Quantiative Social-Cognitive Experimental Pragmatics

MSc Thesis *(Afstudeerscriptie)*

written by

Ciyang Qing
(born September 5th, 1990 in Nanning, Guangxi, China)

under the supervision of **Dr Michael Franke**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the public defense:	**Members of the Thesis Committee:**
June 24th, 2014	Dr Maria Aloni (Chair)
	Dr Raquel Fernández
	Dr Michael Franke
	Dr Henk Zeevat

Institute for Logic, Language and Computation

Abstract

This thesis argues for a quantitative, social-cognitive, and experimental approach to pragmatics. A formal decision-theoretic framework is proposed to model language production and interpretation in the context of social cognition. The framework is applied to several case studies: (1) use of referential expressions, (2) descriptive use of positive forms of gradable adjectives, and (3) use of quantifiers to give quantitative predictions which are tested against experimental data. We discuss design choices in experimental measure and formal modeling to incorporate various cognitive factors involved in pragmatic phenomena. This approach can shed new light on the meaning and use of language.

Chapter 1

Introduction

1.1 Motivation

Natural language is crucial to civilization and is indispensable to human everyday life. Given its great importance and prevalence, a theory of how people use natural languages is needed, in order to describe, explain, and predict various linguistic phenomena.

Intuitively, we use natural languages primarily[1] in two ways: Either we say something for some purpose, or we hear something and (not necessarily explicitly or even consciously) respond to it.

Hence a theory about language use should account for both production and interpretation. Specifically, an account of production describes utterances used under different circumstances for various purposes, and an account of interpretation spells out the effect of an utterance on the listener and its consequences.

Note that this does not imply commitment to Behaviorism (e.g., Skinner, 1957). We by no means deny the legitimacy or need of having mental representations in a linguistic theory. In fact, in our above description, "purpose" is about intention/desire, and we will have no objection to an account where "circumstance" includes the speaker's belief/knowledge and "effect of an utterance" is a change of the mental state of the listener on which subsequent behavior is based.

Neither are we claiming that perfect description or prediction of behavior is all that a theory of natural language is about. The above requirement is taken to be *necessary* rather than *sufficient*. We are merely stating that a linguistic theory that gives no such predictions is at least incomplete, and if it makes wrong predictions on these matters, then revision would be needed.

In this thesis, we argue for a quantitative, social-cognitive, and exper-

[1] One can use language to organize one's thoughts. Although such activities may intuitively be seen as speaking to oneself, this view is not uncontroversial. This type of use is beyond the scope of this thesis.

imental approach to the study of language use (pragmatics) that aims at meeting the above requirement. We will review previous works in this approach that have yielded fruitful results in studying a variety of pragmatic phenomena, most notably the *Rational Speech Act (RSA)* theory (e.g., Frank, Goodman, Lai, & Tenenbaum, 2009; Frank & Goodman, 2012; Bergen, Goodman, & Levy, 2012; Goodman & Stuhlmüller, 2013) and *game-theoretic pragmatics* (e.g., Rabin, 1990; Stalnaker, 2006; Franke, 2011; Jäger, 2013; Franke & Jäger, 2014) and introduce our extensions on several case studies. But first of all, let us explain and argue for the three features of this approach.

1.2 The Need for a Quantitative Theory

Traditionally, pragmatics primarily deals with *categorical* (in most cases, *binary*) or *ordinal* linguistic data. Thus we first need to justify the need of a quantitative theory of pragmatics.

The main reason is that a quantitative pragmatic theory allows us to better capture the subtlety and complexity of how people use language. Take production for example, there can be several possible utterances for a speaker to fulfill a purpose on a certain occasion, but a speaker might use them with different frequencies. Or a listener might feel that an utterance is ambiguous between two readings, but he may have a preference for one reading to various degrees. If our theory only makes categorical or ordinal predictions, it will inevitably lose some of the rich information about how people use language. A categorical theory that only predicts possibilities of utterances is inadequate to explaining preferences among alternatives. An ordinal theory is able to make predictions such as that utterance 1 is preferred to utterance 2, but it may fail to express, for example, the huge preference for utterance 1 to utterance 2, in contrast to the slight inclination for utterance 3 rather than utterance 4. A quantitative theory, on the other hand, has rich enough expressive power which enables us to systematically make any such comparisons, as we can build a quantitative theory that directly models the relative frequency of each utterance (reading). This means that a quantitative theory makes more specific predictions and thus has more commitments, which also makes it more falsifiable.

Hence, we need a quantitative theory to fully capture complex patterns in various phenomena of language use in a systematic way.

1.3 Language Use as Social Behavior

The second feature of the pragmatic theory is to investigate language use in the context of social cognition. This idea can be traced back to several schools in philosophy of language.

A Wittgensteinian view of language (Wittgenstein, 1953) draws the analogy between linguistic activities and games, treating the meaning of an expression as the conventional rule governing its proper conversational use to serve some social function. The speech act theory (e.g., Austin, 1962) demonstrates that one can make an utterance as performing and fulfilling a conventional social action. Last but not least, Grice (1989) illustrates how the meaning of an utterance can be calculated from the expression's conventional meaning and general expectation of cooperativity in communication, and further treats communication as intention recognition.

We shall emphasize here that we do not intend to enter the debate on what is *meaning*. Note that our statement of the goal (i.e., a theory of production and interpretation) in the beginning does not refer to meaning at all. Rather, we want to argue for an approach that focuses on the use of language, especially its social-cognitive aspect in production. In this section we will argue that one of the merits of this approach is that it can provide unique insight into the debate on the nature of meaning.

It helps to reflect on why meaning is such a fundamental issue in philosophy of language in the first place, in the light of our goal. A tentative answer is that meaning seems to be the most straightforward first step for any account of interpretation.

It is a fact that people have different responses to (utterances of) different linguistic expressions, thus any theory of interpretation should at least be able to account for such differences. The most intuitive and straightforward way to achieve this is to state that different expressions have different meanings (and some might have none, in which case the listener will think of the utterance as nonsense), and then proceed to build a theory of meaning that (1) specifies the meaning of each expression, and (2) explains how different meanings contribute to different responses. Traditional truth-conditional semantics is a typical example of a theory of meaning. First, through compositionality it is able to assign meanings to infinitely many expressions that constitute a considerable part of natural language. Second, through logical systems it predicts people's interpretation such as truth judgment, entailment recognition.

This approach has been highly successful to account for a wide range of phenomena concerning interpretation. However, what is missing in this picture is production, which is about utterances of expressions under different circumstances for various purposes. A theory of meaning in the above sense tells us nothing about when and why an utterance would be made in the first place. Hence it is inadequate to explaining production.

Moreover, interpretation is not independent of production. For example, the interpretation of indexicals, such as "I," "here," "now," crucially depends on the circumstances where they are uttered. Hence a theory of meaning in the narrow sense above is inadequate even for interpretation. This observation has led to a revision of the traditional theory of meaning.

According to the new theory, the meaning of an expression is relativized to the circumstances under which it is uttered.

Thus, we have seen how the aspect of circumstances in production affects interpretation and eventually influences the theory of meaning. It is natural to hypothesize that the other aspect in production, i.e., purposes, also plays a role in interpretation. Indeed, this is the aspect that the theories of meaning in the beginning of this section emphasize. Language is mostly used to serve social functional purposes such as conveying information, making requests, and performing actions. Thus it is reasonable to expect that language will be used to best serve these purposes. For example, Grice's theory illustrates how the listener, assuming the speaker is cooperative in conveying relevant information, can learn more information from an utterance than its truth-conditional semantics predicts, by taking the speaker's perspective and reasons hypothetically about production.

However, it shall be emphasized here that informativity is not the only factor that influences language use. Other cognitive factors may also influence both production and interpretation. For instance, Grice's maxim of manner requires that the speaker should normally be brief and orderly, which takes into account the cognitive factors of production and interpretation efforts.

In this thesis, we will further investigate how various social cognitive factors can affect production, which in turn can influence interpretation and eventually shed new light on the meaning of language. We will illustrate how to integrate these cognitive factors into quantitative models whose predictions can be empirically tested, and discuss how the results could complement existing theories of meaning.

1.4 Experimental Pragmatics

Finally, we argue for the need of experimental studies to obtain data for our purposes, since traditionally the primary source of linguistic data in semantics and pragmatics is researchers' own linguistic judgments.

The main reason why pragmatic intuition of a single person is not enough for studying the use of language is the quantitative nature of the phenomena.

First of all, production is usually non-deterministic, and our reflection on such non-determinancy is not accurate. For example, one might use utterance A 70% likely and utterance B 30% likely in a certain situation for a particular purpose, but he does not necessarily know such likelihoods. Hence the judgment of a single person is not reliable and thus inadequate.

Secondly, even though we usually feel highly definite about typical categorical semantic judgements concerning interpretation (such as truth or entailment), it is no longer the case when we are to make context-sensitive, graded pragmatic judgments. It is usually hard to know the exact strengths

of our graded judgments through introspection. This again means that it is not enough to have judgments from a single person.

Last but not least, our pragmatic theory is almost always incomplete to account for all individual differences in language use. For example, people's choices of utterances in the same situation might vary from person to person. Nevertheless, we might observe a general tendency in the population. It is such patterns of language use that our pragmatic theory is aimed at capturing.

In contrast, traditional semantics makes claims about every individual's definite linguistic judgements in a linguistic community. We normally have strong intuitions and feel very certain about those claims. This is for good reasons and such intuitions usually turn out to be correct. Thus it may seem less necessary to conduct experiments for this type of research. However, when we are to deal with more subtle, fine-grained linguistic phenomena where non-determinancy and individual differences add heavy noise to individual judgments, we can only rely on experimental methods to discover and verify the patterns of interest.

Experimental studies also force us to explicitly spell out the link between mental states (if they are part of the theory) and behavior, which helps make pragmatic theories empirically testable. For instance, a theory may predict that after utterance u, the listener would come to believe p. One way to test such a prediction may be to utter u to someone and directly ask him whether he believes p now. What is implicitly assumed here is that people's belief about p is not influenced by the query itself. This assumption may be generally accepted, but the point is that we should still make it explicit, so that when we use different experimental measures, we could compare the underlying assumptions to better understand the (potentially seemingly different) empirical results and what they suggest about the theory.

1.5 Thesis Overview

The rest of the thesis is organized as follows. In Chapter 2, we review the study of referential expressions in a language game, and a previous quantitative model in the literature for predicting production and interpretation, to illustrate the core ideas of the proposed approach and further spell out the general framework.

We develop a probabilistic model under this framework in Chapter 3, to account for the meaning and use of gradable adjectives. It focuses on the production side of vague expressions and shows how (sub-)optimal use of language could shape its meaning. The formal model is based on the previous individual research project, and in this thesis we further test its empirical validity by comparing model predictions against the experimental data we collected.

In Chapter 4 we extend our formal analysis to the meaning and use of the quantifiers such as *many*. We first develop a similar probabilistic model for the meaning of *many* and show how it can account for the meta-linguistic effects of use of quantifiers. Then we argue for the need of other cognitive factors to better capture the complex lexical competition among different quantifiers in people's linguistic judgments. By using both quantitative modeling and experimental studies, we illustrate the effect of two additional cognitive factors and how to use them to improve model predictions.

In Chapter 5 we first analyze the discrepencies between free production and meta-linguistic judgments, and discuss the relation between these experimental measures. In particular, we discuss a proposed hypothesis in the literature that certain types of meta-linguistic judgments can be better understood as the listener's belief about the speaker, which may differ from actual production in terms of the relevant cognitive factors. We revisit the study of referential expressions and illustrate how to incorporate various cognitive factors systematically and use the results of our previous individual research project to provide evidence for the hypothesis about the relation between free production and certain types of meta-linguistic judgments.

Finally, in Chapter 6 we summarize the merits and implications of the proposed approach, discuss potential objections and limitations, and suggest several extensions in the future work.

The main contribution of the thesis is the general framework we propose that summarizes several lines of previous research into a unified framework. In particular, we explicitly argue for the need of considering the purpose of language use in understanding production, which is in turn crucial to interpretation and the theory of meaning.

We also emphasize the need of incorporating various cognitive factors to better capture the complex patterns of language use, especially the need of explicitly specifying the assumptions that link model predictions to experimental measures. This contributes to the recent methodological discussions in experimental pragmatics.

Finally, by using this framework, we also make original contributions to the issues in the case studies that we use to illustrate the merits of the framework's three major features.

Chapter 2

Formal Framework

2.1 Motivating Example: Referential Games

Before we introduce the formal framework, in this section we will use a concrete example to illustrate the kind of problems this thesis is trying to address.

Referential games are confined interactive reasoning tasks used to study pragmatic inference in controlled environments (e.g., Stiller, Goodman, & Frank, 2011; Frank & Goodman, 2012; Degen & Franke, 2012).

A referential game consists of a set of *objects* called a *context*. For example, Figure 2.1(a) is a context which contains different shapes of different colors in a particular arrangement. The speaker in the game is asked to refer to one of the objects (the *target*) by uttering an expression to the listener. The listener, who does not know which object is the target, needs to recover the target based on the speaker's choice of expression.

(a) Example context: green square and circle, and blue circle.

	o_1	o_2	o_3
"Green": e_1	1	1	0
"Circle": e_2	0	1	1
"Square": e_3	1	0	0
"Blue": e_4	0	0	1

(b) Vocabulary and truth table

Figure 2.1: A simple referential game

For simplicity, we restrict the possible expressions (the *vocabulary*) the speaker could use to four words: "Green," "Circle," "Square," and "Blue."

As stated in Chapter 1, we want a theory about both production and interpretation. Production is about utterances used under different circumstances for various purposes. In our example, the circumstance is the context

and vocabulary in the referential game, and the speaker's purpose is to refer to the target object. Interpretation is about the effect of an utterance on the listener and its consequences. In this case, the main effect of a referential expression is to make the listener attend to a particular object in the context (which he believes is the speaker's intended target) in order to fulfill subsequent communication goals.

Since production and interpretation might be non-deterministic, the theory needs to predict the production probability $\sigma^{C,V}(u \mid t)$, i.e., the probability of making utterance u given target t, under context C and with vocabulary V, and similarly the interpretation probability $\rho^{C,V}(t \mid u)$ $(t \in C, u \in V)$.

Let us briefly review the theories in Section 1.3 in the light of this example, and clarify their relations to the goal of this thesis. Truth-conditional semantics states that the meaning of an expression e is its truth condition.[1] In this case it tells us whether the description of object o as e is true (Table 2.1(b)). Clearly, this by itself says nothing about the probabilities that we want, though we shall see later that it can be very helpful.

A Wittgensteinian view essentially equates the meaning of an expression e (in a fixed language with vocabulary V) with the whole collection of $\sigma^{C,V}(u = e \mid t)$ and $\rho^{C,V}(t \mid u = e)$ for every possible context C and target t. However, in this thesis we make no commitment to such a claim. Our goal is a principled theory to derive these probabilities, rather than claiming that this collection of probabilities itself *is* the meaning of an expression.We note that the study of these probabilities is important in its own right, as it helps us understand how people actually use natural language and can have practical applications in, e.g., artificial intelligence.

The speech act theory emphasizes the distinction between an expression e and an utterance of that expression. We can see that this distinction is naturally incorporated in the formula. We also note that we can deal with the various effects of an utterance by replacing t in the interpretation probability $\rho^{C,V}(t \mid u = e)$ with other effects of interest.

Finally, Grice's theory uses truth-condtional semantics together with conversational principle and maxims to account for production and interpretation. For example, the *Maxim of Quality* states that the speaker should not say things known to be false or lacking evidence. Hence in the above example, we have $\sigma(e_2 \mid o_1) = \sigma(e_4 \mid o_1) = 0$,[2] as the intended object o_1 is a green square, which is neither a circle nor blue. In addition, the *Maxim of Quantity* states that the speaker should make the utterance as informative as required. Hence we have $\sigma(e_1 \mid o_1) = 0$ and $\sigma(e_3 \mid o_1) = 1$, as there are two green objects in the context but only one square, which means "Square"

[1] More precisely, the meaning of a sentence is its truth condition, and the meaning of an expression is its contribution to the truth condition.

[2] $\sigma(e_2 \mid o_1)$ is short for $\sigma^{C,V}(u = e_2 \mid t = o_1)$, and similarly later we will use $\rho(o_2 \mid e_1)$ for $\rho^{C,V}(t = o_2 \mid u = e_1)$.

is informative enough while "Green" is ambiguous. Listeners reason counterfactually to interpret an utterance. For example, when a listener hears the utterance "Green," he will consider for which object the speaker can use this utterance according to the maxims. He can rule out object o_3 because of the maxim of quality and object o_1 because of the maxim of quantity. Thus he can conclude that the intended object must be o_2, i.e., $\rho(o_2 \mid e_1) = 1$, even though "Green" is literally true for both o_1 and o_2.

Grice's analysis is very insightful and is the first attempt to meet our goal: predicting $\sigma(u \mid t)$ and $\rho(t \mid u)$. However, since his theory is categorical, its predictions are limited. As we stated earlier, language use is often non-deterministic, which requires predictions about the specific probabilities. For instance, when the target is o_2, "Green" and "Circle" are both true and ambiguous, so there is uncertainty in the choice of word. In this case, Grice's theory cannot predict the production probability $\sigma(u \mid o_2)$. Moreover, Grice's theory is primarily about what a cooperative speaker should say and how a listener could recover the speaker's intentioin, rather than what speakers and listeners actually do. Hence we can expect that even when Grice's theory does predict a probability, which is either 0 or 1, it might not match the probabilities we observe about how language is actually used.

This suggests that we need to further extend Grice's pragmatic theory to meet our goal. In the next section we will introduce such a quantitative extension.

