardinal and the proportional reading of *many*. For *few*, we found evidence for a fixed threshold θ_{few} for the cardinal reading, but further work is required to test CFK semantics' predictions about the proportional reading of *few* as well as further aspects in which *few* differs from *many*.

8.3 Perspectives for Future Research

After having summarized this dissertation's core findings, we want to discuss interesting related phenomena and perspectives for future research. So far, the computational model based on the CFK semantics was used to test the theory's predictions for the cardinal and proportional reading as well as modification by adverbs like *surprisingly* or *incredibly* and *compared to* constructions. But the quantity words appear in more environments, for example in combination with negation. Further investigation and experimentation is recommended to learn whether the model manages to predict the use of *not many* or of *not few* in relation to prior expectations, as exemplified below.

(137) a. Sarah did not go to many restaurants last year.
b. Melanie does not own few pairs of shoes.

Semantically, *few* and *many* are also often treated on par with *little* and *much*. Solt (2009, 2015) assigns them an identical semantics, with the only difference that *few* and *many* are associated with cardinalities whereas *little* and *much* operate on other dimensions.

(138) a. Much land burnt during last year's bush fire season.
b. Little food was eaten at the party.

It would be interesting to see if *little* and *much* also express surprise readings and whether the CFK semantics can be transferred to them. Moreover, a natural next step would be to move from the investigation of the semantics of quantity words to their pragmatics and to extend the model to also capture the fine-grained differences between *few*, *many* and alternative utterances like *a few*, *several* or *lots of*.

When introducing the linguistic background of quantity words, we listed similarities with relative and absolute gradable adjectives like *tall*, *expensive* or *full*. These words are equally context-dependent and their use is also analyzed as being governed by threshold values in the semantics (Kennedy, 2007). Several investigations using probabilistic models assumed that gradable adjectives are dependent on

prior expectations of the context (Franke, 2012; Qing and Franke, 2014a; Lassiter and Goodman, 2015), suggesting the availability of surprise readings. It would be interesting to see if the our model can identify a fixed threshold on prior expectations for gradable adjectives as well and to learn whether the CFK semantics can be transferred to other semantics objects. Furthermore, the model's fit to experimental data on the use of adjectives could be compared with the fit of computational models which are based on other theories. Lassiter and Goodman's (2015) account suggests that threshold values are the result of pragmatic inferences whereas Qing and Franke (2014a) try to explain why particular threshold values are evolutionarily optimal for successful communication. These models could in turn also be applied to the presented data on *few* and *many*. By performing a statistical model comparison, the three approaches could be compared in order to gain further insight into context-dependence.

We have pointed out at several points that the CFK semantics makes predictions for the so-called "surprise reading" which expresses that a cardinality or proportion is lower or higher than expected. But more readings of context-dependent expressions have been attested, which have been claimed not to express surprise. Barker (2002) suggests that the use of gradable adjectives like *tall* affects shared knowledge in a developing discourse. He claims that a sentence like

(139) Feynman is tall.

has more uses than only to convey that Feynman's height is higher than the contextual standard. Instead, a *metalinguistic* use of this sentence would be an answer to a question under discussion of what counts as tall in this country. In this case, (139) would express what the prevailing relevant standard for tallness happens to be.

The metalinguistic use as described by Barker (2002) is also available for quantity words. The sentence

(140) Joe ate many burgers.

is a salient answer to the question under discussion of which numbers of consumed burgers count as *many*. In this case, the listener knows the number n of consumed burgers, and the prior expectations are formed about likely threshold values of which numbers of consumed burgers count as *many*. This example can be related to our discussion in Section 2.4 on the inference of the quantity word's interpretation n and the epistemic state underlying the prior expectations P_E. In the uses of *few* and *many* discussed during this dissertation, the quantity words are used to describe a cardinality. There, the listener makes use of his knowledge of P_E and the threshold values θ_{few} and θ_{many} to learn about the actual degree n. When confronted with the metalinguistic use, a listener would employ his knowledge n to jointly infer

the prevailing relevant standard $x_{\min}$ and P_E. Sentence (140) would then trigger a context update and restrict the set of possible standard values to those which are lower than the number of burgers eaten by Joe, similar to the interpretation rule in Equation 5.3 and illustrated in Figure 5.4b. How exactly the CFK semantics for *few* and *many* could be combined with Barker's (2002) dynamic update semantics for the metalinguistic use would be an interesting area of future research.

