THE UNIVERSITY
of EDINBURGH

The cultural evolution of scalar categorization

How cognition and communication affect
the structure of categories on scalar conceptual domains

Fausto Carcassi

A thesis presented for the degree of
Doctor of Philosophy

The University of Edinburgh

2020

Lay Summary

Languages spoken around the world show enormous variety at all levels of their structure. However, some features appear to be universal, or at least more common than would be expected by chance. In this thesis, I discuss two such universals, namely monotonicity and extremeness. To understand what monotonicity is, consider the concept expressed by the word "tall". That "tall" expresses a monotonic concept (on the scale of height) means that if a person is tall, then any person taller than them, i.e. higher than them on the scale of height, will also be tall. "Short" is also monotonic, because a person shorter than a short person is also short. Crucially, this is not a necessary feature of concepts. For instance, it is easy to imagine a word, "schmall", that expressed a non-monotonic concept: a person is schmall if and only if their height is within 50 centimetres of the population mean. I call a category extreme if it only includes an extreme of a domain. For instance, the word "full" expresses an extreme category because for a container to be literally full it has to be filled to the maximum. The first part of this thesis characterizes which classes of words express monotonic and extreme concepts, and develops a unified cognitive account of these concepts. The rest of the thesis is an attempt to understand why and how monotonic and extreme categories evolve.

Previous work in evolutionary linguistics has shown that languages evolve in response to various pressures, the two most important ones being a pressure for languages to be easy to learn and a pressure for languages to be useful in communication. I rely on this work to argue that monotonicity evolves in response to a combination of these two pressures, and support the proposed picture with a combination of computational modelling and experiments. I also investigate how the pressure from learning could contribute on the evolution of extremeness.

Abstract

Concepts can be thought of as regions of geometrically structured conceptual domains. Of all such possible regions, only very few are lexicalized, i.e. expressed by natural language in a morphologically simple fashion. In the thesis, I discuss lexicalized concepts on conceptual domains that are scalar, more specifically the concepts expressed by gradable adjectives and quantifiers. I consider two generalizations about such concepts. The first generalization is that lexicalized scalar concepts are monotonic, i.e. they can be defined in terms of a single threshold on the scale. The second is that if the conceptual domain has a maximum or a minimum, the threshold is often positioned at one of the extrema. I show that these two properties are non-trivial, in the sense that some scalar concepts, while semantically coherent and cognitively plausible, fail to have these properties.

The main of this thesis is to develop an account of how these two properties of monotonicity and extremeness evolve. I focus first on monotonicity, and show with a computational model that its emergence can be explained as an adaptation of language to two pressures, namely a pressure favouring languages that are easy to learn and a pressure on languages to be useful in communication. This explanation of monotonicity relies on the assumption that language users are pragmatically skilful. Moreover, the model makes assumptions about the cognitive biases of the language users. These assumptions are tested in a series of six category learning experiments. The results of three of these experiments are analysed with a Bayesian cognitive model. Overall, the experimental results are inconclusive. I present an agent-based model where learners are neural networks, which provides evidence that monotonic categories are easier to learn than non-monotonic categories. Finally, I turn to the evolution of extremeness. Previous literature has focussed on the role that communicative accuracy plays in the evolution of extremeness. In contrast to previous approaches, I study the role of learning. I show with an evolutionary computational

model that extreme categories evolve more often than chance even under a pressure from learning alone, as long as the language teachers and learners are pragmatically skilful.

Dedicated to my family:
Elena, Mario, Antonio,
Salvatora, Mariangela,
& myself.

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

Fausto Carcassi
24 March 2020

Acknowledgements

First, I want to thank my supervisors Simon Kirby and Marieke Schouwstra. Without your knowledge, feedback, and constant support, it would have been impossible to steer from philosophy to modelling & experimental cognitive science without getting whiplash. Thank you Wataru Uegaki and Michael Franke for agreeing to examine this thesis and for making the viva an enjoyable experience.

Thank you Jenny Culbertson, Kenny Smith, Christopher Cummins, Stella Frank, Robert Truswell, Matt Spike, Mora Maldonado, Isabelle Dautriche, Alistair Isaac, Brian Rabern, Wolfgang Schwarz, Dariusz Kalociński, Shane Steinert-Threlkeld & Jakub Szymanik for various discussions on the thesis, contributions to its content, and/or feedback on pieces of writing.

I have discussed, interacted, and befriended a lot of PhD students in the CLE, and I would like to thank them all. To keep things orderly, I am going to categorize them in three classes. First, the generation(s) before mine: Jon Carr, Vanessa Ferdinand, Jasmeen Kanwal, Yasamin Motamedi, Carmen Saldaña, Marieke Woensdregt, Cathleen O'Grady. Second, my generation: Asha Sato, Jonas Nölle, Tamar Johnson, Andres Karjus, Fiona Kirton, Svenja Wagner. Last, the generation after mine: Henry Coxe-Conklin, Annie Holtz, Marc Meisezahl, Fang Wang. If I had more time to write acknowledgements, I would try to explain why I am thanking you individually, but as things stand you'll just have to guess. Hint: it's something different for each one of you.

Going beyond the CLE, I want to thank the linguists: Thomas Stephen, Anna Page, Matt King, E Jamieson, Thomas Wood, Jose Segovia Martin, Elizabeth Pankratz, Takanobu Nakamura. Beyond LEL (Linguistics and English Language), the philosophers: Nina Poth, George Deane, Luke Kersten, Abi Thwaites, Jenny Zhang, Lorenzo Spagnesi, James Brown, Liv Coombes, Mara Neijzen, Camden McKenna, Matt Sims, Ian Campbell, Lilith Newton, Maddy Hyde, Lee Wilson, Kate

Nave, Becky Millar, Julian Hauser, Giada Margiotto, Hadeel Naeem, John Dorsch, Giles Howdle, Francesco Ellia, David Malakoff, Guido Tana. And beyond that, the rest of the world: John Cox, Roberta Leotta, Dora Puljic, Ana Puljic, Bruno Savill De Jong, Bede Hager-Stuart, Fran Sučić, Rebecca Holloway. I also want to thank Marco Tamborini and Paolo Savino—if I hadn't met you two, I would still be reading Heidegger.

Infine, voglio ringraziare la mia famiglia—Mario, Elena, Antonio, Salvatora, Mariangela—per avermi supportato anche senza sapere esattamente cosa stessi facendo.

Contents

Chapter 1

Introduction: Two universals of scalarity

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

Lewis Carroll

How *brillig* was it? How *slithy* were the toves? And how *mimsy* the borogoves? Despite the fact that these are not English words, we can say a great deal about their meaning. This thesis is an attempt to explain why we can expect the meaning of some words—including the ones we have never seen before, from languages we have not encountered—to follow certain regularities, called *universals* of language (Christiansen, Collins, & Edelman, 2009). The main aim of the present thesis is to study how some universals might emerge in response to the fact that language is a cultural artefact, one of whose aims is communication.

We will look at two universals. First, if in Carroll's world a tove a is slithier than a tove b, and a is slithy, then b is also slithy. Second, if there is an extreme degree of mimsiness—i.e. a maximum or a minimum of mimsiness–then to check if a borogove is mimsy only requires us to check its mimsiness with respect to an extreme. I call the former generalization *monotonicity* and the latter generalization *extremeness*.[1] Overall, this thesis is an attempt to identify some classes of words that

[1]I am ignoring some subtleties for the moment. I give a fuller explication of monotonicity and

express monotonic and extreme meanings, develop an account of the shared cognitive underpinnings of their meaning, and explain where the two semantic patterns of monotonicity and extremeness come from.

Monotonicity and extremeness are not properties exclusive to adjectives such as "mimsy" and "slithy", but can be found in any word class that expresses categories on scalar conceptual domains. As I argue in section 1.1, monotonicity and extremeness can be observed in the meaning of quantifiers. Despite the fact that the scalar categories that developed in natural language are monotonic and often extreme, it is easy to construct possible scalar categories that are neither extreme nor monotonic, and indeed most such possible categories are neither monotonic nor extreme. Therefore, the fact that these two universals are so common calls for an explanation.

In this thesis, I develop an account of the evolution of monotonicity and extremeness in terms of pressures that act to shape language, and I support the account with a combination of computational modelling and experiments. I consider two evolutionary pressures. First, a pressure that favours languages that are easier to learn. An account of learning requires a theory of the cognitive underpinnings of scalar language. The second pressure favours languages that lead to successful communication.

The main purposes of the present chapter are to define the thesis' explanans and present the state of the literature. First, I analyse monotonicity and extremeness in detail (section 1.1). Then, I argue that monotonic and extreme categories are not necessitated by the semantics of the relevant word classes (section 1.2). This raises the empirical question of where these two universals of scalar meaning come from. I discuss previous attempts to explain monotonicity and extremeness (section 1.3). While the question of how universals of scalarity evolve is approached from various directions in previous literature, some points have not been satisfactorily explained. First, most work on the evolution of monotonicity focuses on quantifiers, and it is unclear how it can be extended to other scalar conceptual domains. Second, previous models of the evolution of extremeness do not deal satisfactorily with the fact that for (at least some) extreme categories, true instances are never encountered in the real world. Moreover, previous explanations do not consider the role that learning might have in the evolution of extremeness of scalar categories, but rather focus on communicative efficiency. In the remaining chapters, I develop an account of

extremeness below.

the evolution of scalar categories that makes progress over previous literature with respect to these points.

1.1 Monotonicity and extremeness

In this section, I introduce the topic that will be investigated throughout the thesis, namely the structure of scalar categories. I call *scalar* any category that can be thought of as a region of a scale. For instance, the category of "tall" can thought of—given a context—as a region of the scale of tallness. What constitutes a scale is a topic of debate in the literature, and I return to the topic in more details in the next chapter. For the purposes of the present section, a scale can be thought of as an ordered set of degrees. For instance, the set of heights ordered by tallness constitute the scale of tallness. This section has two aims. The first is to rigorously define two universals of scalar semantics, namely *monotonicity* and *extremeness*. The second aim is to consider specific classes of words that express scalar, and therefore monotonic and extreme, meanings.

Notably absent from the discussion will be the theories of scales and degrees developed in formal semantics and philosophy. While the concepts of scale and degree are discussed in depth in the philosophical and semantics literature, the discussions are not directly relevant for the present purposes. The aim in much of the semantics literature is to provide a compositional account of *truth conditions*. However, here I am not interested in the conditions that determine the truth value of a sentence, since truth conditions do not have direct causal effect on language. Rather, I am interested in the cognitive underpinnings of scalar language, which influence language evolution. What is needed is therefore a cognitive account of scalarity. In light of this, I do not discuss much of the literature on the metaphysics of scales and degrees. On the other hand, I return to the way scales appear in cognition in chapter 2.

The present section is structured as follows. I start by defining and discussing two properties of Boolean-valued functions—monotonicity and extremeness—that require only that their domains are total orders. I discuss two classes of words where lexicalized meanings on total orders tend to satisfy these two properties, namely gradable adjectives and quantifiers.

1.1.1 Monotonicity

Monotonicity is the first universal of scalar meaning I look at. In this section, I first give a general definition of monotonicity that only assumes the structure of a total order and discuss some properties of monotonic functions. Then, I show data that indicates that terms belonging to some word classes are monotonic.

A general characterization of monotonicity

A binary relation $\leq$ is a *total order* on a set X iff for any $a, b, c \in X$:

$$a \leq b \wedge b \leq a \implies a = b \qquad \text{Antisymmetry} \qquad (1.1)$$

$$a \leq b \wedge b \leq c \implies a \leq c \qquad \text{Transitivity} \qquad (1.2)$$

$$a \leq b \vee b \leq a \qquad \text{Connexity} \qquad (1.3)$$

Antisymmetry says that if two elements are lower than or equal to each other, then they are equal to each other. Transitivity says that if a is lower than or equal to b and b is lower than or equal to c, then a is lower than or equal to c. Connexity says that any two elements of X are comparable with each other in terms of the order.

The concept of monotonicity comes from mathematics. Assume X and Y are two sets ordered by $\leq_X$ and $\leq_Y$ respectively.

Definition 1. *A function $f : X \to Y$ is **monotone increasing (decreasing)** iff for all $x, y \in X$ if $x \leq_X y$ then $f(x) \leq_Y f(y)$ $(f(y) \leq_Y f(x))$.*

A function is monotonic iff it is monotone increasing or decreasing. While definition 1 is the most general characterization of monotonicity, in the following I will mainly consider functions that have truth values as their range, called *Boolean-valued* functions. Moreover, I will only discuss function with a totally ordered domain. I assume, in line with previous literature (Shramko & Wansing, 2018), that the set of truth values is fully ordered with *false* (or 0) as its infimum and *true* (or 1) as its supremum. With these restrictions, the definition of monotonicity above can be simplified.

Intuitively, an element t in the domain of f is a *transition* of f iff f changes truth value at t. More technically,

Definition 2. *An element t of the domain of f is a* ***true transition from true to false*** *iff (1) $f(t) = true$ and (2) there is some $y > t$ such that f is false for all b with $t < b \leq y$.*

Definition 3. *An element t of the domain of f is a* ***false transition from true to false*** *iff (1) there is some $x < t$ such that f is true for all a with $x \leq a < t$ and (2) $f(t) = false$.*

Similarly for transitions from false to true. An element is a *transition from true to false* iff it is a true or false transition from true to false, and similarly for transitions from false to true. An element is a *transition* iff it is a transition from true to false or a transition from false to true. The *type* of a transition is whether the transition is from true to false or from false to true.

Note that the definition of transition is such that a bounded domain can have a transition that affects only the infimum or the supremum. For instance, f might evaluate to true everywhere except for the supremum, where f is false. In this case the supremum would be a false transition from true to false. Moreover, it correctly excludes the possibility of a transition on a boundary to the same truth value, e.g. a true transition from true to false on a supremum. The definition will therefore be useful when discussing the monotonicity of categories that only cover the extrema of a scale. In the following, I assume that f has at most countably infinite many transitions in its domain. This, together with the assumption of density of the domain, implies that between any two transitions there are points in the domain that are not transitions. The definition of transition leaves open the possibility of a point that is surrounded by points with a different truth value. In that case, the same point would instantiate two transitions, e.g. a true transition from false to true and a true transition from true to false.

Next, I show the following intuitive fact about transitions:[2]

Lemma 1. *If there exist two distinct x, y such that neither x nor y is a transition and $f(x) \neq f(y)$, then f has a transition between x and y.*

Next, define the concept of adjacency:

Definition 4. *Any two transitions t_i and t_k are* ***adjacent*** *iff there is no transition between t_i and t_k.*

[2]The proofs for this and all following lemmas are contained in appendix A.

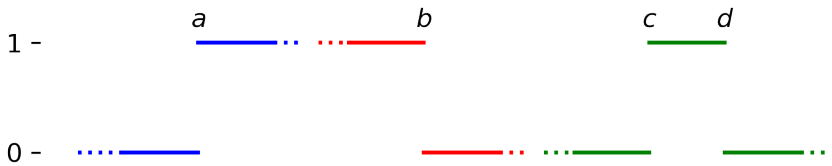


Figure 1.1: An illustration of monotonicity. Each line represents a function from some scale (x-axis) to the ordered set $\{0, 1\}$. The left function (blue) is monotone increasing, with a single transition on point a. The central function (red) is monotone decreasing with a transition on point b. The right function (green) is non-monotonic, as it has two transitions on points c and d.

Lemma 2. *If t_1 and t_2 are adjacent, then they are of different types.*

The important consequence of lemma 2 is that if one knows that there are only two transitions and the type of one of them, then one can infer the type of the other transition. This in turn has consequences, which I explore further below, on the amount of information needed to encode transitions.

Given the lemmas above, it is possible to show a simple and intuitive connection between monotonicity and Boolean-valued functions:

Lemma 3. *A Boolean-valued function f is monotonic iff it has one or fewer transitions.*

Figure 1.1 shows the difference between monotonic and non-monotonic functions in a visual way.

Lemma 4. *The number of transitions, their order, and the type of one of them implies the type of all other ones.*

It is worth considering the relation between two or more monotonic function in the same domain, because some substantial things can be said about their relations. Start by defining the relation of *inclusion* between functions:

Definition 5. *A Boolean-valued function f* ***includes*** *a Boolean-valued function g iff everywhere where f is true, g is also true. If there are x such that $f(x) = true$ and $g(x) = false$, then f* ***strictly*** *includes g.*

Some facts about inclusion can be inferred from type:

Lemma 5. *If two functions f and g with a single transition each are on the same domain and are of the same type, then f includes g or g includes f (or both).*

This lemma will be relevant when considering the role that monotonicity plays in making a system of categories efficient for communication.

Up to this point, I have discussed monotonicity for a Boolean-valued function. Rather than functions, in the following I will mostly be interested in categories, i.e. sets of individuals, expressed by words in natural language. Categories are generally modelled as subsets of a universe set. For instance, the category of people is often defined as a subset of set of individuals. Conveniently, the connection between Boolean-valued functions and categories is easily expressible as follows:

Definition 6. *A Boolean-valued function $f : X \to \{true, false\}$ is the* ***characteristic function*** *for a category $C \subseteq X$ iff for all $x \in X$, $f(x) = true \iff x \in C$*

In words, this means that a Boolean-valued function corresponds to the category of objects for which the function is true. In the following, I will talk interchangeably about categories and Boolean-valued functions, sometimes in a slightly imprecise way. The context should clarify in each case what is meant.

In the discussion of transitions above, I put few requirements on the domain of f. For instance, I did not assume a well-defined notion of distance on the domain of f. Moreover, I made some assumptions that can be weakened while keeping a reasonable definition of transition, for instance the assumption of density of the scale and that the domain of f is totally ordered. In the following, I discuss semantic phenomena that requires scales with sometimes more and sometimes less structure than was put on the domain of f in the discussion above.

Any account of the meaning of natural language categories defined based on transitions has to deal with the phenomenon of *vagueness*. The definitions above do not straightforwardly apply to many linguistic phenomena because they assume a crisp transition in the function, meaning that every point either clearly belongs in the category or it clearly does not. On the other hand, categories expressed by natural language expressions are often vague, meaning that at least some points neither clearly belong nor clearly don't belong in the category. For instance, some individuals are neither clearly tall nor clearly not tall. The model of categories based on transitions as defined above can be made more similar to real linguistic categories

while keeping the proofs substantially unmodified. A classic strategy for doing so is *supervaluationism*.[3] I do not discuss supervaluationism in details, but rather sketch how it could be applied to the situation at hand. Supervaluationism as a strategy will appear again when discussing contemporary versions of prototype theory.

The supervaluationist approach (van Fraassen, 1966; Fine, 1975; Kamp, 2013) starts with the concept of *partial model*, i.e. a model where predicates are assigned true to some objects, false to other objects, and are undefined for other objects. For instance, "tall" will be true in the partial model for the clearly tall individuals, false for the clearly-not-tall individuals, and undefined for the individuals that are neither clearly tall nor clearly not tall. The truth of sentences can be evaluated as usual for the clear cases. To deal with the unclear cases, consider the possible *completions* of the model, i.e. the possible ways of extending the partial model by attributing either true or false to each of the unclear cases, for all the predicates in the partial model. The degree of membership of an individual in a category can be supervaluated in the following way. If the individual belongs to the category in all completions, then it clearly belongs to the category. If it does not in any completion, then it clearly does not belong to the category. If it does in some completions and not in others, its degree of membership can e.g. depend on the proportion of completions in which it belongs to the category. Since categories are, in my model, defined by their transitions, different completions for a category will differ with respect to the position of the transitions. The supervaluation analysis of vagueness for transitions-defined categories has the crucial advantage of being compatible with the definitions and lemmas above. While for crisp categories the discussion above applies straightforwardly, for vague categories it applies for all completions. Therefore, the lemmas above are true in the supervaluation.

From a modelling point of view, instead of starting as I did above with a function f that characterises a category directly, I start with a set $T = \{t_1, t_2, \ldots, t_n\}$ of random variables, distributed as unimodal probability densities such that for any two indices $i < j$, the expected value μ_i of t_i is lower than the expected value μ_j of t_j. Unimodality ensures that the corresponding cumulative distribution functions are strictly monotonically increasing. Each $t \in T$ models the position of one of the category's transitions. Each realization of a $t \in T$ is a completion for that transition.

How is degree of membership in the category defined in this formal framework?

[3]Thanks to Brian Rabern for discussions on the literature on vagueness and for suggesting the connection to supervaluationism.

First, define the set $C = \{c_1, \ldots, c_n\}$, whose ith member is the cumulative distribution function of t_i. The function c_i models the probability that the relative transition is below any specified point. Next, t_1 is associated with a type of transition w, either -1 to define the type from true to false or type $+1$ to define the type from false to true. All other transitions w_i are associated with the same type as t_1 if their index is odd and with the opposite type if their index is even. The function f, which can be interpreted as describing the probability that each point of the scale belongs to a category, is defined as:

$$f(x) = \sum_{i=1}^{n} w_i c_i(x) \tag{1.4}$$

If there is no transition, then f is directly specified as being 1 everywhere or 0 everywhere, representing respectively a category that certainly applies everywhere and one that certainly applies nowhere. Definition 1.4 is effectively approximating the probability that the closest transition below any point is a transition from false to true. If the transitions are far enough from each other (compared to their variance) that there is little chance of them having a different order than their index order, the c_i become close to 1 for all transitions lower than the one immediately below x, and become close to 0 for all transitions above x. Therefore, if the transition immediately below x is from false to true, the cumulative distribution function locally gives the probability of x being true. If the transition immediately below x is from true to false, the cumulative distribution function locally gives the probability of x being false. Figure 1.2 shows the effects of multiple transitions on f.

Note that this way of setting up the model allows for some intuitively implausible categories if the transitions are close to each other. It is however unclear whether the f that are implausible should be excluded by the model itself. A natural way of making f have a more plausible shape would be to restrict the functional form of the distributions in T, and to put constraint on how close their means can be. In the larger context of this thesis, the idea expressed by equation 1.4 will come up again in section 2.2.3. However, the interaction between multiple thresholds will not constitute a problem as discussion will be limited to categories with a single threshold. Therefore, I leave a more complicated elaboration of the model of non-monotonic vague categories to future research.

In the remainder of this section, I present familiar data that suggests words in various classes are monotonic with respect to salient scales. To show that a term is

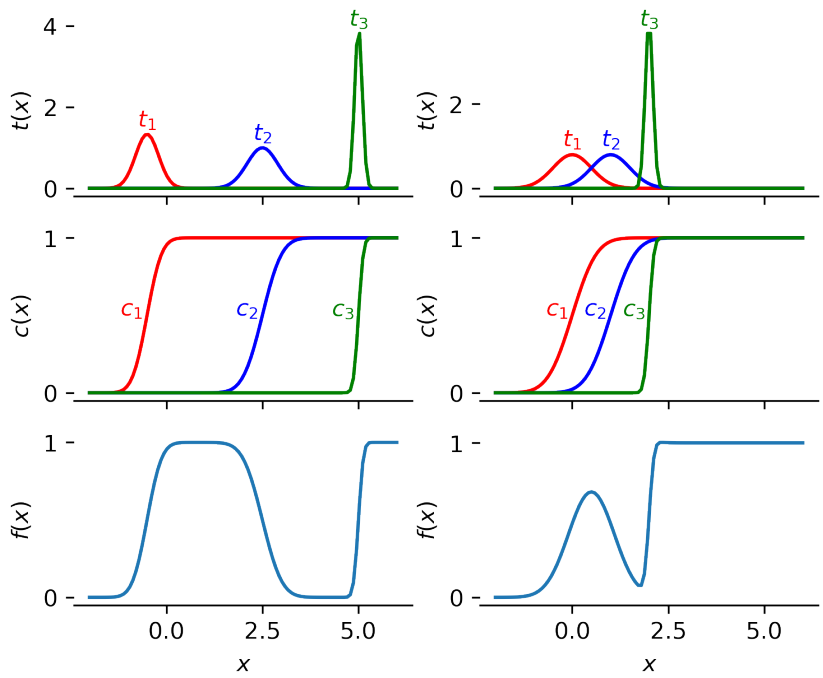


Figure 1.2: Category with vague transitions modelled by a function f. Each column shows the distribution over transitions positions (top), the cumulative probability functions (center) and the function obtained by combining them (bottom). The categories in both the left and right columns have three transitions, t_1, t_2, t_3. In each case, t_1 is a transition from false to true, which implies that t_2 is from true to false, and t_3 from false to true. The left column of plots shows a category where the transitions are far enough that equation 1.4 is a good approximation. Transitions are distributed normally, parameterized by mean and standard deviation. t_3 has a smaller standard deviation than the other two transitions. This has the effect that the increase in the cumulative density function is sharper (middle left), and therefore there is less vagueness about its position on the scale (bottom left). The right column shows a pathological category, where the vague areas of different transitions overlap substantially. As a consequence of definition 1.4, f has a local maximum not at 1.

monotonically increasing, I show that for any two individuals a and b, if a falls in the extension of the term and b is higher on the scale than a, then b also falls in the extension of the term. Similarly for monotonically decreasing. This test corresponds

to definition 1. The only requirement for this strategy is that it should be possible to compare two individuals with respect to a scale. In section 1.2, I show that for each word class I considered independently motivated semantic analyses define a suitable scale.

Gradable adjectives

Gradable adjectives are those adjectives susceptible of gradation with (at least some) modifiers such as "very", "slightly", "almost", "half". They are used to convey information about the degree to which individuals have properties. In the predicative position, they mainly appear in two contexts, so-called *measure* and *bare* contexts. Measure contexts are used to convey the precise degree to which an object has a property. An example of a measure context is "Mary is 181cm tall", which asserts that Mary has a very specific height. In bare contexts, gradable adjectives are used to communicate that an entity has a property to a degree that falls within a certain range. An example of a bare context is "Mary is tall", which conveys information about the height of Mary, but not Mary's exact height.[4]

To find out whether Mary satisfies the bare predication of the adjective "tall", all we need to do is check whether Mary is tall *enough*. For instance, in a certain context we might say that Mary is tall if she is taller than 180 centimetres. The bare predication of a gradable adjective if verified by every individual that has the relevant property to a degree greater than some degree called the *standard of comparison* (with the proviso of contextual sensitivity explained below). I argue below that standards of comparison play the role of transitions in a formal account of adjectival semantics. Consistently with the formal analysis above, the existence of a unique transition/standard of comparison suffices to show monotonicity. The comparison test confirms the conclusion:

(1.5) If N_1 is ADJ and N_2 is ADJ-er than N_1, then N_2 is ADJ.

Where N_1 and N_2 are nouns and ADJ is a gradable adjective. Instances of this pattern are:

(1.6) If John is tall and Mark is taller than John, then Mark is tall.

(1.7) If Russia is cold and Iceland is colder than Russia, then Iceland is cold.

[4]In the following, I will refer to the bare use of gradable adjectives unless stated otherwise.

A sentence of the form "N_1 is ADJ-er than N_2" is true iff the degree to which N_1 has the property referred to by ADJ is greater than the degree to which N_2 has that property.[5] The validity of the inferences in examples 1.6 and 1.7 indicates that having the relevant property to a degree higher than a degree that verifies a bare predication suffices to also verify a bare predication. This in turn implies that in bare contexts every degree greater than a degree that falls in the adjective's extension also falls in the adjective's extension. Bare predications of gradable adjectives therefore refer to monotonic categories on the respective scales.[6]

Quantifiers

Quantifiers are determiners that are used to convey information about magnitudes. Intuitively, quantifiers express a quantitative relation between the sets referred to by two common nouns. For instance, the sentence "Most ducks quack" conveys a quantitative relation between the set of ducks and the set of quacking things. Call the first argument—"ducks" in the example—the *restrictor* argument and the second argument the *scope* argument.

Two important classes of quantifiers are proportional quantifiers and numerical quantifiers. Proportional quantifiers such as *all*, *some*, *none*, *most* convey information about the proportion between the sizes of restrictor and scope. For instance, *all* conveys that the proportion of individuals in the restrictor set that belong to the scope set is 1, while *most* conveys that the proportion is greater than 0.5. Each proportional quantifiers can therefore be expressed as a portion of the scale of proportions. Proportional quantifiers include the Aristotelian quantifiers (*all*, *some*, *no*) as well as the proportional quantifiers sensu stricto (*half*, *most*, *three quarters of*).[7]

Numerical quantifiers such as *more than five* convey information about the number of elements in the restrictor set that also belong to the scope set. For instance,

[5] See Lassiter (2015) for a review of the relevant literature on comparison.

[6] It is debated whether there are adjectives that are monotonically decreasing, or whether all adjectives are monotonically increasing on a scale that is structured by an order inversed with respect to the intuitive direction. For instance, "cold" could be monotonically decreasing if $\geq$ tracks increasing temperature, or monotonically increasing if $\geq$ tracks decreasing temperature. See Heim (2006) and Heim (2008) for a classical proposal to this effect.

[7] Throughout the thesis, I disregard the presuppositional content of quantifiers. For instance, I treat "the", "both", "either", and "neither" all as proportional quantifiers, because once the respective presuppositions are satisfied they reduce to other proportional quantifiers—respectively, "a", "two", "one", and "no". This disregard for presuppositional content works under the assumption that the asserted meaning and the presuppositions are stored and encoded independently.

five cats sleep says that five elements of the restrictor set of cats belong to the scope set of sleeping things. The meaning of each numerical quantifier can be expressed as a set of numbers, namely the sizes of the intersection of restrictor and scope set that verify the quantifier. Numerical quantifiers include the cardinality quantifiers (*one*, *five*) as well the parity quantifiers (*an even number of*, *an odd number of*).

Quantifiers can be divided into morphologically simple and complex. Morphologically simple quantifiers (such as *some*, *all*, and *no*) consist of a single morpheme, while complex quantifiers (such as *an even number of*) have internal morphological structure. A great deal of literature, starting with Barwise and Cooper (1981) supports the claim that simple quantifiers are monotonic. Therefore, I refer to the literature for support of the claim that quantifiers are monotonic. While the classic definition of monotonicity for quantifiers differs from the definition I gave above, I show below that the two are equivalent with some plausible assumptions. This allows the conclusion that simple quantifiers are monotonic according to definition 1 above in the respective scales, i.e. proportional quantifiers are monotonic on the scale of proportions and numerical quantifiers are monotonic on the scale of numbers.

As an illustration of quantificational monotonicity, consider the proportional quantifier "most". The monotonicity of "most" allows for inferential patterns that are analogous to the one in 1.5, as shown by the comparison test:

(1.8) If most N V_1 and there are more V_2ing N than V_1ing N, then most N V_2.

where V_1 and $_2$ are any two appropriate verbs. Note that the comparison is on the scale of proportions. If there are more V_2ing N than V_1ing N, then the proportion of N that V_2 is greater than the proportion of N that V_1.

1.1.2 Extremeness

A general characterization of extremeness

The tools developed in the section on monotonicity allow us to easily give a general definition of extremeness. First, define the concept of an extreme transition:

Definition 7. *A transition t is* ***extreme*** *iff is it a supremum or an infimum of the domain of f.*

Then, define the notion of extremeness of categories:

Definition 8. *A category is* ***extreme*** *iff it has an extreme transition.*

Note that this definition of extremeness for categories does not imply that an extreme category is monotonic. A category with two transitions, one at the supremum and one at the infimum, is extreme but not monotonic. In practice, this will not affect the discussion because I will mostly consider monotonic categories below.

The universal of extremeness I propose is that there is a tendency for transitions in natural language to fall on the extrema of scales when possible. To test whether a category has a true extreme transition at the supremum, I test whether two things can be predicated at the same time of an individual: (1) the individual falls in the category and (2) it is possible for the individual to be above its actual position on the scale underlying the category. If a category is not extreme, an individual falling in the category does not need to be at the scale maximum, and therefore it is possible for both conditions to be satisfied together. If the category is extreme, the two conditions cannot be true together. On the other hand, to check if a category is a false extreme transition at the infimum, I check if an individual can (1) have a degree outside the category and (2) be capable of having a degree that is lower than its current degree. If the only false point for the category is the infimum, the two conditions cannot be verified together.

The universal of extremeness is weaker than the universal of monotonicity. While the vast majority if not all terms discussed above are monotonic, there are many exceptions to extremeness in each of the three groups of terms I consider. A model of the emergence of extremeness should therefore not aim to predict that every category will become extreme. Rather, such a model should aim at explaining the surprising commonness of extremeness in scalar categories. Crucially, extremeness is more surprising than monotonicity. Extremeness is prima facie maladaptive from the point of view of both the learning and the communication pressures discussed above. I return to this issue in chapter 6.

The discussion of vagueness above helps making sense of what is peculiar about extreme categories. When taking transitions to be single points, the difference between vague transitions and sharp transitions cannot be modelled; all transitions are modelled as sharp. However, sharpness is a non-trivial property of extreme transitions, which distinguishes extreme transitions from the generally vague non-extreme transitions.

Examples of extremeness

In this section, I only consider a few examples of extremeness in the word classes introduced above. Below, I give a fuller treatment of what extremeness is in each case. Consider first gradable adjectives, which I analysed as categories on ordered sets of degrees. There are numerous examples of gradable adjectives that are extreme in the sense described. Consider "dry" as an illustration. The following sentence is semantically incoherent:

(1.9) # The shirt is dry, but it could be more dry.

However, the same pattern is acceptable for other adjectives:

(1.10) The stick is bent, but it could be more bent.

This shows that some adjectives refer to extreme categories, while other do not. Adjectives that refer to extreme categories are called *absolute* adjectives. They can be divided into *maximal*, which have a true transition from false to true at the supremum, and *non-minimal*, which have a false transition from false to true at the infimum. Examples of the former are "full", "straight", "closed", as can be checked with pattern 1.9. Example of the latter are "bent", "awake", "visible", "open":

(1.11) # The stick is not bent (at all), but it could be straighter.

There are substantial generalizations that connect the structure of the underlying scale with extremeness, which I discuss in the next section.

Similar patterns showing category extremeness can be constructed with quantifiers:

(1.12) # All the cats in the room are sleeping, but more of the cats in the room could be sleeping.

(1.13) # None of the students failed the example, but fewer of the students could have failed the exam.

1.2 Universals of scalarity in natural language semantics

In the discussion above, I described two striking properties of the semantics of scalar terms. I considered classes of words that, to different extents, seem to verify these

generalizations. However, I have not described in details the semantics of the classes of words and exactly where scalarity is required. In this section, I make the treatment of scalarity more precise by couching monotonicity and extremeness in independently motivated analyses from the formal semantics literature.

The section is structured as follows. I start by discussing gradable adjectives. First, I present a widely accepted account of their semantics. I then show how monotonicity can be defined within this account, and discuss previous literature concerning the monotonicity of gradable adjectives and related concepts. I also discuss extremeness for gradable adjectives. I then move onto quantifiers and introduce *generalized quantifiers theory* as a commonly accepted model of quantificational semantics. I show how monotonicity can be defined in generalized quantifiers theory. Then, I discuss an alternative scalar approach to the semantics of quantifiers and how to define monotonicity in it. I also briefly discuss extremeness in quantifiers. I conclude the section with a discussion of some possible exceptions to monotonicity.

This section has three aims. First, it will show that a scalar analysis of the word classes I discuss is independently motivated. Second, I will argue that the previous literature on formal semantics implicitly assumed the universal—monotonicity for gradable adjectives—or explicitly argued for it—monotonicity for quantifiers, extremeness for gradable adjectives. This is important because the semantic analyses encode considerations of large amounts of linguistic data. The third aim of the section is to show that semantics alone is not sufficient to explain the universals under discussion. This will motivate the search for an evolutionary explanation.

1.2.1 Gradable adjectives

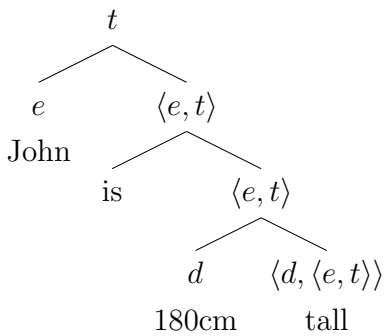
The formal semantics of adjectives

In the rest of this section, I present a widespread account of the semantics of gradable adjectives. According to Kennedy and McNally (2005) the meaning of a gradable adjective makes use of a scale and a function. The scale is a partially ordered set of degrees, like the one I introduced above. The function maps the scale's degrees onto sets of individuals, namely the individuals that have the property to the given degree. Being functions from degree to sets of individuals, adjectives are of type $\langle d, \langle e, t\rangle\rangle$.

For instance, the adjective 'tall' comes with the set of heights ordered according to tallness, and a function *height* which maps degrees onto the set of persons of the given

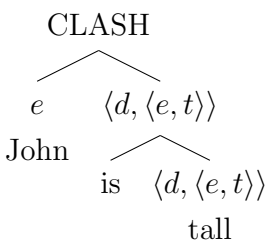
height.[8] Formally, $[\![\text{tall}]\!] = \lambda d \lambda x.[height(x) = d]$. This accounts straightforwardly for measure uses:

(1.14)



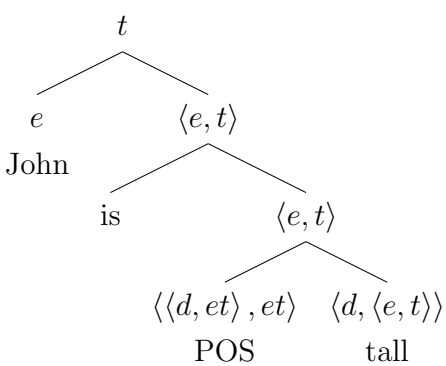
I take the expression "180 cm" as simply referring to a degree on the height scale, which implies that it is of type d. However, we get a type clash when we draw a naïve tree for the bare use of "tall":

(1.15)



Kennedy and McNally (2005), following Stechow (1984), solve the type clash by assuming the existence of a null morpheme POS that takes the adjective as an argument and returns a function of type $\langle e, t\rangle$. POS is therefore of type $\langle\langle d, et\rangle, et\rangle$:

(1.16)



POS takes the adjective as an argument and returns a function from individual to truth values, or equivalently a set of individuals. This solves the type theoretical

[8] To keep explanation simple, I assume that degrees are points on a scale rather than intervals. This simplification does not affect the present argument.

problem but leaves open how to interpret POS. It is easy to see that the set returned by POS should be the set of things that verify the bare predication of the adjective. I consider three additional intuitive desiderata. First, the analysis should account for the context sensitivity of bare predications of gradable adjectives. A given sentence containing an adjective in bare use can be true or false of the same individual in different utterance contexts. For example, consider Mary, a six years old who is quite tall for a child, but still shorter than the average person. In a conversation about people in her school class, the sentence "Mary is tall" is true. However, in a conversation about the average human, the sentence "Mary is tall" is false. The lesson is that the comparison population can, for some adjectives, shift the standard of comparison.[9] Call *relative standard* those adjectives that shift standard of comparison depending on the contextually given comparison population. Examples of relative standard adjectives are "tall", "intelligent", and "big". Not all gradable adjectives are relative standard. For instance, "full", "dry" and "straight" have a standard whose position is insensitive to the comparison population. Call *absolute standard* those adjectives with a fixed standard of comparison. At least for relative adjectives, POS should be sensitive to the comparison population.

The second desiderata for POS is that given a comparison population the bare predication should not discriminate between individuals that have the property to the same degree. This means that whether an object falls or not on the set only depends on the degree to which it has the property. Lastly, I want the analysis of POS to provisionally make bare predications monotonic. The account I consider, which is the one of Kennedy and McNally (2005), satisfies these restrictions.

A relation *standard* is introduced between a degree d, a function G of adjectival type $\langle d, \langle e, t\rangle\rangle$ and a contextually given population C. C is a free variable that can be given a value by the context. The relation $\mathit{standard}(d)(G)(C)$ is true iff d is greater on the scale specified by G than a degree that depends on the comparison population C:

$$\mathit{standard}(d)(G)(C) = d_{G,C} \leq_G d \tag{1.17}$$

In the most typical cases, such as "John is (POS) tall", $d_{G,C}$ is somewhere above the average of property G in the population C.[10]

[9]The comparison population can also be set explicitly by adding "for a C", where C is a set of individuals.

[10]The issue of how $d_{G,C}$ is calculated in actual languages is orthogonal to the problem at hand. Below, I discuss various proposals for how the position of the standard of comparison is set.

POS is a function from a gradable adjective G and an individual x that is true iff the degree d to which the individual has the property satisfies $\mathit{standard}(d)(G)(C)$ and false otherwise:

$$[\![\text{POS}]\!] = \lambda G \lambda x. \exists d[\mathit{standard}(d)(G)(C) \wedge G(d)(x)] \tag{1.18}$$

Given this analysis, "John is (POS) tall" (sentence 1.15) is true iff:

$$\exists d[\mathit{standard}(d)([\![\text{tall}]\!])(C) \wedge [\![\text{tall}]\!](d)([\![\text{John}]\!])] \tag{1.19}$$

A complication in the semantics of gradable adjectives is that not all individuals have every property, and therefore in formal terms taking "tall" as an example:

$$\bigcup_{d \in H} [\![\text{tall}]\!](d) \subset M \tag{1.20}$$

for the set H of degrees of height. M is the domain of individuals. This clarification is needed to make sense of the fact that not all adjectives can be meaningfully predicated of every object. For instance, the sentence "This rock is intelligent" is hard to interpret, because there is no degree to which rocks have intelligence. This type of failure has a particularly stark effect with adjectives that select or fail to select for clauses:

(1.21) It is funny to see a clown.

(1.22) # It is fast to see a clown.

A plausible reason for the different acceptability of the two examples is that while an event can be funny, it cannot be fast.

The monotonicity universal, formalized

I described a standard account of the semantics of gradable adjectives. The next step is to connect it to the discussion of monotonicity in the previous section. I will discuss the adjective "tall" for illustration. I want to characterise the bare use of a gradable adjective as a category on a scale. There are two straightforward ways of doing this, which are equivalent for the present purposes. The first is to focus on the *standard* relation. Partial application of the gradable adjective "tall" and

the context C_1 to *standard* delivers a Boolean-valued function with heights as its domain:

$$standard_{[\![\text{tall}]\!],C_1} = \lambda d.standard(d)([\![\text{tall}]\!])(C_1) \tag{1.23}$$

The partially saturated $standard_{[\![\text{tall}]\!],C_1}$ is monotonic, because for all degrees of height x, y:

$$x \leq_{\text{height}} y \wedge standard_{[\![\text{tall}]\!],C_1}(x) = 1 \implies standard_{[\![\text{tall}]\!],C_1}(y) = 1 \tag{1.24}$$

Expand with analysis 1.17:

$$\iff x \leq_{\text{height}} y \wedge d_{[\![\text{tall}]\!],C_1} \leq_{\text{height}} x \implies d_{[\![\text{tall}]\!],C_1} \leq_{\text{height}} y \tag{1.25}$$

Inference 1.25 is an application of transitivity (equation 1.2), which simply requires that degrees of heights are ordered. Therefore, the monotonicity assertion in 1.24 is verified by the analysis of *standard* above.

More generally, given any G and C, the following is true $\forall d_1, d_2$:

$$d_1 \leq d_2 \wedge standard_{G,C}(d_1) = 1 \implies standard_{G,C}(d_2) = 1 \tag{1.26}$$

Equation 1.26 means that any degree greater than a degree that satisfies *standard* also satisfies *standard*. Equation 1.26 abstracts from specific adjectives and contexts, and therefore asserts that all adjectives are monotonic. This follows from the analysis of *standard* in equation 1.17. In the following, I abuse the terminology and say that POS or adjectives in their bare use are monotonic when I mean that *standard* is.