2.2 The Rational Speech Act Theory

The *rational speech act* (RSA) theory (e.g., Frank et al., 2009; Frank & Goodman, 2012; Bergen et al., 2012; Goodman & Stuhlmüller, 2013) provides a quantitative extension of Grice's theory. In general, there are two main problems that a quantitative extension of Grice's theory needs to address. On the production side, it needs a quantitative rendering of cooperative principle and conversational maxims to explain how non-deterministic use of language could adhere to them. On the interpretation side, it needs a quantitative counterpart of the listener's counterfactual reasoning in Grice's theory.

For the first problem, the RSA model uses information theory to quantitatively measure the informativity of an utterance. Specifically, it starts with a postulated literal listener with uniform prior belief over all objects in the context. Formally, we denote this uniform prior as $\mathcal{U}(\cdot)$, which is a function mapping each object to the same probability $\frac{1}{|C|}$:

$$\mathcal{U}(o) = \frac{1}{|C|}, \quad \forall o \in C \ .$$

When the literal listener hears an utterance u, he does a probabilistic

conditioning on the literal meaning of u:

$$\rho_0(o \mid u) = \mathcal{U}(o \mid [\![u]\!]) = \begin{cases} 1/|[\![u]\!]| & \text{if } o \in [\![u]\!] \\ 0 & \text{otherwise} \end{cases} . \tag{2.1}$$

In other words, after receiving utterance u, the literal listener believes that every object for which u is true is equally likely to be the speaker's intended object, and those objects for which u is not true cannot be the intended object.

Intuitively, the closer the induced literal listener's belief $\rho_0(\cdot \mid u)$ is to the speaker's intended object t, the more informative the utterance u is. Technically, the informativity of utterance u for speaker's intended object t can be measured as the negative Kullback-Leibler divergence (e.g., MacKay, 2003) of ρ_0 from the speaker's own belief δ_t:

$$\text{Info}(u, t) = -\text{KL}(\delta_t \| \rho_0) = -\sum_o \delta_t(o) \log \left(\frac{\delta_t(o)}{\rho_0(o \mid u)} \right), \tag{2.2}$$

where δ_t is a delta distribution with all probability mass on target object t, as the speaker knows her intended referent:

$$\delta_t(o) = \begin{cases} 1 & \text{if } o = t \\ 0 & \text{otherwise} \end{cases} . \tag{2.3}$$

Given the definitions of δ_t and ρ_0, (2.2) can be simplified as

$$\text{Info}(u, t) = \log(\rho_0(t \mid u)) = \log \left(\frac{1}{|[\![u]\!]|} \right) . \tag{2.4}$$

We can see that the fewer objects for which the utterance u is true, the more likely that the literal listener believes in the speaker's intended object t after hearing the utterance, and thus the more informative the utterance is. Thus (2.4) gives us a quantitative measure of the informativity of an utterance.

It remains to be answered how the production probabilities can be derived from the informativity of the utterances. Intuitively, we can weaken the maxim of quantity a little bit and require that the more informative an utterance is, the more likely that it will be used. Technically, we will use a soft-max function (e.g., Luce, 1959; Sutton & Barto, 1998) to capture this intuition in specifying the production probability:

$$\sigma(u \mid t) \propto \exp(\lambda_S \cdot \text{Info}(u, t)), \tag{2.5}$$

where λ_S is a parameter measuring the speaker's degree of rationality, i.e., to what extent the speaker sticks to the most informative utterance. The symbol $\propto$ means "linearly proportional to." For example, suppose there are

two expressions e_1 and e_3 true[3] for target t, then the probability of uttering e_1 is

$$\sigma(e_1 \mid t) = \frac{\exp(\lambda_S \cdot \text{Info}(e_1, t))}{\exp(\lambda_S \cdot \text{Info}(e_1, t)) + \exp(\lambda_S \cdot \text{Info}(e_3, t))},$$

and similarly for e_3 we have

$$\sigma(e_3 \mid t) = \frac{\exp(\lambda_S \cdot \text{Info}(e_3, t))}{\exp(\lambda_S \cdot \text{Info}(e_1, t)) + \exp(\lambda_S \cdot \text{Info}(e_3, t))} \ .$$

It can be shown that when $\lambda_S = 0$, $\sigma(u \mid t)$ is uniformly distributed over all true utterances, meaning that the speaker just randomly selects one of the true utterances without any considerations of informativity. On the other hand, when $\lambda_S \to \infty$, $\sigma(u \mid t)$ is uniformly distributed among the most informative utterances, meaning that the speaker strictly adheres to the maxim of quantity. Thus, we use a finite positive λ_S to capture the intuition that actual speakers are generally as informative as possible, but need not be strictly so.

From (2.1)-(2.5) we obtain the speaker's production rule:

$$\sigma(u \mid t) \propto \exp(\lambda_S \cdot \log\left(\frac{1}{|[\![u]\!]|}\right) = |[\![u]\!]|^{-\lambda_S}, \tag{2.6}$$

which is also referred to as the *size principle*, since the larger the set of objects for which utterance u is true, the less informative the utterance is, and thus the less likely it will be used by the speaker.

So far we have seen how the RSA theory addresses the production problem. For the interpretation problem, the RSA theory adopts *Bayes' rule* as the quantitative counterpart of the listener's reasoning in Grice's theory.

Formally, in the RSA model, the actual listener, who reasons pragmatically, upon hearing utterance u, updates his prior belief $\mathcal{S}(t)$ by applying Bayes' rule:

$$\rho(t \mid u) \propto \mathcal{S}(t) \cdot \sigma(u \mid t) \ . \tag{2.7}$$

The pragmatic listener's prior belief $\mathcal{S}(t)$ comes from perception or other non-linguistic contextual factors. It can be empirically measured. For each object, the listener reasons counterfactually about how likely the speaker would use the utterance if that object were intended and weights these likelihoods by the prior belief, to form a posterior belief about the target object. We can see that this is the same kind of reasoning in Grice's theory, except that it is done quantitatively via Bayes' rule.

[3]Technically, (2.2) implies that the informativity of a false utterance is $-\infty$ and it can be shown that $\exp(\lambda_S \cdot -\infty) = 0$, which means false utterances will never be uttered even if we consider them in (2.5). However, we note that this is not always the case for all the variations we will discuss in Chapter 5. Unless otherwise stated, we generally assume the maxim of quality holds, i.e., the speaker only chooses among literally true utterances.

Frank and Goodman (2012) conducted experiments to collect participants' judgments in referential games, and obtained a highly significant correlation between the model's prediction (setting $\lambda_S = 1$) and the experimental data. This provides evidence of the potential of the approach. In later chapters we will present further evidence of its empirical predictive power on other case studies, but before that, we will abstract away from this case study and summarize the general framework in the next section.

2.3 General Framework

In the previous sections we introduced referential games as a model of referential use of language, and showed how quantitative predictions about both production and interpretation can be derived using the RSA theory, which can be seen as an extension of the classical Gricean theory.

In this section we summarize this process of pragmatic study on an abstract level and describe the general framework.

When we want to investigate a phonomenon of language use (e.g., referential expressions), including both production and interpretation, below are the typical steps we follow.

1. First and foremost, we need to identify the purpose of the type of language use we are interested in. For example, the purpose of using referential expressions is for the speaker to refer the listener to something that she intends to talk about. While it seems rather trivial in this case, the purpose of language use is not always easy to specify. Besides the fact that people can use same type of expressions for different purposes, there can also be conceptual subtleties when the purpose has to be specific enough to allow for a formal quantitative model. Assumptions and simplifications often need to be made, and we will see in Chapter 5 that even in the case of referential expressions, there are different ways in specifying the purpose in detail.

2. For an account of production, we also need to specify the circumstances under which utterances are made and the range of expressions involved. Again, usually we need to make many assumptions and simplifications, to make both formal modeling and experimental study feasible. Usually, we start with many simplifications in an attempt to capture the essence of the phenomenon, and we will then gradually relax the assumptions to give more realistic accounts.

   For example, for referential expressions, we use referential games as a simple characterization of the circumstances. We treat a context as a collection of objects and restrict the vocabulary to one-word features of the objects.

Of course, this formalization is too simple to capture all the relevant phenomena concerning referential expressions. For instance, it does not allow speakers to use compositional expressions and is not able to account for the possibility of *over-specification* (e.g., Gatt, van Gompel, van Deemter, & Kramer, 2013). Thus it is certainly not a fully generative story of referential language use. However, our hope is that, since it succeeds in prediction in a simplified, controlled scenario, the model may capture some mechanism in pragmatic reasoning that is part of the whole picture. Thus we may be able to put together such pieces of knowledge together to better deal with more complex and realistic situations later on, even if adjustments are often needed.

Another concern is that the game is too artificial to be relevant to the actual use of natural language. We will further discuss this issue in the last chapter. Here we note that the response above similarly applies.

3. Production is treated as a decision problem. Given the purpose and circumstance, the speaker needs to choose among alternatives to best serve his purpose. The choice is based on the speaker's belief of the effect of each alternative and her evaluation of such an effect. The speaker's belief of the effect is specified as the behavior of a hypothetical literal listener based on the speaker's semantic knowledge. The evaluation of such an effect, which is technically called its *utility*, is based on the purpose specified in Step 1 and is quantitative. The speaker's choice will be such that maximizes the utility, adjusted by cognitive factors.

   For instance, in the RSA model of referential expressions, the alternatives are literally true utterances. The effect of an utterance is specified as the literal listener's updated belief, and the utility is measured as how far away this belief is from the speaker's own intention, i.e., the informativity of the utterance. The speaker chooses her utterance by soft-maximizing the utility. The parameter λ_S incorporates the cognitive factor that people do not always make strictly optimal decisions.

4. For interpretation, after hearing an utterance, the listener needs to infer the speaker's intention and the circumstance (if it is not observable by the listener).

   As in the classical Gricean theory, the actual (pragmatic) listener takes the speaker's perspective to reason about how likely the speaker would use the utterance, for each possible intention and under each circumstance. The reasoning is mediated by other cognitive factors and we use Bayes' rule to quantitatively integrate all these factors to obtain the posterior belief of the listener.

   For instance, in the RSA model, the listener's posterior belief is an

integration of perspective-taking and perceptual salience.

Here we note that there could be discrepencies between actual production and listener's belief about the speaker. Also, when we conduct experiments, we need to specify the link from the listener's belief to the observed behavior. We will further discuss these issues in Chapter 5.

Chapter 3

Meaning and (Sub-)Optimal Language Use

In the previous chapter, we introduce the general framework and use the study of referential expressions as an example to show how the framework can be applied to account for both production and interpretation and how it extends previous Gricean analysis by incorporating various cognitive factors and giving quantitative predictions.

The use of referential expressions is usually considered as part of *pragmatic pragmatics*, in the sense that it is only about how people use expressions beyond what they literally mean.

In contrast, as pointed out earlier, the study of indexicals is considered to be about *semantic pragmatics*, as we have to consider the circumstances under which they are used to correctly specify their meanings.

In this chapter, we will apply the framework to the study of gradable adjectives, to illustrate that the framework, which focuses on production and particularly (sub-)optimal language use, can contribute to the study of meaning as well.

3.1 Gradable Adjectives: Background

According to the degree-based approach to the semantics of gradable adjectives such as *tall* (e.g., Kennedy & McNally, 2005; Kennedy, 2007), the denotation of a gradable adjective is a function that maps individuals to *degrees* on its underlying *scale structure*, e.g., $[\![\text{tall}]\!] = \lambda x.\mathbf{height}(x)$. The meaning of the *positive form* of a gradable adjective, such as *tall* in the sentence "John is tall," is taken to be the composition of the gradable adjective with a silent morpheme *pos*:

$$[\![pos\ \text{tall}]\!] = \lambda x.\mathbf{height}(x) \geq \mathbf{d}_s,$$

where $\mathbf{d}_s$ is the contextually appropriate standard of comparison (also called *threshold*).

In order to fully specify the semantics of *pos*, we need to address how the threshold $\mathbf{d}_s$ is determined.

First, we know that $\mathbf{d}_s$ is *context-sensitive*. For instance, we use different thresholds for *tall* when we talk about men or trees, which are called *comparison classes*.

Secondly, for many gradable adjectives, $\mathbf{d}_s$ is also vague, in that we can still be uncertain about the threshold even if the comparison class is explicit. However, as Kennedy (2007) observes, some gradable adjectives, such as *full* and *dry*, have positive forms which are arguably not vague.[1] Such gradable adjectives are called *absolute* adjectives, in contrast to *relative* adjectives such as *tall*, whose positive forms are vague.

Thus a theory of the meaning of the positive forms must correctly predict the contextual resolution of the threshold and in particular the difference between absolute and relative adjectives.

The organization of the rest of the chapter is the following. In Section 3.2 we review previous works, especially the evolutionary and probabilistic approaches. In Section 3.3 we combine insights from the evolutionary and probabilistic approaches and apply our framework to derive a speaker-oriented model of the use of positive forms of gradable adjectives and account for the absolute-relative distinction. In Section 3.5 we introduce our replication of the experimental study by Solt and Gotzner (2012) and use the empirical data to test our model's predictions in Section 3.6.

3.2 Previous Works

3.2.1 Scale Structures and Interpretive Economy

Kennedy (2007) illustrates that there is a correlation between the type of scale structure underlying a gradable adjective and the threshold for the its positive form.

Depending on whether or not it has a maximal element and a minimal one, a scale structure can be classified into four categories: (1) totally open (neither), (2) lower closed (only minimum), (3) upper closed (only maximum), and (4) totally closed (both). For example, *tall* has a totally open scale, *wet* has a lower closed one, *pure* has an upper closed one, and *open* has a totally closed one.

Kennedy observes that if a scale structure has an accessible endpoint, then the corresponding gradable adjective by default uses that endpoint as the standard of comparison in its positive form (if it has both, then addi-

[1] In reality we often use these positive forms *loosely*, e.g., a full glass of water may not be absolutely full, but this is a different issue.

tional selection may apply). For instance, *wet*, being lower closed, receives a minimum standard of comparison, i.e., something is wet just in case it has a non-minimal degree of wetness.[2] On the other hand, *tall*, being totally open, has a vague standard of comparison.

From this observation, Kennedy proposes the *Interpretive Economy* principle to address the difference between relative and absolute adjectives. It stipulates that when computing the truth conditions of a sentence, one should maximize the contribution of the conventional meaning of its elements. In the case of gradable adjectives, since the endpoints in the scales are part of the conventional meaning, the corresponding gradable adjectives should use them by default whenever they are available. Only when such conventional standards are not available would people attempt to use vague standards.

This analysis is insightful, but it does not completely solve the problem. First, it does not account for how the thresholds for relative adjectives are contextually determined. Moreover, the Interpretive Economy principle is not fully satisfactory.

3.2.2 Evolutionary Approaches

As Potts (2008) points out, the Interpretive Economy principle (IE) does not really *explain* why scale structures influence the meaning of positive forms of gradable adjectives, but rather *characterizes* the phenomena in a more illuminating fashion. As an explanation, IE is yet "an optimization principle left unsupported by a theory of optimization" (p. 5). What is needed, according to Potts, is an explanation of why it is beneficial to use endpoints, in terms of more basic mechanisms.

Potts himself provides an evolutionary account. He uses a strategic game where speakers and listeners need to coordinate the standard of comparison among various options. Assuming that endpoints are most cognitively prominent, Potts show that they constitute *Schelling points* (Schelling, 1960) which attract coordination and thus are evolutionarily stable (Benz, Jäger, & van Rooij, 2005).

The main problem with this account is the assumption that endpoints are most cognitively salient. Since a scale is only an abstract theoretical construct, we do not have much evidence or justification about why some part of it should be cognitively more salient than others.

Franke (2012) provides an alternative game-theoretic account that avoids this assumption. According to this account, the salience of degrees on a scale is replaced by the probability distribution of the degrees. The scale structure

[2]It is non-minimal because technically the minimal degree in the scale is the one that corresponds to zero degree of wetness, i.e., *dry*. Later we will refer to the corresponding threshold as *the* non-minimal threshold, which technically is the second least degree on the scale.

constrains the type of distribution. For example, the distribution on open and totally closed scales are assumed to be normal and uniform, respectively.

Franke models the referential use of positive forms by using *referential games* in which there are objects possessing various degrees in all types of scales according to the corresponding probability distributions. The speaker tries to convey an intended referent among its competitors to the listener, by choosing a property of that referent and using the positive form to indicate whether it has a high or low degree of that property.

He proposes a heuristic for such referential games. The speaker, knowing the intended referent, always selects a property that makes it the most salient among its competitors, and indicates the direction, i.e., whether the degree is high or low, accordingly. Salience is measured in terms of how far away the intended referent's degree deviates from its competitors, The listener, upon receiving a property and the direction, simply chooses an object that has the most extreme degree of that property in the correct direction. Through simulations Franke shows that using this heuristic leads to a high chance of referential success. Further examination reveals that whenever a closed scale property is selected by the speaker, the corresponding degree is always close to the endpoints, while open scale properties allow for a wider range of degrees to be used.

Essentially, instead of postulating the cognitive salience of degrees, which are abstract semantic entities, Franke's account adopts a more empirically attested view that an individual that has extreme degrees within a group is more cognitively salient. Scale structures influence the thresholds by constraining the probability distribution of degrees, which affects the probability of a degree being extreme in a certain group.

This model provides an account of referential use of gradable adjectives. However, there are other types of use of gradable adjectives this analysis does not address.

3.2.3 Descriptive and metalinguistic use of language

Barker (2002) argues that a dynamic perspective on vagueness and context helps better illustrate the difference between *descriptive* and *metalinguistic* use of language. Consider an utterance of the sentence "John is tall." Usually, it is used to convey information about John's height to someone who does not know him. This is an instance of descriptive use. However, sometimes the utterance is used even if the listener already knows the height of John, e.g., the conversation takes place at a party where John is present. In this case the main effect of the utterance is to convey that the speaker thinks John's height suffices to count as tall. This would be an instance of metalinguistic use, as it concerns information about how language itself is used.