Another example that might be claimed to not express a surprise reading was given by Fernando and Kamp (1996).

(141) As expected, many students arrived today.

(141) could be claimed to be paradoxical under a reading of *many* which describes a number as greater than expected. However, Fernando and Kamp (1996) resolve the contradiction by arguing that "the expectation underlying *many* above might concern arrivals on days other than, or in addition, to *today*; the expectation referred to in *as expected* pertains specifically to *today*" (Fernando and Kamp, 1996, 65). This example shows that more readings than expected at first sight can be accounted for with the CFK semantics.

Even though the assumption of having prior expectations represent the contextual contribution produced good results for *few* and *many*, we have seen that forming expectations can be much more difficult in some contexts than in others. In real-world contexts as in the experiments in Sections 5.4 and 7.2, participants' expectations did not vary a lot, but in abstract contexts as in Section 7.3, participants' judgments were more diverse and some contrasted strongly with the mathematical, normative prior (see Figure 7.14). The effect of great uncertainty about the context on context-dependent expressions and the more psychological question of how prior expectations are formed and can be measured under uncertainty is another aspect worth investigating, I believe.

Bibliography

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

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

Bastiaanse, H. A. (2014). The intensional many-conservativity reclaimed. *Journal of Philosophical Logic 43*(5), 883–901.

Bates, D., R. Kliegl, S. Vasishth, and H. Baayen (2015, June). Parsimonious Mixed Models. *ArXiv e-prints*.

Bates, D., M. Maechler, B. Bolker, and S. Walker (2013). lme4: Linear mixed-effects models using eigen and s4. *R package version 1*(4).

Beck, S. (2006). Focus on *again*. *Linguistics and Philosophy 29*(3), 277–314.

Beck, S. (2009). Positively comparative. *Snippets 20*, 4–6.

Beck, S. (2011). Comparison Constructions. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning*, pp. 1341–1389. De Gruyter Mouton.

Beck, S. (2012). Degp scope revisited. *Natural Language Semantics 20*(3), 227–272.

Beck, S., T. Oda, and K. Sugisaki (2004). Parametric variation in the semantics of comparison: Japanese vs. english. *Journal of East Asian Linguistics 13*(4), 289–344.

Bennett, E. and N. D. Goodman (2015). Extremely costly intensifiers are stronger than quite costly ones. In *Proceedings of CogSci*, pp. 226–231.

Büring, D. (2007a). Cross-polar nomalies. In *Proceedings of SALT*, Volume 17, pp. 37–52.

Büring, D. (2007b). When less is more (and when it isn't). In *Chicago Linguistic Society Meeting, Chicago*.

Bylinina, L. (2014). *The grammar of standards: Judge-dependence, purpose-relativity, and comparison classes in degree constructions*. Ph. D. thesis, Universiteit Utrecht.

Chemla, E. and R. Singh (2014). Remarks on the experimental turn in the study of scalar implicature (part I & II). *Language and Linguistics Compass 8*(9), 373–386, 387–399.

Clark, H. H. (1991). Words, the world, and their possibilities. In G. R. Lockhead and J. R. Pomerantz (Eds.), *The Perception of Structure: Essays in Honor of Wendell R. Garner*, pp. 263–277. American Psychological Association.

Cohen, A. (2001). Relative readings of *Many*, *Often*, and generics. *Natural Language Semantics 9*, 41–67.

Coventry, K. R., A. Cangelosi, S. Newstead, A. Bacon, and R. Rajapakse (2005). Grounding natural language quantifiers in visual attention. In B. G. Bara, L. Barsalou, and M. Bucciarelli (Eds.), *Proceedings of CogSci*, Mahwah, NJ, pp. 506–511. Cognitive Science Society.