The analysis in equation 1.26 expresses the monotonicity of bare adjectival categories with respect to degrees in the scale relative to the adjective. The second way of characterizing bare adjectives as categories on a totally ordered set is with respect to the individuals having the property themselves. This second characterization maps more naturally onto the syntactic structure of the sentence. The category in this case is the set characterized by the phrase [[POS ADJ]]:

$$[\![\text{POS ADJ}]\!] = \lambda x.\exists d[standard(d)([\![\text{ADJ}]\!])(C) \wedge [\![\text{ADJ}]\!](d)(x)] \tag{1.27}$$

$$= \lambda x.\exists d[d_{[\![\text{ADJ}]\!],C} \leq d \wedge \mu_{ADJ}(x) = d] \tag{1.28}$$

where $\mu_{ADJ}(x)$ is the (maximal) degree to which x has the property referred to by

ADJ. The monotonicity property with respect to individuals rather than degrees can be expressed as follows:

$$[\![\text{POS ADJ}]\!](x) \wedge \mu_{ADJ}(x) \leq \mu_{ADJ}(y) \implies [\![\text{POS ADJ}]\!](y) \tag{1.29}$$

for any x and y in M. Equation 1.29 follows from the definition of *standard*:

$$\exists d[d_{[\![\text{ADJ}]\!],C} \leq d \wedge \mu_{ADJ}(x) = d] \wedge \mu_{ADJ}(x) \leq \mu_{ADJ}(y) \implies \tag{1.30}$$

$$\exists d[d_{[\![\text{ADJ}]\!],C} \leq d \wedge \mu_{ADJ}(y) = d] \tag{1.31}$$

While the definition with respect to degrees is a *strict* total order, i.e. there are no ties between non-identical degrees in terms of the order, the definition in terms of individuals is a *weak* order, i.e. there are (in principle) multiple individuals that have the property to the same degrees. Despite the fact that not all individuals have all the properties, the set of individuals should not be modelled as a partial order. Any two individuals that have the property to any extent will be comparable in terms of the property. Therefore, it is better to analyse μ_{ADJ} as undefined for many individuals.

The relation *standard* is ultimately responsible for the semantic behaviour of bare contexts, since it determines which degrees are covered by the adjective in its bare use. I have argued that *standard* is monotonic. The monotonicity of *standard* is however not a semantic necessity. In general, *standard* could be any function from a degree, a comparison population and a gradable adjective to a truth value, and only some such functions satisfy equation 1.26. For illustration, consider a language similar to English, Twinglish, such that $standard_{\text{Tw}}$ is true iff d is between the smallest and the greatest degree (excluded) of G in the set C. "John is (POS) tall" when uttered by a Twinglish speaker means that John is not the shortest nor the tallest person in the contextually relevant set ("John is tall" and "John is short" are synonymous in Twinglish). It follows that $standard_{\text{Tw}}$ is non-monotonic. For any degree belonging to $standard_{\text{Tw}}$, there is one degree higher than it which does not belong to $standard_{\text{Tw}}$ (namely, the maximum of the scale) and one degree lower than it which does not belong to $standard_{\text{Tw}}$ (namely, the minimum of the scale). Twinglish does not differ from English just in the meaning of bare adjectives. Notably, the modifiers of bare adjectives that make use of *standard* in English also get different meanings in Twinglish. An example is "very", which according to Kennedy

and McNally (2005) (following Klein (1980)) is equivalent to predicating the bare adjective after restricting the population C to the individuals to which the simple bare adjective applies. Formally,

$$[\![\text{very}]\!] = \lambda G \lambda x. \exists d[\mathit{standard}(d)(G)(\lambda y.[\![\text{pos(G)(y)}]\!]) \wedge G(d)(x)] \tag{1.32}$$

"John is very tall" in Twinglish means that John's height is between the second shortest and the second tallest person in the contextually relevant set. The crucial point is that Twinglish is a prima facie conceivable and coherent language, which shows that non-monotonic *standard* functions cannot be excluded a priori. Despite languages with a non-monotonic *standard* function being prima facie possible, gradable adjectives in their bare use are dominantly monotonic. The monotonic standard analysis has been applied cross-linguistically (Bogal-Allbritten, 2013; Grano & Davis, 2018; Liu, 2010; Sawada & Grano, 2011; Svenonius & Kennedy, 2006), which give some evidence that adjectives are monotonic cross-linguistically. This is a striking empirical pattern in the semantics of gradable adjectives, which raises the question of where adjectival monotonicity originates. An answer to this question might also shed light on the monotonicity of quantifiers, which I discuss below.

A critic could argue that monotonicity is entailed by the standard semantic account of gradable adjectives, and that the monotonicity pattern has therefore already been explained. This criticism confuses the relation between data and theory. The data tells us that adjectives are monotonic, and the semantic analysis is *designed* to model this behaviour. Therefore, an explanation of monotonicity based on the account's predictions would be circular. The critic could answer that the semantic analysis of gradable adjectives that entails their monotonicity is doing explanatory work thanks to its theoretical virtues, e.g. by being elegant, simple, and parsimonious. This type of argument is central to many areas of linguistics (Chater & Christiansen, 2007). The idea that general semantic principles are explanatory if they account for and predict apparently unrelated phenomena is plausible. More specifically, these general principles are arguably explaining the behaviour of the specific phenomena that fall under them, because if the general principles hold the individual phenomena could not have been different without deep changes in the language.

However, this type of explanation is irrelevant to the present discussion, for two related reasons. The first reason is that we are interested here in an evolutionary and

therefore causal explanation, which is not the type of explanation that theoretical virtues can deliver (even though a causal story can itself have theoretical virtues that make it better or worse). The second reason has to do with the difference between general principles and individual phenomena. The question why a phenomenon is not different from how it is in a way that breaks a general principle can arguably be answered by appealing to the general principle that regulates it. However, it is much less clear whether theoretical virtues can explain why the principles are like they are. In this case, the account with explanatory virtues is the introduction of the POS morpheme. The question why POS is not different from how it is in a way that makes it more complex cannot be answered by appealing to its actual meaning, since its actual meaning is precisely what is being called into question. The explanatory value of theoretical virtues would come into play if we were considering alternatives to POS that require a change in even more general semantic principles. However, I showed above that a local change in the *standard* function is enough to make the POS analysis of gradable adjectives compatible with non-monotonic extensions. The change is confined to the lexical entry for *standard* in a way that does not affect the semantic theory at large, e.g. by introducing new semantic types or compositional rules. Therefore, the non-monotonic account only introduces complexity at the level of the lexical entry for *standard*. In conclusion, the semantic analysis of adjectives cannot explain why gradable adjectives are monotonic.

I have discussed only one of the various analyses of gradable adjectives in the literature, namely the one proposed in (Kennedy & McNally, 2005). It is worth mentioning some alternative analyses in the literature to show that they do not change the space of possible meanings of adjectives in their bare use. This means that the problem of explaining monotonicity remains for these alternative account and that the solutions I will consider are equally applicable.

An analysis takes adjectives to be of type $\langle e, d\rangle$ instead of $\langle d, et\rangle$.[11] Thus, according to this analysis the lexical entry contains a measure that maps individuals onto degrees on the relevant scale. An adjective like "tall" then has lexical entry

$$[\![\text{tall}]\!] = \lambda x.tall(x) \geq d_c \tag{1.33}$$

where d_c is a contextually determined standard of comparison. This analysis avoids

[11] S ee Vennemann (1972), Kennedy (2013). Moreover, see Lassiter (2015, p. 155) for an overview of the debate.

explicit quantification over degrees and instead uses contextual indices to store the information relative to the adjectival thresholds (Lewis, 1970; Barker, 2002). The assumption that can be relaxed in this approach is the comparison to a single standard. Within this picture, a non-monotonic language could add structure to the index and add complexity to the comparison instead of adding structure to the POS morpheme (the details of how this would work are irrelevant for the issue at hand). This way, a comparison with any number of indices can be obtained, and therefore the space of possible meanings is the same as the ones analysed above.

A third analysis avoids degrees altogether and takes both gradable and non-gradable adjectives to have semantic type $\langle e, t\rangle$. One such analysis is developed in Klein (1980).[12] Klein argues that adjectives should satisfy the *Principle of compositionality*: If A is a gradable adjective, then the meaning of [$_{AP}$ A-er than __] is a function of the meaning of A. The degree analyses presented above do not satisfy this principle. Klein proposes a different picture in which adjectives are usual predicates of type $\langle e, t\rangle$, and more specifically context-dependent, partial functions that divide the universe set in three subsets:

- The positive extension of the adjective, $\{x \mid [\![\text{Adj(x)}]\!] = 1\}$
- The negative extension of the adjective, $\{x \mid [\![\text{Adj(x)}]\!] = 0\}$
- The extension gap of the adjective, $\{x \mid [\![\text{Adj(x)}]\!]$ is undefined$\}$

The context determines the universe set. Assume that it is essential to the semantics of gradable adjectives that they do not distinguish between any two objects a and b unless either a or b has the relevant property to a greater degree (Klein assumes with Sapir (1944) that comparison is a psychological primitive and not something that should be explained by the semantic theory). Then, it is easy to get the space of possible meanings that can be encoded with transitions. It is sufficient to lift the assumption that the set of degrees that are instantiated by the individuals in the positive extension of the adjective is monotonic in the sense defined above. This account is particularly compatible with the possibility of non-monotonic adjectives, since it does not encode monotonicity in the formal analysis.

[12] Also see Doetjes, Constantinescu, and Součková (2009), Burnett (2016).

Previous literature on monotonicity in adjectival semantics

It is worth discussing previous literature where monotonicity, sometimes in the sense defined above and sometimes in a different sense, is considered in relation to gradable adjectives. Kennedy (2001) talks about monotonicity of adjectives, starting from slightly different data. Kennedy (2001) is mainly interested in giving an account of *cross-polar anomaly*, namely the phenomenon that some comparative sentences are uninterpretable:

(1.34) # Mike is shorter than Carmen is tall.

In a classic view where the degrees of tallness and shortness are both degrees of height, 1.34 should not be problematic, as it is simply comparing two degrees of height.

Kennedy (2001) explains the phenomenon by analysing adjectives as mapping individuals to monotonic categories on scales rather than, as I have assumed until this point, single degrees. Positive adjectives, such as "tall", deal with monotonically decreasing categories, while negative adjectives, such as "short", with monotonically increasing categories, which Kennedy calls positive and negative *extents* respectively. More specifically, an adjective in bare use asserts that the property extent of some individual is fully contained in some contextually determined (monotonic) extent.

Consider "tall" as an illustrative example. A sentence like "Martha is tall" asserts that the extent to which Martha is tall, a monotonically decreasing category on the scale of heights, is fully contained in some contextually determined monotonically decreasing category $e_{s(tall)}$. To make this work compositionally in the simplest case of bare use "Martha is tall", "tall" is analysed as follows:

$$[\![\text{tall}]\!] = \lambda x.\text{tall}(x, e_{s(tall)}) \tag{1.35}$$

A comparative such as "John is taller than Mary" is true iff there is an extension e on the scale such that the tallness extension of Mary is fully contained in e but the tallness extension of John is not. The semantic oddness of example 1.34 can then be explained by the fact that extensions of tallness and shortness are different kinds of objects, and therefore cannot be compared.

The term "monotonicity" comes up in Kennedy (2001) in the discussion of the relation between positive and negative adjectives. As an example, consider "tall" again and its antonym "short". The same individual can be described in terms both

of tallness and shortness. In a classical view where adjectives map degrees onto sets of individuals, a given degree of tallness and the same degree in terms of shortness map onto the same set of individuals. Therefore, as Kennedy (2001) points out the relation between the scale of height and the orders of tallness and shortness is stipulated, rather than inferred from the lexical meaning of the involved terms. Specifically, it is stipulated that if a is greater than b on the scale of height, then "a is taller than b" and "b is shorter than a" are true:

$$a <_{height} b \implies \iota d.[\![\text{tall}]\!](d)(a) <_{tallness} \iota d.[\![\text{tall}]\!](d)(b) \tag{1.36}$$

$$a <_{height} b \implies \iota d.[\![\text{short}]\!](d)(b) <_{shortness} \iota d.[\![\text{short}]\!](d)(a) \tag{1.37}$$

This inferential pattern is what Kennedy calls monotonicity. In Kennedy's account, monotonicity in his sense is a natural consequence of considering extensions rather than single degrees, which is independently motivated by cross-polar anomaly. If the relation $<$ for an extent is seen as full containment, then

$$a <_{height} b \implies e_{tall,a} <_{tallness} e_{tall,b} \tag{1.38}$$

$$a <_{height} b \implies e_{short,b} <_{shortness} e_{short,a} \tag{1.39}$$

Where $e_{ADJ,x}$ is the extension of individual x with respect to adjective ADJ.

The crucial point for the the present purposes is that Kennedy's analysis of monotonicity does not imply that adjectives must be monotonic in their bare use in the sense I defined above. Kennedy does briefly considers adjectives that can select clauses:

(1.40) To see rain in July is odd. $\implies$ To see snow and rain in July is odd.

With the assumption that to see rain and snow in July is more odd than it is to see just rain in July. In view of the discussion above, this inference pattern can be recognized as a consequence of adjectival monotonicity in my sense. Crucially, there are possible adjectives compatible with Kennedy's semantic analysis that fail to be monotonic in my sense. For instance, Twinglish "Mary is tall" can be defined as:

$$\lambda x.\text{tall}(x, l_{s(tall)}, u_{s(tall)}) \tag{1.41}$$

which is true iff the extent of tallness of x contains $l_{s(tall)}$ but does not contain $u_{s(tall)}$, which are two contextually determined extents such that the former is strictly

included in the latter. Therefore, Kennedy's monotonicity is different from the notion of monotonicity discussed in this thesis.

A second paper where monotonicity is discussed in relation to gradable adjectives is Rett (2015). Rett discusses an asymmetry between monotonically increasing and decreasing adjectives:

(1.42) The car is faster than allowed.

(1.43) The car is slower than allowed.

While the sentence with the positive polarity adjective "fast" is unambiguous, the negative polarity adjective "slow" can be interpreted with respect to a minimum or a maximum. Sentence 1.43 can be interpreted as saying that the car is slower than some maximum allowed speed, or slower than a minimum allowed speed. Rett observes that monotonic categories across different grammatical categories give rise to this asymmetry. Rett's discussion is orthogonal to the discussion at hand, as she does not attempt to explain monotonicity in my sense.

The extremeness universal, formalized

I have shown above that at least for some adjectives, bare predication refers to extreme categories on the relative scale. How would such categories be modelled in the formal analysis of adjectival semantics I reported above? They are modelled with the *standard* relation. When *standard* gets a maximal adjective, it compares its argument with the maximum of the adjective's scale (see section 1.1.2 for the distinction between maximal and non-minimal adjectives):

$$\mathit{standard}(d)([\![\mathrm{ADJ}_+]\!])(C) = \max(\mathrm{scale}_{ADJ_+}) \leq d \tag{1.44}$$

where $[\![\mathrm{ADJ}_+]\!]$ is a maximal adjective and scale_{ADJ_+} is the adjective's scale. Since by definition no point is strictly greater than the maximum, equation 1.44 reduces to asserting that d is the scale's maximum. The definition for a non-minimal adjective ADJ_- is similar:

$$\mathit{standard}(d)([\![\mathrm{ADJ}_-]\!])(C) = \min(\mathrm{scale}_{ADJ_-}) < d \tag{1.45}$$

Note that, coherently with the analysis, the standard for absolute adjectives does not depend on the comparison population C.

Kennedy and McNally (2005) show a remarkable connection between the structure of the adjective's scale and the adjectives that are defined on that scale. Namely,

> gradable adjectives associated with totally open scales have relative standards; gradable adjectives that use totally or partially closed scales have absolute standards.

This means that if the ordered set underlying the adjective has an infimum or a supremum (or both), then the adjective refers to an extreme category. Otherwise, it does not.

There are various additional complications to this picture. First of all, there seem to be exceptions to the generalization above. A classic example is "full". Some containers do not have to be completely filled to count as full, e.g. a glass full of wine might not be completely full. Moreover, this introduces a dependence between the standard of comparison and the comparison population, which were predicted to be independent for absolute adjectives in the analysis above.

A second complication comes from the fact that absolute adjectives are in practice used to refer to non-extreme points. For instance, a stick can be described as straight despite not being completely straight. The fact that this is a pragmatic effect is demonstrated by the incoherence of the following example:

(1.46) # The stick is straight, but it is very slightly bent.

There are two main approaches to describe the pragmatic effect whereby an absolute adjective receives a non-absolute interpretation. First, the interest-dependent *granularity* of the scale might be the reason that non-extreme points are treated as extreme (Gaio, 2009). Second, there can be a *pragmatic halo* at the scale's extremes. I return to the latter approach in chapter 6.

In this section, I reviewed previous literature on the semantics of gradable adjectives, including various accounts of bare forms, and previous analyses of monotonicity. In particular, I concluded that the semantic analyses of gradable adjectives are consistent with non-monotonic and non-extreme categories. In the next section, I review the literature on quantificational semantics and the properties of monotonicity and extremeness for quantifiers.

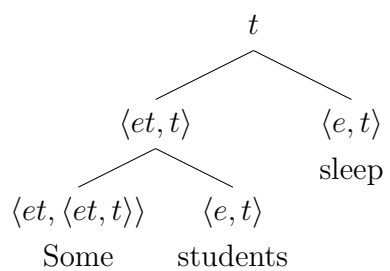
1.2.2 Quantifiers

In the previous section, I discussed a formal analysis of gradable adjectives that analyses bare uses as referring to monotonic and sometimes extreme categories on scales. In this section, I move onto discussing another class of terms, namely quantifiers. Quantifiers have received both set-theoretical analyses and analyses based on degrees and scalar concepts. I briefly discuss both and show how they can formalize monotonicity and extremeness. Both styles of analyses are ultimately compatible with the proposed universals of scalarity.

Generalized quantifiers theory

A common analysis of the semantics of natural language determiners sees them as expressing *generalized quantifiers*.[13] Generalized quantifiers are semantic objects of type $\langle\langle e,t\rangle, \langle\langle e,t\rangle, t\rangle\rangle$. Therefore, they can be thought of as functions that take a set and return a function from a set to a truth value.[14] Alternatively, they can be understood more simply as properties of tuples of sets. I write $Q_{\mathcal{M}}(A,B)$ to indicate the quantifier evaluated on the sets A as first argument and B as second argument, in model $\mathcal{M}$. Attributing this semantic type to determiners in natural language allows a simple compositional analysis of sentences such as "Some students sleep", where ⟦students⟧ and ⟦sleep⟧ are sets:

(1.47)



When the quantifier appears in predicative position, the situation is slightly more complicated. However, the syntactic details are irrelevant for the present purposes. It is worth noticing that the truth of a quantifier is evaluated in a universe of objects M. Features of M, in addition to the quantifier's arguments, can determine the truth of a quantifier. I return to M in more details below.

[13] In the following, I will focus on D-quantifiers. What I say can mostly be easily extended to A-quantifiers. See Lewis (1975) for the origins of the distinction.

[14] This is a specific type of generalized quantifiers. In the general definition, generalized quantifiers can take as arguments any n-tuple of relations. Notation wise, $\langle i_1, \ldots, i_n\rangle$ indicates the type of a generalized quantifier that takes an i_j-ary relation in its jth argument.

Any function from tuples of sets to truth values is a generalized quantifiers. This includes very odd functions, for instance a function that is true iff either of the sets contains a dodo, and false otherwise. Simple determiners in natural language can only express a few of the possible generalized quantifiers. Some substantial restrictions on the meaning of natural language determiners have been noticed, which limit the possible generalized quantifiers that morphologically simple determiners can express. These restrictions are called the *universals of quantification*. I consider a few of them.

First, consider a property of some generalized quantifiers called *conservativity*. For all universes M and all sets $A, B \subseteq M$, a generalized quantifier Q is conservative iff:

$$Q_{\mathcal{M}}(A, B) \iff Q_{\mathcal{M}}(A, A \cap B) \tag{1.48}$$

In intuitive terms, this means that the elements in B that do not also belong to A are irrelevant to the truth of the quantifier. For instance, "all" expresses a conservative quantifier:

(1.49) All bird fly $\iff$ All birds are flying birds.

A second proposed universal of quantification is *extension*. For any two universes M, M' such that $A, B \subseteq M \subseteq M'$:

$$Q_{\mathcal{M}}(A, B) \iff Q_{\mathcal{M}'}(A, B) \tag{1.50}$$

This intuitively means that changes in the universe that do not affect the quantifier's arguments do not affect the quantifier's truth. In combination, conservativity and extension mean that the truth of the quantifier can be established by simply looking at what is inside A.

A third universal is *isomorphism closure*. To define isomorphism closure, the concept of isomorphic models is needed. A model $\mathcal{M}$ is a tuple consisting of a set of objects M and relations over M. Two models $\mathcal{M} = \{M, R_1, \ldots, R_n\}$ and $\mathcal{M}' = \{M', R'_1, \ldots, R'_n\}$ are *isomorphic* iff[15] there is a bijection f from M to M' such that for any element a of M and any $i \leq n$,

$$R_i(a) \iff R'_i(f(a)) \tag{1.51}$$

[15]I only define monotonicity for generalized quantifiers that relate two predicates, i.e. type $\langle 1, 1 \rangle$ quantifiers, for simplicity of exposition. For a more general definition, see Peters, Westerstahl, and Westerståhl (2006, p. 99).

In intuitive terms, an isomorphism from a models to another is a mapping that preserves the structure and has an inverse (it does not lose any structure). In the case of quantifiers, the relevant bit of structure is the set to which each object belongs. Therefore, two models are isomorphic if there is a mapping f from the objects in one to the other model that preserves to which sets each object belongs. Based on the concept of isomorphism, the universal of isomorphism closure can be defined. A quantifier Q is isomorphically closed iff for any two isomorphic models $\mathcal{M}$ and $\mathcal{M}'$,

$$Q_{\mathcal{M}}(R_i, R_j) \iff Q_{\mathcal{M}'}(R'_i, R'_j) \tag{1.52}$$

This says that the quantifier does not distinguish between isomorphic models. The only information that is retained across all isomorphic models is the number of elements in each set. Therefore, isomorphism closure says that all that matters for the truth of a quantifier is the number of elements in the involved sets, i.e. $A, B, A \cap B, M - (A \cup B)$.

A quantifier Q that is conservative, extensive and closed under isomorphism satisfies the following for every sets $A, B \subseteq M$ and $A', B' \subseteq M'$. If $|A \cap B| = |A' \cap B'|$ and $|A - B| = |A' - B'|$, then:

$$Q_{\mathcal{M}}(A)(B) \iff Q_{\mathcal{M}'}(A')(B') \tag{1.53}$$

This means that the truth of the quantifier only depends on the size of two sets, namely the intersection of restrictor and scope set and the restrictor minus the scope set. This is proved in Keenan and Westerståhl (2011, p. 875). It is generally accepted that natural language quantifiers satisfy 1.48, 1.50, and 1.51. Therefore, as an illustration, 1.53 implies that for any natural language quantifier Q, if Q cats sleep and there are (1) as many non-sleeping cats as jumping horses, and (2) as many sleeping cats as non-jumping horses, then Q horses don't jump.

The monotonicity universal for generalized quantifiers

I move next onto monotonicity as a proposed universal of quantification. In the generalized quantification literature, a quantifier Q is *monotonically increasing* [decreasing] in its right argument iff for all $A \subseteq M$ and all $B \subseteq B' \subseteq M$ [all $B' \subseteq B \subseteq M$]:

$$Q_{\mathcal{M}}(A, B) \implies Q_{\mathcal{M}}(A, B') \tag{1.54}$$

Similarly for its left argument. In words, a quantifier is monotonically increasing in an argument if the quantification cannot be falsified by adding elements to the set corresponding to the argument. As an example, consider "most", which is monotonically increasing in its left argument. Since the set of sleeping things is a (proper) subset of the set of living things:

(1.55) Most cats sleep. $\implies$ Most cats live.

While some quantifiers (the Aristotelian quantifiers "all", "some", "no") are monotonic in both arguments, monotonicity in natural language quantifiers is usually found in the right argument. The intuitive reason for this becomes clear when thought in the context of the universals described above. Giving the freedom to add new elements or remove old elements from the left argument A, i.e. the restrictor argument, means that elements can be removed or new elements added both from $A - B$ and from $A \cap B$. One could thus change everything that is relevant to the truth of the quantifier. Such a quantifier would then have to encode information that is robust to removing or adding new elements to both of the sets relevant to its truth. On the other hand, adding new elements or removing existing elements from the right argument B, the scope set, can at most change the size of $A \cap B$, but not the size of $A - B$. It is easier for quantifiers to be robust to such changes to their right arguments.

The properties expressed by equations 1.53 and 1.54 point to a way that quantifiers are monotonic in the sense introduced above in this thesis. If a quantifier is monotone increasing in its right argument and $Q(A)(B)$, it means that while keeping A the same, increasing the size of $A \cap B$ cannot falsify the quantifier. Therefore, quantifiers that are *saturated* in their first argument (i.e. whose first argument has already been given a value) are monotonic on the scale of numbers. Denote a quantifier Q that is saturated by set A in its first argument Q^A. For all sets $A \subseteq M$ and any quantifier Q that satisfies the universals of quantification presented above:

$$|B| \leq |B'| \wedge Q^A_{\mathcal{M}}(B) \implies Q^A_{\mathcal{M}}(B') \tag{1.56}$$

This clarifies the relation between the concept of monotonicity in generalized quantifier theory and the concept of monotonicity I discussed above. For any value of the restrictor set, quantifiers are monotonic in the sense defined above on the total order imposed by cardinality (rather than e.g. the partial order of inclusion).

The definition of monotonicity for quantifiers in equation 1.56 is however somewhat unsatisfying, because it only concerns the truth value of the scope set, once the restrictor set is fixed. This is at odds with the fact that quantifiers have two arguments, namely two sets $\langle A, B\rangle$. Since quantifiers take two arguments, a total order on tuples of sets would be more natural for quantifiers to be monotonic on. The classes of proportional and numerical quantifiers presented above help in this respect. Recall that all the information that a quantifier can exploit after it has been restricted to satisfy the universals discussed above is the sizes of $A - B$ and $A \cap B$. Proportional quantifiers exploit information from both sets, while numerical quantifiers only make use of the latter. In each case, it is possible to define monotonicity without the need to fix the restrictor set. I first define proportional and numerical quantifiers in the framework of generalized quantification theory, and then I define scales on which such quantifiers are monotonic.

Each proportional quantifier can be expressed as

$$\lambda A.\lambda B.\frac{|A \cap B|}{|A|} \circ P \tag{1.57}$$

where $\circ \in \{<, >, \leq, \geq\}$ and $P \in [0, 1]$. $\circ$ is not an argument of the quantifier, but rather part of its lexical meaning. P can be lexically encoded, or context dependent. This is the case for quantity words such as "few" and "many", to which I return below. Many quantifiers can be analysed in this manner. For instance:

Q	$\circ$	P
All	$\geq$	1
Some	$>$	0
Most	$>$	0.5
No	$\leq$	0

Each numerical quantifier can be expressed as

$$\lambda A.\lambda B.|A \cap B| \circ H \tag{1.58}$$

Where $\circ$ has the same meaning as in equation 1.57. The assumption that numerical quantifiers only depend on $|A \cap B|$ implies that "all" is not a numerical quantifier, while "none" might be considered one. A quantifier analysed in these terms is:

$$[\![\text{Five}]\!] = \lambda A.\lambda B.|A \cap B| \geq 5 \tag{1.59}$$

This a common analysis of bare numerals as monotone increasing quantifiers, where the exactness reading comes from a scalar implicature.[16] The exactness implicature is cancellable in appropriate contexts. For instance, if a tax reduction scheme requires a family to have three children, families with four children will qualify too.

Monotonicity for numerical and proportional quantifiers can now be defined in a way that involves both argument sets. As in the case of gradable adjectives, there are two different ways of defining monotonicity. First, quantifiers can be seen as monotonic categories on the totally ordered sets of proportions or integers. In this definition, P in equation 1.57 and H in 1.58 are the transitions of the categories expressed by the quantifiers. Quantificational monotonicity in this sense abstracts from specific sets and deals directly with proportions or magnitudes. Second, quantifiers can be seen as monotonic categories on the set of tuples of sets, ordered by a total order. The total order is slightly different in the case of proportional and numerical quantifiers. For any four sets $A, B, C, D \subseteq M$, the order for proportional quantifiers is:

$$\langle A, B\rangle \leq \langle C, D\rangle \iff \frac{|A \cap B|}{|A|} \leq \frac{|C \cap D|}{|C|} \tag{1.60}$$

Proportional quantifiers can then be seen as monotonic categories on the set of tuples of sets ordered by the relation in equation 1.60. On the other hand, the relevant order for numerical quantifiers is:

$$\langle A, B\rangle \leq \langle C, D\rangle \iff |A \cap B| \leq |C \cap D| \tag{1.61}$$

Numerical quantifiers can be seen as monotonic categories on the set of tuples of sets ordered by the relation in equation 1.61.

The analysis of quantifiers above explains an asymmetry between "all" and "none" when the restrictor set is empty:

(1.62) No blue cat sleeps.

(1.63) Every blue cat sleeps.

If there are no blue cats, there is a sense in which 1.62 is true. This is predicted by the interpretation of "no" as a numerical quantifier. On the other hand, the truth value of example 1.63 is unclear. In the contemporary analysis of quantification, it

[16]This analysis was introduced in Horn (1972), but see Barwise and Cooper (1981) for a classic formal implementation.

is conventionally taken to be *vacuously* true, while in medieval logic it was taken to be false (Parsons, 2014, p. 10). This shows that the intuition on the truth value of proportional quantification with an empty restrictor set are unclear. In my analysis, the expression is semantically defective, since the denominator of the fraction is 0.

Like in the case of adjectives, it is easy to imagine a quantifier that is non-monotonic but satisfies the other universals of quantification. For instance, define the following quantifier:

$$\lambda A.\lambda B.\frac{|A \cap B|}{|A|} \neq 0 \wedge \frac{|A \cap B|}{|A|} \neq 1 \tag{1.64}$$

This is true iff neither none nor all of the objects in A are also in B. The possibility of non-monotonic quantifiers raises the question of why simple determiners in natural language express monotone quantifiers. This question will be the topic of chapter 7 and, to some extent, of chapter 3.

One further advantage of a scalar analysis of quantification is that it can be naturally extended to quantification over mass nouns. I only review the most basic facts. A simple picture of quantifiers with mass nouns is that they bind a variable ranging over portions of stuff. However, this cannot be the right analysis. von Heusinger, Maienborn, and Portner (2011) illustrate this point with the following type of example:

(1.65) All honey is either fluid or crystallized.

The example can be true despite the fact that some jars of honey contain both fluid and crystallized honey. This means that there are some portions of honey of which it is false that they are "either fluid or crystallized", because they are both. Therefore, 1.65 would be false in a stuff-variable analysis. An alternative analysis sees quantifiers as including a mereological sum operator, which takes a predicate and returns the sum of all the stuff that the predicate applies to. I write:

$$\sigma x.P(x) \tag{1.66}$$

to refer to such a sum for the predicate P. $\sigma x.water(x)$ would for instance be the object consisting of all the water. The lexical entry for "all", when its arguments

are mass nouns, receives then a similar interpretation as with count nouns:

$$[\![\text{all}]\!] = \lambda A.\lambda B.\frac{\mu(\sigma x.A(x) \cap \sigma x.B(x))}{\mu(\sigma x.A(x))} \geq 1. \tag{1.67}$$

Where μ is some measure function which plays the same role of mass as cardinality for sets. Much more could be said about quantification with mass nouns, but it would be beyond the scope of the present work.

Monotonicity in a degree analysis of quantification

I have discussed in the previous section the approach that sees natural language determiners as expressing generalized quantifiers. I examined the difference between monotonicity in generalized quantifier theory and monotonicity in the sense defined in this thesis. A different approach to analysing the meaning of quantifiers starts with the terms "many", "much", "few", and "little", called *quantity words* in Rett (2018).

Rett (2018) provides the most up-to-date degree analysis of quantity words. In particular, Rett aims at accounting for the previously unexplained fact that quantity words can modify verb phrases, prepositional phrases, and comparatives, the so-called *non-singular* uses:

(1.68) Mary can't see much.

(1.69) The thesis isn't much below expectations.

(1.70) John is much further away from Rome than us.

I sketch Rett (2008) and Rett (2018)'s implementation of the degree analysis of quantity words. Rett (2018) assumes a null operator M-OP which takes a predicate and relates it to a degree:

$$[\![\text{M-OP}]\!] = \lambda P.\lambda d.\lambda x.P(x) \wedge \mu(x) = d \tag{1.71}$$

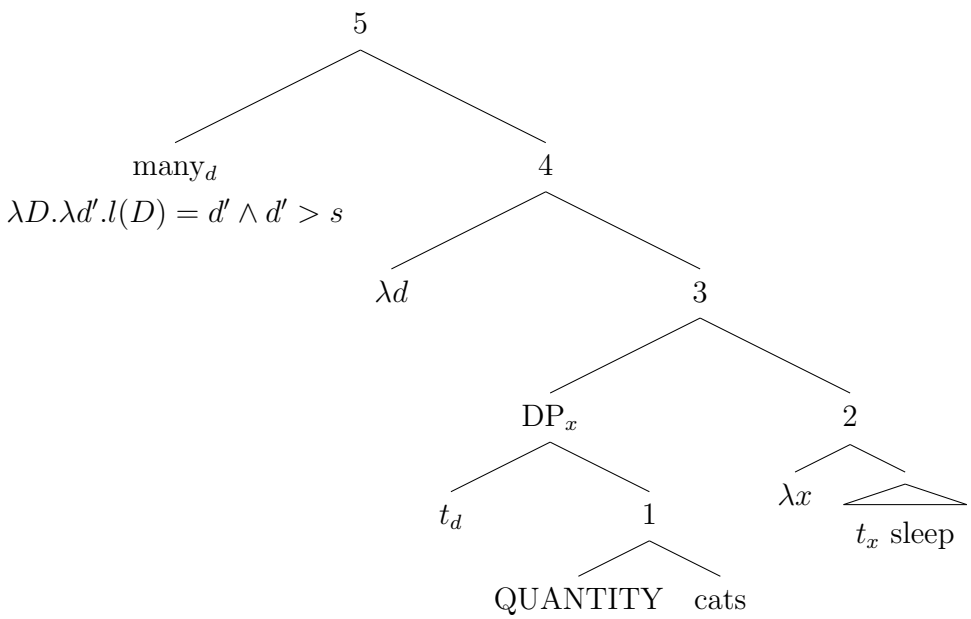
while in Rett (2008) a similar null operator QUANTITY plays essentially the same role:

$$[\![\text{QUANTITY}]\!] = \lambda P.\lambda d.\lambda Q.\exists x[P(x) \wedge Q(X) \wedge \mu(x) = d] \tag{1.72}$$

Where μ is some measure function that is appropriate in the context. The differences between the two accounts are irrelevant for the present purposes.[17]

Consider the sentence "Many cats sleep". In a degree account, the M-OP morpheme is introduced to connect the bare numeral to the rest. I report here the analysis of Rett (2008, p. 42):

(1.73)



D is the characterizing function of a set of degrees. l is a function that returns the measure of such a set.[18] Note that "many" raises leaving a trace t_d. This ensures that the many refers to the cats that sleep, rather than just to the cats. The meaning of the whole sentence can be obtained compositionally:

$$1 = \lambda d.\lambda Q.\exists x[\mathbf{cat}(x) \wedge Q(x) \wedge \mu(x) = d]$$
$$2 = \lambda x.\mathbf{sleep}(x)$$
$$3 = \exists x[\mathbf{cat}(x) \wedge \mathbf{sleep}(x) \wedge \mu(x) = d]$$
$$4 = \lambda d.\exists x[\mathbf{cat}(x) \wedge \mathbf{sleep}(x) \wedge \mu(x) = d]$$
$$5 = \lambda d'.l\,(\lambda d.\exists x[\mathbf{cat}(x) \wedge \mathbf{sleep}(x) \wedge \mu(x) = d]) = d' \wedge d' > s$$

[17] While in M-OP the two predicated P and Q combine with predicate modification in the style of Heim and Kratzer (1998), they are distinct arguments in QUANTITY. Moreover, even though both M-OP and QUANTITY require existential closure over degrees at the sentential level, QUANTITY also explicitly encodes existential quantification over objects.

[18] l is different from μ exactly in that the argument of the former is a set of degrees and the argument of the latter is an individual. l returns the greatest among the degrees that verify its argument.

Finally, existentially close the remaining variables:

$$= \exists d' \left[l \left(\lambda d. \exists x [\mathbf{cat}(x) \wedge \mathbf{sleep}(x) \wedge \mu(x) = d] \right) = d' \wedge d' > s \right]$$

When its argument is a set, μ returns the cardinality of the set. Since any subset of a set of sleeping cats is also a set of sleeping cats, node 4 is verified for any number from 1 to the maximal number of sleeping cats. The function l will return that maximum number. The respective account with M-OP is worked out in details in Rett (2014).

Note that QUANTITY alone does not imply the monotonicity of the quantity words. It is easy to imagine a lexical entry similar to that of "many" that does not require the output of l to be greater than a contextually determined degree, but rather e.g. between two degrees. One would then obtain a non-monotonic category without changing QUANTITY. In Rett's analysis, monotonicity follows from the lexical entry of "many" in a way parallel to the adjectival case. The reason why "many" implies monotonicity is that if the maximal degree d_1 of a set of degrees D_1 is greater than the standard s, then any $d_2 > d_1$ is also greater than s. The fact that monotonicity is not implied by QUANTITY is important for the present purposes, because if monotonicity followed from the way degrees themselves are encoded there would be an explanation for monotonicity that only appeals to semantics. Such an explanation would make the evolutionary approach less appealing.

The extremeness universal for generalized quantifiers

The totally ordered set of proportions has an infimum, 0, and a supremum, 1. Given the generalization from the discussion on gradable adjectives, we should expect categories defined on that scale to have their transitions at the infimum and supremum. However, many quantifiers do not. For instance, "most", "many", "few", "several", "numerous" all have non-extreme transitions.

A possible reason why quantifiers express non-extreme categories is the need to be accurate when communicating about quantities. There are only four extreme monotonic categories on any given totally bounded scale, namely

1. A true transition from true to false at the infimum (corresponding to "none").

2. A true transition from false to true at the supremum (corresponding to "all").

3. A false transition from false to true at the infimum (corresponding to "some").

4. A false transition from true to false at the supremum (corresponding to the morphologically non-simple "not all").

Since there are more than four proportional quantifiers, at least some have to be non-extreme. In the case of numerical quantifiers, only types 1 and 3 are possible because the set of integers $\mathbb{N}$ does not have a supremum. Since there are more than two numerical quantifiers, not all of them can express extreme categories.

It is worth noticing that the quantifiers that do express extreme categories, namely the Aristotelian quantifiers "all", "some", and "no", are remarkable in several respects. Experimental evidence shows that participants verify Aristotelian quantifiers faster than other quantifiers (Szymanik & Zajenkowski, 2009). Moreover, they are used more often than other types of quantifiers (Szymanik & Thorne, 2017). While many quantifiers express non-extreme categories, the ones that do seem to play a special role.

1.2.3 Possible exceptions to monotonicity

In the previous sections, I presented standard formal analyses of gradable adjectives and quantifiers. I have shown that all these analyses allowed the construction of a natural order, and that they referred to monotonic categories on these orders. Moreover, monotonicity is not implied by the semantics of these terms, but is rather a semantic-external fact. At various points, we encountered and discussed various exceptions to extremeness. I now discuss a few terms that are prima facie non-monotonic. This will illuminate possible ways that monotonicity can fail, showing that non-monotonic categories are possible and therefore that the commonness of monotonic categories calls for an explanation.

The first apparent counterexample to monotonicity in the adjectival domain is "chubby". Prima facie, chubby means something like "above average weight, but not fat". The intuition is that chubby shares the same scale as "fat", and refers to a non-monotonic category on this scale. However, there are various indications that "chubby" is not on the same scale as "thin" and "fat". A first indication that "chubby" is not on the same scale as "fat" comes from modification:

(1.74) Mary is slightly chubby.

The modifier "slightly" indicates that "chubby" is a non-minimal adjective, and therefore that the scale of chubbiness is half-open. Second, no amount of intensification of "chubby" seem to force the conclusion that somebody is fat. I conclude that the scale of chubbiness is not the same as the scale of fatness. Lastly, the fact that "chubby" satisfies the following inference indicates that it is monotone:

(1.75) If Roberta is chubby, and Elizabeth is chubbier than Roberta, then Elizabeth is chubby too.

"Lukewarm" is a particularly puzzling case. Clearly, "lukewarm" refers to an interval on the scale of temperatures. However, it is not gradable:

(1.76) # The soup is slightly / very / completely / absolutely / somewhat lukewarm.

There is an interpretation of "the soup is completely lukewarm" where every mereological part of the soup is lukewarm, but this is not the intended temperature interpretation. Moreover, "lukewarm" does not allow comparison:

(1.77) # The soup is more lukewarm than the curry.

Therefore, "lukewarm" does not seem to have a degree argument, despite its connection to the temperature scale. The fact that "lukewarm" is not gradable does not imply that it should not be monotonic on the scale of temperature. I take gradability to be a symptom of scalarity, but not a requirement. This is because the mechanisms of gradation in natural language might require monotonicity, making my explanation circular. I argue below for an account where monotonicity depends solely on the structure of the conceptual domain. Therefore, the fact that "lukewarm" refers to the scale of temperature should in itself suffice to make it monotonic.[19] Whatever the structure of "lukewarm", the existence of non-monotonic scalar categories is compatible with the evolutionary model I develop in chapter 3. Indeed, I conclude that a large enough pressure for communicative accuracy might cause non-monotonic terms to evolve. Temperature is a good candidate for a common topic of communication with high-stakes.[20]

[19] I return in the next section to the problem of differentiating between the semantic phenomenon of scalarity and the grammatical phenomenon of gradability.

[20] See Koptjevskaja-Tamm (2015, p. 341) for a discussion of intermediate temperature adjectives "tiepido" and "fresco" in Italian. Their behaviour is interestingly different from "lukewarm" in that they can be graded.

Categories on conceptual scalar domains are in some cases expressed by nouns. For instance, consider nouns that refer to individual of different ages: "infant", "child", "adolescent" ("teenager"), "adult". These terms are all monotonic on the scale of age except "adolescent". Interestingly, "child" and "adult" are not etymologically related to gradable adjectives, showing that their monotonicity is not derived from adjectival monotonicity. The same considerations apply to "adolescent" as for intermediate gradable adjectives in the previous paragraph. While much more could be said about the semantics of nouns in scalar conceptual domains as well as other alleged instances of non-monotonicity, I leave a more detailed discussion to future work for reason of space.

1.2.4 Conclusion: the role of formal semantics in the debate

In this section, I discussed the role that scales have to play in the semantics of two word classes, as well as two universals of scalarity. It is commonly accepted that a semantics of gradable adjectives requires the use of scales,[21] and I showed that bare uses of adjectives are monotonic on these scales in the sense defined above. The case of quantifiers is slightly more complicated. I found that much discussed universals of quantification imply that quantifiers refer to monotonic categories on the natural (weak) total order imposed by cardinality, once a restrictor set is fixed. Going beyond the universals of quantification to the more specific classes of proportional and numerical quantifiers allowed me to define a sense in which quantifiers refer to monotonic categories on scales that involve both arguments. The independently motivated degree analysis of quantification confirmed this account.

Formal semantics has three roles to play in this work, which justified the extended discussion in this section. First, some classes of words have been given an analysis based on scales, pointing to the fact that the respective conceptual domains are scalar. The relevance of this fact will become clear in the next chapter. Second, the formal semantic accounts of gradable adjectives and quantifiers are based on a large set of examples and linguistic data. Therefore, couching the proposed universals in terms of independently motivated analyses in the semantics literature ensures that the universals indeed apply when a large number of linguistic data is considered.

[21] See the discussion of Klein (1980) on page 40 for an attempt at developing a semantics of gradable adjectives without scales, and how to make sense of the question about the evolution of monotonicity within such a theory.

The third role of formal semantics, which I briefly discussed above, is to show that the explanation for the universals proposed are not internal to semantics. Suppose that monotonicity and extremeness could be explained by some other, more general semantic fact. Then, it might still be an empirical question why the more general semantic fact holds, but there would be no need for an explanation of the universals as such. In other words, formal semantics can show what the right explanandum is.