To capture this distinction between two modes of language use, Barker

develops a dynamic update semantics where every possible world contains metalinguistic information (such as the standard of comparison for *tall*) as well as factual information (such as John's height). In this way both modes of use can be treated as a contextual update by which worlds inconsistent with the current utterance get eliminated.

3.2.4 Probabilistic Approaches

Lassiter (2011) points out that Barker's implementation implies that in each world there is a single precise language being spoken, but such a strong commitment to epistemicism (Williamson, 1996) is not necessary to exploit its crucial idea, i.e., treating vagueness as a result of language users' *imprecise knowledge of language*. He proposes an alternative model where the set of possible worlds and the set of possible languages are separated. We can then distinguish between *factual* and *linguistic common ground* and regard the current languages being spoken simply as those in the linguistic common ground. In this way, the linguistic standard is no longer some external fact in the world that conversational participants passively discover, but rather the result of their active coordination for effective communication.

In addition, Lassiter suggests that the above dynamic account should be naturally extended into a probabilistic model to encode the uncertainty in factual and metalinguistic information in a more fine-grained manner. In the new model, there is a probability distribution over world-language pairs and contextual updates are carried out by probabilistic conditioning. However, although he points out that our world knowledge and linguistic knowledge are in general correlated, Lassiter (2011) does not enlarge on how this relation is established. Thus the model is not yet fully capable of explaining the semantic properties of gradable adjectives.

Later, Lassiter and Goodman (2013) apply the RSA theory to give a quantitative account of how our world knowledge constrains the range of plausible thresholds in the semantics of positive forms of gradable adjectives.

The analysis is based on a game-theoretic scenario of descriptive language use. We will adopt the same scenario in our model and describe it in detail in the next section.

The main difference between the RSA model and the model we will introduce in the next section is that the RSA model puts the contextual resolution of threshold on the interpretation level. As a result, it does not have an production account of how positive forms of gradable adjectives are actually used by the speaker.

In the next section, we will apply the general framework introduced in the previous chapter to provide a quantitative model that predicts both the production and interpretation of positive forms. In particular, we will show how the model predicts the difference between absolute and relative adjectives.

3.3 Formal Modelling

3.3.1 The Speaker-Oriented Model (SOM)

In this section we will apply our formal framework and combine insights from the evolutionary and probabilistic approaches introduced earlier to account for the descriptive use of positive forms of gradable adjectives, and shows how the model sheds new light on the semantics of the *pos* morpheme.

The first step is to identify the purpose of the type of language use we are interested in. We will focus on descriptive use of positive forms. As noted earlier, the purpose is usually to convey information about the degree of the individual mentioned. We assume that the *question under discussion* (QUD) is how tall John is, as a working example.

The next step is to specify the possible range of expressions and the circumstances of production.

Since we want a semantic account, the focus is whether the positive form is *semantically applicable*.[3] Thus we assume that the speaker can only choose between using the positive form (u_1) and saying nothing (u_0).

The circumstance of a descriptive use is the degree of the intended individual and the comparison class that the speaker has in mind. Here, similar to the evolutionary account and the RSA model, a comparison class is taken to be a probability distribution of degrees on the scale, rather than simply a set of individuals, to better capture our background world knowledge. For instance, when we talk about John's height, we not only know that we are comparing him against the set of male individuals, but also have some prior world knowledge about the distibution of adult male heights, $\phi(h)$. In addition, the speaker knows John's height h_0 while the listener does not.

We use the probabilistic production rule proposed by Lassiter (2011), according to which the probability that one would call someone of height h_0 "tall" is the probability that the threshold θ is no greater than h_0:

$$\sigma(u_1 \mid h_0, \Pr) = p(\theta \leq h_0) = \int_{-\infty}^{h_0} \Pr(\theta)\, \mathrm{d}\theta \ . \tag{3.1}$$

where $\Pr(\theta)$ is the probability distribution of θ, which is the result of combining *pos* and the contextual comparsion class $\phi(h)$.

The remaining question is how $\Pr(\theta)$ is derived. In general, this is from the speaker's selection of the (sub-)optimal threshold. Note that since we want to account for the semantics of positive forms, which is part of the linguistic convention, the selection here is on the community level, i.e., how should the speakers in a linguistic community choose the threshold for comparison class $\phi(h)$, so that on average they can most successfully convey the degree of an individual from that comparison class. This is different

[3] More precisely, the focus is *to what extent* the positive form is applicable. Applicability is the generalization of truth to allow for graded or probabilistic judgments.

from selecting referential expressions in the previous chapter, which is on the individual level.

Technically, according to our framework, we need to specify the speaker's belief of the effect of each threshold and the corresponding utility. We specify the effect of using threshold θ via a hypothetical literal listener based on the speaker's semantic knowledge, which states that "tall" is true iff $h_0 \geq \theta$.

If John's height $h_0 < \theta$, the speaker can say nothing, since "tall" is not true. In this case, the literal listener can only use the prior information to infer John's height, so his belief about John's height is the same as the prior distribution:

$$\phi(h \mid u_0, \theta) = \phi(h) \ . \tag{3.2}$$

In particular, the probability of him believing in the correct height is $\phi(h_0)$.

If John's height $h_0 \geq \theta$, the speaker can utter "tall" and the literal listener can do an update via conditioning on its truth, which yields a new distribution:

$$\phi(h \mid u_1, \theta) = \phi(h \mid h \geq \theta) = \frac{\phi(h)}{\int_\theta^\infty \phi(h)\,\mathrm{d}h} \quad \text{if } h \geq \theta, \text{ otherwise } 0 \ . \tag{3.3}$$

In particular, the probability of him believing in the correct height is $\frac{\phi(h_0)}{1-\Phi(\theta)}$, where $\Phi(\theta) = \int_{-\infty}^\theta \phi(h)\,\mathrm{d}h$ is the *cummulative probability* of the prior distribution $\phi(h)$ at θ.

The probability of John's height being h_0 is $\phi(h_0)$, hence on average we have the *expected success* of θ:

$$ES(\theta) = \int_{-\infty}^\theta \phi(h_0)\phi(h_0|u_0, \theta)\,\mathrm{d}h_0 + \int_\theta^\infty \phi(h_0)\phi(h_0|u_1, \theta)\,\mathrm{d}h_0, \tag{3.4}$$

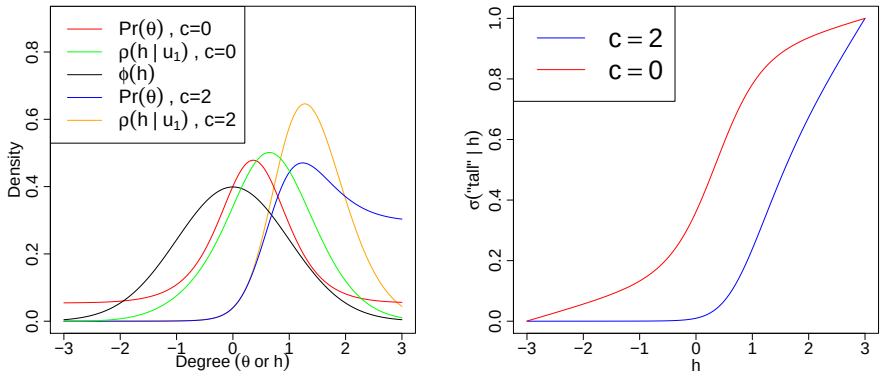
where the left summand corresponds to situations where the speaker has to stay silent because $h_0 < \theta$ and the literal listener can only use the prior knowledge, and the right summand corresponds to heights to which "tall" is applicable to induce a more accurate belief.

Since h_0 is a bound variable in the above formula, for simplicity we will simply rewrite it as h:

$$ES(\theta) = \int_{-\infty}^\theta \phi(h)\phi(h|u_0, \theta)\,\mathrm{d}h + \int_\theta^\infty \phi(h)\phi(h|u_1, \theta)\,\mathrm{d}h, \tag{3.5}$$

Now we will introduce another cognitive factor in specifying the utility. Since it takes some effort to produce an utterance, other things equal, people would prefer a threshold that requires utterances less often. Formally, we introduce a cost parameter of the positive form c, and specify the general utility as the expected success minus the cost:

$$U(\theta) = ES(\theta) - \int_\theta^\infty \phi(h) \cdot c\,\mathrm{d}h \ . \tag{3.6}$$



(a) $\Pr(\theta)$ and $\rho(h \mid u_1)$ (b) $\sigma(u_1 \mid h)$

Figure 3.1: SOM predictions for Gaussian distribution N(0,1), with $\lambda = 4$.

The integral of cost starts from θ because the positive form is used only when $h \geq \theta$.

Now we are finally able to link the linguistic knowledge $\Pr(\theta)$ to the utilities of each threhold, $U(\theta)$. The greater the utility of a convention, the more likely that people are going to use it:

$$\Pr(\theta) \propto \exp(\lambda \cdot U(\theta)) \ . \tag{3.7}$$

Combining (3.1), (3.6) and (3.7), we have a full production model at our disposal, and the corresponding interpretation model can be derived by applying Bayes' rule:

$$\rho(h \mid u_1) \propto \phi'(h) \cdot \sigma(u_1 \mid h, \Pr') \ .$$

Note that $\Pr'(\theta)$ and $\phi'(h)$ are correlated the same way as before, but the listener's prior world knowledge $\phi'(h)$ need not be the same as the speaker's. In the simplest case, prior world knowledge is in the common ground, so $\phi'(h) = \phi(h)$. (Note that this does not mean the speaker and listener will always have common linguistic knowledge, because they might have different λ and c.)

Fig. 3.1 shows predictions by the SOM for the Gaussian distribution N(0,1), with all parameters the same for both the speaker and the listener. We can see from Fig. 3.1(a) that the SOM predicts that the distribution of the threshold $\Pr(\theta)$ peaks slightly to the right of the average height, and that the posteriorof height after hearing "tall" $\rho(h \mid u_1)$ is shifted from the height prior to the right. Also, we can see from Fig. 3.1(b) that the production rule of the SOM does give sensible predictions. The probability of calling someone of height h "tall," $\sigma(u_1 \mid h)$, roughly has an "S" shape.

Note that our model gives reasonable predictions even when the cost $c = 0$.[4]

Since $\Pr(\theta)$ is the core component of the SOM, in the next section we will focus on $\Pr(\theta)$ to better illustrate the SOM's predictions for different prior distributions. We will further show how the SOM predicts the difference between absolute and relative adjectives observed by Kennedy (2007).

3.4 The Absolute-Relative Distinction

The crucial difference between absolute and relative adjectives is whether their underlying scale structures have accessible endpoints. Previous accounts (Franke, 2012; Lassiter & Goodman, 2013) interpret this difference as a constraint on the type of probability distribution of the degrees. More specifically, probability distributions on open and closed scales differ in whether there can be significant probability mass on the endpoint.

For instance, a relative adjective such as *tall*, corresponds to a scale that has no maximal element because the degree of height is unbounded and thus the probability must asymptotically fall to 0. On the other hand, an absolute adjective such as *open*, is associated with a scale that has a maximal element, and the occurrence probability of maximally open objects is usually non-negligible.

We adopt this view and apply the SOM to various distributions within the beta distribution family, which not only has a wide range of distributions that help us explore the exact boundary between absolute and relative adjectives, but also has nice closure properties that facilitate analytic derivations.

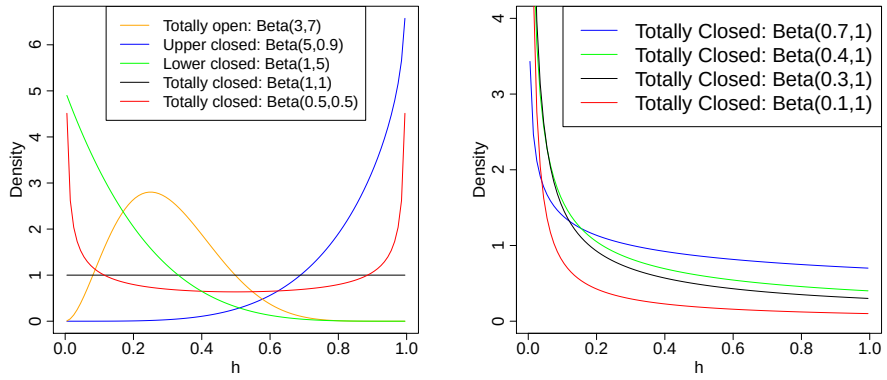
A beta distribution is defined on $[0, 1]$ and has two positive shape parameters α, β. Its densitiy function is defined as follows:

$$\phi(d; \alpha, \beta) = Kd^{\alpha-1}(1-d)^{\beta-1}, \tag{3.8}$$

where $K = 1/B(\alpha, \beta)$ is a normalization constant.

There is a tight correspondence between parameters of the beta distribution and scale structures (Fig. 3.2). If $\alpha, \beta > 1$, both endpoints have zero probability mass, which corresponds to open scales. If $\alpha > 1, \beta \leq 1$, the lower endpoint has zero probability mass and the upper endpoint has nonzero probability mass, which corresponds to upper closed scales. Similarly, $\alpha \leq 1, \beta > 1$ corresponds to lower closed scales. Finally, if $\alpha, \beta \leq 1$, both endpoints have nonzero probability mass, which corresponds to totally closed scales.

[4]Given the prevalence of gradable adjectives, we do not think it is very costly to utter them. In particular, the cost should not be the main factor that drives model prediction. Thus we take $c = 0$ as an approximation of the relatively small cost of the positive form. Later we will use empirical data to estimate the cost.



(a) Beta(α, β) (b) Beta$(\alpha, 1)$

Figure 3.2: Correspondence between Beta distributions and scale structures.

In general, the SOM predicts the following correspondence between endpoint probability mass and the optimal threshold:[5]

(i) If there is a sufficient amount of probability mass at the upper endpoint, then the maximal threshold is always optimal.

(ii) If (i) is not the case and the probability mass at the lower endpoint is sufficiently larger than elsewhere, then the non-minimal threshold[6] is optimal.

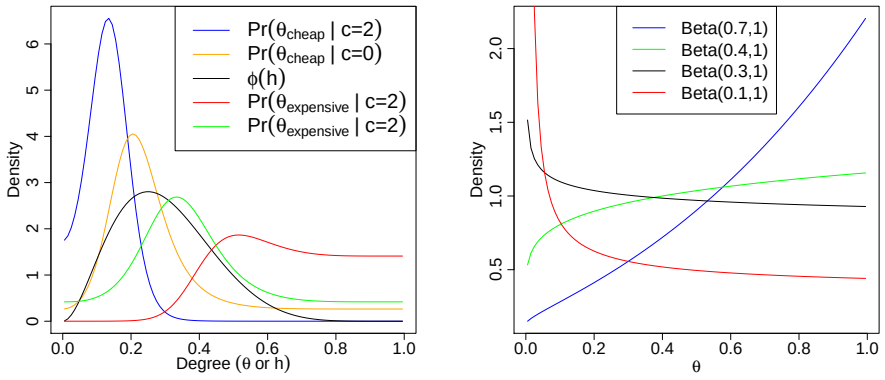
(iii) Otherwise the optimal threhold is sensitive to $\phi(h)$.

Let us first briefly explain how this correspondence correctly predicts the difference between absolute and relative adjectives observed by Kennedy (2007). The critical point is that in our formulation of the model, we assume that prior distribution $\phi(h)$ as the speaker's knowledge about the comparison class. However, in reality, besides the fact that comparison classes are often implicit, the speaker almost always has uncertainty about the exact distribution $\phi(h)$. Typically, the speaker knows the type of probability distribution for each adjective from its degree scale, but they do not know the exact distribution. For instance, the speaker might know the ranges of the shape parameters for a particular adjective, but he is uncertain about the exact values.

However, in the cases of absolute adjectives, the speaker does not need to know the exact $\phi(h)$ in order to know where the optimal threshold is. As

[5]For now we always assume $c = 0$, and in the end we will show that this assumption is not crucial to the prediction.

[6]Recall that the non-minimal threshold is the one that corresponds to the non-minimal reading, rather than an arbitrary non-minimal degree.



(a) Pr(θ) predicted by the SOM

(b) Pr(θ) predicted by the SOM, with $c = 0$

Figure 3.3: Predictions by the SOM for Beta(α,β), with $\lambda = 4$.

long as there is sufficient probability mass at either endpoint, the optimal threshold will be there.[7] This stability of optimal threshold explains why absolute adjectives are semantically not vague.

For open-scale adjectives, on the other hand, since the optimal threshold is sensitive to $\phi(h)$, the speaker cannot be sure where it is when they cannot be sure about $\phi(h)$. Thus, the vagueness of relative adjectives is the result of such sensitivity of the optimal threhold when there is uncertainty about the exact prior.

Now we are going to show such correspondence through representative examples in the beta distribution family.

We start from the relatively simple part, the open scales. The corresponding beta distribution has $\alpha, \beta > 1$. For example, Fig. 3.3(a) shows the SOM's prediction of Pr(θ) for Beta(3,7), which is an open scale that roughly corresponds to *cheap* and *expensive* (Lassiter & Goodman, 2013). [8]

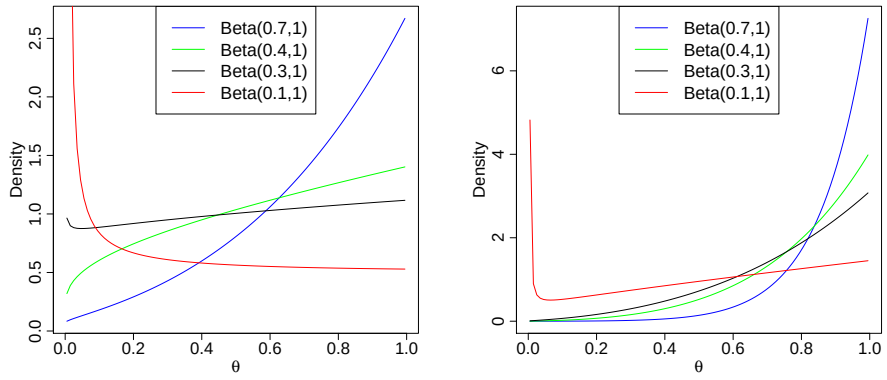
Indeed, we can see that as the prior probability mass shifts to the left in Fig. 3.3(a) compared to Fig. 3.1(a), the optimal threhold also shifts to the left.