Coventry, K. R., A. Cangelosi, S. N. Newstead, and D. Bugmann (2010). Talking about quantities in space: Vague quantifiers, context and similarity. *Language and Cognition 2*(2), 221–241.

De Vries, H. (2012). The syntax and semantics of evaluative degree modification. In D. Lassiter and M. Slavkovik (Eds.), *New Directions in Logic, Language and Computation*, pp. 195–211. Springer.

Degen, J., M. H. Tessler, and N. D. Goodman (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. OUP USA.

Dobrovie-Sorin, C. (2013). Proportional many vs. most and the quantificational status of strong indefinites. *Revue roumaine de linguistiqueRomanian Review of Linguistics*, 401–417.

Eckardt, R. (1999). Focus and nominal quantifiers. In P. Bosch and R. van der Sand (Eds.), *Focus*, pp. 166–187. Cambridge: Cambridge University Press.

Égré, P. and F. Cova (2014). Moral asymmetries and the semantics of *many*. to appear.

Fara, D. G. (2000). Shifting sands: An interest-relative theory of vagueness. *Philosophical topics 28*(1), 45–81. Originally published under the name "Delia Graff".

Fernando, T. and H. Kamp (1996). Expecting many. In T. Galloway and J. Spence (Eds.), *Linguistic Society of America SALT*, Ithaca, NY: Cornell University, pp. 53–68.

Frank, M. C., N. D. Goodman, and J. B. Tenenbaum (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science 20*(5), 578–585.

Franke, M. (2012). On scales, salience & referential language use. In M. Aloni, F. Roelofsen, and K. Schulz (Eds.), *Amsterdam Colloquium 2011*, Lecture Notes in Computer Science, Berlin, Heidelberg, pp. 311–320. Springer.

Franke, M. (2016). Task types, link functions & probabilistic modeling in experimental pragmatics. In F. Salfner and U. Sauerland (Eds.), *Proceedings of Trends in Experimental Pragmatics*, pp. 56–63.

Franke, M., F. Dablander, A. Schöller, E. D. Bennett, J. Degen, M. H. Tessler, J. Kao, and N. D. Goodman (2016). What does the crowd believe? A hierarchical approach to estimating subjective beliefs from empirical data. In *Proceedings of CogSci*, pp. 2669–2674.

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

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical science 7*(4), 457–472.

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

Goodman, N. D. and D. Lassiter (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. In S. Lappin and C. Fox (Eds.), *The Handbook of Contemporary Semantic Theory, 2nd Edition*, Chapter 21, pp. 655–686. Wiley-Blackwell.

Griffiths, T. L. and J. B. Tenenbaum (2006). Optimal predictions in everyday cognition. *Psychological Science 17*(9), 767–773.

Hackl, M. (2000). *Comparative quantifiers*. Ph. D. thesis, MIT.

Hackl, M. (2001). Comparative quantifiers and plural predication. In *Proceedings of WCCFL XX*, pp. 234–247.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: *Most* versus *more than half*. *Natural Language Semantics 17*(1), 63–98.

Heim, I. (1999). Notes on superlatives. MIT lecture notes.

Heim, I. (2006). Little. In M. Gibson and J. Howell (Eds.), *Proceedings of SALT*, Volume 16, pp. 35–58.

Heim, I. (2008). Decomposing antonyms. In A. Grøn (Ed.), *Proceedings of Sinn und Bedeutung*, Volume 12, pp. 212–225. ILOS.

Heim, I. and A. Kratzer (1998). *Semantics in generative grammar*, Volume 13. Blackwell Oxford.

Herbstritt, M. and M. Franke (2016). Definitely maybe and possibly even probably: efficient communication of higher-order uncertainty. In *Proceedings of CogSci*, pp. 2639–2644.

Herburger, E. (1997). Focus and weak noun phrases. *Natural Language Semantics 5*(1), 53–78.

Hohaus, V. (2015). *Context and Composition: How Presuppositions Restrict the Interpretation of Free Variables*. Ph. D. thesis, Universität Tübingen.