The semantic analysis was needed to show the naturalness of analysing the meanings of the categories I discuss in scalar terms. However, the literature I discussed in this section was not concerned with the causal mechanisms that lead to the considered universals. In the next section, I turn to the accounts of the evolution of monotonicity and universals that have been proposed in the literature.

1.3 Previous work on the evolution of scalar universals

In the literature, there have been attempts to explain the linguistic patterns for monotonicity and extremeness I pointed out above. In this chapter, I will review and evaluate the most important of those. The aim for the literature presented in this section is to find a mechanistic explanation of the two universals under discussion across the different word classes I considered.

I start by considering previous work on the evolution of monotonic categories, both in gradable adjectives and in quantifiers. Then, I move onto the evolution of extremeness, which has been studied especially in the case of gradable adjectives. I delineate some problems and gaps in the previous literature that motivate the approach I take in the rest of the thesis.

1.3.1 The evolution of monotonicity

Gradable adjectives

Monotonicity in the sense I defined above is not discussed directly in the evolutionary literature on gradable adjectives. However, discussions of similar phenomena suggest some routes to explore. A first possible explanation of the evolution of monotonicity is based on Kennedy (2007)'s *Interpretive Economy* principle:

(1.78) (Interpretive Economy) Maximize the contribution of the conventional meanings of the elements of a sentence to the computation of its truth conditions.

This principle, to which I return again in the next section, is introduced to explain data in a discussion related to the position of the comparison standard. Intuitively, it says that the computation of truth conditions tends to stick to conventional aspects of the involved meanings, avoiding when possible non-conventional, context sensitive factors. Since comparison standards can be context-sensitive, each additional transition increases the amount of sensitivity to the context. Therefore, the principle predicts a tendency to minimize the number of transitions. Assuming that one standard is the minimum possible number of standards, the principle predicts monotonicity.

The main issue of this principle in an evolutionary context is mentioned in Potts (2008): "It is an optimization principle left unsupported by a theory of optimization". In other words, Kennedy does not propose a mechanism through which context-sensitivity would be avoided in language. Without such a mechanism, Interpretive Economy is not an evolutionary explanation for monotonicity. A second, more substantial issue with explaining monotonicity using the Interpretive Economy principle is the possibility of non-monotonic gradable adjectives that are not context-sensitive. The Interpretive Economy principle cannot account for the absence of such adjectives, since they do not introduce additional context sensitivity. I constructed an adjective with this behaviour in section 1.2. It might be possible to refine the principle and use it to explain monotonicity, but in chapter 3 I take a different more promising direction, and argue that monotonicity is the result of optimization of semantic structure to the requirements imposed by various pressure, but I will give an explicit mechanism through which this optimization happens.

Quantifiers

A possible evolutionary account of quantificational monotonicity is based on the idea that reasoning patterns are simpler with monotonic concepts. Therefore, monotonic concepts would have been preferred in the evolution of conceptual structure. Geurts and Van Der Slik (2005) argue that the psychological complexity of inferences is lower with monotonic than non-monotonic quantifiers, which is particularly important as the meaning of quantifiers are in general complex: "a system for pro-

ducing monotonicity inferences can be very simple, because it requires only a shallow understanding of the representations it operates on". This direction of research is promising, since it makes clear empirical predictions and delivers a clear picture of the connection between semantics and cognition. However, it is unclear through which mechanism the role that quantifiers have in the cognitive complexity of specific inferences would influence the meaning of quantifiers at the language level. Moreover, the literature has mostly attempted to compare the complexity of upward and downward monotonic quantifiers, rather than monotonic and non-monotonic quantifiers. See Szymanik (2016, p. 38) for a review of processing-based explanations for monotonicity in the semantics of quantifiers.

As I mentioned above, learning is an important pressure acting on language evolution. Learnability-based explanations have been developed for quantificational monotonicity. Steinert-Threlkeld and Szymanik (2020) and Steinert-Threlkeld and Szymanik (2019) use neural networks as a model of learning, and show that learning monotonic quantifiers is easier than learning non-monotone quantifiers. They infer that ease of learning is a plausible explanation for why natural language quantifiers are monotonic. One component missing from this explanation is the evolutionary component, namely a story of how individual-level ease of learning explains the development of monotonicity at the language level. I return to this approach again below in chapter 7, which attempts to partially fill this gap in the literature.

The project of explaining universals of quantification in terms of learnability is developed in a different direction in van de Pol, Steinert-Threlkeld, and Szymanik (2019). The authors use the tool of (approximate) Kolmogorov complexity (Chater & Vitányi, 2003) to compare monotonic and non-monotonic quantifiers. Often, a string can be described in a way shorter than the string itself, by exploiting patterns in the string. Kolmogorov complexity is a measure of the shortest description of a sequence of symbols, which is uncomputable but can be approximated. van de Pol et al. found that monotonic quantifiers are robustly less complex than non-monotone quantifiers. This is a valuable result in the context of the present work, as it provides evidence that the right explanation for the universal of monotonicity lies in the simplicity of monotonic quantifiers. However, the relation with the conceptual spaces approach that I will develop in the next chapter is, for the moment, unclear. In particular, the result in van de Pol et al. (2019) cannot readily be applied to the monotonicity of gradable adjectives. An extension of the work on Kolmogorov complexity to other lexical classes constitutes a promising research avenue.

Magri (2015) presents a similar attempt to derive monotonicity from learnability in the PAC (*Probably Approximately Correct*) learnability framework. The result is that while other universals of quantification help with learnability, monotonicity does not. Note that PAC learnability is not concerned with the simplicity of a representation, but rather with the possibility of constructing an accurate classifier from observed data. Therefore, this result further supports the hypothesis that a simplicity bias is needed to explain the evolution of monotonicity.

Another relevant attempt at modelling the acquisition of the meaning of quantifiers is Piantadosi, Tenenbaum, and Goodman (2012) (expanded in Piantadosi, Tenenbaum, and Goodman (2016)). The authors model a rational Bayesian learner, who observes the cardinalities of the relevant sets along with quantifiers produced by a noisy adult speaker. The set of possible meanings is defined in terms of a language of thought (LOT), consisting of a probabilistic context-free grammar whose words are sets along with logical and quantitative relations. The model encodes an important bias for meanings that are encoded with shorter LOT expressions. The modelled learners are capable of acquiring the correct meaning for many quantifiers with just a few hundred to a few thousands observations. This is remarkable as it is comparable to the amount of data based on which a child learns the meaning of quantifiers. A further advantage of this model is that it deals with presuppositions, which I have not explicitly discussed.

Piantadosi et al. (2012) does not explicitly discuss monotonicity or extremeness, but its basic approach is very different from the approach taken in this thesis. The difference in approach—language of thought (LOT) vs conceptual spaces theory—would make a systematic comparison of the approaches difficult. Moreover, I will not give an explicit model of learning until a later chapter. I limit the discussion of this paper to a few considerations. First of all, the choice of basic rules for the language of thought in which meanings are encoded is somewhat arbitrary, and allows the encoding of meanings that are implausible to be considered by learners. Secondly, in the account proposed in this thesis the meanings of "few" and "many" are derived naturally as context-sensitive transitions on the scale of proportions, which are independently needed for gradable adjectives but would not be straightforward to implement in a LOT approach (see chapter 2). Thirdly, unless transitions are encoded in the LOT, the connection between monotonicity in quantifiers and gradable adjectives becomes more mysterious than it is in my model. Lastly, in chapter 2 I present a picture of scalar meanings as transition-based categories on an ordered

conceptual domain. This picture significantly restricts the set of possible quantifiers compared to the quantifiers that can be encoded in a LOT approach, making the learning task as easy as finding the appropriate scale and setting the right number and position of transitions. Whether the picture for which I will argue below is too restrictive in the set of possible quantifiers is an important question which however needs to be investigated empirically.

Chemla, Buccola, and Dautriche (2019) presents a picture of the relation between the property of connectedness (which they name *connectedness*) and monotonicity that is connected in interesting ways to the analysis presented above. While in the next chapter I will emphasize the differences between monotonicity and convexity, Chemla et al. (2019) emphasize their similarities. More specifically, they notice the following important relation between monotonicity and convexity:

Theorem 1. *A quantifier is monotone iff it is connected and its negation is connected.*

An alternative explanation for monotonicity consists in introducing, beyond the usual reasons for convexity, a pressure that pushes negated categories to also be convex. In their words,

> monotonicity is *the* minimal property that ensures both connectedness and stability of connectedness under negation. [...] Therefore, if there were pressures for meanings obtained compositionally (from conjunction and negation) to be connected, then we might expect 'primitive' expressions to generally be monotone.

The pressure for meanings obtained compositionally to be simple is an exciting direction of research in evolutionary linguistics. However, Chemla et al. (2019) do not propose an explicit, mechanistic model of how this pressure should act on language.

My picture differs from Chemla et al. (2019)'s in a few respects. First of all, while Chemla et al. (2019) do not discuss an explicit theory of categorization, I claim that monotonicity and convexity arise in response to two different categorization strategies, namely prototype-based and transition-based strategies (see chapter 2). I argue below in section 2.1.4 that it is difficult for a unified categorization strategy to account for both monotonicity and convexity. A second difference is that while Chemla et al. (2019) focus on the distinction between content words and function words, I focus on the distinction between ordered and non-ordered conceptual

domains. These two distinction are different, as ordered conceptual domains can be found both among content words (gradable adjectives) and function words (quantifiers). My distinction does a better job of carving out the places where monotonicity, as opposed to merely convexity, appears. Until a more explicit evolutionary account of the pressure acting on language to make negated meanings simple is proposed, a direct comparison of the present theory with the one in Chemla et al. (2019) is difficult.

In addition to the theoretical component presented above, Chemla et al. (2019) offer experimental evidence that monotonic quantifiers are indeed more learnable. In an experiment, they show participants a series of screens, each showing 5 coloured circles, and they ask participants to judge whether the shown screen is consistent with a rule. The task consists of a two-option forced choice with immediate feedback. The rules are defined in terms of the number of red circles in the screen. Each rule is true for some number of red circles and false for other numbers, and therefore expresses a quantifier of the type "Q circles are red". Participants are tested on three types of quantifiers in three conditions respectively:

Condition	Rules	"Transitions"
Monotonic	a. 0, 1, or 2 red circles	2-3
	b. 3, 4, or 5 red circles	2-3
Merely connected	c. 1, 2, or 3 red circles	0-1, 3-4
	d. 2, 3, or 4 red circles	1-2, 3-4
Non-connected	e. 0, 1, or 5 red circles	1-2, 4-5
	f. 0, 4, or 5 red circles	0-1, 3-4
	g. 1, 2, or 4 red circles	0-1, 2-3, 3-4

The results were only partially supportive of the hypothesis. Monotonic quantifiers were on average learned the fastest, followed by connected, and finally by non-connected quantifiers. However, only the difference between monotonic and non-connected quantifiers was significant. Within the non-connected rules, (e) and (f) were learned at a similar speed as the merely connected ones, while rule (g) took longer. Chemla et al. (2019) interpret this result by analysing the connectedness of the rule's negation. However, these results are also predicted by the transition-based picture, as shown in the table; the higher the number of transitions needed to encode the meaning, the harder to learn the rule. Which of the two interpretations is closer to the cognitive reality could be investigated in future empirical work.

Brochhagen, Franke, and van Rooij (2018) develop a model of the evolution of monotonicity based on a combination of Iterated Learning and a communicative pressure with RSA agents. Given the similarities between this model and the models I develop below, I delay discussion of this paper until after presenting the models in chapter 3.

In this section, I have discussed previous literature on the evolution of monotonicity. The literature on monotonicity for gradable adjectives is scarce, whereas different proposals have been made for quantificational monotonicity. Most previous proposals to explain the evolution of monotonicity lack a mechanism to connect the level of the individual agent and the level of linguistic structure. The model presented in chapter 3 will address this issue.

1.3.2 The evolution of extremeness

In this section, I discuss previous literature on the evolution of extreme categories. Nearly all the literature on the evolution of extremeness concerns gradable adjectives. The first evolutionary account I consider we encountered already, namely Kennedy's Interpretive Economy principle (statement 1.78 above). The picture explains why *standard* selected extreme degrees when bounds are available in terms of two mechanisms.[22] First, *standard* selects a natural transition, i.e. a transition that stands out. In the case of open domains, statistical properties of the comparison population provide a transition that stands out. In the case of domains with some boundary, the boundary also provides such a natural transition. According to Kennedy (2007), this first part of the picture is insufficient to explain all the data, since in bounded scales there are two natural transitions: the ones provided by statistical properties of the comparison population, and the boundaries. Therefore, we should expect adjectives on bounded scales to be absolute and relative with roughly equal frequency, contrary to the observation that they are in fact mostly absolute. The second part of the picture evokes the Interpretive Economy principle to fix this hiatus between data and theory. Since the comparison population is context-dependent while the domain's boundary is not, the latter minimizes the contribution of the context to the computation of the truth-conditions. This explains why gradable adjectives on bounded domains tend to be absolute.

[22]For a review on the *standard* function and how it contributes compositionally to the meaning of adjectives in their bare uses, see section 1.1.1.

Kennedy's picture explains the data relating to gradable adjectives extremely well. It is capable of explaining why extremeness is a tendency rather than a strict universal, as well as accounting for much of the behaviour of gradable adjectives described in Kennedy (2007). However, as mentioned in the discussion of evolution of monotonicity, it does not provide a mechanistic evolutionary account and is therefore unsatisfactory in this context. Even though the correct evolutionary account of extremeness might involve something like the Interpretive Economy principle, it will do so in the context of a causal explanation. Moreover, Kennedy's theory cannot directly explain the tendency towards extremeness for categories whose transition's position is not context-dependent, such as quantifiers.

Potts (2008) is an attempt to couch the Interpretive Economy principle within a theory of optimization. In Potts' words, the aim is to "show that Interpretive Economy follows from basic assumptions about cognitive prominence and evolutionary stability". Potts notices that what Kennedy's explanation lacks is an account of why boundaries are cognitively salient and evolutionarily stable. Cognitive saliency is analysed in terms of *Schelling points*, i.e. points in a state space that are salient in the common ground: They are salient for me, I know that they are salient for you, I know that you know that they are salient for me, etc. The boundaries of a scale are Schelling points. Saliency alone is however not enough to explain the extremeness of absolute adjectives, as other points on the scale can be more salient in specific contexts. To explain the stability of boundary-transitions across many situations, Potts argues that boundaries are fixed because they become a *convention*. The evolution of the convention is shown by Potts within the framework of evolutionary game theory. The crucial result is informally that a population of speakers that is establishing a convention will eventually converge to boundary transitions—i.e. the Schelling points. In sum, Potts' picture identifies the root of extremeness in the tendency for conventions to stabilize on points that are salient in the common ground, given the fact that boundaries are particularly salient points.

Franke (2012a) and Franke (2012b) criticise Potts' explanation on two grounds. The first is that according to Franke an explanation of extremeness ought to account not only for the boundary position of transitions on closed scales, but also for the extremeness[23] and sensitivity to context of transitions on open scales. For instance, although the transition for "tall" is not on a boundary—since the scale of height

[23]In this paragraph, the word "extreme" is used in its everyday meaning rather than the technical meaning introduced above.

is unbounded—it is on a point that is extreme with respect to tallness compared to e.g. average height. Potts' explanation does not straightforwardly account for the extremeness of open scales, since there are no stable Schelling points on open scales. Moreover, without a Schelling point Potts' account predicts that transitions should be fixed to a context-independent value, contradicting the context-sensitivity of relative adjectives.

The second of Franke's criticisms to Potts' picture is that it fails to confront a prima facie problematic fact about extreme categories, namely their low utility for communication. Under the simple picture of communication which I will discuss in more detail in section 3.1, a communicative event is successful iff the state o observed by the sender is identical to the state g guessed by the receiver. This simple picture cannot be retained when dealing with continuous spaces. If the signal conveys a single point, the sender will never have a chance to use it; on the other hand, if the signal conveys a region of the space, the receiver will never guess exactly the right observation. An alternative picture quantifies success as the similarity between o and g. This way of measuring communicative success is more apt for continuous state spaces, because it allows the receiver to be more or less successful even when not guessing o exactly. This more realistic model of communication however introduces a problem for Potts' account. Namely, extreme meanings are guaranteed to perform communicatively very poorly.

Franke argues that communication with gradable adjectives serves different purposes, and that in important cases communicative success does not depend directly on the similarity of o and g. More specifically, Franke draws a distinction between referential and descriptive language use. To understand the distinction, consider the following cases. You are describing your favourite toothbrush that is in your bathroom, and you say it is "green". Based on this, I can form an accurate picture of the color of this toothbrush. This is a case of descriptive language use, as the aim for the receiver is to construct a representation that matches reality as closely as possible. In particular, notice that the success of this communicative interaction does not depend on whether e.g. there are other toothbrushes in your bathroom that resemble the one you described. On the other hand, consider a case where I am in your bathroom and I have to pick one of the toothbrushes to use. Only one of them is new, and you described it as being green. The aim is not to form an accurate representation anymore—even though forming an accurate representation might help in the task—but rather to pick a specific one of the toothbrushes. This is a case of

referential language use, where the receiver tries to identify one object from a set of possible objects based on information about the intended object's properties coming from the sender. In referential language use, the population of alternative picks for the receiver matters for communicative success. Even if I pick a toothbrush that is exactly the same color as the one you meant to describe, the communication would be a failure if there are two green toothbrushes in your bathroom and I happen to pick the used one.

Franke proposes that we should consider the evolution of extreme meanings in gradable adjectives in the context of referential, rather than descriptive, language use. A pressure on language to be communicatively accurate would then push for languages that allow its users to succeed in tasks like the new-toothbrush-identification task described above. To model referential communication, Franke formalizes the situation as follows. A *referential game* consists of a sender s, a receiver r, and a context $c = \langle o_1, \ldots, o_n \rangle$ consisting of n objects. Sender s wants to have r pick a specific object o_i. Each object is associated with a vector of m features $\in \mathbb{R}^m$. Each feature is distributed in a way that encodes whether it corresponds to a property that is open, bounded, or half-bounded:

Scale	Distribution density
Open	Normal ($\mu = 0, \sigma = 1/3$)
Bounded	Uniform (in $[0, 1]$ interval)
Half-bounded	Truncated Normal ($\mu = 0.1, \sigma = 1/3$)

The whole information can be modelled as an $n \times m$ matrix, whose j, k element represents the degree to which object o_j has feature k. The game is structured as follows. First, o_i is selected. Then, s picks a signal that conveys information about the features of o_i and sends the signal to r. Finally, r picks an object from c based on the information contained in the signal. If r picks o_i, then the communication is successful, otherwise it is not.

In the referential game described by Franke, there are $2m$ many signals, i.e. two signals for each feature. Each signal conveys two distinct elements of information, namely (1) which property s intends to single out, and (2) whether o_i has the property to a high or low degree. For instance, if one of the features is height, there would be one signal that conveys that o_i has a greater than average height, and a signal conveying that o_i has lower than average height.

Franke (2012a) proposes a specific strategy that senders and receivers might be

implementing in communication. Namely, senders pick the signal associated with the property that is most salient among the ones in the intended object. Receivers choose the object that has the signalled property to the most salient degree. In formal terms, sender and receiver behave as follows. First, s selects the most salient property $j*$ among the properties of o_i. Franke proposes a way to measure the saliency of an object with respect to a property based on the difference between the object and all other objects in the context with respect to that property. Then, s selects the signal conveying that o_i has property $j*$ to a high or low degree, depending on the real value of o_i with respect to $j*$. After receiving the signal, r selects the object that is most salient with respect to $j*$ among the objects compatible with the signal, i.e. the object that have a high or a low degree of the property. The focus here is not on how language users arrive at this strategy, but rather on the fact that it is a simple, plausible communicative strategy that leads to on average very successful communication. I implemented the model based on the description in Franke (2012a) and Franke (2012b). Results are shown in figures 1.3 and 1.4.[24]

The model presented by Franke has a number of advantages. However, there are issues with it that lead me to ultimately reject it as an evolutionary explanation of adjectival extremeness. First of all, the space of possible strategies that the language users consider for referential communication is not defined. Rather, a specific strategy is hand-coded. While the picked strategy is a plausible one, a causal, mechanistic evolutionary model needs to simulate how the strategy is picked and spreads. Rather than showing how the strategy evolved, Franke's model shows that a salience-based strategy is a good compromise between communicative success and simplicity.

Second, there are reasons to think that the strategy described by Franke is not the one underlying the meaning of gradable adjectives. Franke's model makes the wrong prediction for minimum-standard adjectives, which have a transition at the scale's infimum. Specifically, Franke's model predicts a transition above the mean of the property's distribution (see bottom plot in figure 1.4). This contradicts the fact that e.g. "wet" does not refer to a (soft) extreme with respect to wetness (i.e. "very wet"), but rather the threshold is at an extreme of dryness. This indicates that, consistently with Potts (2008), there is something special about boundaries that is not explained in terms of communicative utility. I return to this point again

[24]Code for this and the following reimplementations is available at (Carcassi, 2020).

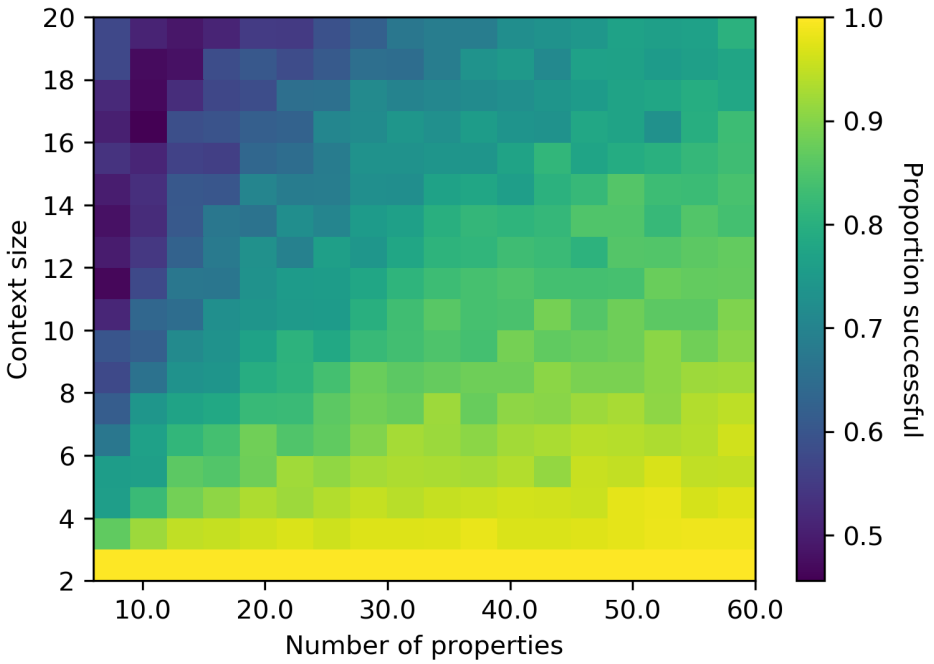


Figure 1.3: Reimplementation of the results from Franke (2012a). The plot shows the proportion of successful communications for various combinations of context size and number of properties for each object. Communication is on average more successful when there are fewer objects and more properties to choose from.

in chapter 6. Franke's model shows a related problem in closed scales. Even if the degree expressed by a signal on a closed scale can get very close to the extremes, it will never refer to a truly extreme degree in the way I defined above. This is because the signals are used for property degrees that are actually instantiated in the context, rather than having a fixed meaning. Since extreme values are never instantiated, they will never be used in Franke's model.[25] It is important to not misunderstand the problem here. The problem is not that signals in the model are never used for extreme values—since extreme values are arguably never observed in the real world, the model should also not rely on observations of extreme values.

[25]Zhao and Cremers (n.d.) argue that in some contexts there might be non-zero probability mass on a scale extreme. I return to this point below and argue against this solution.

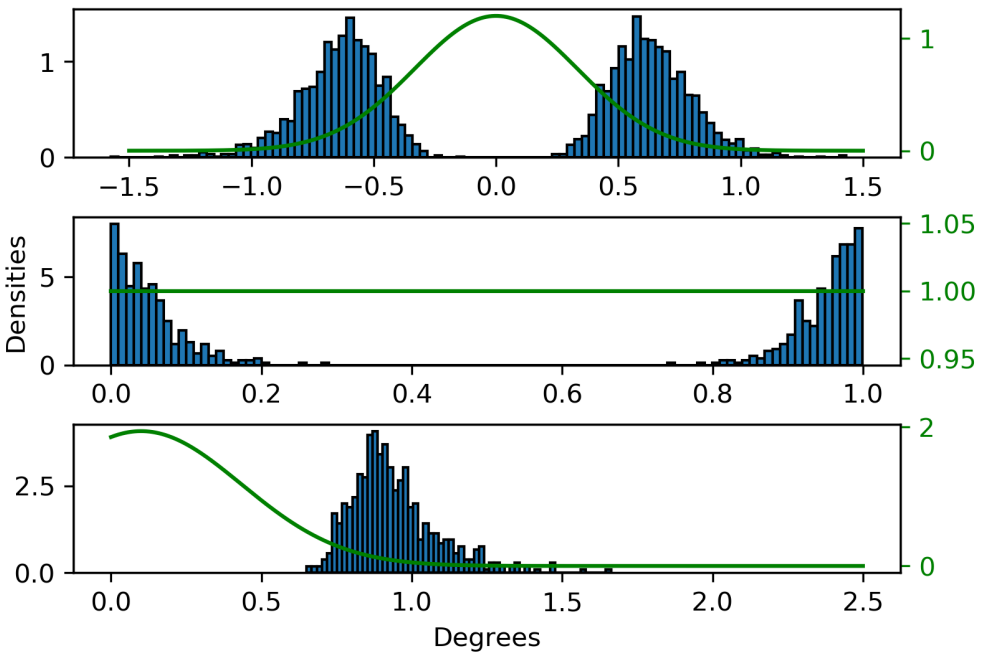


Figure 1.4: Reproduction of results from Franke (2012a). Degrees referred to by the signals on open (top), closed (middle), and half-closed (bottom) scales. The green lines show the distribution densities of the properties on the scale type.

Rather, the problem is that since extreme values are never observed in the world but some signals with extreme meaning are used for non-extreme degrees, usage and semantics have to be kept separate in an evolutionary model of extremeness.

A third problem with Franke's model is the assumption that the greatest communicative pressure for the transition's position comes from referential use. This is a problem because it is difficult to estimate the proportion of referential and descriptive communication in real language, and therefore difficult to evaluate this assumption of the model. More in general, it is unclear what referential language use for quantifiers would consist of, indicating that descriptive language use suffices for the evolution of extremeness.

Lassiter and Goodman (2013) approach the problem of the origins of extremeness

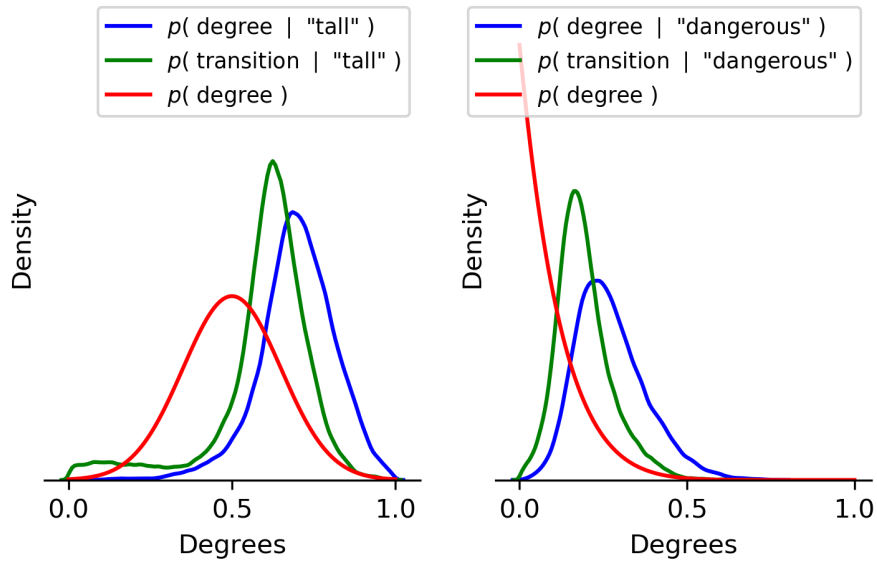


Figure 1.5: Reimplementation of results from Lassiter and Goodman (2013). The plot shows the pragmatic receiver's posterior density for (1) the threshold position and (2) the degree observed by the sender, after receiving "tall" (right plot) and "dangerous" (left plot). The model correctly predicts that "tall" conveys approximately "significantly above average height".

from a different prospective.[26] Rather than investigating the communicative success of various threshold positions in a population, they assume that all that adjectives encode lexically is polarity. The distribution of the property in the comparison population determines whether the scale counts as open, closed, or half-closed. A rational receiver, after receiving a signal such as "tall", calculates a joint distribution over (1) the degree observed by the sender and (2) the threshold's position. The distribution over thresholds and over degrees can then be obtained by marginalizing the joint posterior. The model is implemented in the Rational Speech Act (RSA) framework, which will be explained in more detail below. A reproduction of some of the results from Lassiter and Goodman (2013) is shown in figure 1.5.

[26]While Lassiter and Goodman (2015) is closely related to Lassiter and Goodman (2013), it does not focus on the distinction between absolute and relative adjectives, and therefore I do not discuss it further.

I do not discuss the model in detail in the interest of space. Qing and Franke (2014a) discuss the advantages and flaws of Lassiter and Goodman (2013)'s model in detail. I only remark briefly on features of the model that are relevant here. Lassiter and Goodman (2013) assume that the threshold's position is not encoded in the lexical entry for the term, but rather its distribution is calculated, along with the distribution over degrees, as part of pragmatic inference. The view that threshold position is not encoded lexically, while debated in the literature on gradable adjectives, is much less plausible in the context of explaining the extremeness of quantifiers. I instead will assume that extremeness is lexically encoded, making a unified explanation for the tendency towards extremeness easier.[27]

A second problem is that Lassiter and Goodman (2013)'s model is missing, like Franke (2012a)'s model, a separation between semantics and usage. I argued above that this difference is necessary to make sense of how extreme meanings can be used for non-extreme observations. The conflation is not problematic for Lassiter and Goodman (2013), because they reject the assumption that absolute gradable adjectives express extreme categories. Rather, they work under the assumption that absolute adjectives' thresholds are vague, but tend to be close to the scale's extrema. As pictured in figure 1.5, Lassiter and Goodman (2013) predicts that the thresholds for half-bounded scales are not extreme, although they tend towards extreme values of the scale.

Qing and Franke (2014a) improves in various respects on Lassiter and Goodman (2013)'s model. Qing and Franke (2014a) do not assume that the threshold's position is determined by pragmatic inference. Rather, they rely on the evolutionary perspective that language is optimized for communication. Therefore, they assume that senders employ a threshold that tends to maximise long-term communicative success. The model defines the probability density over position of the transition that the sender will use. Like in the previous models, the language users' behaviour is influenced by the distribution of the property in the comparison population. The case of properties with substantial probability mass around the extremes, which is taken to correspond to bounded (or half-bounded) scales, is particularly interesting. Here, Qing and Franke (2014a) predict that, given enough probability mass at the border, the probability density function increases monotonically as it gets closer to

[27] I do not show that extremeness is encoded lexically, but rather assume it. The point here is that an account of extremeness that does not assume it has to also provide a story for how extremeness evolves in quantifiers.

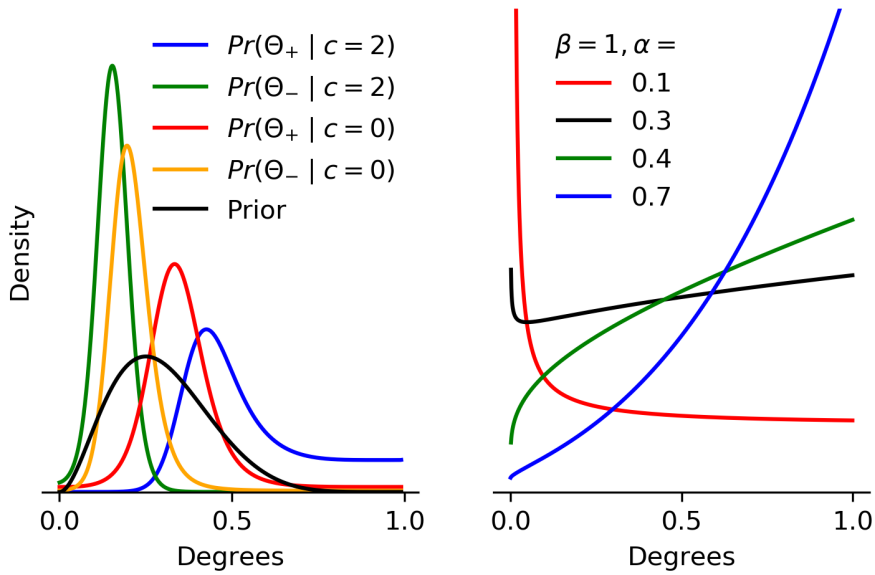


Figure 1.6: Reimplementation of results from Qing and Franke (2014). The left plot shows the probability of the speaker using thresholds Θ for positive (Θ_+) and negative (Θ_-) adjectives, for various production costs (c). The property is distributed as a beta distribution with $\alpha = 3, \beta = 7$. The right plot shows density of threshold position for various distributions of the property. When the beta distribution has more mass concentrated towards the lower extreme ($\alpha \to 0$), the probability density for the thresholds is also concentrated at the bottom.

the border (see right plot of figure 1.6).

Despite the improvement over previous models, Qing and Franke (2014a)'s model is somewhat unsatisfactory if one assumes, as I do in this thesis, that absolute adjectives express truly extreme categories. This point is argued in more detail in Zhao (2019). As I mentioned above, the proposed picture is that senders sample a transition from the distribution in usage. However, since the density puts no probability mass on the extrema of the scale, a truly extreme transition will never be used. This makes sense in the picture of Qing and Franke (2014a), because an extreme transition would lead to low expected communicative success; the resulting language would be equivalent to a language without the adjective. This result is satisfactory in two cases. First, if one works under the assumption that absolute adjectives do

not express truly extreme categories in the sense defined above. Second, despite assuming that absolute categories' transitions are extreme, one might rather be interested in the transition's position insofar it explains usage. We have seen above that since extreme values are never observed, some pragmatic adjustment have to be made for extreme categories to be used. Leffel, Xiang, and Kennedy (2017) and Aparicio, Xiang, and Kennedy (2016) elaborate on the difference between the semantic view of gradable adjectives in Qing and Franke (2014a) and Lassiter and Goodman (2013). A related indication of Qing and Franke (2014a)'s problem with absolute adjectives comes from the experimental data in Qing and Franke (2014b). Model fitting shows that the participants' categorization behaviour when using absolute adjectives is more extreme than predicted by the model. Qing and Franke (2014a) note that the probability of the transition's position is maximised at the extremes of the scale for a wide variety of property's distributions that put some mass at the extreme. However, it is unclear how this feature of the model predicts extreme, crisp transitions, rather than almost extreme, vague ones.

Zhao and Cremers (n.d.) propose a model that is designed to deal with extremeness. This model makes three innovations compared to the models presented above. First, the model is extended to account for complex expressions, i.e. adjectives modified by "very". I will not discuss this innovation of the model as it is not relevant for the present purposes. The second innovation of the paper is that the sender selects a threshold optimally for any given distribution of the property in the comparison population. However, the sender is uncertain about what this distribution is. This is modelled through a hyperprior, i.e. a prior on the parameters of the property's distribution. In sum, albeit the sender selects a threshold optimally, uncertainty about the property's distribution in the comparison population defines a distribution over optimal thresholds.

The third innovation of Zhao and Cremers (n.d.) is to use a prior over the property that is not absolutely continuous, but rather attributes non-zero probability mass to the infimum of the scale. This is modelled as a rectified Gaussian distribution, which concentrates the probability mass of negative degrees on 0. The intuition behind this choice comes from the absolute adjective "late". Intuitively, the adjective "late" is defined on the $[0, \infty]$ portion of the time domain, where 0 corresponds to the present. Since "late" is a minimum-standard adjective, any time that is non-minimal counts as late. However, any arrival before the present is counted as belonging to time 0 for the purposes of "late". The rectified Gaussian distribution models this situation

well.

The strategy of using a mixture of a continuous distribution and a constant in Zhao and Cremers (n.d.) assumes that the adjective's scale is a subscale of a larger scale, and that a portion of the latter is mapped onto a single point of the former. While not many scales can be described in this way, the strategy works for some adjectives. For instance, some processes that over time monotonically move an object in a specific direction of a closed scale can ensure the existence of extrema; since the processes that empty and dry objects can render objects completely empty and dry, there is a non-zero probability of encountering true infima of the scales of wetness and fullness. However, it is hard to see how this strategy can be applied in general. The problem is apparent with scales such that the available processes to change the degree do not easily lead to an extreme. For instance, the process of filling a container will always leave some space, e.g. the space between the molecules of the liquid, that prevents a container from ever being, in its most literal sense, full.

Beyond these three main innovations and a few technical details, Zhao and Cremers (n.d.)'s model of communication is the same as Qing and Franke (2014a). Overall, Zhao and Cremers (n.d.) improves over previous models in that it proposes a picture of how a truly extreme categories can be communicatively advantageous. However, the picture only applies to some adjectives. A full picture should be able to explain how it is possible for extreme categories to evolve, even when the scale's extreme has zero probability mass.

1.4 Conclusions and general plan

In this chapter, I discussed the linguistic data relating to monotonicity and extremeness. A discussion of recent formal accounts of the meaning of these terms showed that categories that fail to be monotonic and extreme are possible, ruling out a purely semantic explanation of the universals. I also considered previous work on the evolution of these two universals, and indicated various respects in which the previous literature only gives a partial explanation of the data. In the case of monotonicity, previous work did not develop an explicit evolutionary mechanism through which monotonicity might have evolved. In the case of extremeness, previous work has treated true extremeness unsatisfactorily, focusing rather on closeness to the border, and moreover focused solely on linguistic adaptation for communication. In this the-

sis, I analyse monotonicity and extremeness, and by taking their scalar properties seriously I give a more complete account for both phenomena in an evolutionary context. I test my ideas using computational modelling as well as experiments.

In chapter 2, I turn to conceptual spaces theory as a cognitive theory of the meaning of scalar terms. I present various problems with the standard account of categorization in conceptual spaces theory, and propose an alternative account of categorization that solves them. My proposal includes an account of how scalar meanings are encoded. While categories on highly-dimensional domains, expressed by nouns, are encoded in terms of prototypes, a prototype-based analysis is implausible for categories on one-dimensional domains, such as the ones expressed by gradable adjectives and quantifiers. Instead, one-dimensional categories are encoded in terms of transitions. The reason why this conclusion about encoding will be crucial is that it will influence the measure of complexity for scalar categories, which as discussed above is fundamental in modelling the evolution of language.

In chapter 3, I develop a model of the evolution of monotonicity for scalar categories. This model shows that neither a pressure for learnability nor a pressure for simplicity alone can explain the evolution of monotonicity. Rather, a combination of the two pressures is needed, in addition to the assumption that agents are pragmatic, rather than literal, language users. Moreover, I model without additional assumptions the evolution of other features that characterize systems of scalar categories in natural language, as presented in chapter 1.

In chapters 4 and 5, I turn to the account of learning of scalar categories that I developed in chapter 2 and that was assumed for the models in chapter 3. Specifically, I attempt to test the prediction that monotonic categories have an advantage in learning over non-monotonic categories for scalar conceptual domains, but not for non-scalar domains. I test this hypothesis in a series of six experiments. I analysed some of the data with a hierarchical Bayesian model and a cognitive Bayesian model. Much of the chapter focuses on the formulation of the learning model and its implementation as a statistical model. The data did not support the hypothesis.

In chapter 6, the focus shifts to the universal of extremeness. I present a computational model showing that learning alone can contribute to the observed preference for extreme categories. This model contributes to the literature in two respects. First, I consider the contribution that learning might make to the evolution of extremeness. Second, I distinguish clearly between the semantic and pragmatic factors at play in the use of scalar categories, which I argue was a problem in previous

computational models of scalar extremeness.

Finally, in chapter 7, written in collaboration with Shane Steinert-Threlkeld and Jakub Szymanik, I return to monotonicity in the specific case of quantifiers. I show how monotonicity can emerge in a population of feedforward neural networks, general-purpose learning algorithms without language-specific biases.

Chapter 2

Scalar meaning in conceptual spaces

In the first chapter, I focused on the universals of scalarity and monotonicity from the point of view of formal semantics. Morphologically simple terms are monotonic and when possible often extreme on the scales that are naturally associated with them. However, a purely formal, truth-conditional treatment of scalarity is not sufficient for the purposes of this thesis. The aim is to give a mechanistic account of the evolution of scalar categories, which depends on how scalar categories are acquired and encoded by the human cognitive system. Therefore, in this chapter I develop a cognitively grounded theory of scalar meanings within the framework of *conceptual spaces*.

This chapter is structured as follows. I start by introducing the framework of conceptual spaces in section 2.1, and an account of how categorization happens in conceptual spaces based on *prototype theory*. Then, Gärdenfors (2017)'s analysis of gradable adjectives and quantifiers is presented. I argue that the prototype picture does not work for scalar categories. Finally, in section 2.2 I propose a new account of scalarity in the conceptual space framework, which solves the problems of the prototype account and supports an appealing picture of categorization in scalar conceptual spaces. I conclude that the proposed extension to conceptual spaces theory is not sufficient on its own to explain the evolution of the two universals I discussed in the previous section, namely monotonicity and extremeness. However, the extension proposed in this chapter provides an important basis for the cognitive component of the model in chapter 3 and the experiments in chapters 4 and 5.

2.1 Conceptual spaces

2.1.1 Some fundamentals

Conceptual spaces theory is a theory of cognition with important applications to semantics. The theory is cognitive because it concerns the way that humans encode various features of the world, as well as providing a model of the concepts used to talk about these features. The simplest features of the world that are encoded, i.e. the fundamental dimensions with respect to which two objects can be different, are called *quality dimensions*. Quality dimensions constitute one of the fundamental building blocks of conceptual spaces theory. Examples of quality dimensions are pitch, width, length, brightness, temperature, sweetness. Quality dimensions are the fundamental features of the world that cognition keeps track of (Gärdenfors, 2011). Some quality dimensions are concrete, such as length, but others, such as boringness, are abstract. A definitive list of quality dimensions in humans cannot be given, because while many quality dimensions are plausibly innately determined others are acquired culturally (Gärdenfors, 2011).

Quality dimensions are endowed with geometric structure. To see what this means, consider as an example the domain of time (Gärdenfors, 2017, p. 22). It is the underlying structure of the time domain that allow us to make comparisons between time points. Time is isomorphic to $\mathbb{R}$, the set of real numbers: the present is the zero point, future is $\mathbb{R}^+$, the past is $\mathbb{R}^-$. Moreover, a time point can be twice as far in the future as another time point, and similarly for the past. Crucially, while these might not be features of physical time, they are an essential part of how humans think about time. The idea of conceptual spaces theory is that geometrical structure is not an exclusive feature of the temporal conceptual domain, but rather is pervasive in human thinking.

Quality dimensions can stand in different relationships to each other. A group of quality dimensions is *integral* iff whenever an object is given a value on one of them, it must be given a value on all of them. For instance, brightness and saturation are different quality dimensions, because their values are independent of each other and they each have their own geometric structure. However, they are integral, because whenever the brightness of a surface is imagined, so is its saturation. Quality dimensions that can be specified independently and are therefore not integral are called *separable*. A *domain* is a set of integral dimensions that are separable from all

other dimensions. As Gärdenfors (2017) notes, many domains, such as temperature, consist of a single quality dimension. The temperature of an object can be imagined independently of any other of its features.