Now we turn to closed scales and we will focus on cases where $\beta = 1$, as other cases will become straightforward after we show the results for $\beta = 1$.

From (3.8) it can be proved that Beta($\alpha, 1$) has density function $\phi(h; \alpha, 1) = \alpha h^{\alpha-1}$, and specifically the probability mass at the end point $h = 1$ is α.

[7] The speaker actually chooses the threshold sub-optimally via soft-max, reflecting the loose use of language, but if they are forced to, they can confirm that semantically the threshold is not vague because it is always at either endpoint.

[8] Technically, we use Beta(3,7) for *degrees of expense* and Beta(7,3) for *degrees of cheapness*, as they have inverse orderings on the degrees, and put the predictions of both models in the same plot.



(a) $c = 0.2$ (b) $c = 2$

Figure 3.4: Predictions by the SOM for Beta(α,1), with $\lambda = 4$.

Fig. 3.3(b) shows the SOM's prediction of $\Pr(\theta)$. We can see that when α is high (0.4 or 0.7), $\Pr(\theta)$ is always increasing, which means the upper endpoint is always the optimal standard. Meanwhile, when α is low (0.3 or 0.1), $\Pr(\theta)$ is always decreasing on (0,1], which means that a non-minimal standard is always optimal.[9]

In fact, it can be shown that $\alpha = 1/3$ is when a "phase transition" takes place, i.e., $\Pr(\theta)$ is increasing when $\alpha > 1/3$, uniform when $\alpha = 1/3$ and decreasing when $\alpha < 1/3$. Hence we know that optimal thresholds for absolute adjectives are stable to across a wide range of priors, which means slight uncertainty in $\phi(h)$ will not affect the speaker's knowledge about the optimal standard.

Fig. 3.4 further shows the robustness of the SOM's predictions with respect to costs. We know in general a higher cost will drive θ to the right, so we only need to focus on thresholds for non-minimal readings. We can see that when the cost is relatively low, the prediction is almost unaffected, as shown in Fig. 3.4(a) for Beta(0.1,1). Meanwhile, when the cost becomes relatively high, the upper endpoint also becomes a local optimum, as reflected by the v-shaped curves in Fig. 3.4(a) for Beta(0.3,1) and in Fig. 3.4(b) for Beta(0.1,1).[10] Nevertheless, the maximal threshold is only a local optimum. In fact, it can be proved that for $\alpha < 1/3$, $U(\theta)$ always goes

[9]The plot does not show the utility of $\theta = 0$. Note that $\theta = 0$ means the positive form is always true, which is effectively the same as staying silent all the time. Thus $\theta = 0$ has very low utility and can never be optimal. Of course, in the continuous case like this, there is no single non-minimal threshold, but in reality we almost always have limited precision with degree scales. We can then effectively make them discrete, and the non-minimal threshold

[10]In fact, this is also true for Beta(0.3,1), but the turning point is too close to 0 to be observable.

to infinity[11] when θ approaches 0, regardless of the cost, so the non-minimal threshold is always globally optimal.

Finally, when $\beta < 1$, the density $\phi(h)$ goes to infinity at the upper endpoint, which means $\Pr(\theta)$ will be driven even faster to the upper endpoint. Hence again we obtain the maximal threshold as desired and the predictions are also very robust.

In sum, we have shown that the SOM correctly predicts the difference between relative and absolute adjectives. We interpret vagueness as the stability of the optimal threshold under uncertainty about the exact prior distribution of degrees in the comparison class and shows how degree scales influence speaker's knowledge about the optimal threshold by constraining the type of priors. Closed scales constrain the priors such that there is always sufficient probability at either end point, which is enough for the speaker to be certain about the optimal threshold, even if she is not sure about the exact prior. This explains why absolute adjectives are semantically not vague. Optimal thresholds for priors on open scales, on the other hand, are sensitive to the exact prior, so the speaker cannot be sure about where they are when she is uncertain about the exact prior. This accounts for the vagueness of relative adjectives.

It should be noted here that the analysis above is based solely on the lexical properties of gradable adjectives. The prior distributions used here are the default priors constrained by the scales. In reality, when the speaker learns more about the degree distribution of the comparison class, he will adjust his prior accordingly, and different gradable adjectives can have constraints of different strengths on to what extent the default prior must be maintained in the light of new information. Our experiment will illustrate this issue in the next section.

3.5 Experiment

We have seen how the SOM can predict the difference between absolute and relative adjectives. Such a prediction is based on a specific mechanism of contextual resolution of the threshold, which can be seen as completing the semantics of the *pos* morpheme. Hence it is important to empirically test the validity of such a mechanism.

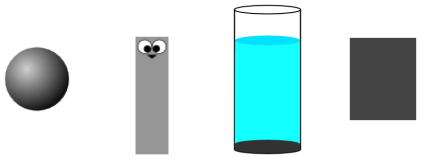
Since the threshold distribution $\Pr(\theta)$ is not directly observable, we test it by comparing the production model, which is determined by $\Pr(\theta)$, to actual speakers' choices. If the production model gives good prediction about the speaker data, then we will have more evidence of the empirical validity of the proposed mechanism.

In order to test the predictive power of the above production model, we collected participants' production of the positive forms of several adjectives

[11] Again, this is the result of using continuous scales.

under explicit comparison classes of varying degree distributions. Our design is that of Solt and Gotzner (2012), with minor modifications. We will introduce our replication first and then mention these minor differences.

Participants, Materials and Methods 96 US participants were recruited via Amazon's Mechanical Turk. Each of them received $0.25 for the experiment.



(a) Example items

Prior	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}
baseline	1	2	3	4	5	6	5	4	3	2	1
left-skewed	2	5	6	6	5	4	3	2	1	1	1
right-skewed	1	1	1	2	3	4	5	6	6	5	2
moved	1	2	3	4	5	6	5	4	3	2	1

(b) Number of items for each degree in each prior condition

Figure 3.5: Stimuli used in our replication of Solt & Gotzner's study

We tested participants' production for four gradable adjectives: *big*, *dark*, *tall* and *full*. For each adjective, we presented contexts of 36 items. Each item instantiated the adjective in question to one out of 14 possible degrees (balls varying in size, grey rectangles varying in lightness, cartoon characters varying in height, glasses varying in water level; see Fig. 3.5(a)). We chose mostly abstract items so as to minimize the effect of participants' background world knowledge. Stimuli were designed to make all 13 differences between adjacent degrees perceptually uniform.

We included 4 kinds of contextual prior distributions in our experiment. Each context consisted of 36 items spanning over 11 out of the 14 degrees. The *baseline*, *left-skewed* and *right-skewed* priors span over the lower 11 degrees with different distributions, and the *moved* prior spans over the upper 11 degrees (4th–14th) and has the same shape of distribution as the baseline. Fig. 3.5(b) shows the number of items for each degree in the 4 distributions.

Each participant finished 4 trials. In each trial they saw a context corresponding to 1 of the 4 adjectives under 1 of the 4 priors and were asked to check all items for which they would use the adjective in the given context (Fig. 3.6). We used a Latin square design for adjective-prior combinations within the 4 trials and counterbalanced the order of adjectives.

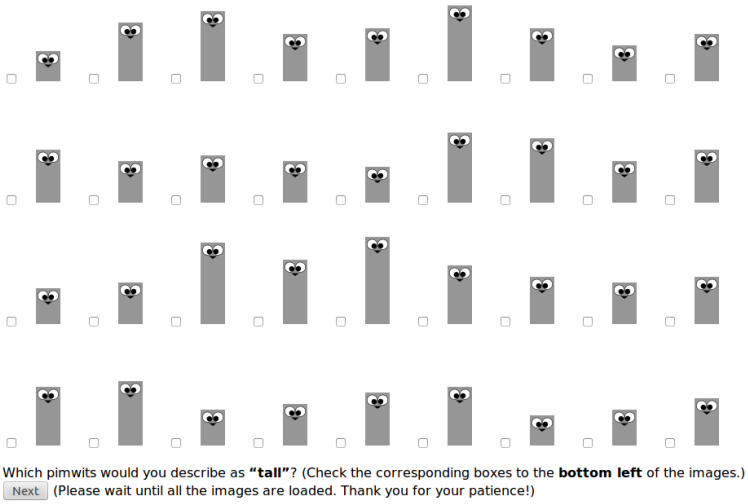


Figure 3.6: A sample trial

General Predictions Since the SOM's prediction of production is based on the distribution of degrees in the comparison class, when we systematically manipulate the distribution in the comparison class, we expect to see that the speakers' choices change correpondingly as predicted by the model.

However, note that the prior distribution is also constrained by the lexical property of gradable adjectives, and the strengths of such constraints can vary. Thus we cannot always directly use the empirical distribution[12] of degrees in the presented comparison class as the prior distribution in the speaker model. Rather, we need to make additional assumptions about how the prior distributions are constrained by each adjective.

The only constraint on degree distributions by open-scale adjectives is that the probability must asymptotically fall to 0 to both sides. The empirical distribution already satisfies this constraint. Also, for the two open-scale adjectives *tall* and *big* in our experiment, since the stimuli were artificially constructed, participants would have very little prior knowledge about the comparison class. Thus we expect that participants would approximately use the empirical distribution of degrees as the prior distribution in the speaker model.

On the other hand, the two closed-scale adjectives have more initial constraints on the degree distribution. Moreover, as pointed out in the previous section, such constraints have different strengths. Hence we expect participants' belief of the degree distribution might be adapted in the light

[12]By empirical distribution we mean a distribution where each degree's probability is just its relative frequency in the stimuli set. For instance, in the baseline condition, the empirical distribution is one such that degree 5 has probability 5/36 and degree 10 has probability 2/36, and so on.

of the statistical information of the presented stimuli.

Qualitative Results The results are shown in Fig. 3.7. As expected, proportions of intuitive applicability judgements followed an S-shaped curve rising from lower to higher degrees. More importantly, the statistical distribution of the contextual comparison class had an apparent influence on the applicability judgements. E.g., when there are many high-degree items such as in the right-skewed condition, smaller proportion of low-degree items were chosen.

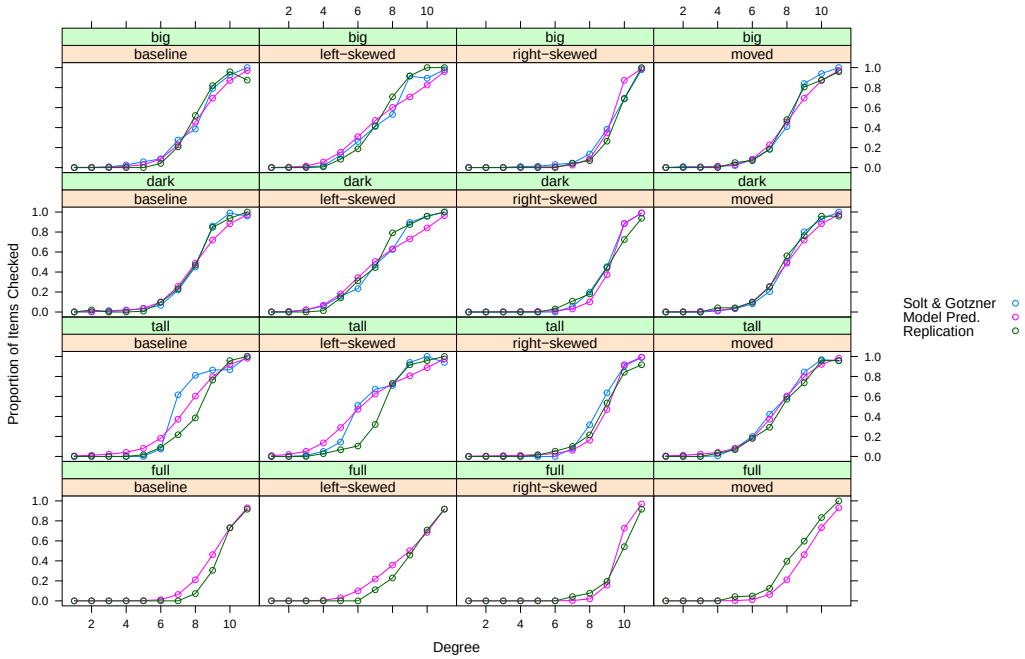


Figure 3.7: Observed and predicted applicability judgements for each degree for each adjective-prior pair. The blue and green curves show the observed proportions of items checked in each condition. The pink curve shows the mean posterior predictive values of the model when condition on the data by Solt & Gotzner.

The Original Dataset The experiment by Solt and Gotzner (2012) had 194 participants in total (47 – 50 participants in each condition). Test items included *big*, *tall*, and *dark*, but also *pointy* instead of *full*. We chose *full* primarily because *pointy* is a rather unusual word and it is hard to construct items with uniformly spaced degrees of "pointiness."

Results of their experiment are shown in Fig. 3.7 (blue lines). We can see that the result of our replication is close to theirs in most conditions,

except for the baseline and left-skewed conditions for *tall*. Since our main purpose here is to use these data to test our model, we do not report further statistical analysis of the data themselves.

3.6 Model Fitting and Validation

Our model has free parameters: λ (rationality) and c_A (cost of adjective A). We will use *Bayesian inference* (MacKay, 2003) to learn likely values of these parameters from the data of Solt and Gotzner (2012), and then test the model's predictions on our own replication.

Since we do not have a theory of the strengths of the initial constraints on priors, we will use the empirical distributions in each stimuli set as the prior distribution p that is used in the SOM.[13] As discussed earlier, we expect participants to use such empirical distributions for *tall* and *big*, so the bottomline is that the SOM should provide reasonable predictions for these two adjectives. In addition, we will compare the SOM prediction and actual production data for *dark* and *full* and try to diagnose possible discrepencies between empirical distribution and that in participants' belief.

We assume the following binomial process that generates data in both experiments: for each adjective A and prior p,

$$n_i^{A,p} \sim \text{Binom}(N_i^{A,p},\ \sigma_p(A \mid d_i;\ \lambda, c_A)), \tag{3.9}$$

where $n_i^{A,p}$ is the number of items of degree d_i checked by participants in the condition with adjective A and prior p, and $N_i^{A,p}$ is the total number of items of degree d_i in this condition.[14]

Intuitively, it means the following. Under prior p, with parameters λ, c_A, for degree d_i, the SOM predicts that with probability $\sigma_p(A \mid d_i;\ \lambda, c_A)$ the participants are going to use the positive form. Now the participants judged $N_i^{A,p}$ items of degree d_i in total, from probability theory we know that the number of items that were described as A, $n_i^{A,p}$, follows the binomial distribution $\text{Binom}(N_i^{A,p},\ \sigma_p(A \mid d_i;\ \lambda, c_A))$.

Hence, for a given adjective, λ and c_A, for each one of the 4 priors, our model makes predictions for all 11 degrees. Thus the model makes 44 predictions for each adjective.

[13] Note that we use p instead of ϕ because we are using discrete degrees.

[14] Note that we allowed participants to check none of the pictures, and some participants did not check all the pictures with the highest degree. In order to take these possibilities into account, we introduce an unobserved maximal degree d_{12} with prior probability $p(d_{12}) = 0$. Since this degree corresponds to a θ according to which the positive form is never used, the utility associated with it is rather small. Nevertheless, the soft-max function will assign a small non-zero probability to it, and hence the model always predicts that d_{11} might have a small probability not to be checked.

Parameters Learning We assume that λ is a constant, while each adjective has its own parameter c_A. This is because λ is the general degree of rationality in our sample population, whereas different adjectives could have different costs c_A depending on their lexical properties. With this, we use the following hyperpriors (initial belief about the parameters):

$$\lambda \sim \text{Unif}(0, 100) \quad c_A \sim \text{Unif}(-1, 0), \tag{3.10}$$

where $A \in \{\text{big, dark, tall}\}$. We draw 8000 samples (after a burn-in period of 9000 samples) from the posterior distribution $P(\lambda, \mathbf{c} \mid \mathcal{D}_{SG})$, i.e., we make a joint inference of λ, c_{tall}, c_{dark}, c_{big} from the dataset $\mathcal{D}_{SG}$ (Solt & Gotzner, 2012). For these posterior samples of parameters, we have $\bar{\lambda} = 48.23$, sd $= 1.14$; $\bar{c}_{\text{big}} = -.064$, sd $= .003$; $\bar{c}_{\text{tall}} = -.024$, sd $= .003$; $\bar{c}_{\text{dark}} = -.054$, sd $= .002$.

Since Solt and Gotzner (2012) did not include *full* in their experiment, we cannot learn the parameters directly from their dataset. Instead, we use the posteriors from their dataset to constrain the parameter λ:

$$\lambda_{\text{full}} \sim \text{Norm}(48.23, 1.14) \quad c_{\text{full}} \sim \text{Unif}(-1, 0)\,. \tag{3.11}$$

We get $\bar{\lambda}_{\text{full}} = 46.67$, sd $= .904$; $\bar{c}_{\text{full}} = -.158$, sd $= .006$.

Model Validation We validate our model in two ways.