Hörmann, H. (1983). *Was tun die Wörter miteinander im Satz?, oder, Wieviele sind einige, mehrere und ein paar?* Verlag für Psychologie.

Kao, J. T., J. Wu, L. Bergen, and N. D. Goodman (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences 111*(33), 12002–12007.

Katz, G. (2005). Attitudes toward degrees. In *Proceedings of Sinn und Bedeutung 9*, pp. 183–196.

Keenan, E. L. and J. Stavi (1986). A semantic characterization of natural language determiners. *Linguistics and Philosophy 9*(3), 253–326.

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

Kennedy, C. and L. McNally (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language 81*(2), 345–381.

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

Krasikova, S. (2011). On proportional and cardinal many. *Generative Grammar in Geneva (GG@ G) 7*, 93–114.

Kratzer, A. (1996). Severing the external argument from its verb. In J. Rooryck and L. Zaring (Eds.), *Phrase structure and the lexicon*, pp. 109–137. Springer.

Kruschke, J. E. (2011). *Doing Bayesian Data Analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press.

Kruschke, J. E. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Burlington, MA: Academic Press.

Lappin, S. (2000). An intensional parametric semantics for vague quantifiers. *Linguistics and Philosophy 23*, 599–620.

Lassiter, D. and N. D. Goodman (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Proceedings of SALT 23*.

Lassiter, D. and N. D. Goodman (2015). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 1–36.

Lee, M. D. and E.-J. Wagenmakers (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Meier, C. (2003). The meaning of too, enough, and so... that. *Natural Language Semantics 11*(1), 69–107.

Milsark, G. (1977). Toward an explanation of certain peculiarities of the existential construction in english. *Linguistic Analysis 3*, 1–29.

Moxey, L. M. (2006). Effects of what is expected on the focussing properties of quantifiers: A test of the presupposition-denial account. *Journal of Memory and Language 55*(3), 422–439.

Moxey, L. M. and A. J. Sanford (1987). Quantifiers and focus. *Journal of Semantics 5*(3), 189–206.

Moxey, L. M. and A. J. Sanford (1993). Prior expectations and the interpretation of natural language quantifiers. *European Journal of Congitive Psychology 5*(1), 73–91.

Moxey, L. M. and A. J. Sanford (2000). Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology 14*(3), 237–255.

Newstead, S. E. and K. R. Coventry (2000). The role of context and functionality in the interpretation of quantifiers. *European Journal of Congitive Psychology 12*(2), 243–259.

Nouwen, R. (2010). Whats in a quantifier. *The Linguistics Enterprise*, 235–256.

Nouwen, R. (2011). Degree modifiers and monotonicity. In *Vagueness and language use*, pp. 146–164. Springer.

Partee, B. (1989). Many quantifiers. In J. Powers and K. de Jong (Eds.), *5th Eastern States Conference on Linguistics (ESCOL)*, pp. 383–402.

Partee, B. H. (2010). Focus and information structure: Semantics and pragmatics. Lecture Notes on Formal Semantics.

Piñón, C. (2005). Comments on morzycki and katz. unpublished manuscript, available at pinon.sdf-eu.org.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, and A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Plummer, M. (2008). Penalized loss functions for bayesian model comparison. *Biostatistics 9*(3), 523–539.

Qing, C. and M. Franke (2014a). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In J. Grieser, T. Snider, S. D'Antonio, and M. Wiegand (Eds.), *Linguistic Society of America SALT*, Volume 24, pp. 23–41. elanguage.net.

Qing, C. and M. Franke (2014b). Meaning and use of gradable adjectives: Formal modeling meets empirical data. In P. Bello, M. Guarini, M. McShane, and B. Scassellati (Eds.), *Proceedings of CogSci*, Austin, TX, pp. 1204–1209. Cognitive Science Society.

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review 60*, 20–43.

Rett, J. (2008). *Degree modification in natural language*. Ph. D. thesis, Rutgers, The State University of New Jersey.

Rett, J. (2016). The semantics of *many*, *much*, *few*, and *little*. unpublished manuscript.