I take quality dimensions as given, whether by fundamental feature of human cognitive making or by acculturation. Once quality dimensions are given, a fundamental way that humans make use of them is in categorization. Gärdenfors (2017) distinguishes between two types of categories, *properties* and *concepts*. A property is a *convex* region of a single domain. An extension is convex iff for any two points p_1 and p_2 that fall in the extension, if a point p_3 is between p_1 and p_2 then p_3 also falls in the extension (See figure 2.2). An example of a property is blue, which is a convex category in the domain of color. Blue is a convex property because any color shade between two shades that belong to the "blue" category is itself blue. I return to the problem of defining betweenness below, and rely on intuition for the moment. While properties are restricted to individual domains, concepts are convex regions on one *or more* domains. An example is the concept of banana, which includes information across many domains such as color, size, and taste. Note that the definition of concept assumes convexity, and therefore assumes that a relation of betweenness is defined for all involved domains. To get an intuition for what convexity means in the case of multidimensional concepts, consider the concept of chair. That the concept of chair is convex means that all the objects that interpolate between two chairs are also chairs. Technically, the *conceptual space* is the set of all domains. However, I will talk about conceptual spaces laxly as any set of quality dimensions and domains.

2.1.2 Convexity, prototypes, & Voronoi tessellations

Gärdenfors (2017) points out a relation between conceptual spaces and prototype theory, originated from the work on categories by Eleanor Rosch and collaborators (E. H. Rosch, 1973; E. Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; E. Rosch & Lloyd, 1978). Prototype theory starts with the observation that the classical theory of concepts, which defines them in terms of necessary and sufficient conditions, gets a crucial feature of human way of thinking about concepts wrong. Namely, the classical theory predicts that all objects that satisfy the necessary and sufficient conditions should count as equally valid instances of the concept. However, humans have a strong intuition that some individuals are better examples of a category than others.

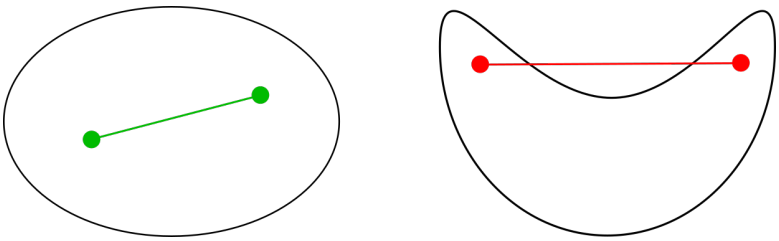


Figure 2.1: Non-convexity is illustrated by the red line: although both extremes belong to the category, some of the points in between fall outside it. For the convex category on the right hand side, it is not possible to construct such a line. Instead, all lines behave like the green line: if the endpoints are within the category, all points will be too.

For instance, robins are better examples of the category of bird than penguins. In some intuitively geometrical sense, robins can be thought of as occupying a more central position in the category of birds than penguins. The most typical instance of a category, which occupies the most central position in the category, is called the category's *prototype*. In prototype theory, as objects get more distant from the prototype, they become worse examples of the category. Eventually, they stop being part of the category and fall in another category.

Gärdenfors (2017) argues that there is a natural way how convexity, a fundamental property of categories in geometrical spaces theory which I introduced above, might emerge from encoding categories in terms of prototypes. Suppose that we start with a set of points $P = \{p_1, \ldots, p_n\}$ on the conceptual space. Then, we associate every point p_i in P with the set of points in the conceptual space that is closer to p_i than any of the other points in P. In this way, we define a set of categories, one for each point in P.[1] More specifically, this algorithm induces a *Voronoi tessellation* on the conceptual space (Gärdenfors, 2004, chap. 3). Figure 2.2 shows an example of a Voronoi tessellation with six prototypes. The points in P are the most central points for their respective categories, and are therefore the categories' prototypes.

An important feature of Voronoi tessellations is that, except for a set of points of measure zero at the borders between categories, they contain no *gaps*, i.e. points

[1]There is a set of measure zero of points equidistant from two prototypes, under the assumption that the conceptual space is dense. This complication will not be problematic for reasons related to vagueness, which I discuss below.

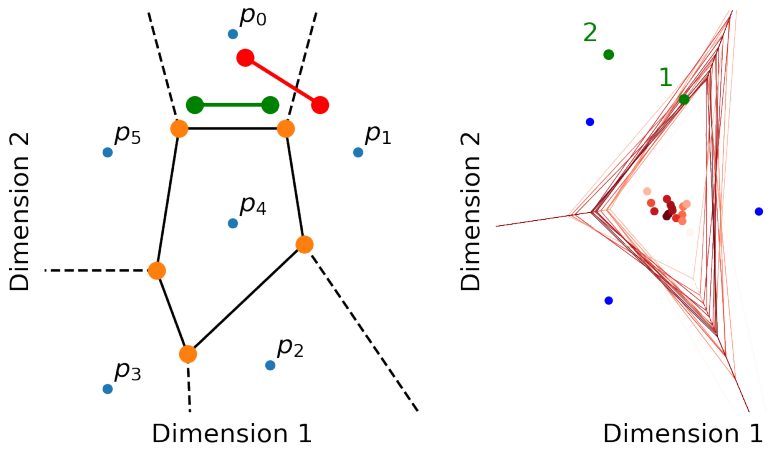


Figure 2.2: The left plot displays a Voronoi tessellation with six prototypes (points p_{0-6}). The lines indicate the borders between categories. Lines that continue indefinitely are dashed. Voronoi tessellations produce convex categories. The green line shows convexity—points between points that belong to a category also belong to the category. In order for the line to contain points that belong to two different categories, the line's extremes have to belong to different categories (red line). The right plot is a visualization of the precisification approach to account for vagueness in Voronoi tessellations. The external prototypes (three blue dots) are fixed, while different possible values of the central prototypes, and the relative tessellations, are shown. The color of the precisified central prototypes (on a red color palette) shows the closeness to the mean prototype. The green points indicate different levels of vagueness. While point 1 belongs to different categories in different precisifications and its category membership is vague, point 2 belongs to the same category in all precisifications and therefore it is clear to which category it belongs.

of the space that do not belong to any category, and no *gluts*, i.e. points of the space that belong to more than one category.[2] There are no gaps because each point is closest to at least one category, and there are no gluts because each point is closest to at most one category. This characteristic of prototype-based categorization will be crucial in showing that the classes of categories discussed in this thesis cannot be based on prototypes.

[2]I borrow the terms gap and glut from many-valued logic, where they are used in a different sense.

While convexity only requires that betweenness is defined on the space, the construction of Voronoi tessellations from prototypes requires a notion of distance. Therefore, something more can be said about the geometrical structure of those conceptual spaces where categories are defined in terms of prototypes. Distance is formalized in terms of a *metric* function. Given a set X, a metric function $d : X \times X \to [0, \infty)$ on X is a function that satisfies the following axioms for all $x, y, z \in X$:

$$d(x, y) \geq 0 \qquad \text{Non-negativity} \tag{2.1}$$

$$d(x, y) = 0 \iff x = y \qquad \text{Identity of indiscernibles} \tag{2.2}$$

$$d(x, y) = d(y, x) \qquad \text{Symmetry} \tag{2.3}$$

$$d(x, y) \leq d(x, z) + d(z, y) \qquad \text{Triangle inequality} \tag{2.4}$$

Since the metric function is meant to model the intuitive notion of distance, it is worth checking that the axioms correspond to intuitive features of distances. Non-negativity says that the distance between two points cannot be negative. The identity of indiscernibles says that points at distance 0 are identical. Symmetry says that distance between two points in one direction is the same as in the other direction. Finally, the triangle inequality says that the distance between two points is smaller than or equal to the sum of the distances of the two points to a third point. While a weaker notion of distance than a metric might suffice for the present aims, Gärdenfors (2017) assumes that the conceptual spaces underlying nouns have a metric structure.

Connecting conceptual spaces theory and prototype theory introduces some further complications. An important distinction, introduced in Kamp (1995) and elaborated e.g. in Hampton (2007) and Douven (2016), is that between *degree of membership* M and *typicality* T. M measures how much an individual belongs to a category, under the assumption argued for by prototype theory that category membership is graded. Individuals that are closer to the prototypes belong more to the prototype's category than ones near the border to other categories. On the other hand, T measures how typical an individual is for a category. While it might be appealing to identify them, M and T can come apart in important cases. For instance, consider the concept of penguins. Penguins are atypical instances of the category of birds, and therefore have a low value for T. However, they clearly belong to the category and therefore have a high value for M. Conceptual spaces theory can account for cases

where T and M come apart as follows. The further a point is from the prototype of a category, the less typical a member of that category it is. For instance, penguins are very far from the prototype of birds. On the other hand, as shown above the membership in a category is also a function of the surrounding categories and the resulting Voronoi tessellation. Therefore, an object can be a very untypical member of the category, while still belonging to the category, as long as there are no other categories close enough to "claim" the object.

The gradedness of M is needed to account for the fact that categories are vague. The simple picture of Voronoi tessellations presented above cannot explain vagueness. An account of vagueness in conceptual spaces is developed in Decock and Douven (2014), which take the supervaluation approach already mentioned in the discussion of scalar categories above. The rough idea is that each category is not associated with a single prototype, but rather with multiple (and possibly infinitely many) prototypes. The Voronoi tessellation obtained by choosing one of the prototypes of each category is called a *precisification*. The complex tessellation that comprehends all the possible precisification is called the *collated* Voronoi tessellation. The degree of membership in a category is a function of the proportion of precisifications in which the point falls in the category. If the point is close to the prototypes associated with the category, it will have a high degree of membership to the category. This is displayed in figure 2.2. This description does not answer the question of how the set of prototypes for each category is determined. A promising option, explored in Verheyen and Égré (2018), is that each category is associated with a probability density that describes the distribution of its prototype. While I do not discuss further how to account for vagueness in a picture of categorization based on prototypes, I will return to the idea of precisification below when discussing the vagueness of scalar categories.

Conceptual spaces theory as presented above is not without critics in the literature. Hernández-Conde (2017) argues that concepts expressed in natural language are in fact not convex (but see Gärdenfors (2019) for a response). Gauker (2007) develops a more general criticisms against similarity-based theories of concepts. I do not discuss these criticisms any further, but rather refer to the literature.

It is important to note that prototype-based categorization trades off the high simplicity of coding categories through prototypes and the requirement it puts on the structure of the involved conceptual domains, i.e. metric structure. For comparison, consider the alternative strategy of simply storing the individual points that belong

to each category. This simple listing strategy does not put any requirement on the structure of the domain. However, it has disadvantages. Learning could not generalize beyond the observed instances of the category, and it would put absurd requirements on memory. On the other hand, prototype-based categorization, while constrained to spaces with a metric structure, only requires learners to identify and encode as many points as the number of categories. The fact that categorization exploits the available structure to gain in efficiency is a point to which I return again below, where I argue that conceptual spaces with a different structure afford other categorization strategies.

2.1.3 Scalar language in conceptual spaces

According to Gärdenfors (2017), conceptual spaces theory can provide a cognitive basis for distinguishing between classes of words. I sketch next the conceptual spaces account of gradable adjectives and quantifiers presented in Gärdenfors (2017). Then, I discuss a previous attempt to apply conceptual spaces to scalar language.

Gärdenfors (2017)'s account of adjectives, previously introduced in Gärdenfors (2014), is based on the notion of property introduced above. Gärdenfors proposes the *single-domain thesis for adjectives*, which says that adjectives express properties, i.e. convex regions on single domains, which themselves consist of one or more integral quality dimensions. Gärdenfors expands his theory of adjectives in various directions. Of particular interest for the present work is Gärdenfors's conclusion that the geometric and topological structure of the domain underlying an adjective has grammatical consequences. This conclusion parallels Kennedy and McNally (2005)'s conclusion about the effect of boundedness on extremeness presented in the previous section. Gärdenfors's own account is missing an account of how adjectives in general are encoded, although the default seems to be prototype theory. I return to this point in the next section and sketch a way to expand Gärdenfors's account.

Gärdenfors (2017)'s analysis of quantification is based on the one in Langacker (2003). Both authors distinguish between two types of quantifiers: *proportional* quantifiers, which I have discussed above, and *representative instance* quantifiers. According to Langacker, proportional quantifiers make reference to a *maximal extension* E of elements of the same type, i.e. the extension that includes all individuals of that type. For instance, the maximal extension for the type of cats would consist of all cats. E corresponds to the restrictor set. Langacker and Gärdenfors (2017)

analyse proportional quantifiers as referring to a proportion of E. Up to this point, Langacker's analysis of quantifiers strongly resembles the scalar analysis proposed above. Langacker however goes further by giving an cognitive embodied analysis of how the comparison between the sets happens. The proposal is that proportional quantifiers are based on the following everyday, bodily process:

- Looking for objects of a particular kind (the elements of the restrictor set, e.g. cats).
- Putting the objects together (defining the restrictor set E, e.g. the set of cats).
- Perceiving an object as having a bounded spatial extension.
- Laying one object on top of another for purposes of comparison or measurement (e.g. comparing the set of cats and the set of sleeping things).

Langacker claims that this forms the basis for the mental process involved in representing proportional quantification. However, Langacker does not explicitly acknowledge the important fact that the overlap of the two sets has to consists of the very same objects. For instance, to verify "Every cat smokes" it is not enough that the set of cats is smaller than the set of smokers, so that an overlap is possible. Rather, the very same objects that are cats have to also smoke.

Gärdenfors (2017) does not explicitly endorse the embodied aspect of Langacker's analysis of proportional quantifiers. However, he analyses proportional quantifiers with the related concept of *multiplex-mass interchange*. The concept is illustrated with the following example:

(2.5) There were soldiers posted all over the hill.

Soldiers are treated here not as a set of individual elements, but rather as a mass, in analogy with e.g. "There was wine all over the tablecloth". Gärdenfors (2017, p. 13) describes the internal process of considering a set of elements as a mass as "squinting with your inner eye". The restrictor and scope sets are compared as two masses. In sum, quantifiers in Gärdenfors (2017) are analysed, similarly to Langacker, as expressing a proportion between two masses. Neither Langacker nor Gärdenfors (2017) discuss numerical quantifiers explicitly, but it is reasonable to assume that they would be analysed in a similar way as expressing the extent of the overlap between two masses.

The representative instance quantifiers are "every", "each", and "any". Gärdenfors (2017, p. 238) analyses their meaning as encoding verification strategies:

- Any: verify by constructing a generic individual in the restrictor set and see that it belongs to the scope set.
- Each: verify by surveying the individuals in the restrictor set sequentially and show for each of them that they belong to the scope set.
- Every: verify by considering the individuals in the restrictor set all at once and see that they all belong to the scope set.

While these quantifiers do encode meanings that are monotonic on the scale of magnitude, is it not clear what the conceptual domain underlying them is, and whether it is scalar. As the topic is quite complex but not directly relevant to the present aims, I do not discuss representative instance quantifiers further.

I have briefly reviewed what Gärdenfors (2017) says about quantificational and adjectival semantics. Crucially, an account of categorization in scalar language is missing. While such an account is developed for nouns, Gärdenfors does not discuss how categories are established for gradable adjectives and quantifiers. Decock, Dietz, and Douven (2013) partially fill this gap by developing a conceptual spaces account of comparative expressions such as "taller than" that goes beyond Gärdenfors (2017). The main idea in the paper is to analyse comparative concepts in terms of the independently defined membership function M_C for any category C:

(2.6) **CC**: For all individuals i and i' and all comparative concepts "C-er than" and corresponding categorical concepts C, i is C-er than i' iff $M_C(i) > M_C(i')$.

Decock et al. (2013) do not give an independent analysis of the root of the comparative, but rather assume the definition of M based on the Voronoi tessellation picture. This picture is unsatisfying for various reasons. First of all, as the authors themselves point out, the account is "limited to comparative concepts whose associated categorical concept has one or more prototypes". Secondly, they notice that "we are unable to think of any concept lacking a prototype that is not representable in a one-dimensional space". As I will discuss in more detail below, multidimensional adjectives provide examples of such concepts. Lastly, the authors propose, without developing the view further, to analyse scalar concepts in terms of the structure of the underlying domain: "age, height, price, all being measured on an interval scale,

the semantics of older than, taller than, and cheaper than can simply be stated in terms of the $<$ relation." However, once such an analysis is developed, as will be done below, the need for an analysis of comparative constructions in terms of prototypes disappears. Crucially, such 1-dimensional conceptual domains constitute the most frequent and important cases of comparison.

In the following, I first argue that the picture of prototypes and the resulting Voronoi tessellations that works for nouns will not work for scalar language. Then, I sketch an extension of conceptual spaces theory that can account for categorization in scalar domains.

2.1.4 The prototype picture won't do for scalar terms

We saw in the previous section that Gärdenfors sees gradable adjectives as extensions on single domains, and proportional quantifiers as extensions on the proportion scale. In this section, I argue that the prototype picture of categorization is inconsistent with categories on scalar spaces by presenting a list of problems. I start with the following problem:

Problem 1. *In a Voronoi tessellation, every point belongs to at least one category. However, systems of gradable adjectives do not cover the whole semantic domain.*

In a system of gradable adjectives, the zone between antonyms that belongs to neither antonym is called the *zone of indifference* after Sapir (1944). For instance, average height is not covered by either of the two antonyms "tall" and "short", nor by any of the other height adjectives such as "towering". A circumlocution such as "of average height" can refer to the zone of indifference, but is not lexically simple. The expression "average" alone, with enough help from contextual cues, can also be used to refer to the zone of indifference. However, the reference to a specific conceptual domain is again not lexically encoded.

A first possible counterargument to problem 1 is that systems of nouns also fail to cover some parts of the conceptual space, and in this respect cannot be modelled by Voronoi tessellations. For instance, consider the systems of nouns used to describe animals. There is no name for an animal that looks like a dog but with a giraffe-like neck. According to this counterargument, 1 does not show that the mechanism of nominal categorization is different from the one for scalar categorization, because it applies to both nouns and adjectives. The success of this counterargument depends

on how it modifies the Voronoi tessellation approach above to account for why some parts of the conceptual space are not covered. There is a fundamental reason to be sceptical that such a modification would succeed. Namely, systems of gradable adjectives neglect to cover the conceptual space in a much more radical way than nouns. While the dog-giraffe chimera described above is never observed, persons of average height—and in general the individuals in the zone of indifference—are the most commonly observed. The zone of indifference does not cover marginal or unusual parts of the conceptual space, but rather the most common. A defence of the uniformity of scalar and nominal language with respect to categorization strategy would have to account for this remarkable difference in how 1 affects adjectival and nominal conceptual domains.

A second counterargument to problem 1 is that the zone of vagueness should be interpreted as a vague threshold, where elements do not belong clearly to either category, rather than a part of the domain that is not covered by any adjective. This counterargument is contradicted by two observations. First, metalinguistic intuition tells us that objects in the zone of indifference clearly do not belong to either category, rather than unclearly belonging to one of them. This explains how "The average person is tall" and "The average person is short" can be both false. Second, variations in the comparison population are hard to explain in a prototype analysis. Consider two comparison populations with different height variances. The standard for "short" and the one for "tall" will get further apart from each other as the variance increases. In a prototype analysis, this has to be understood as an increase in the size of the zone between categories affected by vagueness. It is unclear how the population variance should affect the amount of vagueness in the category.

Problem 1 does not apply to quantifiers, where the whole scale is covered. A second problem however applies to both gradable adjectives and quantifiers:

Problem 2. *In a Voronoi tessellation, no non-zero-measure region of the domain belongs to more than one category. However, there are non-zero-measure overlaps between categories in systems of gradable adjectives and quantifiers.*

No region belongs to more than one category because there is no region of non-zero measure whose points can all be closest to two prototypes. An obvious objection to problem 2 is that there are overlaps between nouns, e.g. the extension of "cat" overlaps with the extension of "animal". Therefore, according to the objection nouns might very well be categorized with the same strategy as adjectives—and neither

would be a Voronoi tessellation. Problem 2 has to be made more precise to counter the objection. The more precise claim is that no overlap should exist between categories at the same level of abstraction (E. Rosch et al., 1976). For instance, the extension of nouns referring to specific animal species or geometric shapes tend to not overlap. The tendency to assume in learning that different nouns refer to different objection is known in psychology under different names: *synonymy avoidance* (Carstairs-McCarthy, 2010), *principle of contrast* (Clark, 1993), and *mutual exclusivity principle* (Markman, 1989).

As opposed to nouns, systems of adjectives show a systematic tendency to overlap. It is useful to distinguish two types of overlap in system of adjectives. The first type is an overlap between a relative and an extreme adjective. Extreme adjectives (not to be confused with adjectives expressing extreme categories as defines above) are adjectives like "gigantic", "fantastic", "gorgeous", which accept modifiers such as "flat-out", "downright", and "positively" (Morzycki, 2009). This type of overlap exists for instance between "big" (relative adjective) and "huge" (extreme adjective). It could be argued that extreme adjectives do not occur often enough to influence the extension of the main couple of antonyms. This criticism is contradicted by the high frequency of extreme adjectives in certain adjectival systems ("freezing"/"cold"/"cool" and "warm"/"hot"/"boiling")[3] and by the fact that in noun systems the extension of frequent nouns can be restricted by rare nouns (e.g. "walk", "trot", "canter" and "gallop", the natural gaits of horses, have widely different frequencies).

The second type of adjective concerns two relative adjectives, e.g. "warm" and "hot". The second type of overlap provides more convincing evidence that adjectives have a systematic tendency to overlap, since extreme adjectives are arguably more specialized words than relative adjectives. Quantifiers also show systematic overlap, e.g. the extension of "all" is contained within the extension of "some".

The third problem I consider has to do with the categories that can be encoded with Voronoi tessellations:

Problem 3. *Extreme categories cannot be defined in a Voronoi tessellation.*

In order to define an extreme category, the border between two categories has to fall at the infimum or the supremum of the scale, and moreover the extreme category

[3]Kranich (2016) found that a surprising proportion (close to a third) of all the instances of positive evaluative adjectives in an English corpus of letters to shareholders were extreme adjectives.

has to be defined as including the border. Consider the case where the border is at the supremum. For the border to fall at the supremum of the scale, both prototypes have to be at the supremum of the scale. However, if the two prototypes coincide, prototype theory would predict that they are in fact the same category. Since couples of adjectives such as "dry" and "wet" refer to different categories, they cannot be encoded in terms of Voronoi tessellations. The same argument applies to "some" and "none" in the semantic domain of quantifiers.

Finally, I discuss a fourth problem with modelling scalar categories in terms of prototypes:

Problem 4. *While prototypes are usually statistically frequent, in order to get the observed adjectival categories prototypes would have to be at points that are rarely observed.*

Prototypes tend to be located in the most frequently observed part of the conceptual space (Ibbotson & Tomasello, 2009). Contrary to this observation, the location of the prototypes that are required to define a category like "tall", i.e. a fairly extreme position on the height scale, are rarely observed.

There are further problems with applying the prototype pictures on scalar categorization. While points further away from the prototypes in nominal categories are judged as being less typical examples of the category, more extreme points in scalar domains are not. For example, increasingly tall persons do not become less typical instances of the "tall" category. Another problem is the metalinguistic intuition that many scalar categories do not have typical instances. For instance, there is no typical instance of "tasty". I do not discuss these further problems in the interest of space.

As we have seen in this section, categorization based on prototypes and resulting Voronoi tessellations will not work for scalar language. Whether a pictures based on prototypes is the correct one for nouns or not, I argued in this section that prototypes offer a better model of nominal categorization than they do for scalar categorization. In the following section, I propose an alternative account of scalar categorization that improves on the picture based on prototypes.

2.2 An expansion of the conceptual spaces account

2.2.1 Coding of scalar categories

I start with the observation that some conceptual domains are structured by orders. This amounts to saying that the relation between points in some conceptual domain is encoded by a relation that satisfies the order axioms as defined above. Often, a single domain is ordered both by a metric and by an order structure. As an example, consider the conceptual domain of size. In order to fully encode what is expressed by specifying a size, the difference in size between any two objects, expressed by a metric, is not enough. This is because a metric is symmetric, i.e. it does not keep track of the direction of comparison. Crucially, when two objects differ in size, one has a smaller size than the other.

Various clarifications are in order. First of all, orders and metrics are compatible. Some conceptual domains are encoded with just a metric (e.g. colors), some with just an order (arguably, tastiness), and some with both an order and a metric (height).[4] To determine the amount of structure that is encoded in a domain, one can check which relations can be asserted to hold between points of the domain. If the domain encodes a metric, it is possible to assert that one point a in the domain is further away from a point b than from a point c. For instance, blue is further away from brown than green is from brown. If the domain encodes an order, it is possible to assert that one point in the domain is greater than another point. For instance, it can be asserted that the temperature of a cup of tea is greater than the temperature of a glass of water. As another example, "height" refers to a certain property, which is not directly tied to either of the adjectives constructed on the height scale, "short" and "tall". Importantly, heights themselves can be compared in terms of an order, e.g. in the expression "greater height". This shows that the domain of heights, rather than some additional structure implied by the adjectives, is thought of as ordered.

In fact, the situation for ordered conceptual domains is subtler than the picture presented in the last paragraph. Some conceptual domains have a natural direction of increase. For instance, height increases in the direction of tallness, temperature increases in the direction of hotness, and luminosity increases in the direction of brightness. This is manifested in the grammar through the distinction between pos-

[4]Other still, such as nationality, arguably lack both an order and a metric structure. A classic classification of possible structures is Stevens (1946).

itive and negative adjectives. "Tall", "warm", and "bright" are positive adjectives, while "short", "cold", and "dark" are negative adjectives. Polarity in adjectives manifests itself e.g. in the bias of questions:

(2.7) How tall is Sam? (no implicature that Sam is tall)

(2.8) How short is Sam? (implicature that Sam is short)

The fact that a domain has a natural direction of increase is sufficient for scalarity, as the notion of increase is only well-defined if there is an order with respect to which the increase happens. However, a natural direction of increase is not a necessary condition for a conceptual domain to count as ordered, which is the feature that is relevant for the present purposes. Given an order $\leq$, is it always possible to define a dual order $\leq^{OP}$ such that $x \leq y \iff y \leq^{OP} x$. For instance, "shorter" defines the dual order of "taller" because "A is shorter than B" $\iff$ "B is taller than A". It is in principle possible that neither of the dual orders defined on an ordered conceptual domain is more natural than the other. What makes the domain ordered is then the fact of encoding the information that determines the two orders. In sum, the claim that a conceptual space is ordered is different from the claim that there is an intrinsic or natural direction of increase.

It is possible in many cases to construct an ordering $\leq_c$ from a metric d and a point c, where $a \leq_c b \iff d(c,a) \leq d(c,b)$. The most familiar case is when c is a prototype. For instance, consider the sentence "This cat is doggier than this one". The adjective "doggy" requires the construction of an order induced by the distance to the prototypical dog. The fact that the individual dimensions of comparison remain available is shown by expression such as "This cat is doggier than this one with respect to its personality".

In sum, I propose that some conceptual domains are structured by orders rather than metrics—i.e. they are scalar—and that they should be a sui generis type of conceptual domain from the point of view of conceptual spaces theory. In the rest of the section, I show how the picture of categorization based on Voronoi tessellations presented above can be modified to make sense of how categorization happens on scalar domains. Moreover, I discuss how the new picture can account for data gathered in previous attempts to explain scalar categories with prototype theory. Finally, I discuss multidimensional scalar domains and the role of monotonicity and extremeness in the picture.

2.2.2 Categorization in scalar domains

Much like nouns exploit the given metric structure to enable complex categories to be encoded with just a handful of points, I propose that categorization in scalar domains exploits the available order structure. The picture is as follows. Each category in an ordered domain is associated with points that play the role of transitions from the category applying to not applying or vice versa, in the sense defined above in section 1.1. Additionally, the type of the first transition is encoded, in the sense of "type" defined on page 21. As shown in lemma 4 on page 22, the type of the first transition is enough to determine the type of all other transitions. The transition coding for categories exploits the order structure because category membership for any point is fully determined by the transition immediately below it. In sum, encoding one category requires one to encode one point for each transition, plus specifying the type of the first transition, as was shown in lemma 4.

It worth remarking on three features of transition-based categorization. First, category membership is independent of the distance of the transition from the point, which implies that while the conceptual domain might have a metric structure, the metric is not exploited when defining categories in terms of transitions. Second, the convexity generalization that Gärdenfors (2017) formulates is satisfied by scalar monotonic categories, given a natural extension of the notion of betweenness to orders. Specifically, given an order point b is between a and c iff $a \leq b \leq c$ or $c \leq b \leq a$. Third, each category defined in terms of transitions varies independently of the other categories defined on the same space. This is different from prototype-based categorization, where the extension of a category depends on the location of the surround prototypes.

The transition-based categorization strategy for scalar terms proposed in this section avoids the problems with prototype categorization of adjectives and quantifiers presented in the previous section on page 89. I consider the problems in turn. Problem 1 is easily avoided by the fact that the transition for the negative antonym lies below the transition for the positive antonym on the scale. Problem 2 is explained by the fact that any two monotonic categories of the same type overlap, as shown above in lemma 5. Problem 3 is avoided by the possibility of extreme categories in the sense of definition 8. Finally, 4 is a more interesting problem to which I return again below when discussing how the position of adjectival transitions are determined.

Since the problems of the last section are avoided, I propose that the word classes

I discussed above, namely gradable adjectives and quantifiers, express categories on conceptual domains structured by an order, and moreover these categories are encoded in terms of transitions. Analysing these classes of words in terms of transitions has advantages beyond avoiding the problems above. First of all, it tracks the metalinguistic intuition that a natural analysis of the meaning of gradable adjectives is in terms of comparison with a threshold. Second, once a prototype-based picture is rejected, a transition analysis gives a unified cognitive account of multiple classes of words. Within the quantificational domain, the transition account illuminates the connection between quantifiers and quantity words such as "many" and "few", as illustrated by the previously developed semantic theories presented in the previous chapter. Lastly, as I argue below, the encoding in terms of transitions plays a crucial role in explaining the evolutionary origins of the universals of monotonicity and extremeness.

It is worth mentioning a prima facie plausible hypothesis that ties the conceptual domain and the type of categories built on it:

(2.9) Hypothesis: Every category on a one-dimensional conceptual domain that is structured by (at least) an order is encoded in terms of transitions.

This in effect amounts to saying that there is a preference for exploiting orders over other structure in one-dimensional domains. If this hypothesis turns out to be correct, it raises the empirical question of why there is a preference for exploiting orders in one-dimensional domains. Note that this hypothesis leaves open the way that categories are encoded for multidimensional domains. I do not argue for this hypothesis in the present work, but leave it as a plausible generalization that can explored further in future work.

It is worth at this point to consider how a transition-based theory of categorization can account for context sensitivity. The relevant phenomenon, already discussed above in the context of formal semantics, is that the same gradable adjective can convey different information depending on the comparison class. For instance, "Jenny is tall" conveys different information about Jenny's height if the topic of discussion is basketball players or 18th century Sicilians. If Jenny was 170 centimetres, the sentence would be true in the latter but not in the former case. This dependency on the context implies that associating a stable transition with each gradable adjective is a partial characterization of adjectival meaning. The challenge therefore is to give a geometrical account of how the meaning is influenced by context.

Gärdenfors (2017, chap. 13) presents a simple geometric picture of context sensitivity based on the concept of *radial projection*. In geometric terms, consider two convex regions A and B that share at least one point. A radial projection is a function from points of A to points of B. Pick (1) a specific shared point o, called the *origo*, (2) a half-line starting from o and intersecting A's boundary at z_A and B's boundary at z_B, and (3) any point x such that $\frac{d(o,x)}{d(x,z_A)} = p$. Then, the radial projection maps x to the point $y \in B$ on the half-line such that $\frac{d(o,y)}{d(y,z_B)} = p$ (where d is the metric function for the domain). Radial projection is pictured in figure 2.3. Gärdenfors (2017) proposes that radial projections account for the sensitivity of the adjective to the modified noun:

> If the region of space representing the head contains a point shared with the space representing the modifier, this point can be taken as the origo of a transformation of the modifier space.

To see how this works, consider the expression "white wine" as an example. First of all, note that white wine is not white, indicating that "white" is sensitive to the meaning of "wine". To compose the two meanings, create a radial projection of the space of colors onto the space of wines. The portion of the projected color space that usually corresponds to "white" is then what is expressed by "white wine".

The approach of Gärdenfors (2017) has various problems. First, radial projections require the region associated with the modifier to be *compact*:

(2.10) (Heine-Borel theorem) For any subset A of Euclidean space $\mathbb{R}^n$, A is compact iff it is closed and bounded.

However, as it should be clear by now, categories expressed by gradable adjectives are not compact when they lie on open domains. For instance, the range of heights covered by "tall" has to upper bound, and is therefore not compact. Despite the systematic lack of compactness, the domain of gradable adjectives can be applied to nouns in a context-sensitive way. More intuitively, it is not clear how one would map the domain of size onto the domain of dogs, since the former does not have an upper or lower bound. A second problem is that the combination between noun and modifier often does not fully specify the comparison population, showing that context sensitivity goes deeper than Gärdenfors (2017)'s account can reach. As an example, consider the phrase "An expensive mistake". If the topic is shopping, an expensive mistake might be a few hundred pounds. On the other hand, if the topic

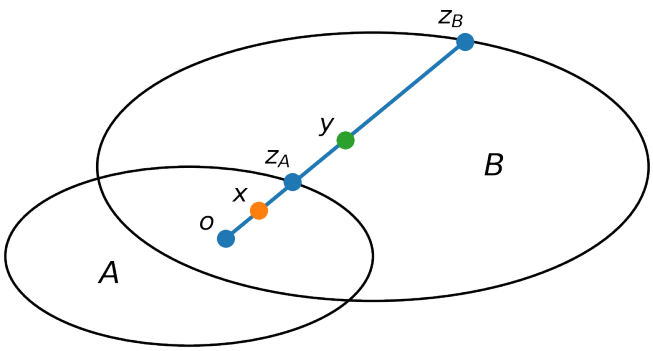


Figure 2.3: The figure shows how a radial projection maps a point x in a set A to a point y in a set B, when o is the origo and the line intercepts the border of A at z_A and the border of B at z_B.

is war an expensive mistake might cost the life of thousands of people. In sum, the head noun is often insufficient to determine the comparison population; the full context is needed. A third problem is that Gärdenfors (2017) does not explain how the origo is picked among the points shared by the modifier space and the noun space. The position of the origo can make big differences in the way the modifier's space is mapped onto the head space.

I do not make a claim about a general mechanism for context sensitivity or semantic composition. Rather, I assume that the process of fixing the position of transitions depends on mechanisms of cultural evolution, to which I return below. Still, I do assume that the transitions' positions can depend on various properties of a contextually given comparison population. For instance, it can depend on the topological structure of the domain, as seen above in the contrast between absolute and relative gradable adjective, or it can depend on the distribution of the property in the comparison population, as seen in the case of relative adjectives. The picture will ultimately have to be more complex than the one presented in Gärdenfors (2017).

I have discussed how the context sensitivity of gradable adjectives can be ac-

counted for. Next, I briefly discuss the work of Ashby and colleagues, who developed an account of categorization based not on prototypes, but rather on the decision boundaries that separate categories (Ashby & Gott, 1988; Ashby & Maddox, 1990, 1993; Maddox & Ashby, 1993). I do not review the empirical evidence in favour of this approach, because its relation to the theory presented in this chapter is generally unclear and a discussion of their relations would be beyond of the scope of the present work. In Ashby's account, the boundaries are not encoded in terms of an underlying order, but rather various possible boundary construction algorithms are considered. The crucial difference between my and Ashby's account is that my account is not meant as a general substitute for the prototype based picture of categorization, but rather as a proposal for domains structured by an order, which are in general one-dimensional (I discuss below the multidimensional case). An advantage of my approach is that it unifies cases where a transition can better account for the data and cases where prototypes do under the single theoretical framework of conceptual spaces.

Kalish and Kruschke (1997) apply the work of Ashby on decision boundaries to single-dimensional conceptual domains. This is particularly relevant because, as I discussed above, there is a special connection between orderedness and one-dimensionality. Kalish and Kruschke (1997) compare exemplar theory and decision boundary theory with respect to how well they can explain data from a series of categorization experiments. Participants were trained on segments of various lenghts, categorized in two overlapping categories. The hypothesis is that if participants are encoding the categories by encoding a decision boundary, the slope of the category membership function should be very steep where the categories change. The main relevant result was that the data is incompatible with a deterministic response model, where participants always respond with the category that is most probable given the training data. Rather, responses in the zone of overlap between categories are probabilistic. The main question this raises is how to make sense of the probabilistic behaviour of participants when the thresholds of two categories are such that the categories overlap. Although I will return to this challenge at various points below, I do not discuss the point extensively and leave it to future work.

In sum, I propose the following picture. Depending on the way a conceptual domain is structured, different way of creating categories are available. Conceptual domains can be structured at least by orders and metrics (or both). Whenever a metric is defined, a prototype-based strategy is available. If an order is defined,

a transition-based categorization strategy can be used. Is is possible that other categorization strategies are available, depending on other conceptual structures. Moreover, I hypothesised a connection between the preferred categorization strategy and dimensionality: for one-dimensional spaces, transition-based categorization is used. This simple picture can be applied to the semantics of gradable adjectives and quantifiers, and it avoids the problem of a prototype analysis presented in the previous section. While the picture of transition-based categorization presented above is motivated both by avoiding problems of the alternative picture and by independent advantages, it leaves much open about categorization in ordered conceptual domains. For instance, it does by itself not explain how the number of transitions, as well as their positions, is decided based on the data observed by the language learner. Therefore, it leaves open what lies behind monotonicity and extremeness. However, the picture provides a crucial component of such an explanation, on which I build in the chapters that follow.

2.2.3 Verheyen & Égré (2018)

Verheyen and Égré (2018) attempt to account for the categories expressed by gradable adjectives in a conceptual spaces framework. The authors ask participants to produce various judgments on the typicality and category membership of degrees in different gradable adjectives. Then, the authors compare different possible ways of weighting the prototypes for a single category in the collated Voronoi tessellation approach which I discussed above. The main aim of the paper is to explain how gradable adjectives can be interpreted in terms of prototype-based categorization. After introducing an extension of conceptual spaces theory meant for scalar categories, an alternative account of the data in Verheyen and Égré (2018) can be developed. I first present the study and then sketch an alternative, scalar account of the data.

The compared couples of adjectives are: (1) short/tall, (2) light/heavy, (3) young/old, (4) low/high, (5) cold/warm, (6) slow/fast, (7) cheap/expensive, (8) thin/thick. For each of these adjectives, participants were asked to generate typical instances (*typicality generation task*), select typical instances from a list of degrees (*typicality selection task*), categorize a list of degrees into belonging to one of the antonyms or an intermediate category (*trichotomous categorization task*), produce degrees of membership in the categories (*continuous categorization task*), categorize degrees into exactly one of the antonyms (*dichotomous categorization task*), and pro-

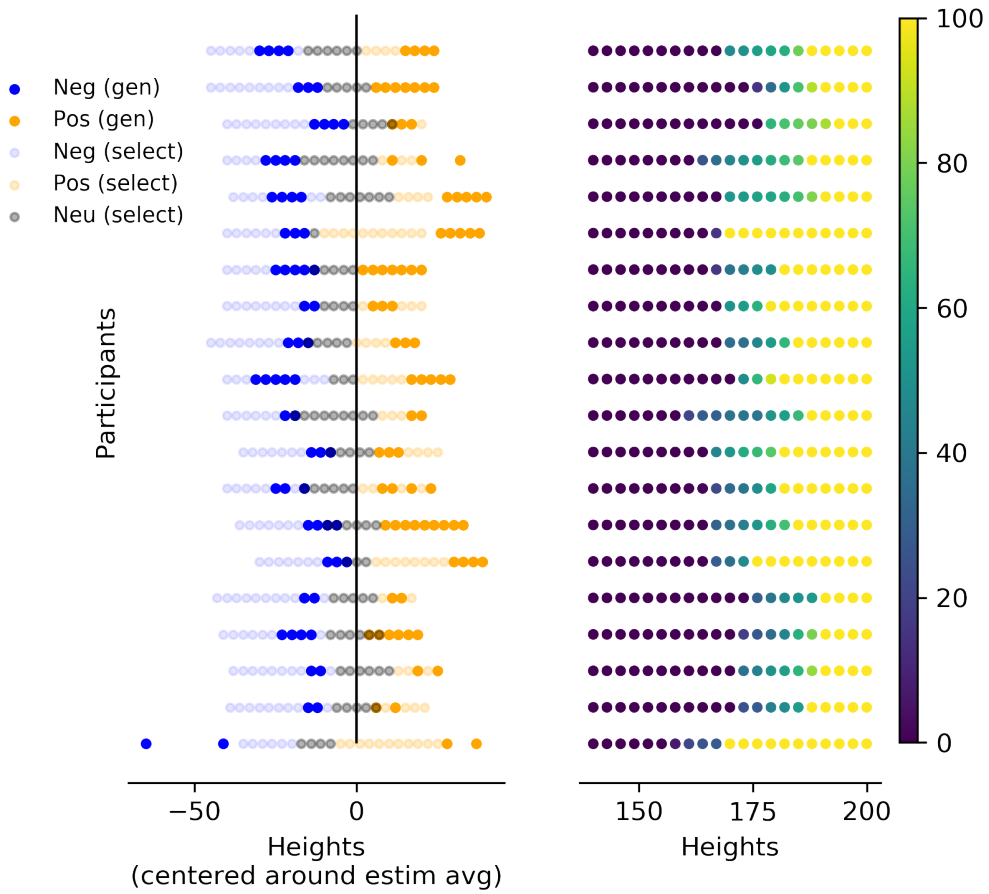


Figure 2.4: The responses for the first forty (out of 80) participants for the section of the experiment concerned with height adjectives. The left plot shows the generated typical instances (gen) as well as the selected (sel) degrees in the trichotomous categorization task. The x-axis does not show the absolute heights, but is instead centered around each participant's judgment of what the average height is. The right plot shows the continuous categorization data for the same participants.

duce average and ideal examples for each category. Some of the data for one of the antonym couple is displayed in figure 2.4, which I generated with the data found in `osf.io/djkdg`.

In the data analysis presented in the paper, the instances produced by participants in the typicality generation task play the role of prototypes. A membership function was obtained by calculating the proportion of precisifications (i.e. proportion of samples from the generated typical instance) where each degree falls in each antonym, after aggregating all data across participants for each antonym couple. Comparison between the membership function thus obtained and the continuous categorization task shows a close fit. The prototype picture can therefore account well for the aggregated data.

A direct comparison to the transition-based account with aggregated data is difficult. Some design choices were influenced by assumption on the author's part that the extension of the two antonyms are interdependent. For instance, in the continuous categorization task participants are asked "How inclined are you to call a male adult of 176 cm SHORT or TALL?", which collapses the vague threshold functions for the two antonyms into a single function. Moreover, throughout the paper membership degree for the negative adjective is calculated as 1 minus the membership degree for the positive adjective. However, I now show how my picture can also account for the data.

Two aspects of the data should be accounted for. First, where do degrees of typicality come from. Second, why do they relate as they do to the position of the transitions. I explain both in a transition picture of scalar terms. I claim that "typicality" in the case of properties referred to by gradable adjectives is a statistical property. Namely, it refers to the most frequently observed degrees in the category. Two factors then are at play. First, since relative adjectives do not cover the most probable degree—e.g. "tall" does not cover average height—the typical individual will be at the threshold. As an example, the typical "tall person" will be the shortest, and therefore most likely to be observed, among the tall persons. The second factor is that since the position of the transition is vague, there is no clearly most typical individual in the category. For example, there is no clearly shortest person among the tall persons. While in the picture of Verheyen and Égré (2018) the transition is determined based on the maximally typical individual, in my picture the maximally typical individual is found based on the transition.

I argued that defining the most typical individual requires balancing the confidence that the individual does in fact fall in the category with the frequency of the individual. How are these two opposing requirements met? The picture of vague transitions discussed in section 1.1.1 can be of help here. The idea was to asso-

ciate each transition not with a single point, but rather with a probability density on the scale expressing the probability that the "true" transition happens at each point. The cumulative density function then expresses the probability that each point belongs to the category.