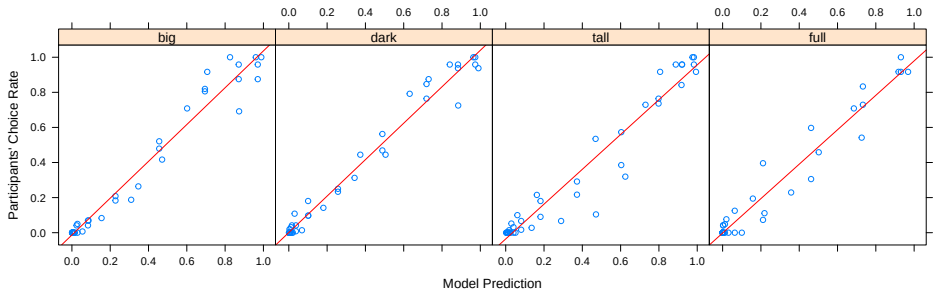


Figure 3.8: The relation between model predictions and participants' choices on the replication dataset

First, we use *Bayes model averaging* (Hoeting, Madigan, Raftery, & Volinsky, 1999)

$$\sigma(u \mid d, \mathcal{D}_{SG}) = \sum_{\lambda, \mathbf{c}} P(\lambda, \mathbf{c} \mid \mathcal{D}_{SG}) \cdot \sigma(u \mid d;\ \lambda, \mathbf{c}), \tag{3.12}$$

to compute the model's predictions after it learns the free parameters from $\mathcal{D}_{SG}$. The idea is that we take the weighted average of model predictions under various parameter configurations, where the weights reflect our strengths of belief about the parameters.

Table 3.1: Posterior predictive credibility values. The left value is for the training data (Solt & Gotzner, 2012), the right for our replication. Values in bold are those where the test values fall below a critical value of .05 for both data sets.

Adj	Prior	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}
big	base	1/1	1/1	1/1	.08/.63	.04/.06	.92/.08	.10/.59	.06/.32	.01/.06	.22/.21	.41/.03
big	left	1/1	.64/1	.33/.19	**.01/0**	.16/.02	.19/.01	.12/.42	.12/.24	**0/.02**	.25/.01	1/1
big	right	1/1	1/1	1/1	1/1	.17/1	.01/1	.06/.23	.01/.66	.24/.04	**0/0**	.10/1
big	moved	1/1	.27/1	.60/1	.52/.64	.58/.16	.92/.64	.12/.20	.21/.68	**0/.05**	.04/.83	.40/.17
dark	base	1/1	1/1	1/1	.57/.42	.87/.02	.05/.89	.19/.53	.30/.67	**0/.04**	0/.37	.39/1
dark	left	1/1	.43/1	1/.08	.65/.01	.37/.34	0/.59	.33/.24	.84/.02	.02/.16	.08/.41	.42/1
dark	right	1/1	1/1	1/1	1/1	1/1	.19/.12	.09/0	**0/0**	.01/.13	1/0	.24/.01
dark	moved	1/1	1/1	.64/1	.39/.07	.62/.80	.30/.89	.07/1	.83/.15	.04/.52	.21/.36	.64/.13
tall	base	1/1	.64/1	.08/.43	**0/.03**	**0/.01**	**0/.01**	**0/0**	**0/0**	.09/.48	.09/1	1/1
tall	left	1/1	.01/.18	**0/0**	**0/0**	**0/0**	.31/0	.36/0	.58/1	.01/.30	.01/1	.12/1
tall	right	1/1	1/1	1/1	.64/1	.19/1	.01/.37	.23/.08	0/.10	0/.19	.41/.01	.12/0
tall	moved	1/1	.65/1	.08/.41	0/1	.91/.74	.46/1	.17/.08	.77/.53	.30/.14	.09/1	.05/.07
full	base	–/1	–/1	–/1	–/1	–/1	–/.45	–/0	–/0	–/0	–/1	–/.68
full	left	–/1	–/1	–/1	–/1	–/.06	–/0	–/.02	–/.07	–/.43	–/.83	–/1
full	right	–/1	–/1	–/1	–/1	–/1	–/1	–/0	–/0	–/.38	–/0	–/.06
full	moved	–/1	–/1	–/1	–/1	–/0	–/.01	–/.02	–/0	–/.05	–/.25	–/.41

The predictions are shown in Fig. 3.7 (pink lines), and Fig. 3.8 shows the relation between model predictions and participants' choices for each adjective on the replication dataset $\mathcal{D}_{\text{rep}}$. We can see that in general the model prediction fits the empirical data.

Specifically, model predictions correlate well with observations ($R^2_{\text{big}} = .97$, $R^2_{\text{dark}} = .98$, $R^2_{\text{tall}} = .94$, $R^2_{\text{full}} = .95$, with overall $R^2 = .96$ and $p < .001$ for all cases). Correlations remain highly significant even when we only keep those data points for which our model's prediction is within the range of $(0.05, 0.95)$ ($R^2_{\text{big}} = .93$, $R^2_{\text{dark}} = .94$, $R^2_{\text{tall}} = .88$, $R^2_{\text{full}} = .90$, with overall $R^2 = .90$ and $p < .001$ for all cases). This suggests that our model does capture the general trend of participants' choices, rather than by simply assigning extreme probabilities to extreme degrees.

Second, in order to better diagnose the model's predictions for each data point, we investigate the posterior predictive distribution (c.f. Kruschke, 2011). Concretely, for each of the 8000 samples of parameters drawn from the posterior distribution described before, we use the binomial generative process (3.9) to generate a new dataset. Thus in the end we have 8000 simulated datasets. Then for each adjective, each prior and each degree, we look at the number of items checked in the actual dataset (either $\mathcal{D}_{SG}$ or $\mathcal{D}_{rep}$) and record the frequency of this actual observation in the simulated datasets. Finally, we calculate the posterior predictive credibility value as the sum of relative frequencies of all observations that occurs no more often than the actual observation in the simulated datasets. This posterior predictive credibility value then captures the estimated maximal threshold on credibility thresholds under which the observed data would not contradict our model. Concretely, a value of .05 means that the observed data falls within a 95% HDI interval of the posterior predictive; a value of 1 means that the observation was the mode of our posterior predictive sampling.

The posterior predictive credibility values are shown in Table 3.1. We can see that the model's predictions generally pass the predictive check. For those degrees where the model fails to meet a critical threshold of .05 on both data sets (marked in bold), we note two possible sources of bad fit: (1) The discrepancy between the two datasets due to noise. As a result, the model fitting the training set can fail to generalize to the test set. We also want to emphasize here that since the model needs to fit all degrees under all priors simultaneously, noise in one degree might influence performance on another degree as well. (2) The discrepancy between the two datasets due to differences in stimuli. For instance, the stimuli for *tall* generally have greater height-to-width ratio in the experiment by Solt and Gotzner (2012) than in our replication. As a result, participants in our replication tended to avoid the use of *tall* when the character's height-to-width ratio failed to meet the precondition of *tall*. This might explain why the model generalizes well in the moved condition but performs poorly on the baseline and left-skewed conditions.

The two validation methods both suggest that the model in general captures participants' applicability judgements well.

Costs and Initial Constraints We mentioned earlier that given the prevalence of relative adjectives, we would not expect them to be very costly. We have seen that the estimated values of costs are indeed very small.

However, whereas *big*, *tall*, and *dark* have relatively similar costs, we can see that *full* has a much greater cost.

One could accept this result and conclude that it is somehow much more costly to produce *full*. However, such a claim is clearly dubious without a justification of why *full* is so costly.

Alternatively, we think it is more plausible to concede that the model predictions for *full* are anomalous. However, this does not necessarily mean that the model itself is wrong. Rather, it is more likely that our additional assumption that participants used empirical distribution as the prior for *full* may be questionable. Note that unlike height, size or darkness, the fullness of a glass is just a *contingent* property, hence it is probably the case that the initial constraint set by the lexical property of *full* on the distribution of fullness cannot be easily overridden by some observations of the stimuli.

Note that *dark*, which semantically has closed scales, nonetheless exhibit properties of relative adjectives very similar to *tall* and *big*. This again suggests that the contextual prior is the result of both world knowledge and the lexical constraint.

Thus future work should aim at estimating the latent degree distribution that the participants believe for adjectives such as *full*, and ultimately we need a theory that predicts how the lexical constraint and world knowledge are combined to obtain such distributions.

Chapter 4

Cognitive Factors

4.1 Meaning and Use of Quantifiers

In the previous chapter, we discussed the use of positive forms of gradable adjectives, and showed that by considering the purpose of descriptive use of positive forms, we could enhance our understanding of the meaning of the morpheme *pos* and give quantitative production and interpretation models.

In this chapter, we try to generalize the analysis to the study of quantifiers such as *many* and *most*.

The classical *Generalized Quantifier Theory* (Barwise & Cooper, 1981), treats denotations of quantifiers as relations between sets. For instance,

$$[\![\text{some}]\!](A)(B) = 1 \text{ iff } A \cap B \neq \emptyset,$$

$$[\![\text{many}]\!](A)(B) = 1 \text{ iff } |A \cap B| \geq \mathbf{n}_s,$$

$$[\![\text{most}]\!](A)(B) = 1 \text{ iff } |A \cap B| > |A - B|,$$

where $\mathbf{n}_s$ is the contextually appropriate cardinality threshold.

This picture, however, does not account for the close relation between *many* and *most*. Subsequent developments (e.g., Hackl, 2000, 2009; Solt, 2009, 2011) try to derive the meaning of *most* compositionally from *many* and the superlative morpheme *-est*, to provide a more unified account of the two quantifiers that captures their relations.

What is left to be answered for any of these theories to be complete is, again, how the cardinality threshold $\mathbf{n}_s$ is contextually derived.

Note the striking similarity between *many* and relative gradable adjectives. From the empirical perspective, in addition to the well-known fact that *many* is both context-sensitive and vague, Yildirim, Degen, Tanenhaus, and Jaeger (2013) conducted a series of experiments and illustrated that people adapt their expectations of how *many* will be used after observing several instances of utterances, exhibiting the dynamic metalinguistic effect in the sense of Barker (2002). From the theoretical side, Solt (2009)

gives a degree-based analysis of *many* in parallel with gradable adjectives. Hence, it is reasonable to expect that our model for gradable adjectives can be similarly generalized to account for the quantifier *many* as well.

The organization of the chapter is the following. Section 4.2 introduces a basic model of *many* using the idea from the previous chapter and shows how dynamic metalinguistic effects could be accounted for. Section 4.3 introduces other cognitive factors involved in the complex lexical competition among alternative quantifiers such as *most* and numerals, and illustrate how they may be used to further address the discrepencies between empirical data and predictions of the basic model. In Section 4.4 we report on experimental evidence of the effects of these factors. Finally, in Section 4.5 we will conclude with a prelude to the discussion of experimental measures and modeling choices in the next chapter.

4.2 Degree-Based Modeling of *Many*

In this section we introduce the basic idea of generalizing the speaker-oriented model we developed in the last chapter to account for the use of quantifier *many*.

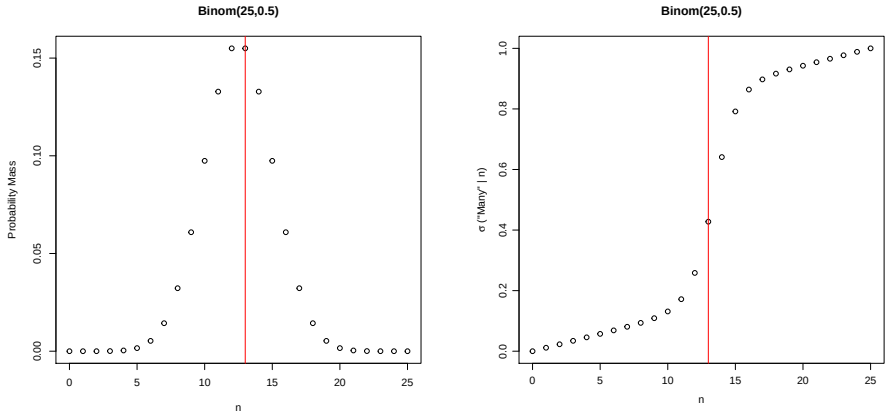
We will focus on the *partitive* construction of *many*, such as in the sentence "*many of* the candies are green." In general, we consider utterances of the form "Many of the A's are B."

As usual, the first step in our framework is to identify the purpose of such language use. We assume that the purpose of this type of sentences is to convey the cardinality of $A \cap B$, i.e., how many things in A are B. For instance, by saying "many of the candies are green" the speaker tries to make the listener have a better idea of the number of green candies.

For simplicity, we will assume that N, the cardinality of A, is fixed and known. Obviously this assumption is too strong in reality, but we should keep in mind that its main purpose is to first give us a good handle of the simple, idealized situations. Its role is similar to degree distributions on comparison classes for gradable adjectives. Later we will discuss how to gradually relax these restrictions.

This assumption also helps make controlled experimental studies feasible. For instance, Yildirim et al. (2013) presented participants with pictures of mixed green and blue candies, with the total number of candies in the picture fixed to be 25, and varied the number of green candies. We use this scenario as a working example and we will describe their experiments in detail after introducing below our basic model of the use of *many*.

Similar to our analysis of gradable adjectives, we will treat the contextual cardinality threshold $\mathbf{n}_s$ as a threhold variable θ and use $\Pr(\theta)$ to denote the speaker's linguistic knowledge about the distribution of the threshold. We use n to denote the cardinality of $A \cap B$, i.e., $n = |A \cap B|$. Note that "many



(a) Binom(25, 0.5), the red line is $n = 13$. (b) $\sigma(\text{“Many”} \mid n)$

Figure 4.1: SOM predictions for Binom(25,0.5), with $\lambda = 48, c = 0$.

of the A's are B" is true iff $n \geq \theta$. Our goal is to derive $\Pr(\theta)$, which will further yield the production model $\sigma(\text{“many”}|n; N)$ that we want.

The previous chapter addresses how $\Pr(\theta)$ can be constrained by the prior belief $\phi(n)$ about the cardinality of $A \cap B$, so our remaining task is to specify an appropriate prior $\phi(n)$.

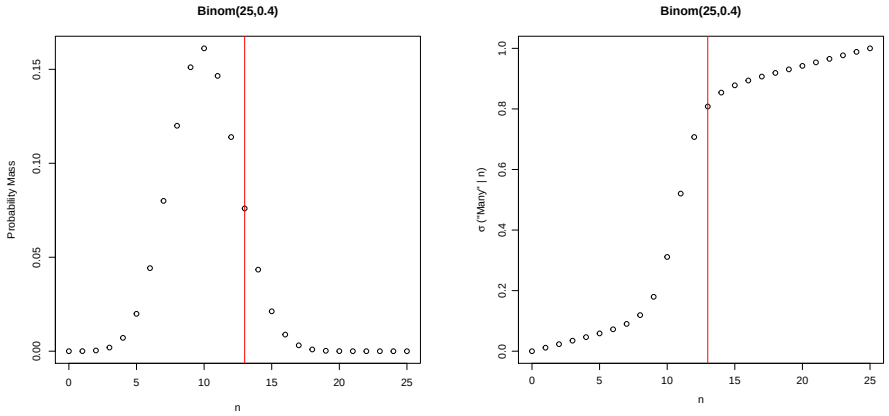
A natural candidate for $\phi(n)$ is the binomial distribution $\text{Binom}(N, r)$, where $r \in [0, 1]$ is the (believed) rate of observing the mentioned property B in the population of A. For example, in our scenario of candies, the $N = 25$ candies are assumed to be randomly drawn from a population of candies. A belief of $r = 0.6$ means that the speaker believes that in the whole population of the candies, 60% are green. Given such a belief about the general frequency of green candies, $\phi(n; N, r)$ is the probability of having exactly n green candies within a group of N candies.

Technically, we have[1]:

$$\phi(n; N, r) = \binom{N}{n} r^n (1 - r)^{N-n} \quad . \tag{4.1}$$

Typically, we do not have much prior knowledge about how colors are distributed in the population of candies, so we might think both colors are equally likely, i.e., $r = 0.5$. Fig. 4.1(a) shows the binomial distribution $\text{Binom}(25, 0.5)$. We can see that if 25 candies are drawn from a population where half of the candies are green, then we are most likely to see 12 or 13 candies are green, and it is quite improbable that we will see a mix of 5 green and 20 blue candies.

[1]Note that since cardinalities are discrete, $\phi(n; N, r)$ is the probability mass for n, rather than the density.



(a) Binom$(25, 0.4)$, the red line is $n = 13$. (b) $\sigma(\text{“many”} \mid n)$

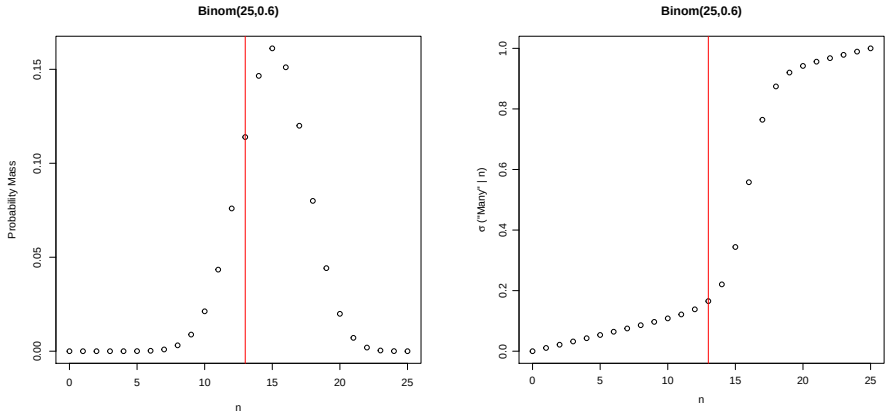
Figure 4.2: SOM predictions for Binom(25,0.4), with $\lambda = 48, c = 0$.

Now that we have the prior $\phi(n; N, r)$ specified, we can use the SOM to derive the production probability $\sigma(\text{“many”} \mid n; N, r, \lambda, c)$. Fig. 4.1(b) shows the production probability for each n, when the prior $\phi(n)$ is Binom$(25, 0.5)$, with $\lambda = 48$ and $c = 0$.[2] We can see that the production probability is around 50% when n is between 13 and 14, and gradually becomes higher. The entire curve is S-shaped, similar to relative adjectives.