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. In J. H. Yoon and A. Kathol (Eds.), *OSU Working Papers in Linguistics: Papers in Semantics 49*, pp. 91–136. The Ohio State University.

Romero, M. (2015). The conservativity of many. In *Proceedings of the 20th Amsterdam Colloquium*, pp. 20–29.

Romero, M. (2017). On the readings of *many*. Handout of a talk given at the University of Tübingen.

Rooth, M. (1985). *Association with Focus*. Ph. D. thesis, University of Massachusetts at Amherst.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics 1*(1), 75–116.

Sanford, A. J., L. M. Moxey, and K. Paterson (1994). Psychological studies of quantifiers. *Journal of Semantics 11*(3), 153–170.

Schöller, A. and M. Franke (2015). Semantic values as latent parameters: Surprising few & many. In S. D'Antonio, M. Moroney, and C. R. Little (Eds.), *Proceedings of SALT*, Volume 25, pp. 143–162.

Schöller, A. and M. Franke (2016). How many *many*s? Exploring semantic theories with data-driven computational models. In N. Bade, P. Berezovskaya, and A. Schöller (Eds.), *Proceedings of Sinn und Bedeutung 20*, pp. 622–639.

Schöller, A. and M. Franke (2017a). Semantic values as latent parameters: Testing a fixed threshold hypothesis for cardinal readings of *few* & *many*. To appear in *Linguistic Vanguard*.

Schöller, A. and M. Franke (2017b). *Surprisingly*: Marker of surprise readings or intensifier? To appear in Proceedings of CogSci.

Schulz, T. (2017). Zuckerbergs Zweifel. *Der Spiegel 14*, 12–21.

Schwarz, B. (2010). A note on for-phrases and derived scales. In *Sinn und Bedeutung*, Volume 15. Handout for talk.

Schwarzschild, R. (1997). Why some foci must associate. unpublished manuscript, Rutgers University.

Schwarzschild, R. (2013). Degrees and segments. In *Proceedings of SALT*, Volume 23, pp. 212–238.

Solt, S. (2009). *The semantics of adjectives of quantity*. Ph. D. thesis, The City University of New York.

Solt, S. (2011a). Notes on the comparison class. In R. van Rooij, U. Sauerland, and H.-C. Schmitz (Eds.), *Vagueness in Communication*, pp. 189–206. Berlin: Springer.

Solt, S. (2011b). Vagueness in quantity: Two case studies from a linguistic perspective. *Understanding Vagueness. Logical, Philosophical and Linguistic Perspectives*, 157–174.

Solt, S. (2015). Q-adjectives and the semantics of quantity. *Journal of Semantics 32*(2), 221–273.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 583–639.

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

von Stechow, A. (2006). Times as degrees: *Früh(er)/Spät(er) 'Early(er)'/'Late(r)'* and phase adverbs. Unpublished manuscript. University of Tuebingen.

von Stechow, A. (2009). The temporal degree adjectives *Früh(er)/Spät(er) 'Early(er)'/'Late(r)'* and the semantics of the positive. In A. Giannakidou and M. Rathert (Eds.), *Quantification, definiteness, and nominalization*, Volume 24, pp. 214–233. OUP Oxford.

Tenenbaum, J. B., C. Kemp, T. L. Griffiths, and N. D. Goodman (2011). How to grow a mind: Statistics, structure, and abstraction. *Science 331*, 1279–1285.

Tessler, M. H., M. Lopez-Brau, and N. D. Goodman (2017). Warm (for winter): Comparison class understanding in vague language. To appear in Proceedings of CogSci.

Umbach, C. (2001). Contrast and contrastive topic. In *Proceedings of ESSLLI 2001 Workshop on Information Structure, Discourse Structure and Discourse Semantics: University of Helsinki*, pp. 175–188.

Vehtari, A. and J. Ojanen (2012). A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys 6*, 142–228.

Westerståhl, D. (1985). Logical constants in quantifier languages. *Linguistics and Philosophy 8*(4), 387–413.

Xu, F. and J. B. Tenenbaum (2007). Word learning as bayesian inference. *Psychological Review 114*(2), 245–272.