When asked to generate typical instances for a positive [negative] adjective, participants will produce the lowest [highest] values that still have a high probability of belonging to the category. On the other hand, in the trichotomous categorization task participants will judge a point as belonging to the positive part of the scale iff the probability of the transition being below it is greater than the probability of it being above it.[5] Therefore, it is to be expected that generated *typical* members of the positive [negative] category are above [below] the ones categorized as belonging to the category. This intuitive explanation is displayed in figure 2.5.

I do not attempt a systematic comparison between the two accounts in terms of how well they fit the data. However, the fact that a transition-based account can in principle explain the data, as I have shown, in combination with the arguments against a prototype-based account presented above leaves us with more support for a transition-based picture of categorization for gradable adjectives.

2.2.4 Gradable adjectives on multidimensional domains

While the examples of scalar conceptual domains discussed above mostly consist of single quality dimensions, orders are also possible for $(d > 1)$-dimensional domains. Examples, which I mentioned above, include multidimensional adjectives such as "healthy" and "intelligent".[6] Gärdenfors (2017) proposes a picture where the quality dimensions of multidimensional adjectives—e.g. temperature, frequency of coughing, tiredness—are merged together to create a single dimension. The single dimensions still remain available, as demonstrated by expressions such as "healthy, except for a slight fever". What defined order-based category coding is the fact of exploiting only the order information to define a category. I sketch two ways that this approach can be extended to multidimensional domains.

[5]The assumption here is that while the extensions for the positive and negative adjectives are each defined with a vague threshold, the zone of indifference is defined negatively as the zone that falls under neither antonym. Therefore, considering only the positive adjective, if a degree has a probability p of being above the transition, it has a probability $1 - p$ of being below it.

[6]For more on multidimensional adjectives, see Sassoon (2013), which argues that not all multidimensional adjectives behave identically in how the constituting dimensions are connected.

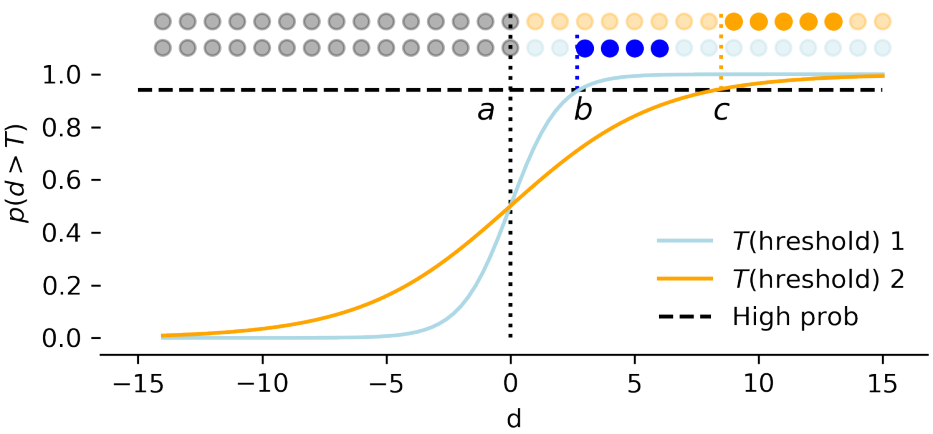


Figure 2.5: The figure shows the behaviour in the trichotomous categorization task as well as in the typical instances generation task for two vague thresholds. The confidence that a degree belongs to T 1 increases rapidly, while for T 2 it is slower. Point a is the point of equiprobability for both thresholds: points above it are more likely to belong to the category than not. Point b is the lowest point with a high probability of belong to category of T 1. Point c is the same for T 2.

The first approach is to build the category as a function of the orders of the individual dimensions of the domain. The order for an individual dimension encodes whether a point is greater or lower than another point in the domain. An order-based category in a multidimensional domain consists then of a set of transitions in each of the dimensions of the domain, plus a way that the dimensions interact. For each individual domain, the transitions play the same role as they did before, specifying whether a point in the domain belongs to the category with respect to that dimension. The relation between the dimensions is then encoded as a function of the Boolean information coming from the individual dimensions. For instance, a point in a domain might belong to the category iff it belongs to it with respect to each dimension, or with respect to at least one dimension. The fact that no order is defined on the domain as a whole prevents a natural definition of monotonicity. However, convexity can be still defined.[7] This first type of definition can only define categories

[7]All that is needed to define convexity is the notion of betweeness. If a metric d is defined on

whose decision boundaries are orthogonal or parallel to the axes. Moreover, it can encode non-convex categories, even when the categories are convex in the individual dimensions, e.g. when the Boolean function connecting the categories is XOR.

The second approach to constructing order-based categorization in multidimensional domains is to first define an order on the whole domain as a function of the single dimensions, and then define a category based on the domain-level order. Depending on the information that is exploited about the individual dimensions to construct the domain-level order, this strategy can construct a great variety of decision boundaries. In this approach, monotonicity can be defined in the usual way with the order that is defined on the whole domain. I present one possible way of constructing an order. Assume we have two points $a = [\, a_1 \quad \ldots \quad a_n \,]$ and $b = [\, b_1 \quad \ldots \quad b_n \,]$ in an n-dimensional domain D, and that each dimension i is structured (at least) by an order $\leq_i$. Then, we can let $a \leq_D b \iff a_1 \leq_1 b_1 \wedge \ldots \wedge a_n \leq_n b_n$. This is a partial order called the *product order*. The strategies in this and the present paragraph are not mutually exclusive. For instance, product order and some strategy from the previous paragraph are interdefinable.

While I do not discuss this in details, product orders on the dimensions of a conceptual domain might account for how multidimensional adjectives can exist on multiple dimensions but retain an intuitive notion of monotonicity. Consider the domain of health as an example. If Marianne is as healthy as Lucy in all respects except coughing more than Lucy, then Marianne is sicker than Lucy. However, if Marianne is sicker than Lucy in some respects and Lucy is sicker than Marianne in others, it is neither true that Marianne is sicker than Lucy nor that Lucy is sicker than Marianne. The bare use of "healthy" can then be analysed as a monotonic extension on the partially (product) ordered domain of health.

2.2.5 Monotonicity & extremeness in categories encoded with transitions

I have argued that the prototype picture of categorization does not work for classes of words expressing categories defined on scalar domains. Then, I presented an alternative picture of how categorization might happen on those conceptual domains. Next, I make a claim about the relation between scalarity and categorization:

---

the domain, a point b is between points a and c iff $d(a,b) + d(b,c) = d(a,c)$.

(2.11) Categories encoded in terms of transitions are monotonic.

This generalization connects the discussion of the universals of scalarity and the discussion of categorization, in a way that parallels Gärdenfors's claim that nouns express convex categories. As shown by lemma 3, 2.11 is equivalent to the claim that when categories are encoded with transitions, they are encoded with a single transition. A tempting idea is that monotonicity follows from transition-based categorization plus a bias from simplicity, which tends to reduce the number of encoded prototypes. However, the picture is more complicated. As we will see in chapter 3, a picture of categorization alone is not enough to explain the evolution of monotonicity, and other ideas from the theory of language evolution are needed.

Claim 2.11 does not make reference to any specific grammatical category. Therefore, if 2.11 is true, we should not expect a priori an alignment between grammatical categories and monotonicity. This contradicts the generalizations from the section on formal semantics, where I claimed monotonicity for whole grammatical classes of words. There are two reasons why such an alignment between monotonicity and grammatical class might come about. First, language seems to keep track of the structure of conceptual spaces, so that differences in type of conceptual spaces are reflected in the grammar, as previously argued in e.g. Culbertson, Kirby, and Schouwstra (2016). For instance, the grammatical class of gradable adjectives requires its members to express categories on scalar domains. Second, whole classes of words share few scales. For instance, quantifiers appear to only require the scale of proportion and the scale of integers. The effect of scalarity would then apply to all words on those scales.

For what concerns extremeness, as I have shown above, transition-based encoding makes extreme categories possible in contrast to categories defined in terms of prototypes. As shown in the discussion of the linguistic data, there is a tendency for categories to transition at the domain's boundaries if the boundaries exist. I return to the effects of transition-based encoding of scalar categories in chapter 6.

As discussed above, an important distinction can be drawn between absolute and relative gradable adjectives with respect to their context-sensitivity. I have briefly discussed how vagueness could be dealt with in a conceptual spaces account of transition-based categories, by using the supervaluation approach. We encountered a similar approach when discussing vagueness in prototype-based theories of categorization, in the concept of collated Voronoi tessellation. These scattered remarks

cannot explain the extremely complex data on gradable adjectives. Specifically, I do not attempt to explain the sometimes vague behaviour of absolute adjectives (Burnett, 2014). For reasons of space, I also do not discuss the relations between distributions of a property, scale structure, and vagueness in a conceptual spaces framework.

2.3 Conclusions

I started this chapter by describing Gärdenfors (2017)'s version of the theory of conceptual spaces, and its associated theory of categorization based on prototypes and Voronoi tessellations. Then, I presented some problems with applying this theory of categorization to scalar language. I developed an alternative theory based on orders and transitions, which solves the problems of the prototype account. Finally, I discussed various complications with the transition account and how to make sense of experimental data and discussions from the previous literature. Much more could be said about the details of a transition-based picture of scalar semantics in a conceptual spaces framework, but for reasons of space I leave further details to future work.

In sum, this chapter discussed the cognitive foundations of scalar language in a conceptual spaces framework. The cognitive account left two important features of the semantics of scalar terms undetermined, namely the number of transitions and their location. The two universals of scalar language I discussed in the previous section, namely monotonicity and extremeness, correspond to these two undetermined features respectively: monotonicity fixes the number of transitions per lexical entry to one, and extremeness consists in a relation between domain structure and transition's location. Since neither monotonicity nor extremeness have a semantic-internal explanation, as I argued in 1.2, nor a cognitive-internal explanation, as I will argue in the next chapter, the rest of the thesis will bring in other factors to explain their origin and evolution. In the next chapter, I show that the picture of learning of scalar categories developed in this section is not sufficient to account for the evolution of monotonicity, and propose a more complex evolutionary picture based on previous work in evolutionary linguistics.

Chapter 3

Modelling the evolution of adjectival monotonicity

In chapter 1, I presented two universals of scalar categories, namely monotonicity and extremeness. In chapter 2, I proposed an extension of the theory of conceptual spaces to account for linguistic data concerning categories on scalar conceptual domains. This extension consists of a proposal for how scalar categories are encoded, namely in terms of transitions rather than prototypes. While this extension of conceptual spaces theory provides an explicit cognitive framework for scalar categorization, I have argued that it is by itself insufficient to explain the universals of monotonicity and extremeness that I discussed in chapter 1. This is because some transition-encoded categories are non-monotonic, and some are non-extreme. In the current chapter, I will zoom in on the monotonicity of gradable adjectives, and present an account of the evolutionary mechanisms that favour languages with monotonic scalar categories, building on the account of scalar categorization developed in chapter 2. I will support this account with computational models of the evolution of monotonicity.

The structure of the chapter is as follows. I first discuss the Iterated Learning (IL) model of language evolution, a theoretical and modelling framework that will form the basis for my account of the evolution of monotonicity (3.1). Then, I develop three agent-based simulations, each building on top of the previous one. In the first model (sec. 3.2.1), I implement a learning bias for simplicity acting on language structure through an IL model. Under the simplicity pressure alone, the signals end up expressing degenerate meanings, covering either all or none of the scale's degrees. In the second model (sec 3.2.2), I add a pressure for communicative accuracy on the

agents. In the second model, the agents communicate literally, i.e. they produce and interpret signals assuming that their communicative partner is truthful, but not necessarily cooperative. The signals do not evolve to express degenerate meanings, but in order to maximize communicative accuracy they express non-monotonic meanings. Lastly, in the third model (sec 3.2.3) I implement more sophisticated agents, capable of doing pragmatic inferences. In combination with the other two pressures, from learning and communication, this causes monotonic meanings to emerge. On the basis of the third model, I conclude that monotonicity emerges as the best trade-off between learnability and communicative accuracy given human pragmatic reasoning.

3.1 The Iterated Learning model of language evolution

In chapter 1, I discussed previously proposed evolutionary explanations for monotonicity and extremeness. Some of these focus on cognitive biases for monotonic or extreme categories, and some focus on the advantage of such categories in communication. This section discusses the theoretical, experimental, and computational framework of IL (Kirby, Griffiths, & Smith, 2014) as a way of unifying and modelling the effects of communicative and cognitive pressures on language. I discuss the two pressures in turn, starting with the pressure from learning in section 3.1.1 and then moving onto the pressure for communicative accuracy in section 3.1.2. The rest of this chapter will argue that a full picture of the evolution of monotonicity needs to consider both of these pressures.

3.1.1 The pressure from learnability

One fundamental feature of language is that it is a cultural object, whose existence is dependent on the cognitive systems that learn it and use it. Since language has to be learned anew by each generation, and learning is a noisy process, we should expect languages to change over time. Crucially, the language learned by new users are not unbiased[1]. While the biases in the learned language might be small, and

[1]Here, we use *unbiased* in the following technical sense. Consider the teacher's language L_t and the acquired language L_a, which are both random variables. L_a is an *estimator* of L_t because the former is an estimate of the latter based on the data observed by the learner. L_a is a *biased estimator* of L_t iff the expected value of L_a is different from L_t.

therefore the changes they induce might be imperceptible for a single transmission, over many generations they can have a large impact on language. IL is therefore a causal mechanism that connects individual learners' biases—which affect the changes they introduce in the process of language learning—and the structure of language as a whole. The linguistic universals caused by the process of IL depend on specific individual biases. Very general biases, which are shared by all humans, will have repercussions on the structure of all languages through IL.

Computational and experimental methods have been developed to model and study the way that cultural transmission affects language (Kirby et al., 2014). A standard IL model consists of a number of *generations* $h_0, h_1, \ldots, h_n$. In turn, each generation $i \leq n$ consists of a number of agents $a_{i,1}, a_{i,2}, \ldots, a_{i,m}$. The life of agents in each generation $h_{0<i<n}$ has two stages. In the first stage, agents in h_{i-1} are selected to be *cultural parents* of the agents in h_i, and proceed to teach the language to their *cultural children*. In the second stage, the agents in h_i become the cultural parents of agents in h_{i+1}. In the case of h_0, the languages of the agents are picked at random from the set of all possible languages. The data to be analysed is the set of languages learned by all agents in generations after a burn-in period. Given this general structure, the essential missing ingredients are a language model and a learning model. Language model and learning models depend on the specific aspect of real language that one is focusing on.

Iterated Learning with Bayesian agents

The learning models that I develop in the following assume *Bayesian* learners. In general terms, a Bayesian agent that learns a language calculates a distribution p encoding the probability of a language l being the one that produced a set of observed linguistic data D. Moreover, $p(l \mid D)$ is calculated by using the Bayes theorem, which expresses the relation between the probability of a language given the observed data, the probability of the data given the linguistic data, and the probability of each language prior to observing the data:

$$p(l \mid D) = \frac{p(D \mid l)p(l)}{p(D)} = \frac{p(D \mid l)p(l)}{\int_{l_i \in L} p(D \mid l)p(l_i)dl_i} \tag{3.1}$$

where L is set of all possible languages.

Bayesian learners are a natural model of learning in IL models, because they

factorize the posterior (up to normalization) into two components: the likelihood, expressing the information coming from the linguistic data from the previous generations, and the prior, expressing the cognitive biases of the agents thought of as expectations about what the true language is. Calculation of the likelihood requires a picture of how the language produces the data.

Since agents in IL models have to pick a language to produce data to train the generation after them, there has to be a decision procedure to pick a language based on the posterior distribution. In the following, I will consider two decision procedures. First, *Maximum a Posteriori* (MAP) agents always pick the language that has the greatest posterior probability. Second, *sample* agents sample a language with a probability proportional to its posterior probability. While the MAP decision rule is deterministic given a dataset, the sample decision rule is stochastic. However, note that the language ultimately picked by a learner is always a random variable, whose distribution depends on (1) the true language spoken by the cultural parent, (2) the likelihood function, (3) the prior distribution, (4) the decision procedure (MAP or sample).

Markov Chains

In the previous section, I introduced Bayesian learning as a convenient formalism for modelling agents in IL models. In this section, I introduce one further modelling tool, which I apply to IL in the next section. *Markov chains* are mathematical objects that can be used to gain insights into IL models with Bayesian agents. A Markov chain is a time-discrete stochastic process, i.e. a set of random variables $\{X_i \mid i \in I\}$ indexed by a totally ordered countable set I. In the following, I will always be some set of integers $\{0, \ldots, n\}$. Markov chains have the *Markov property*, i.e. for every $i \in I$:

$$P(X_{i+1} \mid X_0, \ldots, X_i) = P(X_{i+1} \mid X_i) \tag{3.2}$$

In words, equation 3.2 says that the distribution of the random variable at any index only depends on the immediately preceding variable. This is sometimes expressed as "the future is independent from the past, conditional on the present". The Markov property allows us to think of Markov chains as a series of transitions from a state to another, each happening according to some probability distribution that only depends on the immediately previous state.

If the transition probabilities remain the same over time and the random variables

have discrete support, the transition behaviour of the variables can be encoded in a *transition matrix* P, where $P_{i,j}$ is the probability of transitioning from state s_i to state s_j. For instance, imagine that the support of the random variables is the set $\{s_1, s_2\}$, and that the Markov chain has probability 0.8 of going from s_1 to s_2, and probability 0.5 of going from s_2 to s_1. Then, the transitions happen with the following probabilities:

	s_1	s_2
s_1	0.2	0.8
s_2	0.5	0.5

and the transition matrix P is:

$$\begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix} \tag{3.3}$$

Each row of the transition matrix expresses a distribution, and therefore for all j, i:

$$P_{i,j} \geq 0, \sum_j P_{i,j} = 1 \tag{3.4}$$

The advantage of expressing the transition behaviour of a Markov chain with a transition matrix is that it helps formulating interesting generalizations about the behaviour of the chain. Note that P allows us to easily calculate the probability of moving from each state to any other state by using matrix multiplication. First, we encode the current state s_i as a one-hot vector $\vec{v}$[2] that is 1 at index i and 0 everywhere else. The probability of going from s_i to any other state is then encoded by the vector $\vec{v}^T P$. For instance, imagine that the chain is in state s_2:

$$\vec{v}^T P = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix} = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \tag{3.5}$$

Moreover, imagine we start with a distribution over states encoded by a vector $\vec{g}$, where the ith component of $\vec{g}$ is the probability of starting in state s_i. Then, $\vec{g}^T P$ gives the probability distribution over states after one step. Imagine we start with

[2] In the following, when not specified I will assume column vectors.

distribution over states $\begin{bmatrix}0.3 & 0.7\end{bmatrix}$. Then:

$$\vec{g}^T P = \begin{bmatrix}0.3 & 0.7\end{bmatrix} \begin{bmatrix}0.2 & 0.8\\ 0.5 & 0.5\end{bmatrix} = \begin{bmatrix}0.41 & 0.59\end{bmatrix} \tag{3.6}$$

Combining equations 3.5 and 3.6, we can calculate the distribution after any number of steps n starting with any distribution $\vec{v}$. All we need to do is multiply $\vec{v}$ by P n times. By the associativity of matrix multiplication:

$$\Big(\dots\Big(\big(\big(\vec{v}^T P\big)_1 P\big)_2 P\Big)_3 \dots P\Big)_n = \vec{v}^T PPP\dots P = \vec{v}^T P^n \tag{3.7}$$

The *stationary distribution* of a Markov chain with transition matrix P is a vector π such that:

$$\pi^T P = \pi^T \tag{3.8}$$

According to equation 3.8, the stationary distribution is (up to normalization) the left eigenvector of P that has eigenvalue 1. Therefore, the stationary distribution can be found in any of the usual ways to calculate eigenvectors.

The most important result we will need below connects the long-term behaviour of a Markov chain with its stationary distribution. If P is the transition matrix of an (ergodic[3]) Markov chain and π its stationary distribution, then for any initial vector $\vec{v}$:

$$\vec{v}^T \lim_{l\to\infty} P^l = \pi^T \tag{3.9}$$

Equation 3.9 tells us that for the specified Markov chains, the long-term distribution over states is independent of the starting vector, and moreover it is identical to the stationary distribution. In practice, this allows us to calculate the distribution over states given the transition matrix in the long term.

[3]A Markov chain is called *ergodic* iff it is both aperiodic and irreducible. The *period* of a state s_i for a Markov chain is the greatest common divisor of the set J of numbers such that the probability of starting at s_i and returning to it after $j \in J$ transitions is non-zero. A Markov chain is *aperiodic* iff the period of every state is 1. A Markov chain is called *irreducible* iff every state communicates with every other state. Two states of a Markov chain s_j and s_k *communicate* with each other iff there exists some n, m such that the probability of going from s_j to s_k in n steps is non-zero and the probability of going from s_k to s_j in m steps is non-zero.

Markov chains and Iterated Learning

To see the connection between Markov chains and IL, consider the simple case of an IL chain with generations comprising a single agent, and assume that there is a finite set of possible languages. The language learned by the agent in generation h_i only depends on the language of its cultural parent in generation h_{i-1}. In other words, it is conditionally independent of the languages spoken by the agents in generations before h_{i-1}. Therefore, the languages spoken by the agents in all generations are random variables that form a Markov chain indexed by the generation index. Call $p_{i,j}$ the probability that the cultural child of an agent speaking language l_i ends up using language l_j. Then, the transition probabilities between languages can be encoded by a transition matrix P such that $P_{i,j} = p_{i,j}$.

Recall from equation 3.6 that if the distribution of the language spoken in the first generation is expressed by vector $\vec{f}$, the probability of the agent in the second generation speaking each language is $\vec{f}^T P$. Moreover, recall from equation 3.7 that the distribution over languages in generation n is given by $\vec{f}^T P^n$. Finally, equation 3.9 tells us how to find the stationary distribution over languages from the transition matrix, assuming ergodicity. The stationary distribution in an IL model describes the probability that an agent will use a language, after so many generations that the distribution over languages is mostly independent of the initial distribution.

Griffiths and Kalish (2007) show a remarkable fact about stationary distributions of IL with Bayesian learners. Namely, they find that the stationary distribution for a chain of Bayesian sampling agents is simply the agents' prior distribution over languages. This phenomenon is called the *convergence to the prior*. Moreover, convergence to the prior also holds of an infinitely large population of agents in continuous time. Kirby, Dowman, and Griffiths (2007) shows that, given additional assumptions, the stationary distribution for MAP agents is a more peaked version of the prior. These results show that, under certain conditions, the long term behaviour of an IL model can be predicted without running it.

The bias for simplicity

In the discussion of IL up to this point, I have discussed learners' biases in general, abstracting from biases specific to the human cognitive system. However, IL by itself can only explain observed linguistic universals in conjunction with facts specific to the human cognitive system. In the rest of the thesis, I will work with a specific

bias, namely a bias for simplicity. The fundamental idea of a bias for simplicity is that humans attribute a greater prior probability to simple hypotheses over complex hypotheses, an idea is supported by previous work (Chater & Vitányi, 2003; Feldman, 2016; Culbertson & Kirby, 2016). The bias for simplicity has an important role to play in IL models. Since the stationary distribution fundamentally depends on the prior of the Bayesian agents, and the simplicity bias is encoded in the prior, the stationary distribution will depend on the simplicity bias.

The assumption of a bias for simplicity has to be implemented with assumptions about how complexity is measured by the cognitive system. Prima facie, many measures of complexity seem to be equally valid alternatives. A common way to measure the complexity of an object is as the length of the shortest description of the object. This captures the intuition that the string "000111" has a lower complexity ("Three 0s and three 1s") than the string "100101" (No obvious way to compress it). Such a measure seems to be sensitive to the language in which the description is given. However, the concept of complexity has been studied from a mathematical point of view, and a measure of complexity that is independent of the description language, called *Kolmogorov complexity*, has been shown to exist. However, Kolmogorov complexity is uncomputable (Li & Vitányi, 2008), and therefore approximations have to be used that are sensitive to the language in which hypotheses are described. In particular, the coding language for hypothesis will depend on details of human cognition.

The relevance of the discussion of categorization strategies from chapter 2 should be clear at this point. As argued above, scalar categories are coded with transitions. Therefore, categories that can be expressed with fewer transitions will be simpler to encode. Cognitive biases for easily coded categories are reflected at the language level through IL. In conclusion, scalar categories encoded with fewer transitions should be more common in language. However, as I will argue below this cannot be the full story. To understand why, a second pressure on language evolution, coming from communication, should be considered.

3.1.2 Pressure for communicatively accurate languages

When an IL model is used to study the effects of cognitive biases on language structure, the cultural parents are sampled uniformly from the population. However, in reality not all languages are equally likely to survive the test of usage. A second pressure on the evolution of language, the pressure for communicative accuracy, in-

fluences which languages are more likely to survive and be transmitted to successive generations. The pressure for communicative accuracy follows from the fact that language is used for conveying information. Language is used with many different goals, not all of which are prima facie reducible to the goal of conveying information. However, the transmission of information is arguably one of the fundamental goals of linguistic communication, and therefore we can expect it to put a pressure on language to better adapt to it. In order to implement selection by communicative accuracy in an IL model, a model of communication is required. In this section, I discuss two approaches to modelling communication and communicative accuracy, namely literal communication and pragmatic communication.

Literal communication

In the following, I only consider a simple model of communication that can nonetheless track differential accuracy of different languages in communication. Specifically, communication in the model will be between two agents, a sender S_0 and a receiver L_0.[4] A single communicative event unfolds as follows. First, the sender makes an observation o. Second, the sender selects a signal s and sends s to the receiver. In the simplest model, s is selected with uniform probability among the signals compatible with the observation. Third, the receiver receives signal s and tries to infer the content of the speaker's observation, o, finally making a guess g. In the simplest model, g is selected uniformly from the states compatible with s. In a model of *literal* communication, the receiver's guess of the sender's observation does not depend on the receiver's beliefs about the sender's mental state. However, g might depend on the sender's prior beliefs about the state of the world as well as the sender's knowledge of the meaning of s. The communication's degree of success is calculated by comparing o and g in terms of a chosen utility function. For instance, if the set of observations is categorical, the utility function might just track whether the receiver's guess of the world state is identical to the observation made by the sender. If the set of observations has a metric structure, success might be the distance between

[4]The index 0 indicates that we are dealing here with literal agents. This notation will be convenient when dealing with non-literal agents below. I use the letter L for consistency with the previous literature. L was chosen to refer to receivers as an abbreviation of "listener". However, using the term "listener" as an umbrella term for receivers reinforces a view where the prototypical production in a communication event is verbal. Therefore, I will try to use the term "receiver" instead. While "observation" is often used to talk about the visual modality, it does not necessarily refer to sight, and I will use it throughout the thesis.

the receiver's guess and the sender's observation. I return below to the question of how to calculate the utility of a language for communication, which is a function of communicative success.

In practice, in the models below observations do not have internal structure.[5] Each signal is associated with a set of observations that models its *meaning*. The set of meanings is the powerset $\mathcal{P}(O)$ of subsets of O. For instance, if the set O of possible observations is $\{o_1, o_2, o_3\}$, possible meanings include $\{o_1\}$, $\{o_2, o_3\}$, and $\{\}$. How does a set of observations model the meaning of a signal? Each signal conveys to the receiver that the sender's observation is one of the observations in the set associated with the signal. In the simplest model of literal communication, after receiving a signal the receiver first renormalizes its prior over the world states compatible with the received signal, and then picks the world state with the greatest probability.

The task described above captures the fundamental principle of communicative use of language, namely conveying information about the state of the world. However, the model is of course a simplification of how information is transmitted in real interactions. First of all, the described communicative event is not interactive and lacks the possibility for feedback. Secondly, not all transmission of information has a clear loss function associated with it. For instance, it might not make a difference if information conveyed during small talk is true or false. Thirdly, the conveyed information in real interactions does not have to be even in principle available to the receiver. For instance, it might be impossible to determine the truth of anecdotes about something that happened a long time ago. Many other aspects of real information transmission are left unmodelled. Whether the simplifications made by the model of communication above are reasonable depends on whether they affect the way that the pressure for accuracy shapes language. For reasons of space, I do not discuss further the effects of these ulterior factors on the phenomena studied below.

I presented a simple model of communication for literal agents, how its utility for communication can be modelled, and some respects in which it fails to model real linguistic interactions. While some simplifications can be put aside for the models developed below, a more complex model will be required for the phenomena studied below. Specifically, a model of pragmatic communication will be needed.

[5]Note however that while observations themselves are unstructured, in the models below the set of observations is structured.

Pragmatic communication: the Rational Speech Act model

In interpreting and producing signals, literal language users do not exploit the knowledge that they are communicating with *cooperative* agents, i.e. agents that share their goals in communication. To see an instance in which communication with a cooperative vs an uncooperative agent makes a difference, consider the following case. Sandro and Roberta went to Edinburgh for Christmas, and they are planning to visit the Christmas market. Sandro asks Roberta, who was outside, what the temperature outside is. One way of characterizing Roberta's thought process before she answers Sandro's question (albeit possibly not a cognitively realistic account) is that Roberta considers various possible answers. Some of these answers are true, and some are false. However, even among the true questions, some are more useful to Sandro and some are less useful. If Roberta is cooperative, i.e. she wants to answer Sandro's question in the way that helps him the most, she will tend to pick not simply a true, but also a useful answer. For instance, imagine it is in fact freezing outside, and possible answers are "It is cold" and "It is freezing". While "It is cold" is true, it is less helpful than "it is freezing" for Sandro. Therefore, Roberta will tend to say that it is freezing. From Sandro's point of view, the assumption that Roberta is cooperative can therefore be used to extract more information from the answer than would be possible given the answer's literal meaning alone. Since if it was freezing Roberta would answer "It is freezing" and not "It is cold", if Roberta does in fact answer "It is cold", Sandro can infer that it is not freezing.

The surplus of information that can be extracted from a sentence through the cooperative assumption is called an *implicature* (Grice, 1989). Implicatures can be grouped according to the contexts in which they arise. If there is a set of contextually relevant assertions which only differ in terms of how many world states they are compatible with—their strength—then the choice of one of them by a cooperative sender implicates the falsity of all of the stronger ones. This type of implicature is called a *scalar implicature*. In the following, pragmatics will mostly enter my models in the form of scalar implicatures. The *Rational Speech Act* (RSA) modelling framework offers a convenient way to simulate scalar implicatures in computational models.

In the type of RSA model I consider, there are three agents, a literal receiver L_0, a pragmatic sender S_1 and a pragmatic receiver L_1. L_0 is used by S_1 as part of the calculation of the probabilities of producing each signal. S_1 produces signals, but

also appears in L_1's calculation of the probability of each state given the received signal. When a literal receiver L_0 receives a signal with a meaning m, they attribute to each state d a probability equal to 0 if the meaning is not compatible with d and a probability proportional to the prior probability of d if m is compatible with d. This corresponds to the simplest situation where L_0 has a uniform prior distribution over world states. I assume here that the world is in each state with uniform probability.

Pragmatic speaker S_1 observes a world state d and calculates the utility that each signal s has for L_0:

$$\mathcal{U}_{S_1}(s; d) = -\left(-\log(p_{L_0}(d \mid s))\right) \tag{3.10}$$

were $P_{L_0}(d \mid s)$ is the probability that the literal receiver attributes to d after having received signal s, and $-\log(p_{L_0}(d \mid s))$ is the surprisal for that probability.[6] This equation identifies the utility of a signal with the negative surprisal for the literal agent of the world state observed by the speaker after the literal agent has received the speaker's signal. Signals that maximize the listener's posterior for the world state observed by the speaker have higher utility. Pragmatic speakers then choose the signal to utter with a probability proportional the utility:

$$p_{S_1}(s \mid d) \propto \exp(\alpha \mathcal{U}_{S_1}(s; d)) \tag{3.11}$$

α determines the strength of the increase in the probability of picking an utterance given an increase in utility. When $\alpha = 0$, the left-hand side of 3.11 becomes 1 for the true signals and all true signals are therefore uttered with equal probability. As $\alpha \to \infty$, the signal(s) with the highest utility gets a greater and greater advantage over the other signals.

Finally, L_1 performs Bayesian inference on the basis of the behaviour of S_1. After receiving a signal s with meaning m from a speaker, L_1 calculates the probability of each world state:

$$p_{L_1}(d \mid s) \propto p_{S_1}(s \mid d) p_{L_1}(d) \tag{3.12}$$

where $p_{L_1}(d)$ is the prior probability that the listener attributes to the world state being observed.

To see how L_1 is capable of calculating implicatures, consider the following simple case. Imagine there are two states, d_1, d_2, and two signals, s_1 and s_2. Signal s_1 has

[6] I simplify the standard RSA model by assuming that the utterance cost is the same for all adjectives.

meaning $\{d_1, d_2\}$, while signal s_2 has meaning $\{d_2\}$. Assume also that L_0 has a uniform prior over the world states, i.e. each state takes probability 0.5. L_0 can therefore be modelled with the following matrix:

	d_1	d_2
s_1	0.5	0.5
s_2	0.	1.

In this setup, if d_2 is observed S_1 would tend to pick signal s_2, even though from a semantic point of view s_1 is equally acceptable as s_1, because s_2 increases the probability that L_0 will guess the right world state. Therefore, we get the following approximate production probabilities for S_1, where $\alpha = 4$.:

	s_1	s_2
d_1	1	0
d_2	0.06	0.94

Therefore, if the speaker sends s_1, the observed signal was probably not s_2, but rather s_1. This means that L_1 calculates the scalar implicature:

	d_1	d_2
s_1	0.96	0.4
s_2	0.	1.

The three agents are plotted in figure 3.1.

Measuring communicative accuracy

To implement selection dependent on communicative accuracy in an IL model, a way to quantify communicative accuracy is required. An intuitive measure of communicative accuracy for a combination of sender language $S_{\mathcal{L}}$ and receiver language $L_{\mathcal{L}}$ is the *expected utility* of that combination of languages. This is the expected value of the utility function for the two agents communicating. Consider the simplest case of a utility function which equals to 1 when the receiver guesses the sender's observation and 0 otherwise:

$$CS(S_{\mathcal{L}}, L_{\mathcal{L}}) = \sum_o \left(p(o) \sum_s p_{S_{\mathcal{L}}}(s \mid o) p_{L_{\mathcal{L}}}(o \mid s) \right) \tag{3.13}$$

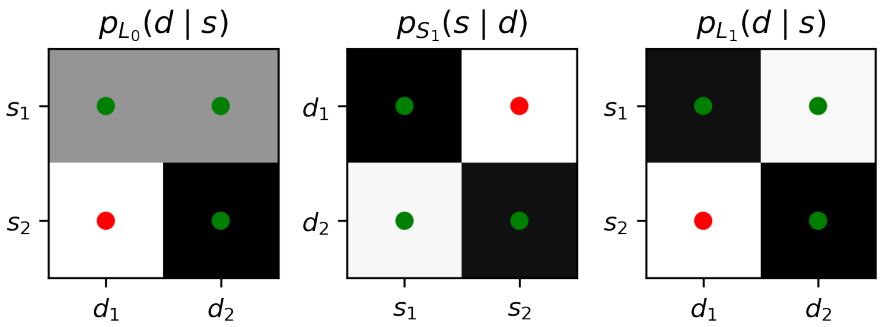


Figure 3.1: The figure displays L_0, S_1, and L_1 for a simple RSA model. The dots show whether a signal is compatible with a state (green) or not (red), and the color shows probability (lighter being more probable). While the probabilities of the two world states given signal s_1 are uniform for the literal receiver (left), the probability of state d_2 given signal s_1 is much greater than that of d_1 for the pragmatic sender. This shows that the pragmatic receiver calculated a scalar implicature.

This is the expected value across all observations of the probability that the sender sends a signal for that observation and that the receiver guesses the right observation given the sent signal.

The expected utility for a combination of languages can be calculated in a more efficient way than as a sum, namely with multiplication between matrices. Define two matrices $\mathbf{S}$ and $\mathbf{L}$, the first representing the sender and the second the listener. For both matrices, each row represents a signal and each column an observation. Moreover:

$$\mathbf{S}_{i,j} = p_{S_\mathcal{L}}(s_i \mid o_j)$$
$$\mathbf{L}_{i,j} = p_{L_\mathcal{L}}(o_j \mid s_i)$$

The value $\sum_s p_{S_\mathcal{L}}(s \mid o_k) p_{L_\mathcal{L}}(o_l \mid s)$, i.e. the probability that a sender makes observation o_k and that the receiver guesses observation o_l, is the k, l index of the matrix:

$$\mathbf{S}^T\mathbf{L} \tag{3.14}$$

The only important cases for communicative accuracy are the cases where the listener guesses the right observation, i.e. the elements of the diagonal of matrix 3.14.

Therefore, define the vector $\vec{a}$ as the main diagonal of matrix 3.14. The ith element of $\vec{a}$ is the probability that the receiver will guess observation i, given that the sender makes observation i. Finally, the success probability for each observation have to be weighted by the probability of the observation. Call $\vec{p}$ the vector whose ith element is the probability of the sender making observation i. The expected communicative success is the dot product of $\vec{a}$ and $\vec{p}$:

$$CS(S_{\mathcal{L}}, L_{\mathcal{L}}) = \vec{a} \cdot \vec{p} \tag{3.15}$$

This way of calculating expected utility makes for a much more efficient implementation.

Below, I will consider other ways of calculating the expected utility of a combination of speaker's language and receiver's language. The main difference will be in the measure of success for a single communicative event. In particular, success might not be categorical—either a communication is successful or it is unsuccessful—but rather graded, and based on the similarity between the observed and guessed world states. How good a language is for communication will still be calculated as the expected success interaction-wise.

3.1.3 Combining the pressures in an IL model

The pressures for simplicity and communicative accuracy always act together to shape language structure. However, they often pull in opposite directions. Simple languages, which can be learned easily, are often incapable of conveying much information. On the other hand, languages that draw many distinctions about the world and are therefore capable of conveying world states to a high degree of precision are generally cognitively complex.

The joint effect of the pressures for simplicity and communicative accuracy can be studied by modifying the IL model presented above in section 3.1.1. In an IL model plus communication, the selection of agents to become cultural parents is not random, but rather happens with a probability proportional to the agents' expected communicative success with other agents. Two choices are necessary to define selection based on communicative success. First, which agents is each prospect parent communicating with? In the models below (chapter 3), I assume that they are communicating with their cultural parents, but other choices are in principle possible.

For instance, they can be modelled as communicating with the other agents in their generation. This latter choice is more computationally expensive and more difficult to interpret. Secondly, is communicative accuracy taken with the prospect parent as a speaker, as a hearer, or both? In the models in chapter 3, I take the prospect parents as speakers, but the other options are also valid possibilities. The expectation is taken over many events of communication, and the measure of communicative success depends on the specific model. A pressure for communicative accuracy added to an IL model in the way described in this section causes the emergence of languages that find the best compromise between simplicity and expressiveness.

Pragmatic skills play a special role in the interaction between language simplicity and communicative accuracy. Communicative accuracy partially depends on the amount of information that is conveyable by a language. A pragmatic agent is capable of extracting more information from a language than a literal agent. Therefore, pragmatic agents will have greater communicative success with simpler languages. Populations of pragmatic agents can evolve language that better adapt to their simplicity bias, without losing in terms of communicative accuracy. This phenomenon will appear again in the models developed in this chapter.

In this section, I presented the framework of IL as a way to study the evolution of language in response to the pressures coming from individual cognition and communication. I focused on the modelling aspect of IL. In the next section, I apply the techniques presented in this section to the problem of the evolution of monotonicity.

3.2 A model for the evolution of monotonicity

3.2.1 Model 1: Pressure from learning

Language model

The first model only includes the pressure acting on scalar categories from cognition. The model consists of:

1. An ordered set $O = \{D, \geq_D\}$ modelling a scale, where $D = \{d_1, \ldots, d_n\}$ is a set of n degrees ordered consistently with their indices.

2. A uniform probability distribution P_D over D modelling the probability of specific degrees being observed.

3. A set M of meanings, which is the set of sets of degrees, i.e. the powerset $\mathcal{P}(D)$.

4. A set of signals $S = \{s_1, \ldots, s_o\}$ that is identical for all languages.

5. A set of languages. Each language $l : S \mapsto M$ is a (total) function from each signal to a corresponding meaning.

This way of defining the languages implies that there is no homonymy. However, l might map two signals onto the same meaning, which means that synonymy is possible. I added the further restriction that in every language there is at least one meaning to refer to each degree. The languages are holistic in the sense of Kirby, Tamariz, Cornish, and Smith (2015) because signals have no internal structure.

Each language l can be represented as a Boolean matrix L. Each row of L, $L_{i,\bullet}$, corresponds to signal i and each column $L_{\bullet,j}$ to state j. $L_{i,j}$ has value 1 iff $d_j \in l(s_i)$, and 0 otherwise. For instance, the following language with 4 signals and 4 meanings:

	d_1	d_2	d_3	d_4
s_1	1	1	1	0
s_2	0	1	1	0
s_3	0	1	0	1
s_4	1	0	1	0

is represented by the corresponding matrix L:

$$\begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \tag{3.16}$$

Each language models a system of scalar categories. In the following, I consider first languages with three signals and three degrees. Since a meaning is a set of degrees, a scale with three degrees implies that there are $2^3 = 8$ meanings. Each way of attributing any three of the possible meanings to the three signals defines a language, with the restriction that for each language each degree is contained in at least one meaning. With three degrees and three signals, there are therefore a total of 343 languages. I also discuss below the case where there are four signals and four degrees, and therefore a total of $2^4 = 16$ possible meanings and 50625 languages.

Relation to adjectival meaning

The models in this section implement a model of the semantics of gradable adjectives I presented in section 1.2. The meanings in M model *standard* functions by modelling the set of degrees that verify $\lambda d.standard(d)(G)(\mathbf{C})$ given a fixed G and $\mathbf{C}$, for a simple scale consisting of only few degrees. The degrees covered by *standard* is sufficient to model the effects of a bare assertion of a gradable adjective for what concerns the degree argument. It will become clear below that the argument that the models illustrate do not in principle depend on the number of scale degrees.

Much like for real scales, a transition in real adjectival meaning corresponds in the model to a transition from m applying to not applying or viceversa when going through the degrees in D in order. A meaning $m \in M$ models a monotonic *standard* function iff m has zero or one transitions. A further distinction is useful to connect the model to real world adjectives. Meanings with zero changes are either compatible with all or with none of the degrees. I call meanings with zero changes *degenerate* meanings. Degenerate meanings can be interpreted in two ways. First, they can be understood as *standard* functions that are true for all degrees or false for all degrees. Alternatively, a signal compatible with all states can be understood as silence, i.e. a signal that does not exclude any possibility.[7]

Meanings with one transition are still monotonic while being compatible with a proper subset of the set of degrees. Finally, meanings with two (or more) transitions are non-monotonic. I will call *degenerate* those languages with only degenerate meanings, *monotonic non-degenerate* those language that only have monotonic meanings but no degenerate meanings, and *non-monotonic* those languages with at least one non-monotonic meaning.

Particular care is needed in interpreting how model meanings represent real language antonyms. As I discussed in 2, antonyms can be modelled as using two different scales with the same degrees but opposite ordering relations, or the same scale with transitions of different types (false-to-true for positive antonyms and true-to-false for negative antonyms). In the model, both monotone increasing and monotone

[7]One must be careful with the latter interpretation, because differently from other signals, silence as a signal that conveys no information is recognizable as such previous to learning. It is prima facie appealing to interpret a model signal that is everywhere true as conveying that an individual has the property to some degree, and a signal that is everywhere false as conveying that the individual does not have the property to any degree. However, this interpretation is meaningless in the model presented in this chapter, because by constructions senders directly observe a degree, rather than an individual.

decreasing terms can exist on the same scale. Nothing substantial hangs on this interpretation. The choice of how to analyse the relation between antonyms is important in a related debate regarding the naturalness of increasing monotonicity as opposed to decreasing monotonicity. If antonyms exist on two different scales in a sense more substantial than I have discussed here, POS is technically not only monotonic, but always increasing monotonic.