The production probability $\sigma(\text{“many”} \mid n; r)$ is sensitive to the population rate r. Fig. 4.2 shows the binomial distribution Binom$(25, 0.4)$ and the corresponding production probability $\sigma(\text{“many”} \mid n; 0.4)$, and Fig. 4.3 shows those when $r = 0.6$.

We can see that for a lower population rate, the probability of using “many” generally increases, and vice versa for a higher population rate. This is expected. For example, suppose someone sees a group of 25 people in an Asian country where the majority of the population has black hair (r is low for blond), if 10 people in that group are blond, it seems fairly likely that he will use “many of the people in that group are blond” as a description, at least much more likely than he would if he saw the group in Northern Europe, where blond hair is not uncommon (higer r).

[2]For simplicity, we will use $\lambda = 48, c = 0$ throughout the chapter to illustrate the concept. Further experimental work is needed and more information about these parameters can be learned from the empirical data in the way introduced in the previous chapter. Since N is often fixed as well, later we will only write $\sigma(\text{“many”} \mid n; r)$ instead of $\sigma(\text{“many”} \mid n; N, r, \lambda, c)$.



(a) Binom(25, 0.6), the red line is $n = 13$. (b) $\sigma(\text{“many”} \mid n)$

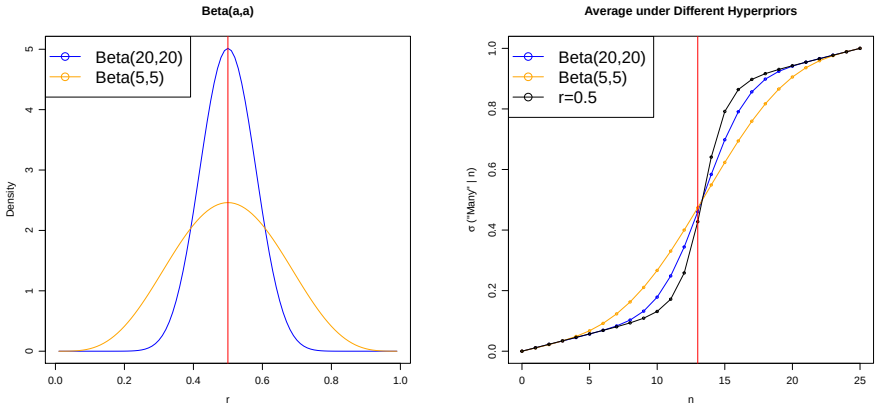
Figure 4.3: SOM predictions for Binom(25,0.6), with $\lambda = 48, c = 0$.

Uncertainty about the Population Rate r

So far we have been assuming that the population rate r is known definitely by the speaker. In reality, this assumption is of course too strong. The population rate r is a latent parameter of the speaker's belief about the population. As was discussed in the previous chapter about comparison classes for gradable adjectives, even if our world knowledge and contextual information provides us with some prior preference about the latent parameter, we usually cannot be entirely sure. Moreover, when we consider conversations, which involve two individuals, it becomes even more evident that we generally cannot expect to know exactly the latent parameters corresponding to what other conversational participants believe about the population. The best we can assume is that people general share the same type of latent parameters, with some noise introducing variability. Hence the latent parameter r is always associated with some uncertainty and we will adopt the beta distribution family again to encode such uncertainty.

We introduced the beta distribution family in Section 3.4. Recall that a beta distribution is defined on $[0, 1]$ and has two positive parameters α, β controlling its shape. While we used it there only because of its variability in shape and nice closure property, the shape parameters of beta distribution in fact have deeper conceptual meanings in Bayesian statistics. For the sake of our modeling, here we only state the mathematical fact that Beta(α, α) is symmetric and has maximal probability density at 0.5. In addition, the greater α is, the more centered at 0.5 the distribution becomes. In the limit case where $\alpha \to \infty$, the distribution effectively becomes a vertical line $r = 0.5$. Fig. 4.4(a) illustrates examples of such distributions.

Since we already know the production probability $\sigma(\text{“many”} \mid n; r)$ for



(a) Beta(α, α), the red line is $r = 0.5$.

(b) $\sigma(\text{“many”} \mid n; \text{Beta}(\alpha, \alpha))$

Figure 4.4: Average SOM predictions under hyperpriors Beta(α, α).

each r, now with the hyperprior quantifiying our belief about r, we can again use the Bayes model averaging technique introduced in Section 3.6 to calculate the average production probability as the weighted average of $\sigma(\text{“many”} \mid n; r)$ using the density $\psi(r)$ as weights:

$$\sigma(\text{“many”} \mid n; \psi(r)) = \int_0^1 \psi(r)\sigma(\text{“many”} \mid n; r)\, \mathrm{d}r \quad .$$

Fig. 4.4(b) shows the average production probability when $\psi(r)$ is either Beta(5,5) or Beta(20,20). We can see that as $\alpha \to \infty$, the average production probability converges to the production probability when r is exactly 0.5.

Metalinguistic Effects of Use of Quantifiers

Incorporating uncertainty about the latent parameter r not only gives us a more realistic production model, but also is the key to accounting for metalinguistic effects of use of vague terms (Barker, 2002).

Let us return to the experiment by Yildirim et al. (2013) as an example to illustrate the effects and show how to formally account for them.

In the experiment, participants were first presented with 10 pictures[3] in which 13 out of 25 candies were green, together with a video of a speaker constantly using “some of the candies are green” or “many of the candies are green” to describe these pictures. Participants were then presented with pictures containing 25 candies. The pictures had various numbers of green candies and participants were asked to judge the production probability of “many” and “some” by the speaker.

[3]The pictures were mixed with 10 other filler pictures to make the manipulation less obvious.

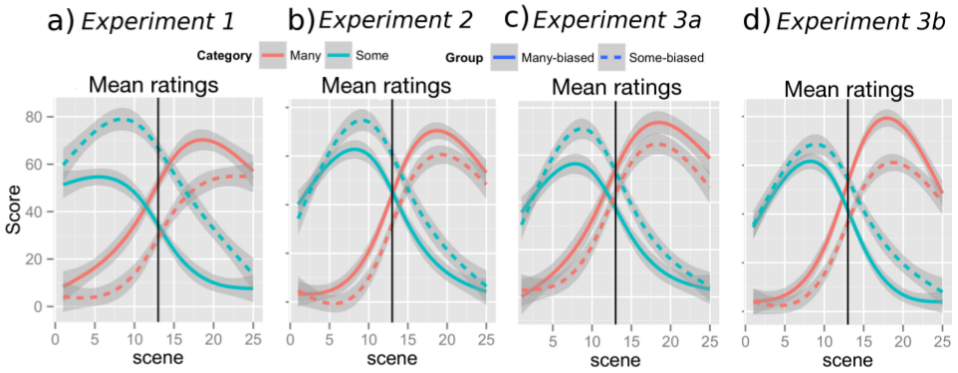


Figure 4.5: Results of the experiments by Yildirm et al. (2013)

They used a series of experiments to test the robustness of the phenomenon and obtained similar patterns in participants' responses. The results of their experiments are shown in Fig. 4.5. We can see that participants' expectation of the speaker's use of "many" changed according to the previous observations of use of quantifiers. If the speaker constantly used "some" when $n = 13$ (*some-biased*), then participants would judge "many" (red dashed line) generally as less likely to be used by the speaker for each cardinality, and vice versa for participants' judgments after observing constant use of "many" when $n = 13$.[4]

We should note that this phenomenon cannot be simply explained in terms of adaptation to the superficial relative frequencies, as differences in judgments may not be constant for each cardinality. Our analysis is that participants use Bayes' rule to update their hyperprior of the population rate r from the observed use of quantifiers:

$$\psi'(r) \propto \psi(r) \cdot \sigma(u_{n:k/K} \mid r),$$

where $u_{n:k/K}$ means that utterance u in total is used k times out of K observations for cardinality n.

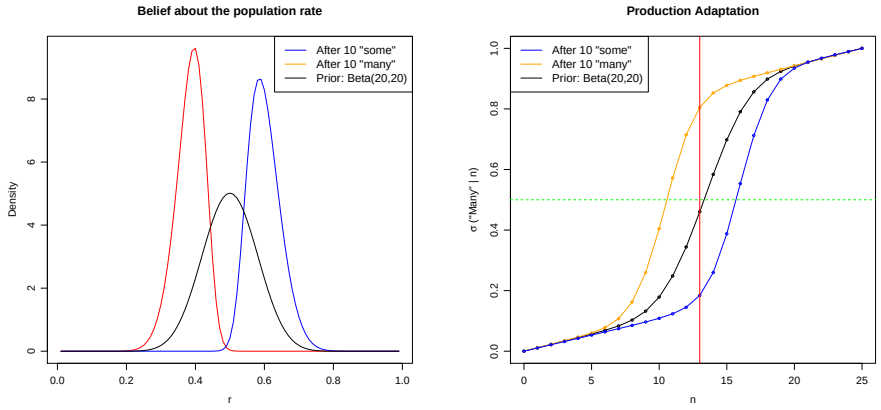
The intuition is that participants were revising their belief about the speaker's latent parameter r in the light of the observations of his use of quantifiers. Note that for each of the K observations, we know the probability of using utterance u is $\sigma(u \mid n; r)$.[5] Hence, the probability of using utterance u k out of K instances of language use, is given by the binomial distribution:

$$\sigma(u_{n:k/K} \mid r) = \text{Binom}_k(K, \sigma(u \mid n; r)),$$

where Binom_k is the binomial distribution's probability mass at k.

[4] Yildirim et al. (2013) did not report the baseline in detail. They only mentioned that participants were indifferent between "many" and "some" when $n = 13$.

[5] For now we will assume "some" is always used when "many" is not. Note the similarity between "some" here and saying nothing as the alternative to positive forms.



(a) Posteriors

(b) $\sigma(\text{“many”} \mid n)$

Figure 4.6: Posterior average SOM predictions.

Hence we can compute the posterior distribution of population rate r, as shown in Fig. 4.6(a). We can see that if participants observed constant use of “some” by the speaker, he would think that the speaker has a belief that correponds to a higher population rate, and vice versa for “many.”

Using model averaging again, we obtain the average posterior production probability of “many,” as shown in Fig. 4.6(b). We can see that the production probability shifts to the intended direction after either type of observations.

4.3 Cognitive Factors and Language Use

It is easy to see that the predictions in 4.6(b) can not fully account for the use of quantifiers “many” and “some.” In general, they capture the borderline cases reasonably well, but they give bad predictions near the ends of the cardinality scale.

Towards the lower end, the account predicts that “some” will almost always be used when n is small (since “some” is assumed to be used whenever “many” is not) and the production probability monotonically decreases as n grows greater. However, this is not the case in the empirical data shown in Fig. 4.5, where participants gave low ratings for “some” on such cardinalities. The judged production probability first increases from $n = 1$ to somewhere between 5 and 10 before it gradually drops as n increases.

Similarly, towards the upper end, the SOM predicts “many” will almost always be used and the production probability always increases as n grows. In contrast, participants’ judgments in the empirical data peak at some n around 20 and drops down afterwards.

There are many plausible explanations for the discrepencies. One intuitive account is that people probably would use exact number terms instead when n is small, and when n moves towards the upper end of the cardinality scale, it becomes more likely for the speaker to use "most" and maybe eventually expressions such as "almost all" and "all but one."

This explanation points out that there are many other linguistic expressions that can serve the purpose of conveying the cardinality. Hence we would expect a much more complex competition and interaction among linguistic alternatives in the domain of quantification than most domains underlying gradable adjectives. For example, when talking about a person's height, we do not have many common, simple linguistic alternatives to "tall" other than "short."[6] For many other gradable adjectives it is even hard to find their antonyms.

As mentioned earlier, the SOM in this and the previous chapters is a "semantic pragmatic" account, in the sense that the analysis only considers the use of an expression *in isolation*. The primary consideration is that, if the expression of interest is the only available expression, how it should be used in order to (sub-)optimally serve its purposes. It is because of this isolation that the resulting model presumably captures some property of the *linguistic expression*, which can then be seen as part of the semantics of that expression.

However, as we have seen in Chapter 2, language use can and arguably should also be considered on the level of "pragmatic pragmatics," where speakers have to choose among several linguistic expressions which are all semantically true but differ in other respects such as informativity.

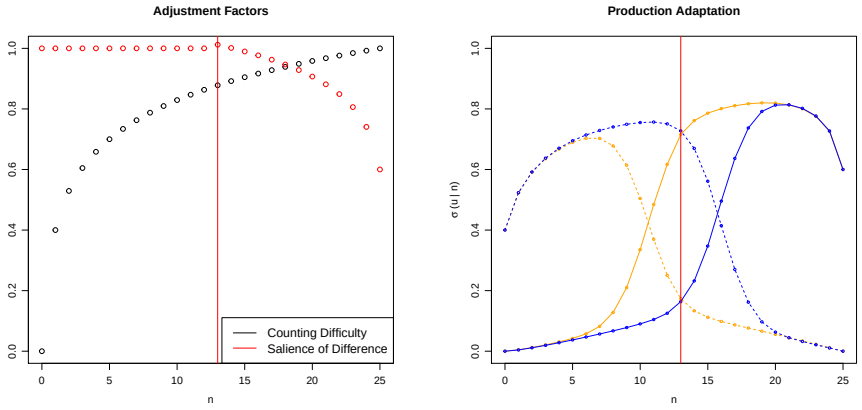
This section will focus on this aspect of competition among alternative expressions to better explain participants' judgments.

There are at least two approaches. A fully generative approach is what we have seen in Chapter 2. It completely spells out the range of alternatives and make them freely compete with each other. This would be ideal in helping us fully understand the phenomena. However, so far it is unclear how expressions which involve latent parameters would fit in such an account. The major difficulty is that if we simply take all latent parameters into account on the same level, the interactions will become too complex to be tractable.

Hence, in this section we will adopt a *semi-generative* approach, according to which lexical competitions can happen on different levels, so that we can treat some level as the basic and adjust model predictions according to considerations from other levels.

More concretely, in our analysis of the use of "many" and "some," we

[6] However, if we take into account modifiers such as *very*, we might similarly observe complex interactions. For instance, someone who is 2.2m is probably not described as just "tall," but as "very, very tall."



(a) Adjustment factors.

(b) $\sigma(\text{“Many”} \mid n)$

Figure 4.7: Adjusted SOM predictions for *many* and *some*.

will treat the SOM as the basic model that captures the informative use of “many,” and additionally consider two other cognitive factors.

The first cognitive factor is the difficulty in counting the exact cardinality n, which involves the competition between non-specific quantifiers such as “many,” “most,” and “some” and number quantifiers such as “one,” “two,” and “three.” Since number quantifiers are more specific and thus more informative, they should generally be used. However, when n grows larger, it takes more effort to count it and it becomes easier to make mistakes, which means number terms become less appealing. Hence we will first assume that the overall probability of using non-specific quantifiers such as “many” is proportional to the difficulty of counting the cardinality n correctly.

Concretely, we have

$$p_{\mathrm{d}}(n) = a_{\mathrm{d}} + b_{\mathrm{d}} \log(n),$$

where $a_{\mathrm{d}}, b_{\mathrm{d}}$ are constants. The reason that we use $\log(n)$ instead of n is that human perception is often linear on the log-scale, known as *Fechners law* in psychophysics. For simplicity, we will specify it by considering endpoint values. When $n = 1$, $p_{\mathrm{d}}(1) = a_{\mathrm{d}}$, which according to empirical data is around 0.4. Thus we know $a_{\mathrm{d}} = 0.4$. When $n = 25$, for simplicity we will assume that $p_{\mathrm{d}}(25) = 1$, from which we know $b_{\mathrm{d}} = 0.186$.

The second factor we consider is the strengthening from “many” to “most.” We will adopt the analysis of “most” as the composition of “many” with the superlative morpheme *-est* (e.g., Hackl, 2000, 2009; Solt, 2009, 2011). We will take the meaning of *many* as a function that maps sets to cardinalities, just as *tall* maps individuals to heights. The meaning of *-est* is to assert that the individual has the greatest degree among all alternatives. For instance, x is tallest means $\mathbf{height}(x) > \mathbf{height}(y)$ for any other y in

the comparison class. In the case of *most* as *many-est*. When we use "most of the candies are green," the superlative morpheme involves a comparison between the set of green candies and the alternative set(s), which in this case is the set of blue candies. Thus, G, the set of green candies, is *many-est* when $|G| > |Y|$ for any other alternative set Y in the comparison class, which in this case is only B, the set of blue candies.

Again, if the sizes of these two sets are obviously different, then one should strengthen "many" to "most" more likely. When there are n green candies, there are $25-n$ blue candies, so the difference is $n-(25-n) = 2n-25$. We can similarly derive the probability factor of maintaining "many" as a function of n. An example is shown in Fig. 4.7(a) and Fig. 4.7(b) shows the adjusted production probabilities of "many" and "some" (dotted lines) after observing "some" (orange lines) or "many" 10 times for $n = 13$.

We can see that even though the predictions are not perfect, it generally captures the trends towards the endpoints.[7]

4.4 Experiments

In the previous sections we generalized the SOM to account for the use of quantifiers, and introduced two additional cognitive factors, i.e., the difficulty of counting and the salience of difference between alternative sets. As noted earlier, our semi-generative model is a primitive one, and more work needs to be done to systematically learn from the empirical data and design better models.

In this section we will report on the results of two experiments adapted from Degen and Goodman (2014) to empirically justify the introduction of these two cognitive factors. The aim is to establish the plausibility of the choices, rather than to fully validate the model prediction, which we leave to future work.

4.4.1 Experiment 1: Salience of Difference and Strengthening of *Many* to *Most*

Our first experiment is about the salience of difference bewteen alternative sets and its effect on the strengthening of *many* to *most*. In the previous sections, we assume that the alternative set is perceived at the same time along with the main set. For instance, in a picture of candies, the alternative set (blue candies) and the main set (green candies) are in the same picture and can be directly compared against each other.