Likelihood and simplicity based prior

Agents are Bayesian learners of the type described in section 3.1.1. To define their learning behaviour, a prior and a likelihood function have to be specified. The prior is calculated as an inverse function of the language's complexity, consistent with previous work showing that learners prefer simpler languages (see 3.1.1 above for a discussion of the relevant work). More specifically, the prior for each language is calculated as follows. First, the *description length* of the language's meanings, i.e. the number of bits needed to encode the meaning of each signal, is calculated. The description length of a language l, $\mathcal{L}(l)$, is the sum of the description lengths of the meanings expressed by the language's signals. Description length is my measure of complexity. Finally, the prior for each language l is defined as:

$$p(l) \propto 2^{-\gamma \mathcal{L}(l)} \tag{3.17}$$

Where γ is a value modelling the strength of the bias for simpler languages. Note that languages with longer descriptions have lower prior probability. I use equation 3.17 for the prior in the agents' Bayesian updating.

Equation 3.17 leaves open how to compress each meaning, which is needed to calculate $\mathcal{L}$. I argued in chapter 2 that scalar categories are encoded by specifying the positions of all transitions from the meaning applying to not applying or from not applying to applying. One extra bit must be added to specify whether the first step is from applying to not applying or viceversa, since a list of the changes' position leaves it unspecified and is therefore compatible with both a meaning and its negation. For a discrete scale with n degrees, transition-based encoding allows to compress any meaning in $1 + c \log(n-1)$ bits, where c is the number of changes and $n-1$ is the maximum possible number of transitions. The more degrees there are, the more convenient it becomes to exploit the ordering on degrees rather than simply encoding the single degrees that fall in the category. Notice that if a meaning

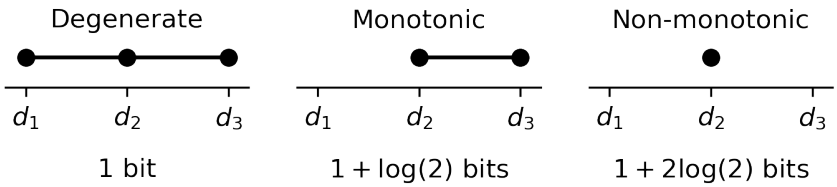


Figure 3.2: The x-axis shows the underlying scale, e.g. the scale of height degrees for "tall". It is a fully bounded scale, consisting of degrees d_1, d_2, and d_3. A meaning is represented as a line extending over the degrees that belong to the meaning—e.g. the degrees of height for which someone counts as "tall". The complexity of a meaning is estimated as the size of a lossless encoding of the meaning. If meanings were coded by simply memorizing the degrees that belong to them, the description length of each meaning would depend only on the number of degrees covered. Transitions-based coding specifies the position of each change and whether the first change is from belonging to not belonging to the meaning or viceversa. A natural result of coding meanings based on transitions is that monotonic meanings have shorter descriptions than non-monotonic ones.

has zero changes, it can be compressed in terms of transitions in one bit, which says whether the meaning contains all the degrees or is empty.[8] This compression method is displayed in figure 3.2.

The consequence of compressing meanings by encoding their transitions is that the meanings modelling monotonic *standard* functions are attributed a higher prior probability than meanings modelling non-monotonic *standard* functions. This follows from two facts. First, an increase in the number of changes always results in a lower prior probability, and therefore degenerate monotonic meanings have the lowest prior probability, followed by non-degenerate monotonic meaning. Second, as was shown above, meanings with zero or one changes correspond to monotonic *standard* functions.

[8]Technically, once a degree has been specified it can be excluded from the set of degrees that still need to be specified. Therefore, at every new degree the number of degrees to pick from decreases by one. Given a set of n degrees to choose from, the description length of a meaning covering k degrees is $1 + \sum_{i=n-k}^{n} \log(i-1)$, which is smaller than $1 + k \log(n-1)$ for $n > 2$. However, the informational gain resulting from being able to exclude a value becomes smaller as the number of degrees to pick from increases. I assume that for normal gradable adjectives the number of degrees is big enough and the realistic number of degrees to specify small enough that this complication makes no practical difference.

The description length of a language in the model—and therefore its prior probability according to equation 3.17—depends on the description length of its meanings but not in the way they are attributed to the signals. For languages with three degrees, there are exactly three complexity levels for the meanings, corresponding to zero, one or two changes. If the language moreover has three signals, there are then at most 10 different distinct levels that the prior probability of a language can take, i.e. the number of combinations with replacement of three complexity levels for three available signals. In fact, there are fewer than 10 different monotonicity levels. All that matters for the total description length of a language is the total number of changes in the three meanings in the language, regardless of how they are distributed across signals. For instance, language

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

which contained three total transitions has the same coding complexity as

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Overall, there are exactly 7 different possible values for the prior probability of a language, one for every way of summing a combination of 0, 1 and 2 changes. This is shown in figure 3.3.

The way the prior above was defined influences how agents learn languages. Specifically, the greater the total number of changes in the language, the longer it takes agents to correctly guess the language, since the prior for the language is smaller. Figure 3.4 shows this phenomenon for languages with four degrees and four signals, for two levels of bias for monotonicity.

Once the prior is fixed, the other component needed for Bayesian updating is the likelihood $p(s \mid d, l)$, which is the probability of signal s being sent by a speaker of language l after observing d. Calculating the likelihood therefore requires a model of production, describing how agents pick a signal given a degree. There are two relevant cases to model. If the observed degree is not in the meaning expressed by the signal s in language l, the likelihood of the signal having been produced is 0. If

Total # changes	# changes by meaning	Colour legend:
0	[0, 0, 0]	degenerate mon
1	[0, 0, 1]	non-deg mon
2	[0, 0, 2], [0, 1, 1]	non-mon
3	[0, 1, 2], [1, 1, 1]	
4	[0, 2, 2], [1, 1, 2]	
5	[1, 2, 2]	
6	[2, 2, 2]	

Figure 3.3: The ten types of language by prior probability and monotonicity level. Since the prior depends only on the total number of changes in the meanings of the language, each language has one of seven different possible prior probabilities, one for each row. Languages are subdivided in ten types in the second column by the number of changes in each meaning. $[a, b, c]$ names the type of language with a meaning with a changes, a meaning with b changes and a meaning with c changes. A language is degenerate mononotonic if each of its meanings has 0 changes, non-degenerate monotonic if it contains at least one meaning with 1 change and no meaning with 2 changes, and non-monotonic if it contains at least one meaning with 2 changes.

one or more available meanings are compatible with the observed degree, the agents need to choose which one to use for communication. For the simple agents in this model, the only semantic criterion for choosing between meanings is compatibility, so all compatible signals are semantically equally apt. Moreover, the agents are literal, which means that they do not have pragmatic criteria to prefer one of two meanings that are semantically equally compatible with a given degree. Lacking a reason to prefer a compatible meaning over any other compatible meaning, agents pick with uniform probability among the compatible meanings. This behaviour can be modelled as:

$$p(s \mid d, l) = \begin{cases} 0, & \text{if } d \notin l(s) \\ \dfrac{1}{|\{h \mid h \in B_l \wedge d \in h\}|} & \text{if } d \in l(s) \end{cases} \tag{3.18}$$

The production model in equation 3.18 has some desirable consequences. If a language only has one signal to refer to the observed degree, then the production probability is 1. A language that could not have produced a combination of signal

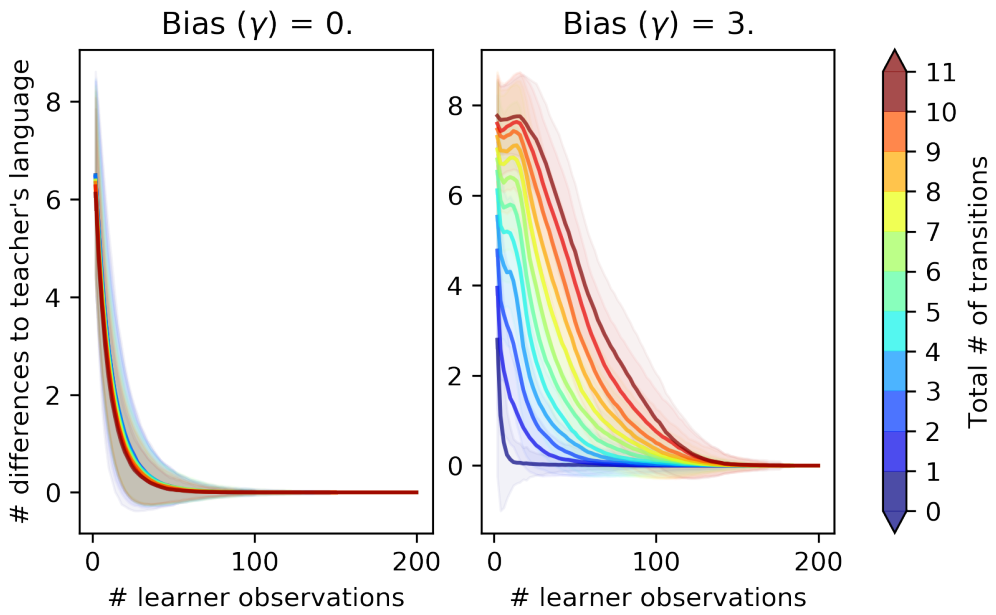


Figure 3.4: The plot shows the difference between the teacher's language and the language with the highest posterior probability for the learner, for a variety of languages, number of observations, and two levels of simplicity bias ($\gamma = 3$ in equation 3.17). The languages of the agents in the plot have four degrees and four signals. Each line corresponds to an averaged sample of 1000 languages with a given total number of transitions. When agents do not have a bias for more compressible languages (left plot), the total number of transitions does not influence the number of observations needed to acquire the language. On the other hand, when agents have a bias for simplicity (right plot), languages with more changes are attributed smaller priors and therefore agents need more observations to acquire them correctly.

and degree is judged impossible. Moreover, if two languages are equally probable but one only has the received signal to refer to the observed degree while the other has multiple signals, the former language is considered more probable.

The probability of each language is evaluated by the learner on a sequence of tuples $\langle \text{degree}, \text{signal} \rangle$. Given equation 3.18, the probability of a sequence $G = \langle \langle s_1, d_1 \rangle, ..., \langle s_n, d_n \rangle \rangle$ being produced by a speaker of language l is:

$$p(G \mid l) = \prod_{\langle s_i, d_i \rangle \in G} \frac{1}{|D|} p(s_i \mid d_i, l) \tag{3.19}$$

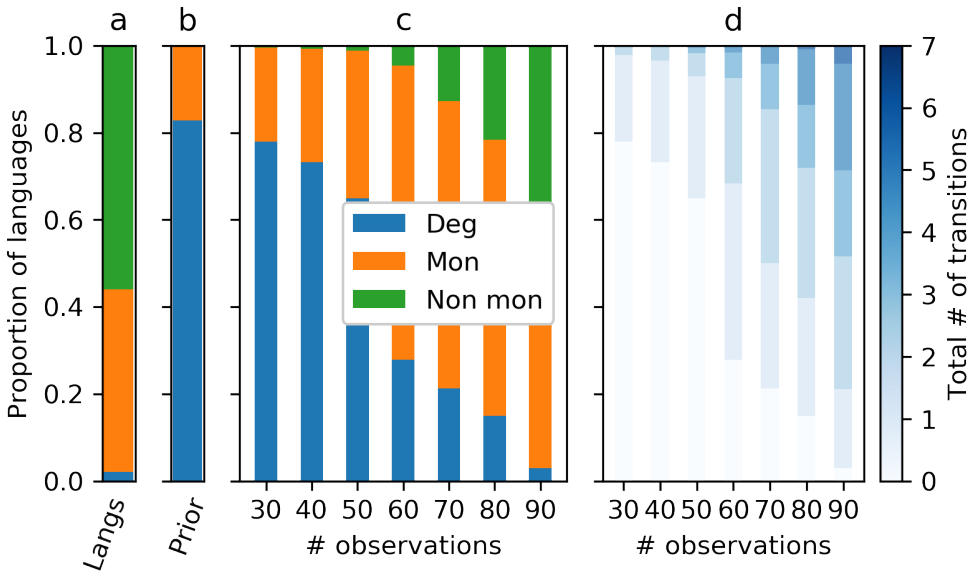


Figure 3.5: Plot a: proportion of types across all languages. Plot b: prior probability that the language will be of each type. Plot c: Frequency of monotonicity types in model 1 over 50 runs of the simulation for each number of observations by learner. Right plot: languages in same data by total number of transitions. The main conclusion of the first model is that learning alone induced the evolution of languages proportionally to their representation in the agents' prior. Therefore, a prior that favours simple languages will induce an over-representation of degenerate languages.

Results

A pure IL condition was ran 50 times for 3000 generations with a population of 10 sample agents, for different numbers of observations by learners. Figure 3.5 shows the frequency of the languages spoken for the different combinations of parameters across all runs of the simulation, after excluding the first 1000 generations as a burn-in. In a simple IL condition, agents across all generations mostly learn degenerate languages for small numbers of observations. When learners make a lot of observations before sampling a language, however, the number of non-degenerate languages increases. This is a consequence of the fact that languages are preserved very accurately across generations (see figure 3.4), and therefore the learner's biases

only influence the distribution of languages spoken in the population over very many generations. Eventually, as discussed in 3.1.1, the distribution over languages will converge to the prior distribution. However, if the languages are learned accurately, it might take such as large number of generations before convergence to the prior that it is infeasible to run the simulation. The convergence problem displayed in figure 3.5 should make us suspicious of results that seem to depend on the number of observations. However, all the results discussed below are stable even for high number of observations.

This result is consistent with the larger framework of IL that was introduced in 3.1.1. IL alone creates a pressure for languages to get increasingly simple, i.e. conform to the prior expectations of the agents, since every new generation has a tendency to learn languages that are simpler than the one spoken by the previous generation.

However, this first model makes the wrong prediction with respect to the distribution of language types. Degenerate languages, while monotonic, are not what we observe in real world adjectival systems. Real world adjectives have non-degenerate meanings and they allow speakers to convey information about amounts of properties beyond whether an object has or not any degree of the property. The crucial advantage of real-world non-degenerate adjectives over the predicted degenerate meanings in model 1 lies intuitively in the fact that non-degenerate adjectives can be used to convey information about the world, while degenerate meanings do not convey any information. This gives a reason for thinking that a pressure for communicative accuracy, as discussed above in section 3.1.1, might be involved in the evolution of monotonicity. The second model, presented in the next section, makes this intuition formal.

The conclusion of model 1 is that the pressure for simplicity coming from iterated learning is not enough to account for the evolution of non-degenerate monotonicity in adjectival meaning. Model 2 is a natural extension of model 1 with the addition of a pressure for communicative accuracy.

3.2.2 Model 2: Communicative pressure on literal agents

Model 1 analysed the consequences that learnability has on the semantic structure of gradable adjectives. The picture that emerged from model 1 is that pressures from learnability drive gradable adjectives to be monotonic, but also to become

degenerate and non-expressive. This is at odds with the observation that real world adjectives tend to strike a balance between learnability and expressivity. Model 2 adds a pressure on languages to perform well for communication by selecting the agents that communicate accurately to be cultural parents.

In model 2 a stage is added to the life of agents between the learning stage and the teaching stage. In this middle stage the agents communicate with each other as described on page 118. The pressure for language to be communicatively accurate is implemented by selecting each agent a (with replacement) to be a cultural parent for the following generation with a probability proportional to the expected communicative success of a with its parent p:

$$p(a \text{ is selected}) \propto e^{\epsilon CS(a,b)} \tag{3.20}$$

where $\epsilon > 0$ determines the strength of the selection. In the models below, ϵ is set to 4, which allows for the emergence of different patterns with different values of the other parameters. The intuition behind equation 3.20 is that an agent which is communicatively successful with the other agents in its population is selected more often to be a cultural parent. The consequence of this way of calculating fitness is that languages that often provoke a failure in communication are taught less often to the following generation.

Results

Model 2 was ran 50 times for 3000 generations with a population of 10 sample agents, for different numbers of observations by learners. Figure 3.6 shows the frequency of each language type being acquired by each generation after their learning phase, again with a burn-in period of 1000 generations.

When pressure for languages to be communicatively accurate is implemented, the number of monotonic languages decreases. This means that the communicative pressure introduced in model 2 acts against the spreading of monotonic meanings. The reason for this is that monotonic languages perform communicatively worse than non-monotonic languages, and therefore the agents with the former languages get fewer opportunities to be cultural parents to the following generation. The communicatively poor performance of fully monotonic languages is a consequence of two facts. First, while how much overlap the meanings have depends on where the changes lie for each meaning, a language with three or more monotonic meanings

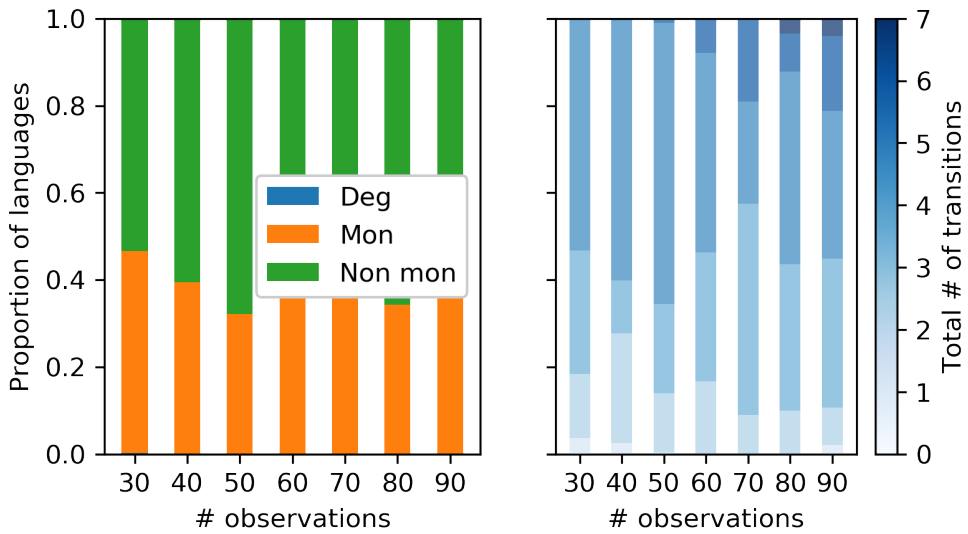


Figure 3.6: Left plot: Frequency of monotonicity types in model 2 over 50 runs of the simulation for each number of observations by learner. Right plot: languages in same data by total number of transitions. The main conclusion of the second model is that a pressure for communicative accuracy leads to non-monotonic languages evolving.

necessarily has intersecting meanings (as long as no meaning is empty), as shown in lemma 5. Second, a language with intersecting meanings is always suboptimal for communication.

That a language l has intersecting meanings means that there exist two meanings m_i, m_j in the range of l such that $m_i \cap m_j \neq \emptyset$. To see why a language with intersecting meanings cannot be optimal, notice that a consequence of the interpretation behaviour described above is that for any two meanings m_1 and m_2, if $|m_1| > |m_2|$ and the speaker and the hearer use the same language, then a communicative event in which m_2 is used has a higher chance of being successful than a communicative event in which m_1 is used, just in virtue of m_1 covering more degrees than m_2. If two languages l_1, l_2 are identical except for one signal which means m_1 in l_1 and m_2 in l_2, l_2 has higher communicative accuracy when speaking with itself than the l_1 when speaking with itself. Moreover, every language with two intersecting mean-

ings m_i, m_j (and possibly other non-intersecting meanings) can be transformed into a language with no intersecting meanings by substituting m_i with a new meaning $m_k = m_1 - (m_i \cap m_j)$. Notice that $|m_i| > |m_k|$, and that the two languages are identical modulo using m_i or m_k. A language with intersecting meanings can always be transformed into a communicatively more accurate language, and is therefore suboptimal. This has an effect on the model if the agents get enough data for their languages to be similar to their parents' languages, therefore approximating the accuracy of the language speaking with itself.

The fact that any system of three monotonic meanings is communicatively suboptimal given the assumptions of model 2 also shows that the evolutionary pressures behind monotonicity can be studied independently of where the threshold of applicability of adjectives in their bare use falls. Wherever the threshold is, three monotonic adjectives will be communicatively suboptimal and the model will give similar result.

In model 1, degenerate monotonic meanings prevail in virtue of their simplicity. In model 2, non-monotonic languages prevail in virtue of their greater accuracy. However, both models fail as an evolutionary account of adjectival monotonicity. The communicative suboptimality of completely monotonic languages is tied with the way that literal agents produce and understand signals. Since humans are not literal agents, I give up the simplistic agents in model 2 and in model 3 study the effects of implementing a more realistic model of communication that takes into account human pragmatic skills.

3.2.3 Model 3: Communicative pressure on pragmatic agents

The agents in models 1 and 2 are literal in the sense that they base their linguistic behaviour purely on the semantics of their language, without exploiting the additional information that comes from interacting with cooperative rather than merely truthful agents. As speakers, literal agents pick with uniform probability among the meanings compatible with the observed degree, without considering the hearer's reasoning process. As hearers, literal agents guess with uniform probability among the degrees compatible with the meaning of the received signal. In model 3, I implement RSA agents as discussed in section 3.1.2. In this model, agents cooperate with each other and take into account their being cooperative when determining their linguistic behaviour.

The result of implementing recursive mindreading is that the agents in the model

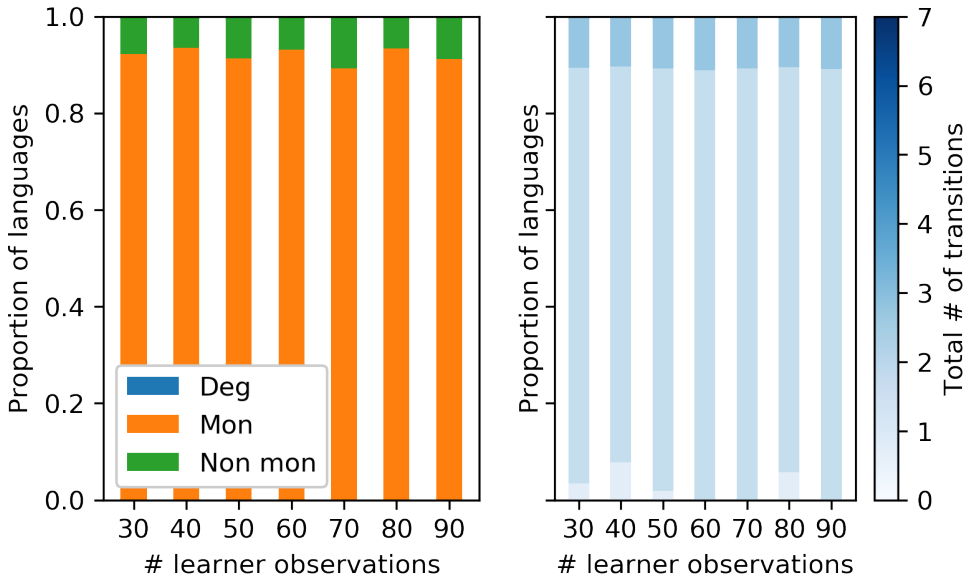


Figure 3.7: Left plot: Frequency of monotonicity types in model 3 over 50 runs of the simulation for each number of observations by learner. Right plot: languages in same data by total number of transitions. The conclusion from the third model is that once pragmatically skilful agents are introduced, monotonicity evolves often as a compromise between the pressure from the simplicity prior and communicative accuracy.

become capable of calculating the model's equivalent of real world implicatures. The listener can in general assume that the speaker has chosen the signal that maximizes the probability of communicative success among the ones compatible with the observed degree. Since the signal that maximises the chances of the listener guessing the correct degree is the one that covers the fewest degrees, after hearing a signal s the listener will not guess any degrees d compatible with s but also compatible with a signal h more specific than s, because if the speaker has observed d, they would have used h.

Results

Model 2 was run 50 times for 3000 generations with a population of 10 sample agents, for different numbers of observations by learners. The frequency of each language type being acquired by each generation after their learning phase was calculated, with a burn-in period of 1000 generations (figure 3.7). Most of the spoken languages are non-degenerate and monotonic.

Model 3 makes the correct prediction, namely that systems of adjectives evolve to be non-degenerate and monotonic. Implementing pragmatic skills, which gives artificial agents the ability to calculate scalar implicatures, allows agents to accommodate the prior to a greater extend than in model 2 without losing in terms of communicative accuracy. Monotonic, non-degenerate languages are the best trade-off between communicative and learnability pressure only if agents are pragmatically skilful.

3.3 Previous model of the evolution of monotonicity: Brochhagen et al (2018)

The models above provided an evolutionary account for the monotonicity property of scalar adjectives: monotone adjectival meanings constitute the best solution for learnability and communicative accuracy, under the assumption that language users are capable of pragmatic reasoning. Brochhagen, Franke, and van Rooij (2016) develop an account that is similar to mine but differs in some crucial respects. The meanings are less structured than in the models above. The structure of each meaning is a function of its relation to an upper bound; each meaning can cover what is below the upper bound, what is above, both, or neither, and is therefore encoded with two bits. This modelling choice has two consequences. The first is that there are no degenerate meanings in the sense used above. A meaning that is true for both states in Brochhagen et al. (2016) is not degenerate, but rather simply one that lacks an upper bound, e.g. the meaning of English "some". In the models above, I concluded that a simplicity pressure alone was insufficient because it resulted in degeneracy. On the other hand, Brochhagen et al. (2016) exclude a pressure for simplicity alone because it results in all the signals getting the same meaning, i.e. the monotonic meaning. The two models offer therefore different arguments for the

insufficiency of a simplicity pressure alone: avoidance of degeneracy in my model and of synonymy in Brochhagen et al. (2016).

Additionally, since Brochhagen et al. (2016) focus on a very general sense of scalarity, there is no obvious way to add structure to the model in a way that encompasses all the relevant semantic structures. As a consequence, the higher complexity of non-monotonic meanings and the size of this difference are stipulated in the model. While this is an explicit modelling choice to avoid introducing assumptions, it makes it harder to extend the measure of complexity beyond two signals. Since the model in Brochhagen et al. (2016) only has two states, there is not much need for an explicit functional form for complexity. One can simply specify how much more complex the non-monotonic meaning is than the monotonic one, and a great variety of complexity measures could fit the two picked complexity values for some parameters specification. On the other hand, calculating the complexity level of three or more meanings requires a decision about their relative complexity, ideally as a function of the differences in their semantic structure. Using scales to model the meaning structure of scalar adjectives allowed us to model the relations between the different meaning structures and their complexity. In sum, having a simple semantic model makes the model in (Brochhagen et al., 2016) suitable to discuss different cases of scalarity, but working with only two states and two signals implies that their model cannot detect differences between degeneracy and monotonicity. Conflating degeneracy and monotonicity is problematic, given that experimental work shows that under a pressure for learning only, people prefer degenerate systems (Kirby et al., 2015).

In a related and more recent paper, Brochhagen et al. (2018) narrow their focus to quantifiers, and provide an explicit measure of complexity based on the set-theoretic analysis of generalized quantification. In this paper, the semantic model has more complexity, and includes three states, i.e. a representation of meaning that is more similar to the one in the model presented in this chapter. However, other differences from the model above are introduced. Crucially, in Brochhagen et al. (2018) degenerate meanings are excluded from the set of possible meanings. The spread of degenerate languages was my reason for introducing a pressure for communicative accuracy. Since there is no degeneracy in Brochhagen et al. (2018), this argument is not available. Brochhagen et al. (2018) however have a different reason for introducing a communicative pressure in their model, namely to explain how the population converges to a single language.

A second difference between my and Brochhagen et al. (2018)'s paper is how communicative fitness is calculated. In Brochhagen et al. (2018) what matters is how well agents can speak with the other agents in the same population. Letting agents interact within their generation is useful when studying convergence to a single language. On the other hand, I calculate communicative accuracy between cultural parent and offspring, implementing communication in a more restricted manner (and allowing for more variability in the population). Calculating expected communicative accuracy of an agent with their cultural parent only is also computationally more efficient than calculating expected communicative accuracy of an agent with respect to all the other agents in their generation. Spike, Stadler, Kirby, and Smith (2017) compare various implementations of communication pressure in a computational model, and their results suggest that the direction of communication (horizontal vs. vertical) does not make a huge difference for obtaining a conventional signaling system. Overall, the two models may produce very similar results. However, how and whether the differences work out in the case of scalar meanings is still an open question.

3.4 Discussion

Above, I considered three combinations of the following parameters:

1. Whether agents have or not a bias for simple languages.
2. Whether there is or not a selection for communicatively successful languages.
3. Whether agents are literal or pragmatic.

It is worth considering the other combinations of parameters to determine what the contribution of each is. The results for other combinations of parameters across different number of observations for learners is shown in figure 3.8.

When literal agents are not biased for simple languages (plot a, b, c, d in figure 3.8), direct selection of communicatively successful languages determines a high proportion of non-monotonic languages, which, as argued above, can convey more information about the parent's observations.

The most interesting results occur when the model has pragmatic agents with a preference for monotonicity, but no selection for communicatively successful languages (plot e in figure 3.8). In this case, monotonicity evolves very often, to levels

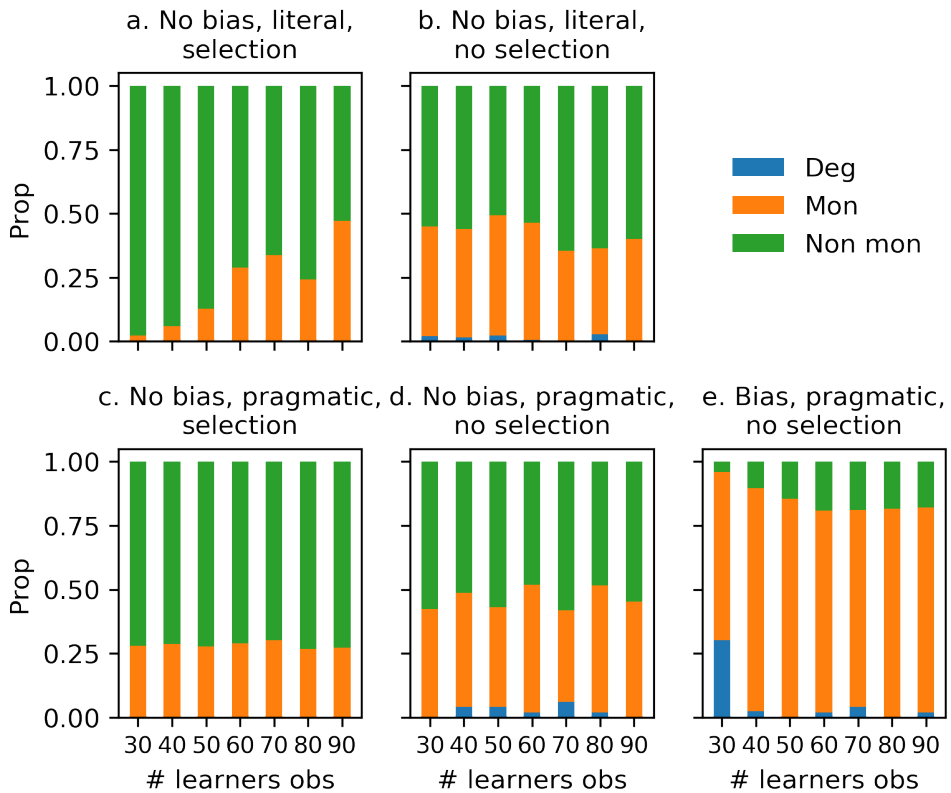


Figure 3.8: Proportion of languages spoken for other combinations of parameters. Agents can have or lack a bias for simple languages, they can be literal or pragmatic, and there can be or not be a selection for communicatively successful languages.

comparable with the third model above. This results is coherent with the results in Kirby et al. (2015), where a pressure for communicatively accurate languages was implemented by having pragmatic agents rather than direct selection of languages. Care is needed when interpreting this result. Without direct selection, the crucial effect of pragmatic agents is not to increase the proportion of communicatively successful (i.e. non-monotonic) languages irrespective of bias, but rather to avoid degenerate languages *when there is a bias*. Indeed, with no bias and no selection, pragmatic agents (plot d in figure 3.8) do not develop more non-monotonic languages than literal agents (plot b). Looking at the proportion of individual languages developed by pragmatic agents with and without direct selection for communicatively

successful languages shows that indeed almost degenerate languages often evolve in the latter, but not in the former (figure 3.9).

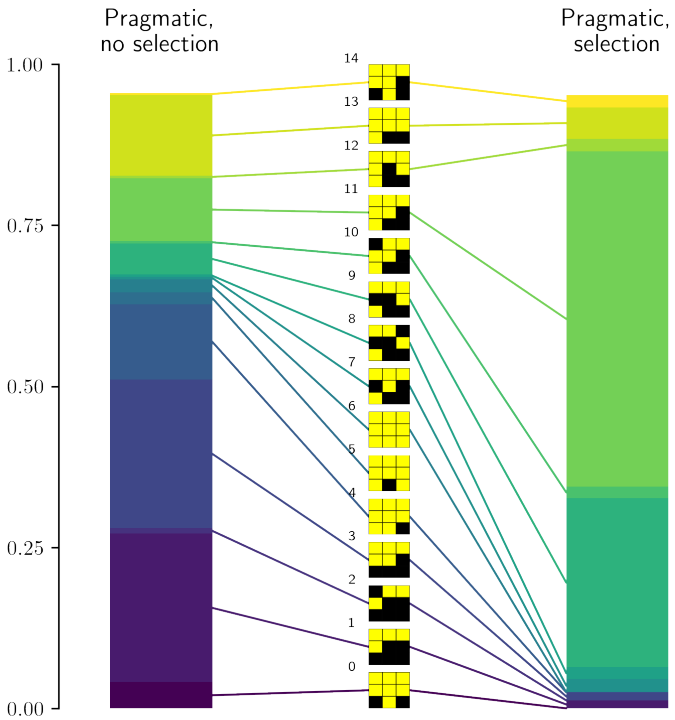


Figure 3.9: Most frequently unique spoken languages in the IL models. Languages are unique if they are not reducible to each other by shuffling the meanings across signals (i.e. shuffling the rows of the matrix representing the language) and reversing the order of the degrees. Each language is displayed as a 3 by 3 matrix as explained above in 3.16. Yellow corresponds to 1 and black to 0. I selected for each condition the 10 most frequently spoken languages (approximately 90% of the total for each condition). The central column shows the selected languages, while the lateral plots show the proportion of each of the languages in the condition with (right) or without (left) direct selection by communicative accuracy (both conditions have pragmatic agents and 40 observation by learners). While both conditions lead to a great majority of monotonic languages (Lang 0, 5, 7, 14 are non-monotonic, but relatively rare), the monotonic languages that evolve are different in the two conditions: languages with two degenerate meanings—i.e. languages 0, 1, 3, 4, 13—often evolve in the condition with pragmatic agents but without direct selection for communicatively successful languages (right plot), but are rare when selection is added (right plot).

Why do pragmatic agents prevent degenerate languages from evolving? The reason is that it is easier to distinguish between degenerate and non-degenerate languages for pragmatic agents than it is for literal agents. In other words, non-degenerate languages will be mistaken more often for degenerate languages by literal than by pragmatic agents, as can be seen in figure 3.10. However, the difference in distinguishability does not depend on how communicatively successful the languages is.

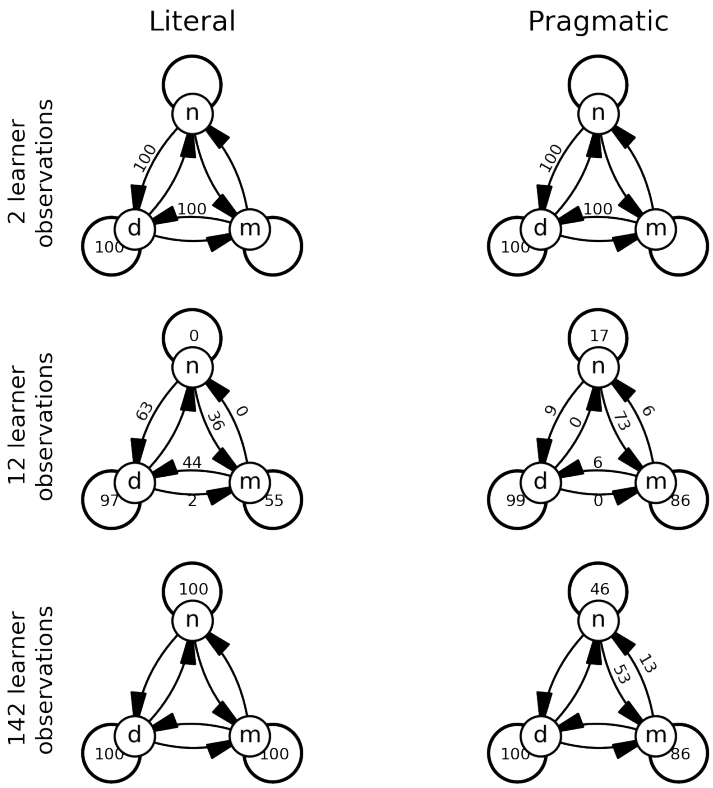


Figure 3.10: The plot shows the transition percentages between degenerate ('d'), monotonic ('m'), and non-monotonic ('n') languages. Top row: with few observations, both pragmatic and literal agents with a bias only learn degenerate languages. Central row: with more observations, literal agents mistake non-monotonic and monotonic for degenerate languages more often than literal agents. However, pragmatic agents mistake non-monotonic for monotonic more than literal agents. Bottom row: With even more observations, literal agents become capable of recognizing each language with high accuracy, but pragmatic agents still mistake non-monotonic for monotonic languages.

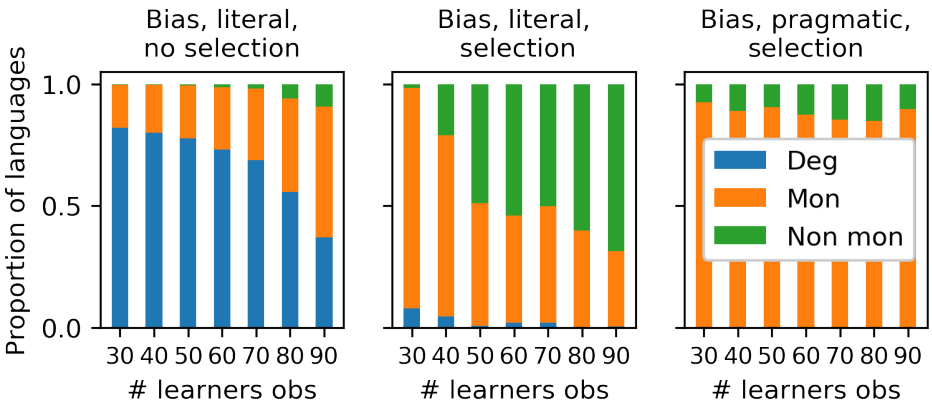


Figure 3.11: Results for the main three conditions with larger languages with 4 signals and 4 degrees. The results presented in the previous section for smaller languages are reproduced for the larger languages. IL with literal agents and bias alone leads to a majority of degenerate languages (left plot). Adding direct selection for communicative accuracy leads non-monotonic languages to emerge often (central plot). Adding pragmatic skills to selection based on communicative accuracy, however, leads to a large majority of monotonic languages (right plot).

Finally, the reason why non-monotonic languages do not evolve with pragmatic agents is that data produced by non-monotonic languages looks similar to data produced by monotonic languages, and therefore the likelihood function will be almost uniform (figure 3.10). The prior will then play a large role in determining the posterior. In sum, adding pragmatic skills makes the difference between agents speaking degenerate and non-degenerate languages greater, and the difference between agents speaking monotonic and non-monotonic languages smaller.

Until now, I have discussed the case of a language with three signals and three degrees. However, it is worth considering the languages that evolve in the three parameters regimes above when the scale contains more degrees and more signals are available. I ran the models with languages consisting of four signals, and with four degrees. Each combination of parameters was ran 50 times for 3000 generations. Figure 3.11 shows that the main results presented above are reproduced for the larger languages.

Larger languages are particularly interesting because they show how strategies to

keep meanings monotonic while preserving communicative accuracy combine. With this in mind, it is useful to observe which languages specifically evolve commonly in the IL condition with pragmatic agents and direct selection for communicative accuracy. In particular, two patterns emerge that I discussed in section 2.1.4. First, two non-overlapping signals of opposite polarity develop on top of a degenerate signal covering the whole space, in such a way that some degrees can only be expressed by the degenerate signal. This is an example of such a language:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

Language 6 in figure 3.12 is an example of such a language. Pragmatic agents will infer from the use of the degenerate signal that none of the other signals were available. If the degenerate signal is interpreted as silence, insofar as it is compatible with all possible world states, the central space that is left uncovered by all other signals can be interpreted as the zone of indifference discussed in 2.1.4. This corresponds to the intuition that a specific signal is not needed for the zone of indifference: when the sender is silent about a certain feature in a descriptive task, the receiver can imagine a degree that is not covered by all other non-silence signal.

The second patterns that emerges is two of the signals partially overlapping with each other, e.g. the second and third signals in the following language:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Language 15 in figure 3.12 is an example of this pattern. This type of language generates, as mentioned above, scalar implicatures for pragmatic users. For instance, when the second signal is used, a pragmatic receiver can infer that the sender probably observed the third degree. If the sender had observed degrees one or two, they would have send the third or fourth signal. This pattern corresponds to the existence, in systems of gradable adjectives, of multiple signals with the same polarity, such as “warm”, “hot”, and “boiling”. Both patterns can be recognized in the

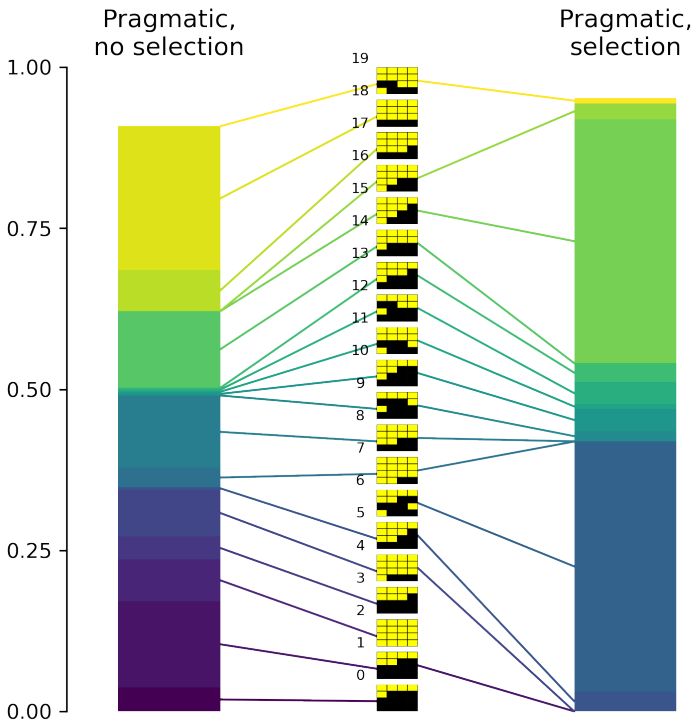


Figure 3.12: Most frequently unique spoken languages in the IL model, with larger languages. For a fuller explanation of the plot, see figure 3.9.

model as strategies to reach high communicative accuracy while retaining monotonic meanings, when language users are pragmatic.

While the models above deliver a picture of how monotonic but not degenerate signals might have evolved, various criticisms could be raised. A first problem is whether the set of possible meanings could be simplified. If the degenerate meanings are excluded from the set of possible meanings, the monotonic languages have the highest prior. Therefore, a model with IL alone naturally converges to monotonic languages. A critic could argue that the possibility of degenerate languages should be given up, remove them from the hypothesis space in the model and think of monotonicity as a direct consequence of simplicity. This move would simplify the

model without loss in explanatory power. However, this move is difficult to justify for at least two reasons. First, degenerate languages are clearly options in the space of adjectival meanings. While degenerate meanings do not give information about the degree to which an object has a property, they do give 1 bit of information saying whether an object has the property to any degree or not. Second, non expressive languages similar to what I defined as degenerate emerge in pure IL experimental conditions. This shows that degenerate meanings applying to all or to none of the objects are not only theoretically possible, but can also emerge in linguistic systems created by human agents when there is no pressure to make the meanings useful in communication. (Kirby, Cornish, & Smith, 2008).

A second objection that could be raised against the model concerns the selection of linguistic forms according to communicative accuracy. Often, it does not make a practical difference to communicate whether the degree to which an object has a property falls within the extension of the positive or the extreme adjective for that scale. Indeed, often scalar implicatures are not calculated for scalar adjectives (Doran, Baker, McNabb, Larson, & Ward, 2009), and there is substantial variation even among gradable adjectives with respect to the frequency of calculation of scalar implicatures (Van Tiel, Van Miltenburg, Zevakhina, & Geurts, 2016). A critic might remark that the mechanism presented in the present chapter, where scalar implicatures allow monotonicity without a loss in communicative accuracy, is therefore in general not acting to shape the semantics of gradable adjectives. Therefore, the communicative pressure in model 2 would be strong enough to prevent degenerate meanings, but not strong enough to overcome the preference for monotonic extensions. This argument could be further developed in a computational model, but I leave this to future work.

In this chapter, I developed an evolutionary account of monotonicity, and argued that the mechanism underlying the spreading of monotonicity rests on a combination three facts. Namely, monotonic meanings are simpler than non-monotonic meanings, language is shaped by both IL and a pressure for accurate communication, and human beings are capable of pragmatic reasoning.

The account of the evolution of monotonicity presented in this chapter makes some assumptions about cognition and communication. These assumptions make specific empirical predictions that can be operationalized, providing support for the models above. A crucial assumption of the models presented in this chapter, which was first introduced and argued for in chapter 2, is that scalar categories are coded

in terms of transitions rather than prototypes, and that therefore the structure of the conceptual domain determines which categories are simpler to code. In the next two chapters, I test empirically the picture I developed about the encoding of scalar categories.

Chapter 4

Testing a bias for monotonicity

4.1 Introduction

In the previous chapter, I discussed a model of the evolution of monotonicity in adjectival semantics. The model made a crucial cognitive assumption, namely that single bounded scalar categories are easier to learn than ($n \geq 2$)-bounded categories in scalar conceptual domains. This assumption was based on a picture of categorization in scalar conceptual domains presented in chapter 2. In this chapter, I present experimental work that attempts to test this account of categorization, and more specifically the hypothesis that the bias for monotonicity is be stronger in scalar domains than domains structured by metric relations. The general strategy I adopt for testing the hypothesis is as follows. First, participants are familiarized with a group of stimuli in two conditions. In one condition, the stimuli are framed as forming a scale. In the other condition, the focus is on the relations of similarity holding between the stimuli. Then, participants are told that one or more of the stimuli belong to an alien category. Finally, participants are asked to guess which other stimuli belong to the category. The structure of the category inferred by the participants can be used to make inferences about their learning biases.

The experiments in this chapter ultimately do not find support for the hypothesis. The lack of a significant result is compatible with the hypothesis being true, because the hypothesized differences in behaviour are difficult to measure. Therefore, in the next chapter I develop a more complex statistical and cognitive model and run more experiments to study the hypotheses of interest.

4.2 Experiment 1: Selection of individual stimuli

4.2.1 Materials and Methods

Participants

28 participants were recruited via Amazon's Mechanical Turk (AMT) crowdsourcing platform[1], a website where users (called *Workers*) can perform tasks (called *HITs*) in exchange for monetary compensation. Participant location was restricted to the United States. One participant was excluded because they said in the feedback that they could not use the interface properly, and one was excluded because they did not complete the experiment. All participants received 0.75 USD for participation in the experiment.

Materials

The experiment was coded in JavaScript and PHP and hosted in a server owned by the University of Edinburgh. Stimuli consisted of 36 images described as alien plants in the narrative frame presented to the participants. In order to keep the conditions as similar as possible and to avoid introducing confounds, the same stimuli are used for both conditions. The stimuli then have to satisfy the following conditions. First, the stimuli have to approximate a continuous change, since the inferred conceptual domain, whether a scale or a metric space, should be continuous. Second, the change cannot be more natural in one direction than the other, to avoid the participants from thinking about the stimuli as a scale when they are supposed to think in terms of similarity. Third, in this experiment the change is not one-dimensional. Instead, various features of the stimuli change, such as color and shape. Stimuli were drawn in the Scalable Vector Graphics (vga) format with the Inkscape software.[2] They are shown in figure 4.1.

Procedure

The experiment consists of two conditions, a *property* condition and a *similarity* condition. The experiment is visualized with a flowchart in figure D.1 of appendix D. 13 participants were allocated to the similarity condition and 15 were allocated

[1] `https://www.mturk.com`
[2] `https://inkscape.org`

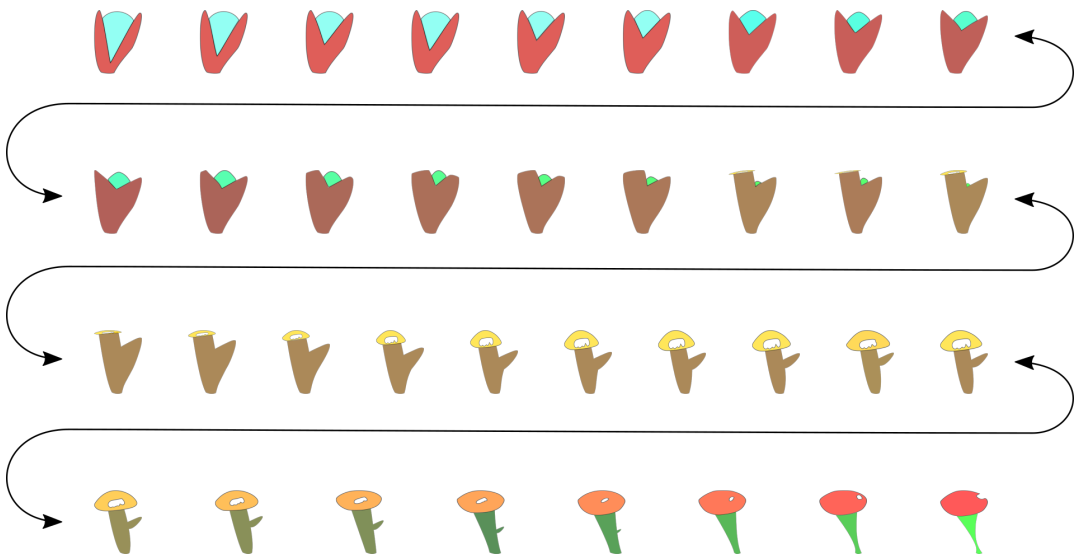


Figure 4.1: Stimuli for the first experiment, arranged in four rows for ease of visualization. The stimuli consist of 36 images of alien plants. The stimuli change continuously and neither direction of change is more natural than the other.

to the property condition. Each condition has two phases, a *familiarization* and a *testing* phase, plus a familiarization form at the beginning and a feedback form at the end. The narrative frame and training were different in the two conditions, but the testing phase was the same in each condition.

The familiarization phase in the property condition starts with a screen giving the narrative frame of the experiment:

> You just landed on an alien world. You are going to see a series of plants from that world. These plants take different shapes depending on the concentration in the environment of blagardium, an alien chemical. Try to learn how different amounts of radiation impact the shape that the plants took. You will be shown pairs of plants and asked which one develops with the highest amount of blagardium.

The second screen of the familiarization phase asks participants to familiarize themselves with the stimuli, which are displayed on the screen. Since the stimuli are designed to lack an intrinsic direction of increase, their order is flipped randomly

across participants. An arrow below the stimuli indicates the direction of blagardium increase. The arrow always points to the right.

The 3rd to 18th screen (a total of 15 screens) of the property condition train the participants to think about the stimuli in a way appropriate to the condition. Each screen shows two stimuli, and the participant is instructed to click on the plant with the highest amount of blagardium. The whole set of stimuli with bottom arrow is visible throughout at the bottom of the screen. The two stimuli are picked with uniform probability and without repetitions. Participants got feedback for every trial, and incorrect trials were repeated until correct

The similarity condition is mostly identical to the property condition. The familiarization phase in the similarity condition is:

> You just landed on an alien world. You are going to see a series of plants from that world. These plants can take different shapes. Try to learn what shapes the plants can take and how similar different shapes are. You will be shown two pairs of plants. You will be asked which of the two pairs has the most similar plants.

In the second screen, the only difference is the absence of an arrow below the stimuli.

The 3rd to 18th screen in the similarity condition train the participants to think about the stimuli as being structured by an order. The arrow below the stimuli at the bottom of the screen is missing. Instead of two stimuli, participants are presented with two pairs of stimuli, and they are instructed to select the pair with the most similar plants. To avoid the two pairs having equally similar stimuli, one pair is picked so that the two stimuli are at a distance of between 0 and 2 (included) stimuli, and the other pair at a distance of 6 or more stimuli.

The testing phase is identical in the two conditions and consists of three screens. In the first screen, the participant receives the following instruction:

> The aliens speak an alien language, and they have words to talk about the plants you observed earlier. Now you are going to observe plants that the aliens call ‘meeb’. You will be asked to select the plants that you think the aliens use the word ‘meeb’ for.

In the second screen, one of the stimuli is selected as belonging to the “meeb” category. The pre-categorized stimulus is chosen randomly from the stimuli that are more than 10 stimuli away from the border. The participant is asked to select the

other plants that they think also belong to the alien category. Selection happens by clicking and dragging over the plants. The plants selected by the participant are surrounded by a black box.

4.2.2 Results and discussion

For each participant, the plants selected as belonging to the 'meeb' category in the testing phase were recorded. Participants behaved in substantially the same way in the two conditions (see fig 4.2). In each condition, two participants picked the monotonic category and all other participants the non-monotonic category. Therefore, the H_0 cannot be rejected based on the data from experiment 1. However, there are reasons to be sceptical of the data from experiment 1. A number of participants only selected one stimulus, that was moreover different from the one shown as categorized (fig 4.2). Since this choice is inconsistent with the given instructions, these participants either used the interface incorrectly or they did not understand the instructions. Third, monotonic categories are the largest categories, which makes them extreme in a certain respect. Participant might have a resistance to picking monotonic categories in virtue of their extremeness. While the preference for monotonicity might still cause a difference in the behaviour across the two conditions, the difference would not show up simply a difference in the number of monotonic categories picked in the two conditions. A second reason to look at the results of experiment 1 with scepticism is the fact that 'meeb' looks more like a noun than an adjective. As discussed in chapter 2, the grammatical class might affect participants' expectations about the structure of the category. In the new design, I try to move away from expectations participants might have on the basis of the form of the word. Experiment 2 is designed to solve these problems through a more intuitive interface.

4.3 Experiment 2: Choice among scalar categories

Experiment 1 tested the participants' bias for monotonicity by having them select the stimuli that they thought belong to an alien category. In experiment 2, participants do not select the plants directly. Instead, they are give a forced choice task, choosing between a monotone and a non-monotone category.

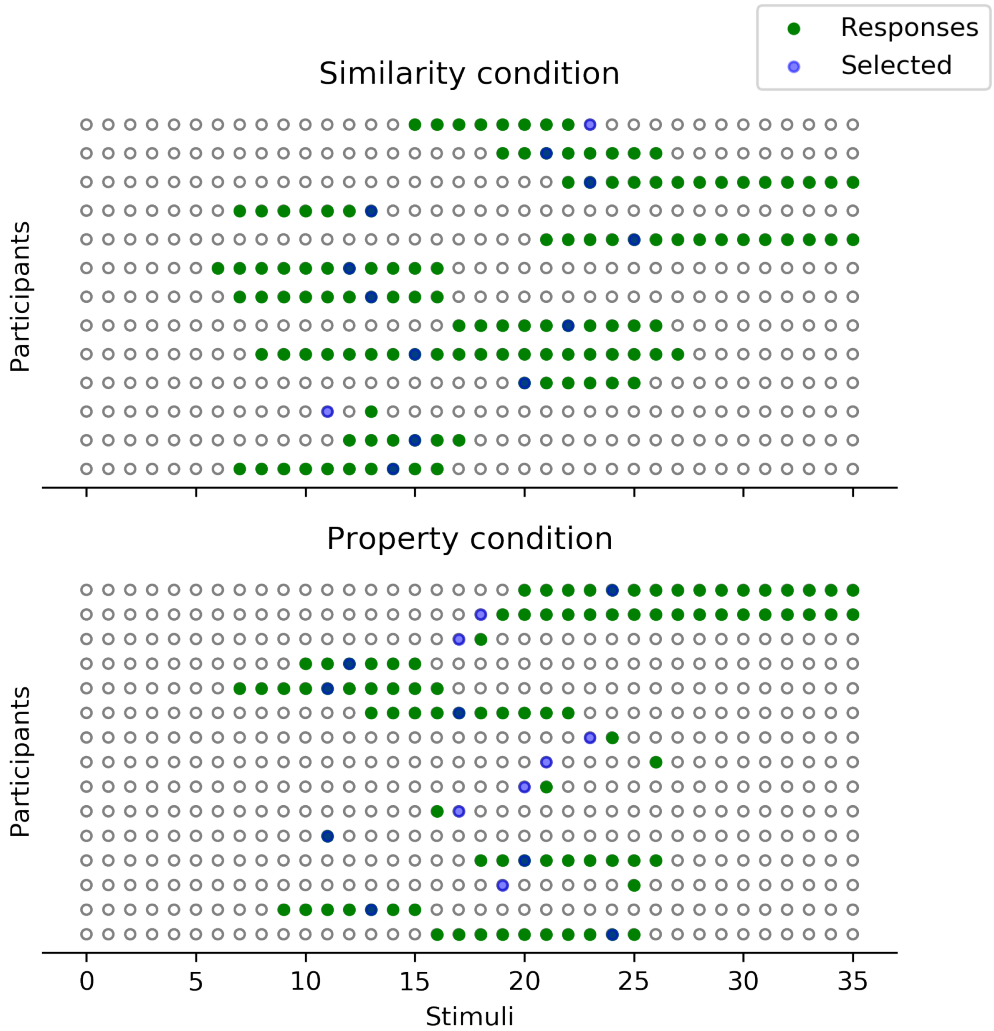


Figure 4.2: The plot shows participants' responses for experiment 1. For each participant, the figure shows the stimulus that was presented as belonging to the category (*Selected*) in translucent blue (transparency $\alpha = 0.5$), as well as the stimuli selected by the participants (*Responses*) in green. The order of the stimuli was randomly flipped across participants, but the stimuli are brought back to the same order in the plot.

4.3.1 Materials and Methods

Participants

82 participants (41 per condition) were recruited via AMT. The same restrictions, payments and AMT description apply as with experiment 1.

Materials

The stimuli are identical to the stimuli in experiment 1 (See fig 4.1).

Procedure

The training phase in the similarity condition is identical to experiment 1, except for small changes in the instruction which can be seen in figure D.2. The testing phase is quite different from experiment 1. The novel alien term 'meeb' becomes '  '. The reason for using a word written in unknown characters is that it would not afford a reading of the word as a noun or an adjective.

In the second screen of the testing phase, participants are presented with one stimulus that they are told belongs to the category, as well as two possible categories, and are asked to identify one of the categories as the alien category by clicking on the corresponding row of stimuli. The categories are picked randomly (see algorithm 2 in the appendix for more details).

4.3.2 Results

In the similarity condition, more participants picked monotonic categories than non-monotonic categories. Moreover, more participants picked the monotone category in the property than in the similarity condition. Fig 4.3 shows the pre-categorized stimuli seen by each participant and the category they picked, and fig 4.4 summarizes the data.

The data was analysed with a simple logistic model, regressing the participant's pick of monotonic category (a boolean variable) on the condition (a categorical variable with two levels, *property* and *similarity*). The monotonic variable is boolean, and the condition variable has values *similarity* (the intercept) and *property*. 25 participants out of 41 picked the monotone category in the similarity condition, while this rose to 27 out of 41 in the property condition. The regression shows that this

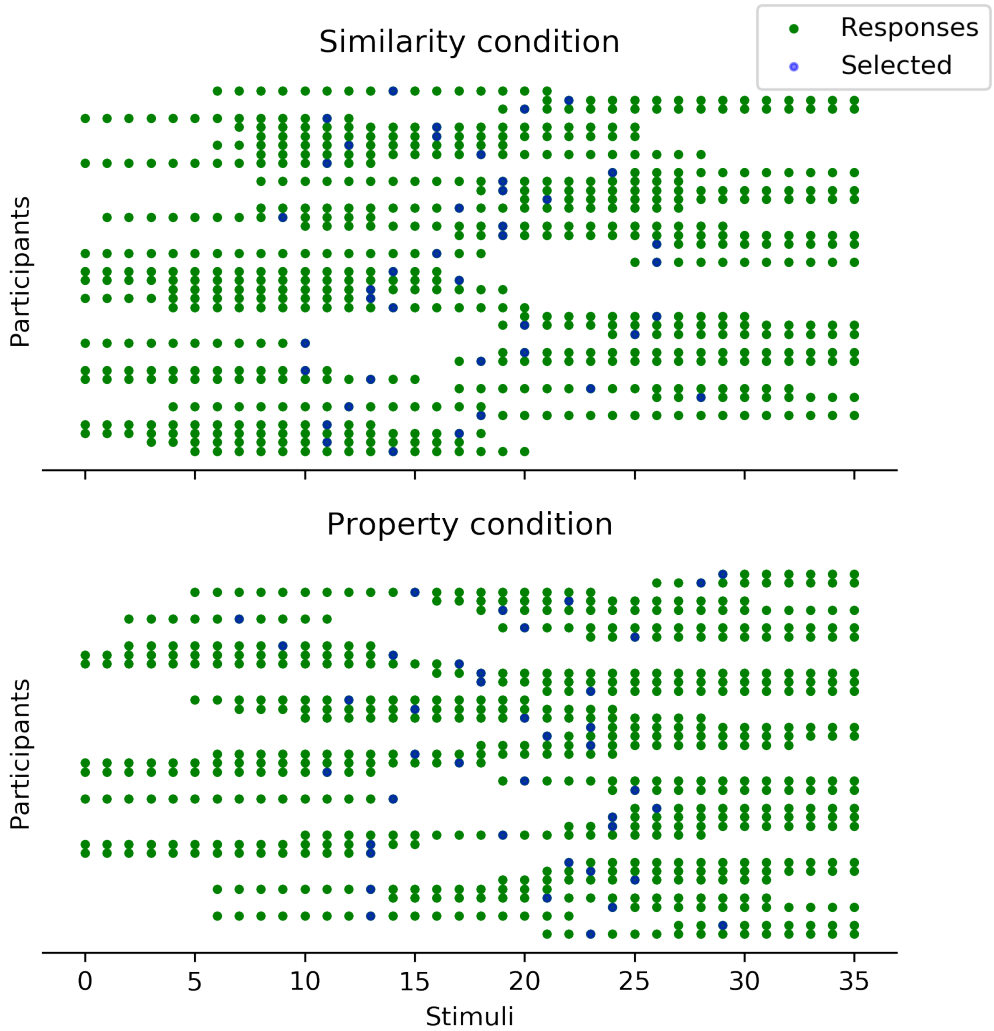


Figure 4.3: The plot shows participants' responses for experiment 2. For each participant, the figure shows the stimulus that was presented as belonging to the category (*Selected*) in translucent blue (transparency $\alpha = 0.5$), as well as the stimuli selected by the participants (*Responses*) in green. The order of the stimuli was randomly flipped across participants, but the stimuli are brought back to the same order in the plot. The monotonic categories are the ones that go up to the maximum (e.g. second row in plot of similarity condition) or the minimum (e.g. last row in plot of property condition) of the scale.

increase in number of participants choosing the monotonic category in the property condition is not significantly different from no increase (See 4.1). More specifically, the probability of the observed or a more extreme increase, under the assumption that there is no difference between the two conditions (p-value), is ≈ 0.647. The estimated log-odds of picking the monotonic category in the similarity condition is 0.4463, and therefore the estimated probability is $\frac{\exp(0.4463)}{1+\exp(0.4463)} \approx 0.61$, which is the proportion of participants that picked the monotone category in the similarity condition. The estimated increase in log-odds when going to the property condition is 0.21, and therefore the total predicted probability of picking the monotone category in the property condition is $\frac{\exp(0.4463+0.21)}{1+\exp(0.4463+0.21)} \approx 0.658$, which is the proportion of participants that picked the monotone category in the property condition.

Table 4.1: Regression for experiment 2. SEs are shown in brackets below the estimated parameters. The slope parameter (*Property* line) did not reach statistical significance.

	Dependent variable:
	monotonic
Property	0.210
	(0.459)
Constant	0.446
	(0.320)
Observations	82
Log Likelihood	−53.745
Akaike Inf. Crit.	111.490
Note:	*p<0.1; **p<0.05; ***p<0.01

4.3.3 Discussion

No clear pattern emerged in the way that participants pick monotonic categories. A post-hoc analysis shows that, surprisingly, the distance of the categorized stimulus from the center seems to have little effect on the participants' choice (right plot in fig 4.4). The model

```
monotonic ~ condition + distance.from.center
```

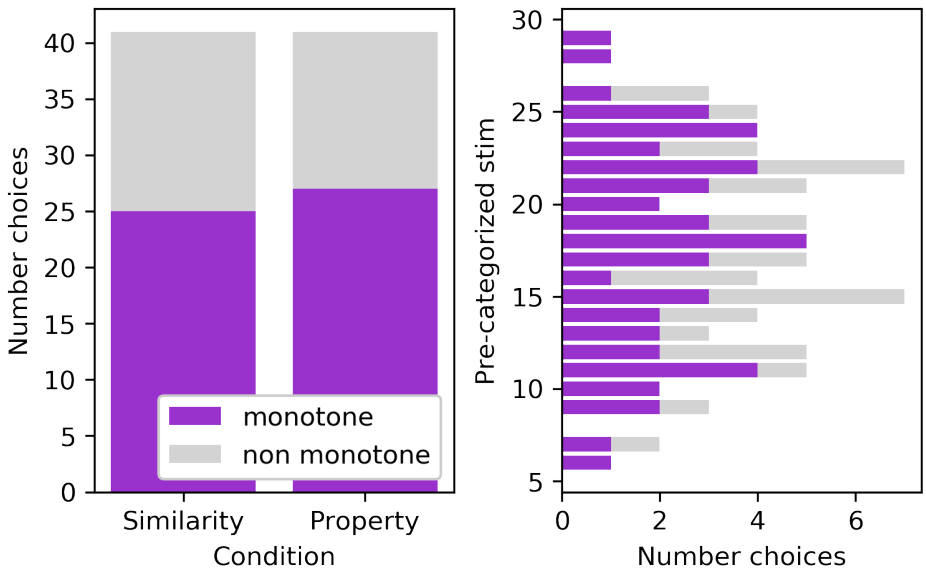


Figure 4.4: Summary plots for experiment 2. The left plot shows the number of participants that picked monotonic and non-monotonic by condition. The right plot shows the same information broken down by position of the pre-categorized stimulus. The closeness of the pre-categorized stimulus to the border does not have a large impact on the guessed category.

shows that the distance from center was not a significant predictor of the choice of category. This means that participants were not more likely to pick the monotone category if the observed stimulus was close to the extremes of the scale.

Participants did not behave significantly differently in the two conditions. One possible reason for this is that a conceptual space with a scalar structure does not in fact induce a preference for monotonic categories. Another possible reason for why conditions were not significantly different is that the experiment failed at producing different mental encoding of the stimuli set. While there is substantial change of the stimuli across the scale, arguably no plant stands out as a potential prototype. Since the hypothesis was that the preference for convex category comes from thinking in terms of prototypes, in experiment 3 we increase the affordance for thinking in terms of prototypes by changing the set of stimuli.

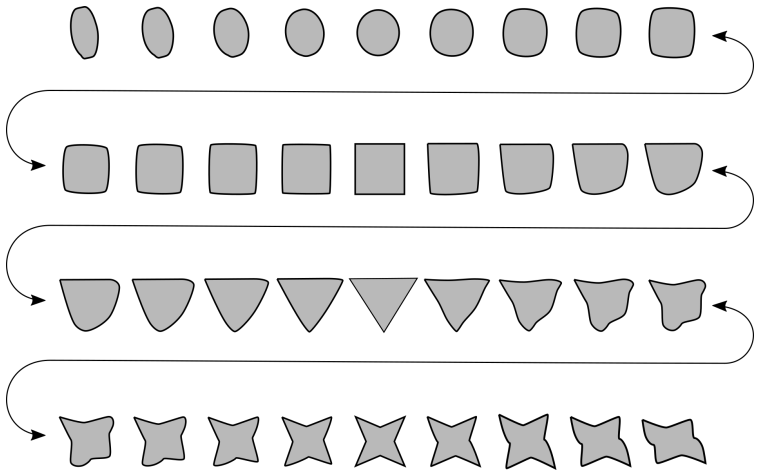


Figure 4.5: Stimuli for experiment 3, arranged in four rows for ease of visualization. The stimuli consist of 36 images of alien plants. Like in the first experiment, the stimuli change continuously but no direction of chance is more natural than the other. The stimuli have some perfect shapes (central column).

4.4 Experiment 3: Affordance for prototypical interpretation

The stimuli set used in experiments 1 and 2 did not afford a prototype based interpretation. Therefore, in experiment 3 we introduce a new set of stimuli (fig 4.5). The new stimuli are simple geometrical shapes morphing into each other in a series. The crucial difference with respect to the stimuli in fig 4.1 is that geometric shapes afford a prototype structure. In particular, the stimuli for experiment 3 include a circle, a square, a triangle, and an isotoxal square star.

4.4.1 Materials and Methods

Participants

40 participants were run for each condition. The same restrictions and MT description apply as in experiment 2.

Procedure

While the design is mostly identical to experiment 2, there are a few differences. The alien word ⍓ ⍀ ⊙ ⍀ is substituted by  . This is because the geometric shapes in ⍓ ⍀ ⊙ ⍀ afford analogies with the geometric shapes in the new stimuli set. For the same reason, in the testing phase the expression "red rectangle" becomes "red frame". Lastly, the way categories are picked is different (see algorithm 3 for details).

4.4.2 Results

Fig 4.6 shows for each participant which stimulus they saw as pre-categorized and which of the possible categories they picked. Fig 4.7 shows the data in a more synthetic form. First of all, more participants picked the non-monotonic categories. This indicates that the change in stimuli produced, to some extent, the intended effect. We ran a simple logistic model, regressing the participant's choice of monotonic or non-monotonic category on the condition. Results are shown in table 4.2. The model confirms the pattern with a significant negative intercept ($p = 0.001$)[3] in table 4.2. Secondly, the difference between the two conditions increased, from 0.21 in experiment 2 to 0.726 in experiment 3.[4] Despite the increased difference, the property condition is not significantly different from the similarity condition in experiment 3 ($p = 0.146$).

Like in experiment 2, a model including a random intercept by the distance of the categorized stimulus from the center is singular, indicating that the distance from the center does not explain the data better than noise.

4.4.3 Discussion

Participants in experiment 2 did not behave significantly differently in the two conditions. However, changing the stimuli had an effect in the intended direction. This is a first indication that participants are sensitive to the affordances of different types of stimuli, and that the design of the experiment might be incapable of detecting the relevant difference in participant behaviour between the conditions. More specifically, a worry about the first three experiments is that they can only detect a large,

[3]I report p-values only up to three decimal points because the amount of data makes more precise estimations meaningless.

[4]Note that the magnitude of this difference is hard to interpret directly because it is in the logit space.

Table 4.2: Regression for experiment 3.

	Dependent variable:
	monotonic
Property	0.726
	(0.500)
Constant	-1.237^{***}
	(0.379)
Observations	80
Log Likelihood	-47.789
Akaike Inf. Crit.	99.578
Note:	*p<0.1; **p<0.05; ***p<0.01

qualitative difference in the behaviour of participants, i.e. the choice between fully monotone and non-monotone categories. However, a preference for monotone categories might have subtler effects on behaviour, for instance determining preferences among the non-monotonic categories.

In the following chapter, I develop a Bayesian model to explore the effects of a preference for monotonicity on categorization behaviour. Having an explicit model of categorization will allow the analysis of more complex categorization data, and therefore of subtler patterns in the behaviour of participants.

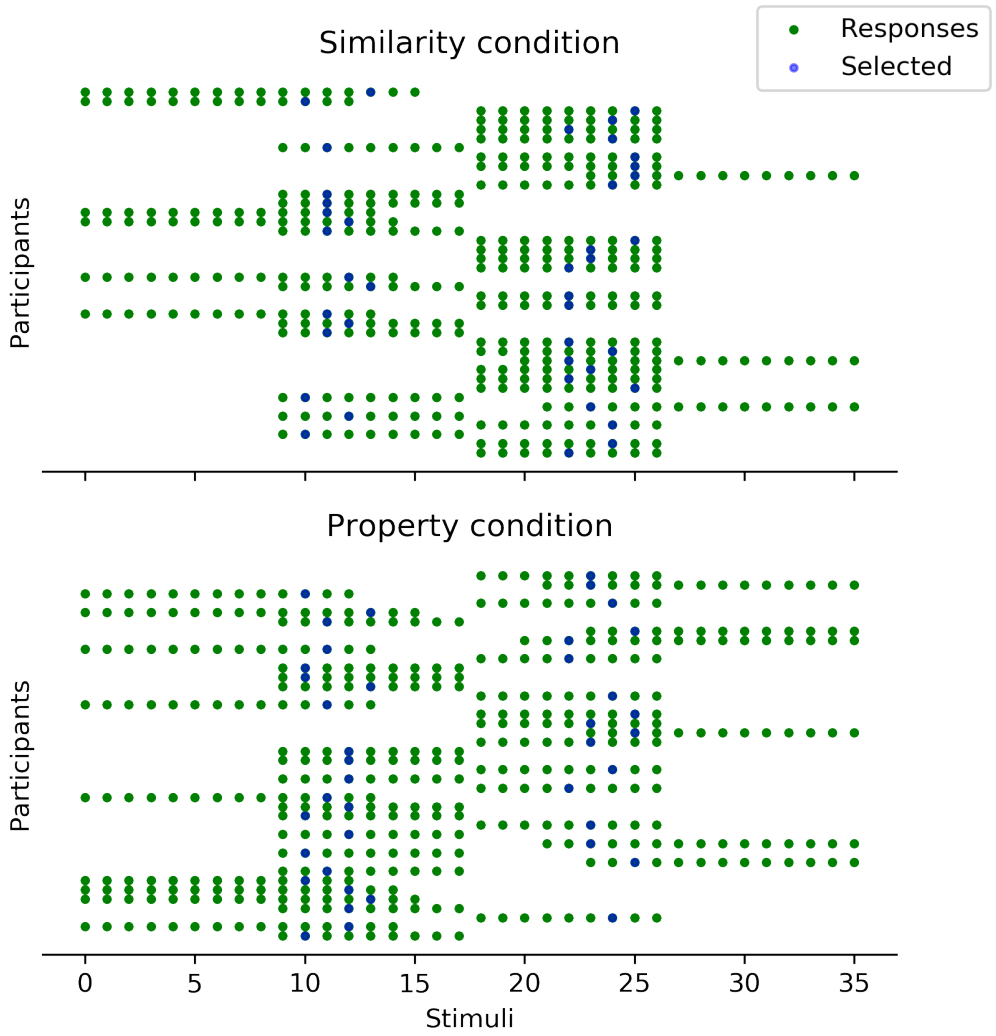


Figure 4.6: The plot shows participants' responses for experiment 3. For each participant, the figure shows the stimulus that was presented as belonging to the category (*Selected*) in translucent blue (transparency $\alpha = 0.5$), as well as the stimuli selected by the participants (*Responses*) in green. The order of the stimuli was randomly flipped across participants, but the stimuli are brought back to the same order in the plot. The monotonic categories are the ones that go up to the maximum (e.g. second row in plot of similarity condition) or the minimum (e.g. last row in plot of property condition) of the scale. No evidence was found that participants were more likely to pick the monotonic category in the property condition than in the similarity condition.

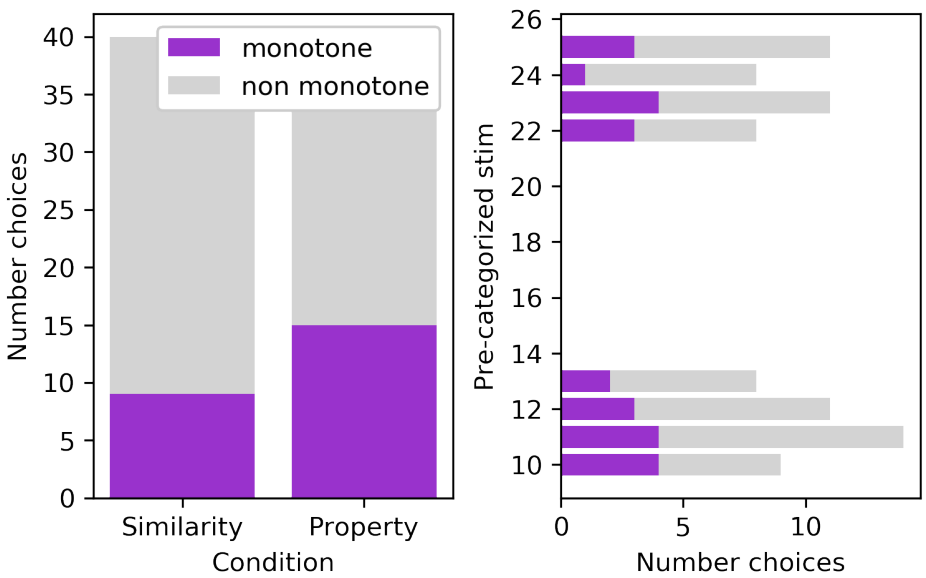


Figure 4.7: Summary plots for experiment 3. The left plot shows the number of participant that picked monotonic and non-monotonic by condition. The right plot shows the same information by position of the pre-categorized stimulus.

Chapter 5

Bayesian modelling and more experiments

In the previous chapter, I discussed three experiments aiming to test the hypothesis that thinking in a scalar way induces a preference for monotonicity. The data did not support the hypothesis. However, it prompted the realization that even if the hypothesis is true, the resulting pattern in the data might be subtler than the statistical model used above is capable of detecting. Therefore, in this chapter I present a more sophisticated cognitive Bayesian model of categorization (section 5.1), and then wrap around it a statistical Bayesian model capable of fitting experimental data (section 5.2). Finally, I present three experiments and the results of model fitting on the data (sections 5.3, 5.4, and 5.5). While the experimental data does not support the hypothesis, the process of developing a model and analysing the data deepens our understanding of what predictions the theory is making exactly, and what is difficult about testing them.

5.1 A cognitive model of categorization

The aim of this section is to define a family H of probability densities h modelling, at the computational level of analysis, the portion of the participants' cognitive apparatus causing the behaviour that is measured in the experiment. Once a set of parameters modelling cognitive variables and observations are specified, H produces a distribution over possible behaviours. The next section embeds H in a Bayesian model that can be used to fit experimental data directly.

H models the following situation. An individual π_i observes a set Σ_i of n stimuli: $\{\sigma_{i,j}\}_{j=1}^{n} \subseteq \Sigma$, where Σ is the set of all stimuli in the experiment. The stimuli in Σ_i are presented to π_i as belonging to some unknown category c. Based on Σ_i, π_i calculates a posterior probability distribution $f(\kappa|\Sigma_i)$ over possible categories $\kappa \in K$. The distribution $f(\kappa|\Sigma_i)$ encodes the probability that π_i attributes to each possible category being c, after having observed the stimuli Σ_i sampled from c. Crucially, f does not have to depend only on Σ_i, but can also depend on π_i's cognitive biases. Ultimately, π_i's biases are what the model will try to estimate. Finally, π_i uses $f(\kappa|\Sigma_i)$ to guess whether other stimuli also belong to c. More specifically, she accepts a stimulus σ as belonging to the same category that produced Σ_i with a probability $h(\sigma \mid \Sigma_i)$, which I call the *acceptance* probability.

Various aspects of the model need to be specified. First of all, what belongs to the set K of possible categories. Secondly, how π_i finds the posterior distribution $f(\kappa|\Sigma_i)$. Lastly, how $f(\kappa|\Sigma_i)$ determines the π_i's behaviour as encoded in $h(\sigma \mid \Sigma_i)$. I address these three questions in turn.

5.1.1 Set of categories

The first question is how to define the set of possible categories K considered by π_i. I assume that π_i knows the underlying set of possible stimuli Σ. In the case of the experiment, the participants know what the stimuli are because they are always visible on screen. Σ is not an unstructured set, but rather a finite totally ordered set $\{\sigma_i\}_{i=1}^{|\Sigma|}$, where $\sigma_i \leq_\Sigma \sigma_j$ iff $i \leq j$.[1] The order relation reflects the fact that while there is a discrete number of experimental stimuli, the stimuli are designed to show an underlying continuous, gradual change. While the is no natural direction to this gradual change, so as not to strongly afford a scalar interpretation, a total order can encode which stimuli are closer together with respect to the change.

The categories are sets of stimuli, i.e. subsets of Σ. One natural way of defining the set of all possible categories K is as the set of subsets of Σ, the powerset $\mathcal{P}(\Sigma)$. Any combination of stimuli would then count as a category. This way of defining K has the advantage of being natural and not making assumptions about the category that participants entertain. However, having $\mathcal{P}(\Sigma)$ as the set of categories poses a computational problem. The number of possible categories increases exponentially with the number of stimuli, as there are $2^{|\Sigma|}$ categories. This would make fitting

[1] I omit the subscript of the order relation when it is clear from the context.

experimental data to the Bayesian model impractical. A solution to this problem is to choose a smaller set of categories. This is not as problematic as it might first seem, as it is a priori plausible that some categories are not considered by participants. Examples of likely ignored categories include the category of alternate stimuli $\{\sigma_i \in \Sigma \mid i \mod 2 = 0\}$ and the category of extremes $\{\sigma_1, \sigma_{|\Sigma|}\}$. Based on this consideration, I assume that not all possible sets of stimuli need to be considered by the model, i.e. $K \subset \mathcal{P}(\Sigma)$.

Exactly which subset of $\mathcal{P}(\Sigma)$ can be assumed to be considered by the participant? In line with previous literature[2], I define K as the set of convex sets of stimuli. Convexity here is defined as follows:

Definition 9. *A subset B of a totally ordered set $\Omega = (O, \leq_\Omega)$ is **convex** iff for all $b_1, b_2, b_3 \in \Omega$ if $b_1, b_3 \in B$ and $b_1 \leq_\Omega b_2 \leq_\Omega b_3$ then $b_2 \in B$.*

Note that this definition exploits the choice above to make Σ a totally ordered set. The set of convex categories contains $\frac{1}{2}(|\Sigma|+1)|\Sigma|$ categories.[3] While the number of all categories increases exponentially with $|\Sigma|$, the number of convex categories only increases quadratically, which solves the computational problem mentioned above.

A final advantage of choosing convexity as the criterion for including a category in K is that the set of convex categories contains all monotonic categories, and can therefore be used to test a preference for monotonicity. To show that the set of convex categories contains all monotonic categories, it needs to be shown that monotonicity implies convexity. Assume ad absurdum that a category κ is monotone but not convex. Then there are three stimuli b_1, b_2, b_3 such that (1) $b_1, b_2 \in \kappa$, (2) $b_1 \leq b_2 \leq b_3$, and (3) $b_2 \notin \kappa$. κ is either monotone increasing or decreasing. If it

[2]See the discussion of convexity in conceptual spaces from chapter 2.

[3]To see why, note that each non-empty convex category can be defined as the set of points between and including its two extremes σ_i and σ_j (with $i, j \leq |\Sigma|$), in two equivalent ways: (σ_i, σ_j) and (σ_j, σ_i). The case where $i = j$ is the case of a category consisting of a single stimulus. Therefore, there are as many non-empty convex categories as combinations of two elements of Σ with repetitions:

$$\begin{aligned}\binom{|\Sigma|+2-1}{2} &= \frac{(|\Sigma|+1)!}{2(|\Sigma|-1)!} \\ &= \frac{(|\Sigma|+1)|\Sigma|(|\Sigma|-1)!}{2(|\Sigma|-1)!} \\ &= \frac{(|\Sigma|+1)|\Sigma|}{2}\end{aligned}$$

where $\binom{n+r-1}{r}$ is the number of r-sized sets out of n elements with repetitions, and I have used the equivalence $n! = n(n-1)!$.

is increasing, since $b_1 \in \kappa$ and $b_1 \leq b_2$, $b_2 \in \kappa$. Since this contradicts assumption (3), κ must be monotone decreasing. Then, since $b_3 \in \kappa$ and $b_2 \leq b_3$, $b_2 \in \kappa$. This generates a contradiction. Therefore, κ must be convex. The left plot of figure 5.1 shows the convex categories for a set of 7 stimuli.

5.1.2 Posterior distribution over categories

As discussed above, $f(\kappa|\Sigma_i)$ is the probability distribution over categories of agent π_i after observing a set of stimuli Σ_i sampled from κ. I assume that π_i is a Bayesian agent and calculates[4] the posterior probability by applying Bayes' theorem:

$$f(\kappa \mid \Sigma_i) = \frac{p(\Sigma_i \mid \kappa)p(\kappa)}{\sum_{\kappa_j \in K} p(\Sigma_i \mid \kappa_j)p(\kappa_j)} \tag{5.1}$$

Where $p(\Sigma_i \mid \kappa)$, the *likelihood*, is a conditional distribution over observations Σ_i given that the observed stimuli are being sampled from a category κ, and $p(\kappa)$, the *prior*, is the probability that π_i attributes to a category κ before observing any data. I explore prior and likelihood in turn.

Agent's prior

Some categories might be preferred by π_i over other categories, independently of the observed data. These preferences, which I will call *biases*, can be modelled as parameters that increase the prior probability of certain hypotheses over others. The biases of π_i might in principle favour any aspect of the category to be guessed. However, here I focus on two biases. Firstly, a bias PM_i (Preference for Monotonic) for monotonicity, which is the parameter that I ultimately want to estimate from the experimental data. Secondly, a bias PL_i (Preference for Large) for larger or smaller categories than would be expected by likelihood alone.

The preference for monotonic categories is modelled because the present aim is ultimately to evaluate the parameter in the two experimental conditions. On the other hand, the reason for adding a preference for large categories is slightly subtler. As we will see, both PM and PL affect the size of the categories that agents tend to infer. Assuming no PL is equivalent, as we will see, to assuming a fixed value of PL

[4]The model in this section is meant as a computational, rather than algorithmic level theory, in Marr, Ullman, and Poggio (2010)'s levels of analysis. I do not make claims about how the quantities in this section are computed or implemented.

for all agents. If different agents have in fact varying preferences for category size, a model with PM but without PL would attempt to explain all variation across agents in terms of PM alone. As we will see below, this might introduce a bias in the estimation of PM. To prevent this from happening, I model PL explicitly. Other aspects of agents' cognition might influence behaviour in a way that biases the estimation of PM, a problem to which we return below.

The prior of participant π_i is then $p(\kappa) = p(\kappa|PM_i, PL_i)$. I assume that the two biases act independently on the preference for a category, and I factorize the prior as the (renormalized) product of two density functions:

$$p(\kappa) \propto p(\kappa|PM_i)p(\kappa|PL_i) \tag{5.2}$$

Note that if a category is certain according to one bias, it will be certain for π_i independently of the the (dis)preference coming from the other bias. To see why, consider that if $p(\kappa_j|PM_i) = 1$, then $p(\kappa_{l\neq j}|PM_i) = 0$, which means as just shown that every category but κ_j will be a priori impossible for π_i, leaving κ_j as the only option. This is under the assumption that the agent does not have contradictory biases, i.e. assuming that $p(\kappa_j \mid PL_i)$ is not also 0.