[7] We pointed out that the phenomenon cannot be explained simply as adaptation to relative frequencies, but such an adaptation could still be one of the factors. Also, we assume the QUD is how many green candies there are, but if participants observed "some" constantly, they might assume that the QUD could be whether there are green candies. These factors are not in the current model.

However, in some situations the alternative set is not present in perception, which means a direct comparison is impossible. For example, suppose we know that 25 candies were distributed in two boxes. When we open the first box, see 20 candies, and say "most of the candies are in this box," we do not directly observe the alternative set (candies in the other box). In situations like this, the speaker can only obtain the cardinality of the main set directly from perception, and has to calculate the cardinality of the alternative set for comparison.

There are two modes of perception of the cardinality of a set, either the speaker counts and believes in a specific cardinality n, or the speaker just forms a coarse representation of the numerosity of the set, resulting in a belief across a wider range of n's.

In the first case, if the speaker counts correctly, the salience of difference will be similar to that in the previous section where the alternative sets are available.

In the second case, since perception is fuzzy for large cardinalities, the coarse representations the speaker forms there will be close to each other (e.g., the coarse representation for $n = 18$ will not differ much from the one for $n = 21$). This means that the difference between the main and alternative set will be less variable (more flat) over large cardinalities.

Thus, when the cardinality n of the target set is fixed, if the verification strategy plays a role in the strengthening of *many* to *most*, we would expect to see an interaction between the proportion n/N and whether exact counting takes place, on the judgment of "most." When the proportion n/N is low, if participants count the exact number correctly, then they will judge "most" as less likely than those who do not count. When the proportion n/N is high, if participants count the exact number correctly, then they will judge "most" as more likely than those who do not count.

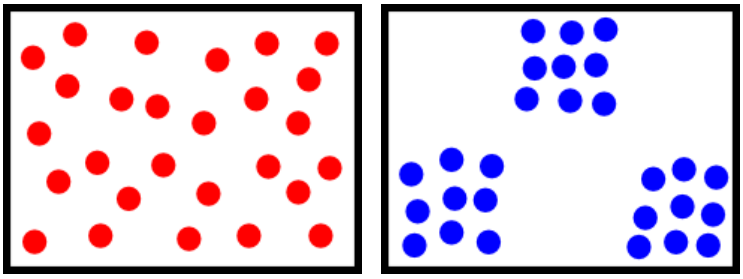
Participants

We recruited 144 US participants via Amazon's Mechanical Turk. Each participant received a payment of $0.25.

Materials and Procedure

Participants first read a brief story about a character losing her marbles in shoe boxes due to the visit of a relative. The story introduced N, the total number of marbles which are relevant in the context, as either 37 or 53, and evoked a *question under discussion* (QUD) that the speaker needs to find all of the lost marbles. An example of the story is as follows:

> Your friend, Kate, is really into collecting marbles. One afternoon, you run into her near her house and she invites you home for tea. On the way she talks enthusiastically about the



(a) Randomly Distributed (b) Orderly Distributed

Figure 4.8: Target Pictures used in Experiments 1 and 2.

special edition of 53 marbles that she recently added to her collection. "I put them on the shoe cabinet and they look perfect!" she says proudly.

When you enter her house, however, there are shoe boxes everywhere on the floor and the new set of marbles is gone. "It must have been my five-year-old nephew," Kate says, "Oh he is always so naughty!"

Kate immediately starts opening one box after another. You know her well enough to tell that she is upset and is determined to find every last one of her new marbles. Wanting to help, you pick up a shoe box and open it...

Participants were then asked to answer the following two questions to make sure that they did pay attention to the story: *How many marbles are there in Kate's missing new set?* and *When will Kate be satisfied?*. The two options for the first answer were the correct N and an incorrect one, and the two options for the second answer were *If she finds every missing marble.* (correct) and *If she finds at least one missing marble* (incorrect). They were allowed to proceed only after they had answered the questions correctly.

Then participants were shown a target picture containing a box of $n = 27$ marbles randomly distributed (Fig. 4.8(a)). We chose $n = 27$ for two reasons. First, previous pilot studies suggested that roughly half of the participants would count and the other half would not when n is around 27. Second, for $n = 27$, we can construct pictures of orderly distributed marbles whose total number can be easily counted (Fig. 4.8(b)). They will be used in Experiment 2. Also, the total numbers of the missing marbles (53 or 37) were chosen to make $n = 27$ has proportions of slightly greater than 50% or slightly less than 75%.

Participants were told that they called out to the friend upon seeing the box: *"I found _____ of the missing marbles!"* They were asked to fill in the blank with (only) one word that they thought they would most likely use.

After that the picture disappeared and participants were asked how many marbles there were in the picture as a memory check. They were instructed that if they knew the answer, then they should report it, otherwise they should indicate that they did not know.[8]

Finally, the participants were asked to adjust sliders to rate how likely that they would use each of the four words *some*, *many*, *most*, *all* to complete the sentence. The endpoints of the sliders were marked as *very unlikely* and *very likely* and numerically the ratings ranged from 0 to 100 (but participants did not know the number).

Names in the story, the character and the visiting relative's genders, marble colors[9] and positions were all randomized.

Results

We classify participants' responses to the memory check into 3 categories: Dunno (when they stated they did not know the exact number n), Correct (when they gave the correct $n = 27$) and Incorrect (when they gave an incorrect n). Table 4.3 shows the number of participants in each category.

Table 4.1: Responses to the memory check in Experiment 1

Condition	Dunno	Correct	Incorrect	Total
$N = 53(50\%)$	34	28	10	72
$N = 37(75\%)$	24	37	11	72

We can see that a small proportion of participants incorrectly counted the number of marbles. This is understandable, considering there were so many randomly distributed marbles in the picture. In fact, most of the incorrect numbers are 26 or 28.

In this chapter we focus on participants' slider ratings for the quantifiers. We will report on and discuss participants' free choices of words in the next chapter.

Table 4.2: Average slider ratings in Experiment 1

Group	Some	Many	Most	All
$N = 53$ (50%), Dunno	83.59	51.47	57.53	13.65
$N = 53$ (50%), Correct	87.82	53.61	37.50	5.54
$N = 37$ (75%), Dunno	65.83	54.54	70.54	18.00
$N = 37$ (75%), Correct	67.59	67.49	84.78	10.65

[8]They were instructed explicitly that stating that they did not know would not affect the payment.

[9]There is only one color in each picture, as different colors may have some grouping effect which can make counting easier.

Table 4.2 shows the average slider ratings for each quantifier in each condition given by the Dunno and Correct groups. [10] We can see that *all* receives very low ratings, which is not surprising since it is false in both conditions. We can also see that within each condition Dunno and Correct groups gave *some* very close average ratings. This suggests that judgment of *some* does not need to involve counting, which is also unsurprising given its semantics.

We are mainly interested in the judgments for *many* and *most*.

We can see that when the proportion is a little more than 50%, both groups gave similar ratings for *many* but the participants who counted correctly seemed to rate *most* as less likely than those who did not count.

In contrast, when the proportion is slightly less than 75%, participants who counted correctly seemed to give higher ratings for both *many* and *most*.

Statistical Analysis

We will now test whether the above observations are statistically significant. Since participants' ratings were not normally distributed and we do not want to make ad-hoc assumptions about the distribution, we adopt *non-parametric bootstrapping*, a resampling method, to test whether the differences between the two groups' ratings are significant.

Concretely, we will construct the 95% confidence interval of the difference between average ratings of the Dunno group and that of the Correct group for each condition. First, we sample with replacement from our original data to get another dataset of the same size (72 for each condition), and (for each quantifier) we compute the difference in average ratings, diff_1, between the Dunno and Correct groups in the new dataset. Next, we repeat this procedure 1000 times to get 1000 new datasets and correspondingly $\text{diff}_1, \text{diff}_2, \ldots \text{diff}_{1000}$. Last, we take the 95% percentile interval of these 1000 differences and treat it as the confidence interval of the difference between the two groups' ratings. If 0 falls out of this confidence interval, then we conclude that the difference is significantly.

Table 4.3: Bootstrap: Dunno − Correct ($N = 53$ (50%))

	Some	Many	Most	All
Mean Diff	-4.20	-2.39	20.14	7.97
95% CI	$[-12.75, 4.91]$	$[-17.08, 11.76]$	$[\mathbf{6.01}, \mathbf{33.62}]$	$[-0.18, 16.95]$

[10] Since there were only a few participants who incorrectly counted the number of marbles, and they had different such false beliefs, we just exclude their responses from the analysis.

Table 4.4: Bootstrap: Dunno − Correct ($N = 37$ (75%))

	Some	Many	Most	All
Mean Diff	-1.78	-12.96	-14.11	7.28
95% CI	$[-16.99, 13.47]$	$[-28.72, 2.45]$	$[\mathbf{-29.33}, \mathbf{-0.36}]$	$[-3.22, 19.37]$

The bootstrap 95% confidence intervals of the differences between two groups' ratings in each condition are shown in Table 4.3 and Table 4.4.

We can see that only the differences between two groups' ratings for *most* are significant. When the proportion is a little more than 50%, participants who counted correctly rated *most* significantly lower than those who did not count, and when the proportion is slightly less than 75%, participants who who counted correctly rated *most* significantly higher than those who did not count.

Discussion

The results suggest that the salience of difference affects the strengthening of *many* to *most*. The larger the difference between the cardinalities of the main and alternative sets is, the more likely that *most* would be used.

4.4.2 Experiment 2: Effect of Difficulty in Counting

Our second experiment is about the effect of difficulty in counting. The natural prediction is that if counting is easy, then speakers are supposed to use vague quantifiers less often.

Participants

We recruited 48 US participants[11] via Amazon's Mechanical Turk. Each participant received a payment of $0.25.

Materials and Procedure

The only difference is that the pictures presented to the participants had marbles that are orderly distributed (Fig. 4.8(b)), making it very easy to count correctly.

Results and Statistical Analysis

Table 4.6 shows participants' responses to the memory check. We can see that clearly the pictures with orderly distributed marbles made almost all participants count and count correctly.

[11]The size was chosen such that roughly the same number of participants would count correctly as that in Experiment 1.

Table 4.5: Responses to Memory Check in Each Condition in Experiment 2

Condition	Dunno	Correct	Incorrect	Total
$N = 53$	0	23	1	24
$N = 37$	2	22	0	24

Table 4.6 shows the average slider ratings for each quantifier by the participants given by the participants who counted correctly in Experiment 1 (Random) and Experiment 2 (Orderly).

Table 4.6: Average slider ratings by the Correct group

Group	Some	Many	Most	All
$N = 53$ (50%), Random	87.82	53.61	37.50	5.54
$N = 53$ (50%), Orderly	86.52	56.17	33.83	8.09
$N = 37$ (75%), Random	67.59	67.49	84.78	10.65
$N = 37$ (75%), Orderly	68.82	53.91	73.18	16.95

We can see that participants' ratings for *some* and *all* are again very stable. This suggests that for large cardinalities, *some* might be judged truth-conditionally, i.e., at least one, which does not involve counting the exact number.

When the proportion is slightly more than 50%, there is no much difference in participants' ratings for any quantifier. This could be a basement effect, since in general this is a borderline case for *many* and *most*, which means the base ratings might already be low and thus cannot drop too much.

When the proportion is slightly less than 75%, the ratings for *many* and *most* seem to drop when the marbles in the pictures were orderly distributed. Using the bootstrap method again, we can see that only the difference for *most* is statistically significant.

All the bootstrap 95% confidence intervals are shown in Table 4.7 and Table 4.8.

Table 4.7: Bootstrap: Random − Orderly ($N = 53$ (50%))

	Some	Many	Most	All
Mean Diff	1.26	-2.38	3.76	-2.61
95% CI	$[-6.60, 9.04]$	$[-17.42, 13.90]$	$[-11.37, 18.73]$	$[-11.30, 5.33]$

Discussion

The experiment suggests that the difficulty in counting has some effect on the use of quantifiers, but they might have different effect sizes for different

Table 4.8: Bootstrap: Random − Orderly ($N = 37$ (75%))

	Some	Many	Most	All
Mean Diff	-0.97	13.75	11.73	-6.42
95% CI	$[-15.34, 13.68]$	$[-4.36, 32.01]$	$[\mathbf{1.03}, \mathbf{23.43}]$	$[-20.94, 5.31]$

quantifiers, because the participants might use different verification strategies in their ratings. This means that the formal model may be improved if we introduce different effects of difficulty in counting to different quantifiers. For instance, for quantifier *some*, it could be that as long as the cardinality exceeds a threshold, participants would stop counting, so the difficulty factor $p_{\mathrm{d}}(n)$ may reach 1 faster for *some* than for *many* or *most*.

4.5 Discussion

In this chapter, we further apply the framework and generalize the analysis of gradable adjectives to the study of quantifiers. We show that the basic idea still applies and derives a quantitative model for the meaning of *many* and show how its meaning can be adapted after exposure to meta-linguistic use. However, we note that the use of quantifiers is more complex because of the competition among various lexical alternatives and some other cognitive factors are involved. As a first step, we use a semi-generative approach to incorporate two cognitive factors in the use of quantifiers: difficulty in counting and salience of difference between alternative sets. Our experiments provide evidence that these factors do play a role in the use of quantifiers.

Chapter 5

Design Choices

In previous chapters we have seen applications of the framework to several case studies of the meaning and use of language. Even though these studies share the main features of the framework such as careful considerations of the purposes of language use, incorporation of cognitive factors, and emphasis on empirical validation via experimental studies, there are nonetheless various design choices in both formal modeling and experimental setup made in each study. In this chapter, we will review and discuss some of these design choices and their implications.

We will start with a discussion on various experimental measures, and then turn to formal modeling.

5.1 Experimental Measure

There are many ways in which one can observe language use in an experiment. Commonly used experimental measures include categorical judgment/forced choice, graded Likert scale/slider ratings (which can be seen as either ordinal or quantitative), and (possibly restricted) free production.

We classify experimental measures to two main categories: free production and metalinguistic judgments.

5.1.1 Free Production

Free production is closest to how people actually use language and is one of the most important goal of our theory and models of language use.

However, clearly it is usually unrealistic to allow participants to use whatever utterance they want without any constraint. A pratical solution is to use a *restricted free production* paradigm, where participants' choices of expressions are under restrictions. Admittedly, restricted free production is a somewhat contradictory notion, but it is intended to show the balance between freedom and feasibility of subsequent analysis.

For example, in our experiments introduced in the previous chapter, we let participants to freely choose one word to fill in the utterance template *"I found _____ of the missing marbles!"* This test is known as *cloze test* (Taylor, 1953) and is used in the assessment of language skills. Thus we adopt it to measure native speakers' disposition to use language.

Participants' choices of words are shown in Table 5.1.

Table 5.1: Free Production (27 Marbles in the Picture)

N	Picture	Group	Number	Some	Many	Most	All	Other
53	Random	Dunno	6	17	2	3	3	3 (1 half)
53	Random	Wrong	5	2	0	0	0	3 (half)
53	Random	Correct	15	6	0	0	0	7 (half)
53	Ordered	Correct	12	6	1	0	0	4 (half)
37	Random	Dunno	2	9	2	9	1	1 (half)
37	Random	Wrong	6	2	1	2	0	0
37	Random	Correct	19	4	1	11	0	2 (several)
37	Ordered	Correct	13	2	0	6	1	0

When $N = 53$, we can see that if participants counted the number (correctly or wrongly), around half of them will use the number term which they believed to be the number of marbles in the picture, and the other half split between *some* and *half*. Only one participant used *lots*, which we classified as *many* for simplicity.

In contrast, around half of the participants who did not count the number used *some*, the other half of the responses are scattered among other quantifiers, round numbers, and others such as *bunches*.

There is no doubt that the data contain a lot of noise and the size of the dataset is too small to draw strong conclusions. Nevertheless, there are already noteworthy patterns in the data.

First, *half* was used loosely by the participants. Semantically speaking, *half* is never true for any of the cardinalities when the total size N is 53, but it was indeed used by some participants. Note that those who wrongly counted the number as 24, 26 or 28 were also willing to use *half*.

Of course, it is far from a novel finding that people can use language loosely. However, this does inform us of what might be missing in our previous model for the use of quantifiers. Moreover, it also posts a theoretic challenge to our modeling techniques, i.e., how we could allow for loose use of language in the model.

Second, *many* and *most* are seldom used by participants who counted, but they received around 50 in slider ratings. This suggests that additional theory of how slider ratings correspond to production probalities is needed, even though participants were instructed to directly evaluate the production probability. A possible method is to normalize the slider ratings (e.g., Degen

& Goodman, 2014). However, we note that this is far from ideal, because even if the slider ratings were normalized, the result would still suggest that using *some* could not be more than twice likely as using *many* or *most*.

Similarly, when $N = 37$, we can see that participants used *many* much less often than *most*, but the large difference was not directly reflected in the slider ratings (See Table 4.2).

These discrepencies between free production and slider ratings prompt us to reflect the nature of metalinguistic judgments.

5.1.2 Metalinguistic Judgments

People not only use language, but also *talk about* langauge. In fact, speakers can have very strong opinions about language. They often agree with each other, but this is not always the case.

That people have certain patterns of linguistic or metalinguistic judgments is by itself a very fascinating cognitive phenomenon that deserves a lot of research. However, as we have already seen, speakers may have judgments different from their actual use of language.

Thus, a theory needs to be explicit about which type of phenomena it intends to address and spells out the assumptions that link theoretical predictions to experimental measures. Of course, since linguistic and metalinguistic judgments are often closely related to actual use of language, normally a theory can explain both types of phenomena, as long as it has a proper mechanism that links them together.

Since the focus of classical semantics and pragmatics is on interpretation, linguistic and metalinguistic judgments are traditionally explained from the interpretation side. Recently, Degen and Goodman (2014) tested a series of different measures and argued that some of the phenomena about metalinguistic judgments, including *sentence verification* might be better explained from the perspective of production, particularly in the light of formal, probabilistic models of language production and interpretation.

It might seem that this account would fail to explain the discrepencies between slider ratings for quantifiers and how speakers actually use quantifiers, but such discrepencies can actually be resolved in the framework.

Recall that the framework accounts for interpretation in terms of listeners' perspective-taking integrated with other cognitive factors. The crucial observation is that, the listener's perspective-taking is always with respect to his belief about the speaker, which may not be fully accurate. Thus, as Degen and Goodman (2014) speculated, some metalinguistic judgments such as sentence verification or slider ratings really measure listeners' underlying speaker models.

Let us try to use this hypothesis to explain why *many* was seldom actually used, but still received considerable ratings. One explanation is that the partitive constrution *many of* may be less acessible to some of the speak-

ers. In other words, speakers might have some production bias against the partitive construction of *many*. Meanwhile, listeners[1] could be unaware of such a bias, and thus they underlying speaker models do not incorporate this bias. This would explain why the slider judgments do not reflect the scarcity of *many of* in free production.

Clearly, this is only one of the many possible explanations of the phenomenon we observed, and it needs to be further rigorously tested. However, the main point here is the realization that the speaker and the listener may be influenced by different cognitive factors, and in addition that the speaker and listener may not be aware of each other's biases.

In the next section, we will revisit our first case study of referential expressions and illustrate how to incorporate the above observation into the formal models to enrich the framework to better capture the use of language.

5.2 Formal Modeling

In previous sections and chapters, we have seen the need for various cognitive factors in our model to better capture various phenomena of language use. In particular, such factors may be different for production and interpretation and may not be shared by speakers and listeners. Finally, we need to specify the link between models and experimental measures.

In this section we discuss variations in modeling the use of referential expressions.

For production, there are three factors we will consider.

Recall the soft-max production rule (2.6), repeated here:

$$\sigma(u \mid t) \propto \exp(\lambda_{\mathrm{S}} \cdot \log \mathcal{U}(t \mid [\![u]\!])) \ .$$

Firstly, as hypothesized in last section, the speaker might have lexical preferences.

In our case, the speaker might prefer features of shapes, which are nouns, to color terms, which are adjectives.

Technically, we can introduce a cost term that encodes the such preferences.

$$\sigma(u \mid o) \propto \exp(\lambda_{\mathrm{S}} \cdot (\log \mathcal{U}(o \mid [\![u]\!]) - \mathrm{Cost}(u))) \ .$$

We define the cost function using a constant $c \in \mathbb{R}$

$$\mathrm{Cost}(u) = c \text{ if } u \text{ is an adjective and 0 otherwise} \ . \tag{5.1}$$

If $c > 0$ it means there is a preference for nouns and if $c < 0$ then the preference is for adjectives. No preference exists if $c = 0$.

[1] We should emphasize here that the speaker and listener here could even be the same person, since the bias may be subconscious.

Secondly, RSA measures the informativeness of an utterance u, in terms of how close the induced belief of the literal listener is from the speaker's own belief. We will call it a *belief-oriented* view.

However, in our case of using referential expressions, it seems more important that the literal listener actually attends to the intended target. Hence an *action-oriented* view would directly use probability $\rho_0(t \mid m$ instead of the KL divergence.

$$\sigma_{\mathrm{a}}(u \mid t) \propto \exp(\lambda_{\mathrm{S}} \cdot (\rho_0(t \mid u) - \mathrm{Cost}(u))) \ . \tag{5.2}$$

The technical difference is whether the probability or its logarithm is used in the formula.

Finally, since we introduce perceptual salience as a cognitive factor for the pragmatic listener model, we want to test whether the speaker will also take it account in production. Thus, we replace the literal listener's uniform distribution $\mathcal{U}(t)$ with the salience prior $\mathcal{S}(t)$. This leads to the alternative production rule:

$$\sigma_{\mathcal{S}}(u \mid t) \propto \exp(\lambda_{\mathrm{S}} \cdot (\log \mathcal{S}(t \mid [\![u]\!]) - \mathrm{Cost}(u))) \ . \tag{5.3}$$

Hence we have four types of speaker models that differ in either the speaker's belief about the literal listener, or the speaker's goal of communication. We now introduce a uniform notation $\sigma_{xy},\ x \in \{\mathrm{a}, \mathrm{b}\}, y \in \{\mathcal{U}, \mathcal{S}\}$ for them:

$$\sigma_{\mathrm{a}y}(u \mid t) \propto \exp(\lambda_{\mathrm{S}} \cdot (y(t \mid u) - \mathrm{Cost}(u))), \tag{5.4}$$

$$\sigma_{\mathrm{b}y}(u \mid t) \propto \exp(\lambda_{\mathrm{S}} \cdot (\log y(t \mid u) - \mathrm{Cost}(u))), \tag{5.5}$$

where $\mathcal{U}$ is the uniform prior and $\mathcal{S}$ is the salience prior. For example, in the original RSA model, the speaker does not take listener's salience prior into account and he has a belief-oriented goal of communication. Thus it will be denoted as $\sigma_{\mathrm{b}\mathcal{U}}$.

Now we turn to the listener model (2.7),

$$\rho(t \mid u) \propto \mathcal{S}(t) \cdot \sigma(u \mid t) \ .$$

As pointed out previously, the likelihood term $\sigma(m \mid t)$ is the listener's belief about how the speaker behaves and it is possible that it might differ from actual production. We thus treat the speaker's production term $\sigma(m \mid t)$ in the listener's model as a parameter, making the listener's belief about production explicit:

$$\rho(\sigma_{xy})(t \mid u) \propto \mathcal{S}(t) \cdot \sigma_{xy}(u \mid t) \ . \tag{5.6}$$

Note that the speaker's production rule has two parameters λ_{S}, c which are also included in the above specification of the listener model.

Next, even though our intuition suggests different objects have different perceptual salience and thus might affect our interpretation, it is after all an empirical question whether it is a relevant cognitive factor. Hence we consider a variation where the listener does not take into account the perceptual salience in his reasoning, which means he has a uniform prior over the referents:

$$\rho_{\mathcal{U}}(\sigma_{xy})(t \mid u) \propto \mathcal{U}(t) \cdot \sigma_{xy}(u \mid t) \ . \tag{5.7}$$

Finally, in the original experiment by Frank and Goodman (2012), listeners were asked to bet over possible referents. We have argued that such an introspective measure may not be very accurate, and in the context of referential expression resolution it makes more sense for the listener to actually decide which is the indended referent. (Consider the case in which the listener is asked to pass something.)

Hence in such cases the listener will choose an object by, essentially, (soft-)maximizing over his posterior beliefs. Formally, the action-oriented listener model becomes:

$$\rho_{\mathrm{a}v}(\sigma_{xy})(t \mid u) \propto \exp(\lambda_{\mathrm{L}} \cdot \rho_{\mathrm{b}v}(\sigma_{xy})(t \mid u)), \tag{5.8}$$

where $v \in \{\mathcal{U}, \mathcal{S}\}$, λ_{L} is the parameter measuring the listener's degree of rationality, and $\rho_{\mathrm{b}v}$ is the belief-oriented model that does the Bayesian update and chooses the referent linearly proportional to the posterior:

$$\rho_{\mathrm{b}v}(\sigma_{xy})(t \mid u) \propto v(t) \cdot \sigma_{xy}(u \mid t) \ . \tag{5.9}$$

For instance, the original RSA listener model is a belief-oriented one with the perceptual salience as prior, whose belief about the speaker is $\sigma_{\mathrm{b}\mathcal{U}}$, hence it is denoted as $\rho_{\mathrm{b}\mathcal{S}}(\sigma_{\mathrm{b}\mathcal{U}})$.

Qing and Franke (2013) adapted the original experiment by Frank and Goodman (2012) and systematically compared the predictions of the above models to the new experimental data. The results suggest that lexical preference is a cognitive factor in production and perceptual salience is one in interpretation. However, the results suggest that the speaker does not take perceptual salience into account in production, and similarly the listener does not take lexical preference into account in interpretation. Moreover, an action-based view might better reflect the goal of communication and link model prediction to the experimental measure in a forced-choice setting where the listener was forced to select one object as the intended referent.

These results provide evidence for the hypothesis that some types of meta-linguistic judgments correspond to listeners' belief about production, which may be influenced by cognitive factors different from actual production. Of course, more work needs to be done to fully understand this relation, but it shall be clear that it is important to explicitly spell out the design choices in both formal modeling and experimental measure.

Chapter 6

Discussion and Conclusion

6.1 Challenges

We have seen that the framework has provided novel insights into a range of phenomena of language use and many of its predictions have been compared to empirical data and shown to have good predictive power. However, there may be methodological concerns about the framework. In this section we discuss the main objections and challenges.

First, some might be skeptical about whether a particular case study has implications for the study of meaning. The answer is that it depends. While the study of referential expressions might not be seen as about semantics *per se*, the studies of positive forms of gradable adjectives and quantifiers do provide new perspectives on issues in semantics.

Secondly, one might doubt whether formal modeling and experimental studies can be generalized to actual language use. The main objection is that most of the studies presented in this thesis seem too artificial to have real implications for our understanding and use of natural language.

Our response is that imperfections in case studies need not undermine the general merits of the framework. Abstraction, simplication and experimentation are crucial parts in the progress of any science. There is no doubt that generalization from a highly controlled environment to the real world, which is inductive in nature, is always risky and might turn out to be wrong. However, this does not mean that we should give up the entire process. Most of the time, the challenge is that the reality is far more complex to be modelled directly. Making simplifications and using controlled experimental methods help us tackle the huge problem piece by piece.

Of course, as always one should be cautious in interpreting the results of model predictions and experiments, especially when one tries to generalize them into more realistic, complex scenarios. Again, we need to repeat the process, do formal modeling and empirical investigations (including other methods such as corpus studies).

6.2 Conclusion

In this thesis, we propose a quantitative, social-cognitive and experimental framework for the study of the meaning and use of language, and use 3 case studies, i.e., referential expressions, positive forms of gradable adjectives, and quantifiers, to illustrate how the framework can be applied to provide unique contributions to different semantic and pragmatic problems.

The main advantages of the proposed framework are the following.

First, by focusing on production, especially the purpose of language use, we are not only able to provide an account of how language is used, but can also shed new light on the study of meaning, which is traditionally investigated primarily from the perspective of interpretation.

Second, by using quantitative models, our framework can better capture the subtlety of meaning and use, especially the graded or probabilistic linguistic judgments that are common in language use in reality.

Finally, by incorporating various cognitive factors and using experimental methods to inform model designs, the framework can better deal with the three-way relation between language, mind and world.

We shall emphasize that the framework proposed here obviously cannot solve all the problem. Rather, we propose to use it as a complementry approach such that, when combined with insights from other approaches, e.g., the rich tradition of formal semantics and pragmatics, can yield a better theory of the meaning and use of language. This is exactly what we plan to explore in the future work.

Acknowledgements

This thesis would have been impossible without the guidance of my supervisor, Michael Franke. No words can fully express my gratitude. I was lucky to take his course in the first semester of the program and it shaped my entire study. I never forgot the hours that he generously spent discussing various topics in semantics and pragmatics with me in his office or at a cafe. He taught me how to come up with interesting ideas, elaborate them, do rigorous works, and write and present them clearly and with style. Obviously I am far from mastering these, but at least I know what I should aim at, and he himself has set a perfect example for me to learn simply by imitation.

Many thanks go to my academic mentor, Raquel Fernández. She has been a great source of information and support in the past two years. I consulted her on course selections, individual research projects, and so on, and received tremendous help throughout the whole program.

I wish to furthermore thank Maria Aloni and Henk Zeevat, also members of my committee. Thanks to Maria for chairing my committee, the inspiring introductory course on formal semantics, and the help during PhD applications. Thanks to Henk for organizing the Bayesian Natural Language Semantics and Pragmatics workshop, which not only broadened my perspective of the field, but also provided me with the opportunity to present and publish in academia for the very first time.

I would like to thank Tanja Kassenaar and Ulle Endriss for being so helpful with various practical issues and patiently answering my (probably too many) questions about details.

I am also very grateful to those who helped make my study here possible in the first place. Special thanks to Yue Yang, without whom I would not have become interested in logic, let alone being aware of the Master of Logic program here. Thanks to the EMEA committee for funding my study. Thanks to Fenrong Liu and Yanjing Wang for the constant help and support before and throughout my study in Amsterdam, to Haitao Cai and Zhiguang Zhao for the practical information and advice during the application.

I would like to thank my housemates, Yuning Feng, Jingting Wu, and Jon Mallinson, and my fellow MoL students for the wonderful atmosphere in my life and study.

Life in another country sometimes is not easy, but my Chinese friends

have made it much smoother. Besides already mentioned, I would like to thank Zhenhao Li, Tong Wang, Shengyang Zhong, Tingxiang Zou, Fangzhou Zhai, and special thanks to Cihua Xu for the help during his visit here.

Last but definitely not least, I wish to thank my parents, for their unconditional love and support.

References

Austin, J. L. (1962). *How to do things with words*. Oxford: Clarendon Press.

Barker, C. (2002). The dynamics of vagueness. *Linguistics and Philosophy*, *25*(1), 1–36.

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

Benz, A., Jäger, G., & van Rooij, R. (2005). An introduction to game theory for linguists. *Game Theory and Pragmatics. Houndsmills, Basingstoke, Hampshire: Palgrave Macmillan*.

Bergen, L., Goodman, N. D., & Levy, R. (2012). That' s what she (could have) said: How alternative utterances affect language use. In *Proceedings of the thirty-fourth annual conference of the cognitive science society.*

Degen, J., & Franke, M. (2012). Optimal reasoning about referential expressions. In S. B.-S. et al. (Ed.), *Proceedings of SeineDial* (pp. 2–11).

Degen, J., & Goodman, N. D. (2014). Lost your marbles? The puzzle of dependent measures in experimental pragmatics. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society.*

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998. doi: 10.1126/science.1218633

Frank, M. C., Goodman, N. D., Lai, P., & Tenenbaum, J. B. (2009). Informative communication in word production and word learning. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 206–211).

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics & Pragmatics*, *4*(1), 1–82.

Franke, M. (2012). On scales, salience & referential language use. In M. Aloni, F. Roelofsen, & K. Schulz (Eds.), *Amsterdam colloquium 2011* (pp. 311–320). Springer.

Franke, M., & Jäger, G. (2014). Pragmatic back-and-forth reasoning. In S. Pistoia Reda (Ed.), *Semantics, pragmatics and the case of scalar implicatures* (p. forthcoming). Palgrave MacMillan.

Gatt, A., van Gompel, R. P. G., van Deemter, K., & Kramer, E. (2013). Are we bayesian referring expression generators? In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of CogSci 35.*

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling lanuage understanding as social cognition. *Topics in Cognitive Science*, *5*, 173–184.

Grice, H. P. (1989). *Studies in the way of words.* Cambridge, MA: Harvard University Press.

Hackl, M. (2000). *Comparative quantifiers.* Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: most versus more than half. *Natural Language Semantics*, *17*(1), 63–98.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, 382–401.

Jäger, G. (2013). Rationalizable signaling. *Erkenntnis*. doi: 10.1007/s10670-013-9462-3

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

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, *81*(2), 345–381.

Kruschke, J. E. (2011). *Doing bayesian data analysis.* Burlington, MA: Academic Press.

Lassiter, D. (2011). Vagueness as probabilistic linguistic knowledge. In R. Nouwen, R. van Rooij, U. Sauerland, & H.-C. Schmitz (Eds.), *Vagueness in communication* (pp. 127–150). Springer.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In T. Snider (Ed.), *Proceedings of the 23rd semantics and linguistic theory conference (SALT 23)* (pp. 587–610).

Luce, D. R. (1959). *Individual choice behavior: A theoretical analysis.* New York: Wiley.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms.* Cambridge university press.

Potts, C. (2008). *Interpretive Economy, Schelling Points, and evolutionary stability.* (Manuscript, UMass Amherst)

Qing, C., & Franke, M. (2013). Variations on a Bayesian theme: Comparing Bayesian models of referential reasoning. In H. Zeevat & H.-C. Schmitz (Eds.), *Proceedings of workshop on Bayesian natural language semantics and pragmatics (BNLSP 13).* (to appear)

Rabin, M. (1990). Communication between rational agents. *Journal of Economic Theory*, *51*, 144–170.

Schelling, T. C. (1960). *The strategy of conflict.* Cambridge, MA: Harvard University Press.

Skinner, B. F. (1957). *Verbal behavior*. Appleton-Century-Crofts New York.

Solt, S. (2009). *The semantics of adjectives of quantity*. Unpublished doctoral dissertation, The City University of New York.

Solt, S. (2011). How many mosts. In I. Reich, E. Horch, & D. Pauly (Eds.), *Proceedings of the 2010 Annual Conference of the Gesellschaft für Semantik (Sinn und Bedeutung 15)* (pp. 565–579).

Solt, S., & Gotzner, N. (2012). Experimenting with degree. In A. Chereches (Ed.), *Proceedings of the 22nd semantics and linguistic theory conference (SALT 22)* (pp. 166–187).

Stalnaker, R. (2006). Saying and meaning, cheap talk and credibility. In A. Benz, G. Jäger, & R. van Rooij (Eds.), *Game theory and pragmatics* (pp. 83–100). Hampshire: Palgrave MacMillan.

Stiller, A., Goodman, N. D., & Frank, M. C. (2011). Ad-hoc scalar implicature in adults and children. In *Proceedings of the thirty-third annual conference of the cognitive science society.*

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1) (No. 1). Cambridge: MIT Press.

Taylor, W. L. (1953). "cloze procedure": a new tool for measuring readability. *Journalism quarterly*.

Williamson, T. (1996). *Vagueness*. Routledge, New York.

Wittgenstein, L. (1953). *Philosophical investigations, trans. G. E. M. Anscombe*. Oxford: Basil Blackwell.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2013). Linguistic variability and adaptation in quantifier meanings. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 3835–3840).