PM_i, the preference for monotonic categories of agent i, is a single parameter bounded in $[0, 1]$. PM_i defines a distribution over categories as follows:

$$p(\kappa|PM_i = \alpha) \propto \alpha \mathbb{1}_{\text{mon}}(\kappa) + (1 - \alpha)(1 - \mathbb{1}_{\text{mon}}(\kappa)) \tag{5.3}$$

where $\mathbb{1}_{\text{mon}}$ is the indicator function which equals 1 when its argument is a monotone category and 0 otherwise.[5] It is worth noting some intuitive features of this way of modelling the bias for monotonicity. First, the bias for monotonic categories increases strictly monotonically as PM_i increases. This ensures that no two values of PM_i convey the same information about the preference for monotonicity. Second, when $PM_i = 0$, monotonic categories are a priori impossible. Third, when $PM_i = 1$, it is a priori certain that the true category will be monotonic. Finally, when $PM_i = 0.5$, π_i has no a priori preference for either monotonic or non-monotonic categories.

The second bias I implement in the model is the bias for category sizes, PL_i. The size of a category is the number of stimuli it contains. PL_i is defined in $(-\infty, \infty)$,

[5]In the following, I use the $\mathbb{1}$ notation in two more ways. When p is an expression that is true or false, $\mathbb{1}p$ is 1 if p is false and 0 otherwise. If c is a set, $\mathbb{1}_c(a)$ is 1 if $a \in c$ and 0 otherwise. The context will make it clear how $\mathbb{1}$ is meant in each case.

and influences the prior probability of categories as follows:

$$p(\kappa|PL_j = \beta) \propto e^{\beta|\kappa|} \tag{5.4}$$

Note that the cardinality function can be used on a category because categories are sets of stimuli.[6] Bigger categories have an advantage over smaller categories if $PL_i > 0$, a disadvantage if $PL_i < 0$, and neither if $PL_i = 0$.

Likelihood based on the size principle

The likelihood $p(\Sigma_i \mid \kappa)$ of a set of observations Σ_i given a category κ to infer is the probability that that κ would have produced Σ_i. The likelihood therefore depends on how categories produce data. There are two main alternatives, *weak* and *strong* sampling (Navarro, Dry, & Lee, 2012). In weak sampling, the likelihood simply expresses whether the data is compatible or not with κ. Therefore, the likelihood function under weak sampling is:

$$p(\Sigma_i \mid \kappa) \propto \mathbb{1}(\Sigma_i \in \kappa) \tag{5.5}$$

As a consequence, all hypotheses that are fully compatible with the data will have the same probability.

In strong sampling, on the other hand, two hypotheses that contain all the observed stimuli can still end up having different probabilities. This is because the assumed data generating process is different from weak sampling. In strong sampling, the stimuli are sampled with uniform probability from the category. Therefore, the probability of a specific stimulus being picked depends on the size of the category. The large the category, the smaller the probability that a specific stimulus in it will be sampled. This behaviour is described by the likelihood function for strong

[6]The symbol κ is somewhat ambiguous, referring to the event of κ is the true category when it appears as the argument of a probability density function, and referring to a set of stimuli otherwise. The context should clarify what is meant in each case.

sampling:

$$p(\Sigma_i \mid \kappa) = \prod_{\sigma_{i,j} \in \Sigma_i} \begin{cases} \frac{1}{|\kappa|} & \text{if } \sigma_{i,j} \in \kappa \\ 0 & \text{if } \sigma_{i,j} \notin \kappa \end{cases} \tag{5.6}$$

$$= \mathbb{1}(\Sigma_i \subseteq \kappa) \frac{1}{|\kappa|^{|\Sigma_i|}} \tag{5.7}$$

where $\mathbb{1}(\Sigma_i \subseteq \kappa)$ is 1 if all the observations belong to the category and 0 otherwise. The consequence of strong sampling is that among the categories that are fully compatible with the stimuli, smaller categories are preferred over larger categories (right plot of figure 5.1). This captures the intuition that it would be a remarkable coincidence for observations to cluster in a certain portion of the stimuli space, if the category covers a larger portion. This phenomenon is called the *size principle*. There is empirical evidence that people assume that the observed data is strongly sampled in categorization tasks.[7] Therefore, I assume strong sampling in the Bayesian model.

5.1.3 Using posterior to categorize stimuli

In the previous section, I presented a model of how an agent π_i forms a posterior probability distribution $f(\kappa \mid \Sigma_i)$ over all the possible categories K after having observed data Σ_i coming from the true category. The last component missing for a full specification of the model is $h(\sigma \mid \Sigma_i)$, which describes how π_i uses the posterior distribution to categorize stimuli as belonging to the category or not. h is a function of both the participant's sampling style, which determines how she chooses a category given the posterior f, and the production error. I first discuss three possible sampling styles, *MAP* (Maximum A Posteriori) sampling, proportional sampling, and hypothesis averaging, and finally turn to production error.

The first type of sampling style I discuss is the MAP. A MAP sampler simply picks the hypothesis with the highest posterior probability. Therefore, once a MAP sampler has picked an hypothesis her behaviour is fully determined. Short of production error, she will accept or reject a given stimulus as belonging to the category iff it belongs to the picked category. In the model presented in the last sections, a MAP π_i with $PM = 0.5$ and $PL = 0$ (i.e. a uniform prior) will always pick the smallest among the hypotheses that contain all the observed stimuli Σ_i. This can be

[7] See Tenenbaum and Griffiths (2001) for a classic discussion of the size principle and related approaches.

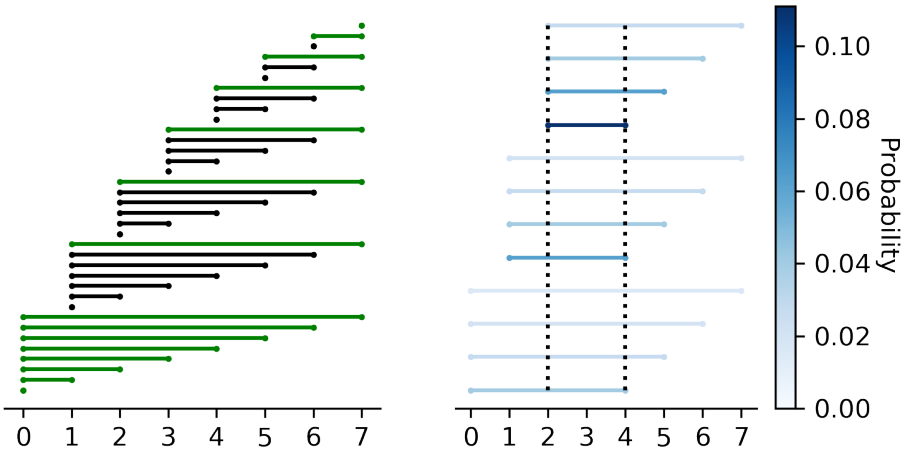


Figure 5.1: The left plot shows all convex categories for a space of 7 stimuli. The monotonic categories are green. The right plot shows the convex categories compatible with observations $\{2, 4\}$. The color indicates the likelihood of each category producing the observed data, under the strong sampling hypothesis. Importantly, larger categories have a lower likelihood.

proved in the following way.[8] Call κ_s and κ_n two hypotheses such that $\mathbb{1}(\Sigma_i \subseteq \kappa_s)$, $\mathbb{1}(\Sigma_i \subseteq \kappa_n)$, and $|\kappa_s| < |\kappa_n|$. As defined in equation 5.6, $p(\Sigma_i \mid \kappa_s) = \frac{1}{|\kappa_s|^{|\Sigma_i|}}$ and $p(\Sigma_i \mid \kappa_n) = \frac{1}{|\kappa_n|^{|\Sigma_i|}}$. The posterior probabilities for κ_s and κ_n are then, from equation 5.1:

$$f(\kappa_s \mid \Sigma_i) = \frac{1}{|\kappa_s|^{|\Sigma_i|}} \frac{p(\kappa_s)}{p(\Sigma_i)}$$
$$f(\kappa_n \mid \Sigma_i) = \frac{1}{|\kappa_n|^{|\Sigma_i|}} \frac{p(\kappa_n)}{p(\Sigma_i)}$$

Since I assumed a uniform prior and therefore $p(\kappa_s) = p(\kappa_n)$, the proportion between

[8] For a similar derivation for likelihood only rather than the posterior see Tenenbaum and Griffiths (2001).

the posteriors is:

$$\begin{aligned}\frac{f(\kappa_s \mid \Sigma_i)}{f(\kappa_n \mid \Sigma_i)} &= \frac{\frac{1}{|\kappa_s|^{|\Sigma_i|}}\frac{p(\kappa_s)}{p(\Sigma_i)}}{\frac{1}{|\kappa_n|^{|\Sigma_i|}}\frac{p(\kappa_n)}{p(\Sigma_i)}} \\ &= \frac{\frac{1}{|\kappa_s|^{|\Sigma_i|}}}{\frac{1}{|\kappa_n|^{|\Sigma_i|}}} \\ &= \frac{\left|\kappa_n\right|^{|\Sigma_i|}}{\left|\kappa_s\right|^{|\Sigma_i|}} \\ &= \left(\frac{|\kappa_n|}{|\kappa_s|}\right)^{|\Sigma_i|} > 1\end{aligned}$$

since by assumption $|\kappa_s| < |\kappa_n|$ and Σ_i is non-empty. Since the posterior probability of a category strictly decreases with its size, the smallest category will have the highest posterior probability. Note that the smallest category compatible with the observed stimuli is unique, consisting of all and only the stimuli between the greatest and the smallest observed stimuli. The smallest category compatible with the observations will then necessarily be sampled unless prior preferences for other categories counterbalance the advantage of smallness coming from the likelihood. In sum, the MAP categorization strategy, without taking production error into account, can described as:

$$p(\text{Accepting } \sigma|\Sigma_i) = \mathbb{1}(\min(\Sigma_i) \leq \sigma \leq \max(\Sigma_i))$$

The second type of sampling strategy is to sample a category with a probability proportional to its posterior probability for π_i. This is in reality not a single strategy, but a family of strategies, parameterized by the frequency of sampling. It is useful to distinguish between *stable* sampling, where π_i samples once and keeps the sampled category, and *unstable* sampling, where π_i samples a new category every time she is asked to categorize a stimulus. This is an important difference if the participant is asked to categorize the same stimulus repeatedly. If she is an unstable sampler, she will in the long term categorize stimuli as belonging to the true category in a way described by the following density, again without taking production error into account:

$$p(\text{Accepting } \sigma|\Sigma_i) = \sum_{c \in K} f(c \mid \Sigma_i)\mathbb{1}_c(\sigma)$$

Equation 5.1.3 can be seen as the expected probability of categorizing a stimulus as belonging to the category—i.e. 1 if the stimulus belongs to the sampled category and 0 otherwise—under the posterior over categories f. If she is a stable sampler, she will always give the same answer, modulo production error.

The third strategy, hypothesis averaging, has π_i average the probability of a stimulus across all hypotheses, and accept a stimulus according to the averaged probability. Behaviourally, hypothesis averaging is indistinguishable from unstable sampling, and it can also be described by equation 5.1.3. However, there is a conceptual difference between the two strategies. While in unstable sampling a specific category is picked, and therefore at least until the successive sampling event π_i accepts the sampled category as the true category, in hypothesis averaging π_i never commits herself to any specific category. Since this difference is not important for the model's aim, i.e. capturing differences in behaviour caused by a preference for monotonicity, I do not take a stance on whether unstable sampling or hypothesis averaging is best. In the following, I describe the participants' behaviour with equation 5.1.3. A final feature of the sampling strategies that is worth noticing is that π_i's judgement of whether a stimulus belongs or not to the category is independent of the choice on the other stimuli, conditional on the posterior $f(\kappa \mid \Sigma_i)$.

Up to this point, π_i is modelled as a perfectly rational agent that applies her probabilistic machinery without errors. However, perfect rationality is an implausible assumption when dealing with data produced by real participants. I therefore introduce a parameter PE_i (Production Error) for the error in production. The production error is implemented as a function $g : [0, \infty) \times [0, 1] \to [0, 1]$ from PE_i and the perfectly rational probability of accepting a stimulus as described in 5.1.3 to the real, noisy probability of accepting the stimulus. For instance, imagine that for some value y of PE_i, $g(y, 0.5) = 0.0$. This would mean that if the participant should in theory accept some stimulus with probability 0.5 according to the model above, she will actually always reject the stimulus, i.e. accept it with probability 0. Note that I assume that the production error is independent of the specific stimulus to categorize, given the probability of the stimulus prior to applying the production error. While this assumption might not be completely accurate, it greatly simplifies the estimation of PE_i from the data.

I define the function that applies a production error to the probabilities as follows:

$$g(y, x) = x + (\frac{1}{1 + e^{-y}} - 0.5)(1 - 2x) \tag{5.8}$$

Visually, this is a straight line rotating around the point $(0.5, 0.5)$, and having slope 1 when $PE_i = 0$ and tending towards slope 0 as $PE_i \to \infty$. Function g has various intuitive features. First, g is monotonically increasing with respect to the second argument, for any value of the first argument. Second, when $PE_i = 0$, $g(0, x) = x$ for all $x \in [0, 1]$. This means that if the PE_i parameter is 0 there is no production error, as g leaves the input probabilities unchanged. Third, as $\lim_{y\to\infty} g(y, x) = 0.5$ for all $x \in [0, 1]$. This means that as PE_i increases, the categorization behaviour depends less and less on the observations, but rather becomes maximally entropic—agents decide whether to include stimuli in the category at chance. Fourth, for all values y of PE_i, $g(y, 0.5) = 0.5$. This means that the production error by itself does not introduce a bias towards accepting or rejecting a stimulus as belonging to the category. Figure 5.2 shows the relation between input and output probabilities for various values of PE_i.

Note that production error complicates the connection between MAP, stable, and unstable samplers. Without production error, if π_i gave different responses when queried about a given stimulus, π_i could not be a MAP or stable sampler. However, if the production error is > 0, π_i might give different responses about the same stimulus even with a stable or MAP sampling strategy. Therefore, even after observing different responses for the same stimulus, none of the three sampling strategies can be excluded in principle. The data still gives information about which type of samplers participants are.

It is worth mentioning an alternative approach to modelling error, where noise is introduced in the participant's estimation of the likelihood of the data rather than in the categorization itself. Noise in likelihood estimation would cause noise in the distribution over categories, and therefore in the posterior. This in turn would influence the categorization behaviour. A possible implementation of noise in the likelihood would be to make the likelihood function tend to 0.5 for every set of data as the noise increases. Disregarding the contribution of the prior, a large amount of noise would therefore determine a more uniform distribution over categories. In production, this implies that stimuli contained in more categories would tend to be accepted more often. The stimuli contained in more categories are the central stimuli in the scale. In sum, a great noise in the likelihood would increase the probability

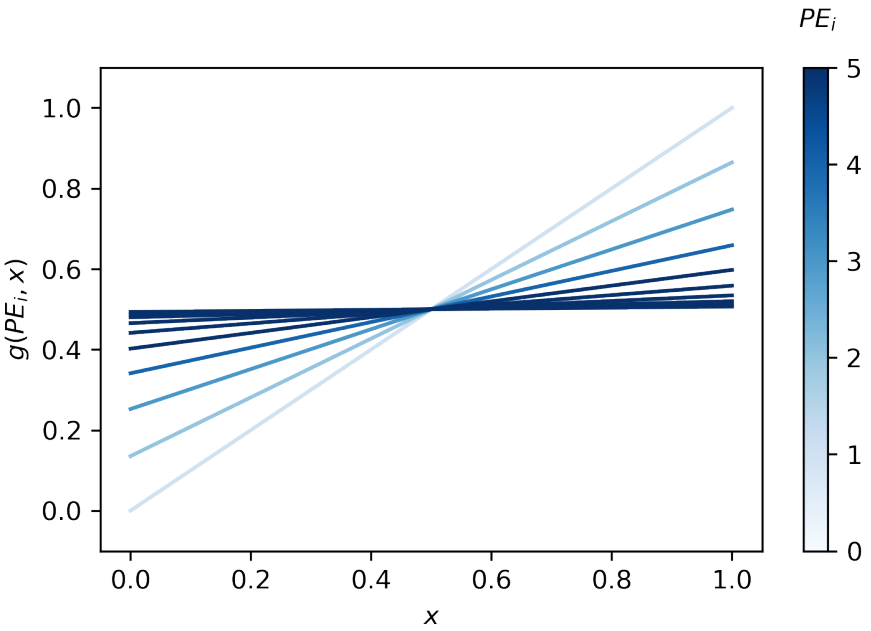


Figure 5.2: Error function $g(PE_i, x)$ as defined in equation 5.8 for various combinations of x and PE_i. When $PE_i = 0$, $g(PE_i, x) = x$. As PE_i increases, $g(PE_i, x)$ becomes more uniform across xs. However, $g(PE_i, x)$ always has value 0.5.

of accepting stimuli that are central on the scale. This effect does not track what we intuitively consider noise in the model, and is less interpretable than the noise in production we implement. Therefore, I did not add noise in the estimation of the likelihood.

5.1.4 Getting a feeling for the model

The last few sections presented a model of categorization behaviour. In this section, I consider what the model predicts for various combinations of parameters, to make sure that the model works as intended. The first thing to notice is that the closest a stimulus is to any of the stimuli in Σ_i, the more likely it is that s_i will accept it as belonging to c. Moreover, the more stimuli are observed, the steeper h decreases around the observed stimuli (See top left plot in figure 5.3). This behaviour comes from the assumption of strong sampling encoded in equation 5.6, and is independent from the prior.

The parameters PM_i and PL_i can model a large variety of behaviours (See top right plot in figure 5.3). A high preference for monotonicity leads to an overall increase in the probability of accepting stimuli around the observed stimuli. The increase is moreover asymmetrical, namely it is steeper in the direction of the closest border. A high preference for large categories also increases the probability of accepting stimuli that were not observed as belonging to the category, but does not privilege the scale's extreme that is closest to the observations. Very high and very low levels of PL_i lead to π_i that tend to accept every stimulus or reject every stimulus except for the observed ones respectively (see bottom left plot in figure 5.3). Therefore, the PM_i parameter cannot influence the data produced by a participant with extreme levels of PL_i, making it impossible to estimate a value for PM_i. Finally, the PE_i parameter regulates the amount of production noise for π_i. When $PE_i \to \infty$, π_i becomes equally likely to accept and reject each stimulus (bottom right plot in figure 5.3).

5.2 Embedding the cognitive model in a statistical model

The previous section presented a cognitive model of how a preference for monotonicity might influence categorization behaviour. The aim of developing this model was to analyse behavioural data produced in a categorization experiment similar to the three experiments presented above. In order to use it to analyse experimental data, the cognitive model has to be embedded in a statistical model.

Statistical Bayesian models are slightly different from the Bayesian models we encountered in previous chapters, where the aim was to model learners attempting to learn a language. In the Bayesian model I develop in this section, the posterior is over experimental hypotheses and the data is the participant's data. In our case, the posterior will over over values of PM, PL, and PE for each participant, as well as parameters describing how those values are distributed in the population of participants. The likelihood component of the Bayesian update is defined by a behavioural model of the participants, which was developed in the last section. The prior component encodes, in a conservative way, expectations about hypotheses previous to observing the experimental data. I first explain the functioning of Bayesian models in general and then develop a Bayesian model to estimate a preference for monotone

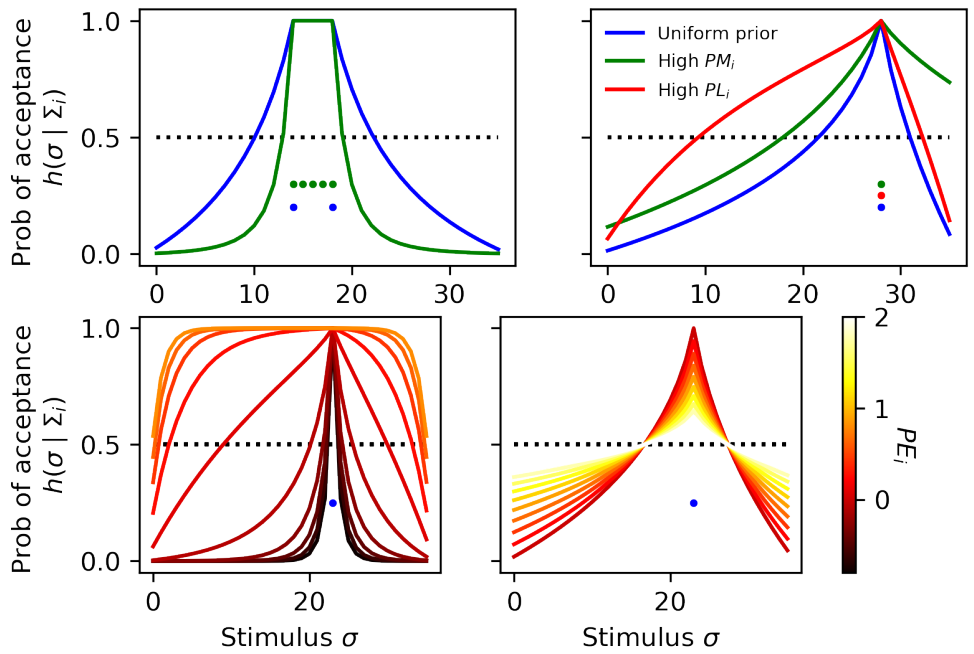


Figure 5.3: Probability of π_i accepting each stimulus (lines) given the observations (dots). Each line has the same color as the corresponding dots. The dotted line indicates $h(\sigma \mid \Sigma_i) = 0.5$. In the plots in the top row, each coloured line shows the probability of accepting each stimulus given the observations indicated by the dots with the same colour as the line. In the bottom row, the color indicates varying values of the parameters. Top left: $h(\sigma \mid \Sigma_i)$ for two cardinalities of Σ_i with uniform prior over categories. The more stimuli are observed, the steeper the decrease in probability of accepting a stimulus as it gets further away from the observed stimuli. Top right: Effects of high values of PM_i and PL_i. A higher preference for monotonic categories confers an advantage to stimuli towards the scale's extreme that is closest to the observed stimuli. Bottom left: Effects of different values of PL_i in the $[-0.8, 0.8]$ interval. As $PL_i \to -\infty$, π_i only accepts the stimuli between the greatest and the lowest observed stimuli as belonging to the category. As $PL_i \to \infty$, π_i tends to accept all stimuli as belonging to the category. Bottom right: Effect of different values of the production error parameter PE_i on the categorization behaviour. The shown values range in the $[0.0, 2.0]$ interval. As $PE_i \to \infty$, π_i accepts or rejects all stimuli with probability 0.5.

categories from the data.

5.2.1 A toy Bayesian model

Understanding the notation and main idea

Suppose we are interested in the growth of carrots in two different types of ground. We plant n carrots in one type of ground (type 0) and n carrots in another type (type 1), at a certain point we harvest them and measure their sizes Y_i (where i is the index of the carrot). We are interested in whether the carrots grew more in one ground type than the other type. In other words, we are interested in whether there is a difference between the mean carrot growths α_0 and α_1.

The Bayesian approach to this estimation problem is not to answer the question based on a single statistic, e.g. the maximum likelihood estimate of the difference between the growth rate in the two ground, or the probability of observing the observed difference between the mean carrot growths assuming that there is no difference between the true means. Rather, in a Bayesian approach a whole posterior distribution over differences is found. The question of whether we should accept or deny that there is a difference between the grounds can then be answered by taking into account the posterior distribution over differences and possibly other practical considerations.

We would like to find a posterior distribution over relevant uncertain aspects of the world. Two uncertain aspects of carrot growth are relevant for the estimation problem at hand. First, there is uncertainty about the mean growth for each type of ground. Second, there is uncertainty coming from the fact that not all carrots within each ground type are identical to each other. Therefore, the variation in growth also has to be estimated. To simplify, we assume that the variance in growth rate is identical for the two ground. In sum, we want a joint posterior distribution over combinations of mean growth and growth variance. This is encoded in the following model:

$$Y_i \sim \mathcal{N}(\mu_i, \sigma) \tag{5.9}$$

$$\mu_i = \alpha_{\text{GROUND}[i]} \tag{5.10}$$

$$\alpha_j \sim H\mathcal{N}(1) \qquad \text{for } j = 0, 1 \tag{5.11}$$

$$\sigma \sim H\mathcal{N}(1) \tag{5.12}$$

Where, again, i is an index over carrots, Y_i is the size of carrot i at harvest, and GROUND$[i] \in \{0, 1\}$ is the ground where carrot i grew. $H\mathcal{N}$ indicates the half-normal distribution, parameterized by a scale parameter. Understanding the notation for this model will be important to understand the model in the next section.

First of all, $a \sim b$ says that a random variable a is distributed according to some distribution b. Distribution b might be a parametric distribution, meaning that to uniquely specify a distribution the value of some parameters has to be fixed. Suppose for instance that $b = b(c)$. Then, c is the parameter, and depending on the value of c a might be distributed in different ways. The value of c might be known, or c might itself be a random variable. Line 5.9 is an example of the latter case. Y_i is distributed according to $\mathcal{N}(\mu_i, \sigma)$. $\mathcal{N}$ is the normal distribution parameterized by a mean parameter μ_i and a variance parameter σ. μ_i is the mean size of carrot i, and σ is the variance of all carrot sizes. Crucially, μ_i is itself a random variable, distributed as described on line 5.11.

Note that there as many random variables Y_i as there are carrots. This indicates that each individual carrot's size, and not only the mean carrot size of each ground type, is a random variable. It is prima facie unclear what it means for the size of a single carrot to have a distribution. Since the size has been observed, the distribution is not expressing uncertainty about the carrot's size. One interpretation is that line 5.9 describes the process that generates the carrot size, which is inherently stochastic. A Bayesian model like the one above is said to encode a *generative model* of how the data is produced. The generative model describes the relevant portion of the world with a combination of stochastic (e.g. 5.9) and deterministic (e.g. 5.10) functions.

The fact that there is a random mean parameter for each individual carrot seems in contrast with the fact that we are interested in estimating parameters across carrots. The solution to this apparent contradiction is in line 5.10, which connects the mean growth parameters of the individual carrots. This allows a parameter value to be shared by a group of carrots. To see how, consider the following example. Assume that carrots 4 and 43 both belongs to ground 0, i.e. GROUND$[4]$ = GROUND$[43]$ = 0. By line 5.10, μ_4 and μ_{43} equal $\alpha_{\text{GROUND}[4]} = \alpha_{\text{GROUND}[43]} = \alpha_0$. In other words, the mean sizes of carrots 4 and 43 will be identical and equal to α_0. In short, line 5.10 imposes that the mean growths of the single carrots be identical within each ground type, so that the realization of each μ variable gives information about the value of the other μ variables in the same ground type through the relevant α variable. The model above is therefore equivalent to the following model, where the distribu-

tion of individual carrots' heights are parameterized directly by the population level parameters:

$$\begin{aligned} Y_i &\sim \mathcal{N}(\mu_{\text{GROUND}[i]}, \sigma) \\ \alpha_j &\sim H\mathcal{N}(0, 1) && \text{for } j = 0, 1 \\ \sigma &\sim H\mathcal{N}(0, 1) \end{aligned}$$

In the following, we use both of these notations depending on which one is clearer in the context.

The expression "for $j = 0, 1$" on line 5.11 is a way to compactly express the distribution of multiple indexed random variables. Therefore, the model above is equivalent to:

$$\begin{aligned} Y_i &\sim \mathcal{N}(\mu_i, \sigma) \\ \mu_i &= \alpha_{\text{GROUND}[i]} \\ \alpha_0 &\sim H\mathcal{N}(0, 1) \\ \alpha_1 &\sim H\mathcal{N}(0, 1) \\ \sigma &\sim H\mathcal{N}(0, 1) \end{aligned}$$

Writing the distribution of indexed variables for the individual indices can become impractical when the index takes on many values.

Not only the prior distribution parameters, but also the families are chosen based on prior knowledge. For instance, the halfnormal distribution $H\mathcal{N}$ is chosen for the prior over means because of the prior knowledge that sizes cannot be negative. In theory, the carrot sizes in line 5.11 also cannot be negative, but we know that carrot sizes are far enough from 0 and have a variance small enough that a normal distribution is a good approximation.

Bayesian models are generative models of how the world produces data. When specifying a Bayesian model, we encode our expectations about how the world produces data before the data is observed. These expectations are encoded both in the model itself, e.g. the choice of distributions, the conditional dependencies, and in the values of the prior distribution parameters. In complex models, it can be difficult to see from the model specification what the prior distributions look like. Therefore, it is part of a Bayesian statistical analysis pipeline to plot the predictions of a model

based on the chosen prior probabilities, a process called *prior predictive checks*. Prior predictive checks are possible because Bayesian models are generative models, i.e. they encode a way of producing data. A prior predictive check for the carrot height model with the prior specified in 5.12 and 5.11 is plotted in the left plot of figure 5.4.

Model fitting and hypothesis testing

Once a generative model has been specified, the model parameters can be fit to the data, giving a posterior distribution over parameter values. I show an example of this procedure in figure 5.4. First, artificial carrot data is produced by simulating growth with known parameter values. Then, the model is fit to the artificial data to obtain a posterior distribution over parameters values. The posterior is not given in parameteric form. Instead, the fitting algorithm returns a set of samples from the unknown posterior (The fitting algorithm is described in much more detail in section 5.2.5). If enough samples are drawn, various statistics can be calculated on the samples that closely approximate the ones that would have been calculated from the true posterior. Once an approximated posterior is obtained, a first check that the fitting was successful is to simulate artificial data, which should look roughly like the observed data. This process is called the *posterior predictive check*. Figure 5.5 shows posterior predictive checks for the carrot growth model fitted on the simulated data.

Once the model is fit to the simulated carrot growth data and posterior samples for the mean parameters in the two ground types are obtained, the original question is still not answered. Does one ground type produce on average bigger carrots than the other type? While the question cannot be answered with complete certainty, the posterior can help answer it to a chosen level of credibility. To reiterate, the question is whether the difference between two means is different or not from 0. To answer this question, a convenient property of posterior distributions can be exploited. Namely, posterior samples of a function of the estimated variables can be obtained simply by applying the function to the posterior samples. This means in practice that we can obtain the posterior of the difference between two variables by taking the difference of the variables for each posterior sample. We can therefore easily estimate the posterior over differences between mean carrot growth in the two ground types. As a final step, we can consider the hypothesis that there is a difference between the

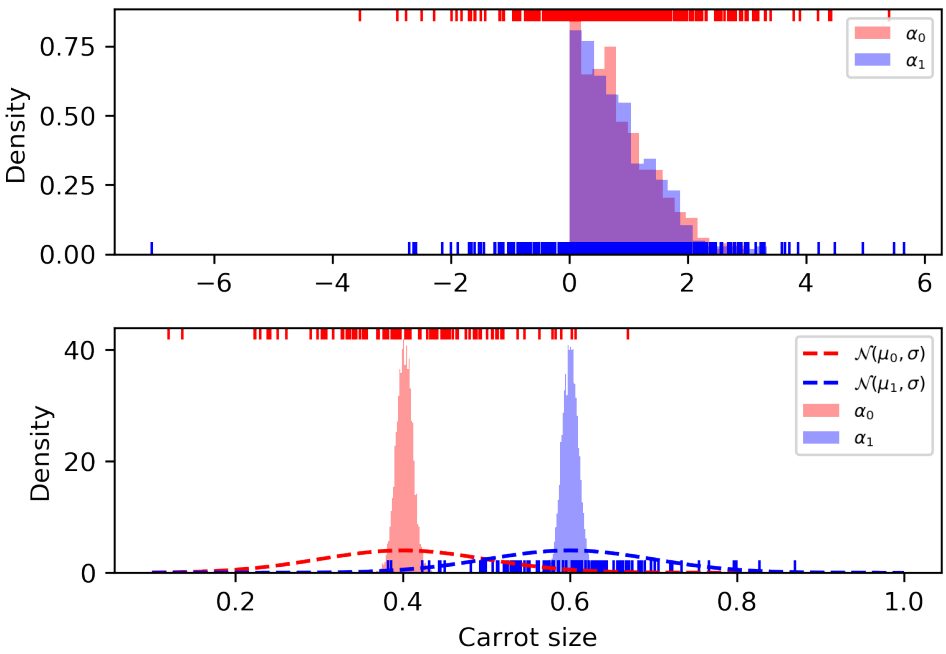


Figure 5.4: Top: prior predictive checks. The histograms show a variety of mean parameters drawn from the prior. An observation is drawn from a normal distribution with the sampled mean and variance parameter. The rugplots show the single observations for each ground type (red for ground 0 and blue for ground 1). The prior simulated data can get values that are known to be impossible, for instance negative carrot sizes in the plot. However, the crucial requirement is that the prior can accommodate a variety of parameter values. Bottom: The rugplot shows the simulated observations. The dashed lines show the true distribution that the data are sampled from. The filled histograms show the estimate for the mean parameter of each ground type. Note that the estimations are mostly close to the true means.

two ground types confirmed if the value 0 does not lie in the 95% Highest Posterior Density (HPD) interval. The HPD interval is the shortest interval containing 95% of the mass of the posterior distribution. Figure 5.6 shows the posterior distribution over differences between ground types along with the 95% HPD interval. We discuss how to test an hypothesis in the context of Bayesian modelling in more detail below in section 5.2.3.

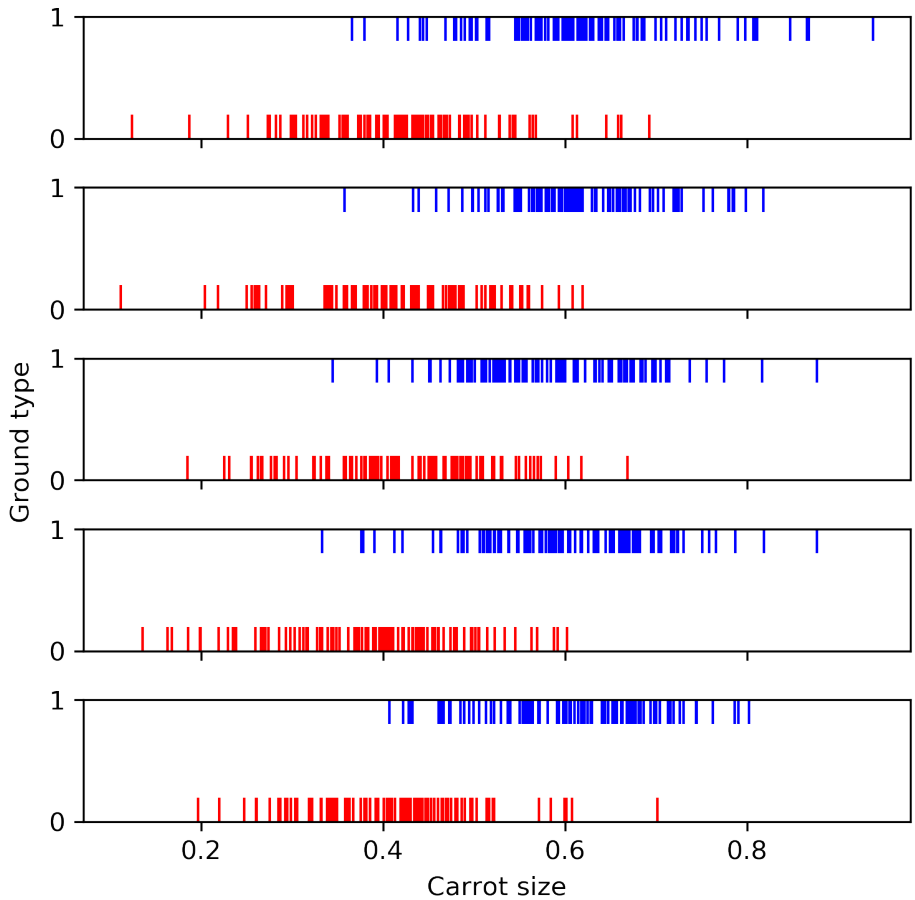


Figure 5.5: Posterior predictive checks for the carrot size model. Each plot shows a dataset of carrot heights for the two ground types. The simulated heights are produced by the Bayesian generative model specified above with the parameter values of the posterior distribution.

This section introduced some fundamental concepts and notation for understanding Bayesian models. We turn next to developing a statistical model to analyse the experimental data by exploiting the cognitive model developed above.

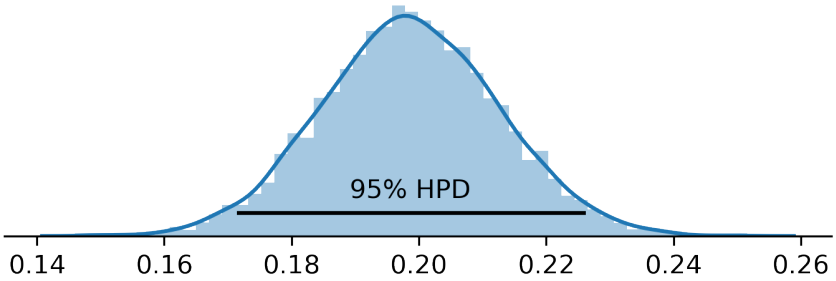


Figure 5.6: Posterior distribution over differences $\mu_1 - \mu_0$ between the two ground types along with the 95% HPD interval.

5.2.2 Statistical model for categorization data

In this section, I develop a model to fit to experimental data from new experiments, presented below. The design of the new experiments is similar to the first three experiments presented above. The new experiment has same two conditions as in the experiments presented in chapter 4, a similarity and a property condition. The training phase is, from the point of view of the statistical model, identical (more details on the differences below). The crucial difference is in the testing phase. While in the previous experiments participants had to either select a category by click-and-drag or by picking one of two options, in the experiments below participants are shown a series of stimuli and asked to categorize them individually.

From a modelling point of view, the response variable $\alpha_{i,j}$ is the number of times a participant π_i judged that a presented stimulus j belongs to the category, out of the total number of times they were asked $\omega_{i,j}$. The predictor variables are the condition k (which equals 0 for the similarity and 1 for the property condition), the stimuli presented as belonging to the category that the participant should infer Σ_i, and $\omega_{i,j}$. The relation between predictor and response variables for each participant is determined by the cognitive model developed in section 5.1 and the estimated parameters.

Posteriors are estimated for variables both at the level of individual participants and at the population level. Seven population-level parameters are estimated:

1. The first four are the mean θ_h and variance δ_h of the population-level logit-

normal distribution over PM for each condition h (the logit-normal distribution is described in more detail below). The (logit-)preference for monotonicity of each participant π_i in condition h is drawn from the distribution $\mathcal{N}(\theta_h, \delta_h)$.

2. The fifth and sixth parameter are the mean μ^{PL} and variance σ^{PL} of the population-level distribution of PL. The preference for large hypotheses PL_i for each participant is drawn from $\mathcal{N}(\mu^{PL}, \sigma^{PL})$. Note that I assume that PL is distributed identically in the two conditions.

3. The seventh and last population-level parameter is the variance σ^{PE} of a population-level distribution over production errors. The production errors of the individual participants are drawn from $H\mathcal{N}(\sigma^{PE})$.

At the level of single participants, a posterior distribution is found for PM_i, PL_i, and PE_i. The bottom level is the level of single judgements by each participant (line 5.13). At the bottom level, participants' judgements are distributed binomially. The binomial distribution has two parameters, a success probability parameter p and a number of trials parameter n. In the model, n corresponds to the number of times that the participant was asked to categorize that stimulus $\omega_{i,j}$. The p parameter for stimulus j and participant i is the probability $\phi_{i,j}$ of i accepting j as belonging to the category, as described by the cognitive model developed in the previous section.

The full model specification is as follows:

$$\alpha_{i,j} \sim \text{Binomial}(\omega_{i,j}, \phi_{i,j}) \tag{5.13}$$

$$\phi_{i,j} = h(\sigma_j | \Sigma_i, PL_i, PM_i, PE_i) \tag{5.14}$$
$$PL_i \sim \mathcal{N}(\mu^{PL}, \sigma^{PL}) \tag{5.15}$$
$$PE_i \sim H\mathcal{N}(\sigma^{PE}) \tag{5.16}$$
$$PM_i = \text{logit}^{-1}(\eta_i) \tag{5.17}$$
$$\eta_i \sim \mathcal{N}(\mu_i^{PM}, \sigma_i^{PM}) \tag{5.18}$$
$$\mu_i^{PM} = \theta_{\text{CONDITION}[i]} \tag{5.19}$$
$$\sigma_i^{PM} = \delta_{\text{CONDITION}[i]} \tag{5.20}$$

$$\mu^{PL} \sim \mathcal{N}(0, 2) \tag{5.21}$$
$$\sigma^{PL} \sim H\mathcal{N}(2) \tag{5.22}$$
$$\sigma^{PE} \sim \Gamma(1, 2) \tag{5.23}$$
$$\theta_h \sim \mathcal{N}(0, 1.5) \qquad \text{for } h = 0, 1 \tag{5.24}$$
$$\delta_h \sim H\mathcal{N}(1) \qquad \text{for } h = 0, 1 \tag{5.25}$$

Where I have called the standard deviation parameters σ, superscripted by the specific parameter they describe. Since three parameters are fit for each participant and seven parameters at the population level, if 80 participants are run a total of 247 parameters is fit. In traditional non-hierarchical statistics, it would be difficult to justify fitting such a large number of parameters with few datapoints. However, in hierarchical models the value of population-level parameters informs the plausible values of each individual parameter, and the total tendencies of individual-level parameters inform the population-level estimation. Sharing information across parameters at different levels reduces the effective number of estimated parameters.

Priors and prior predictive distribution

I defined a prior over the seven population-level parameters. The priors over parameters at the population level are called *hyperpriors*, because sampling from them produces priors over lower-level parameters. While hyperpriors can be used to encode

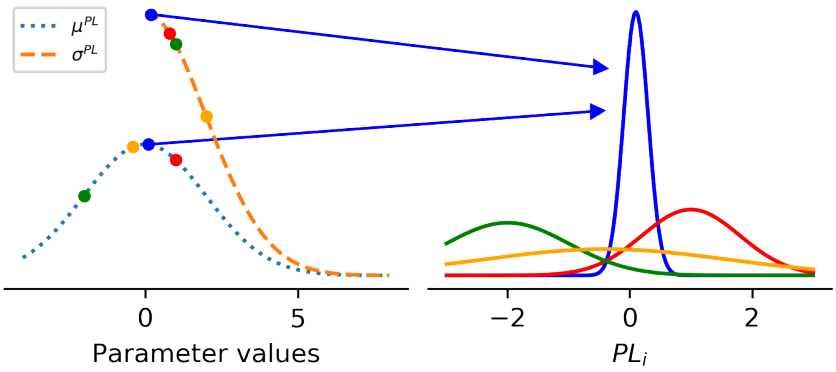


Figure 5.7: Hyperprior parameters for PL. The left plot shows the hyperprior distribution over μ^{PL} and σ^{PL}. Some values are sampled for both of them (dots). Each combination of sampled values defines a population-level distribution over PL, from which the PL parameter of individual participants is drawn. The left plot shows the distribution corresponding to the color-coded dots sampled in the right plot. For example, the sampled mean and variance parameter indicated by the blue dot define the blue population-level distribution (arrows).

domain specific knowledge about the modelled phenomenon, we use them simply to help model convergence. Therefore, we define regularizing hyperpriors whose purpose is mainly to exclude values that would produce nonsensical behaviour. Visualizing the effects of specific hyperprior parameters on predicted individual behaviour is difficult, because the impact of an hyperprior parameter depends on the value of the other variables. It is however possible to get an idea by considering typical values that the model parameters take according to the hyperprior parameters.

The hyperprior over μ^{PL}, the mean of the population-level distribution of PL_is, is a normal distribution with mean 0 and variance 2. The hyperprior over σ^{PL} is a halfnormal distribution with variance parameter 2. A single sample from these two hyperprior distributions defines one population-level distribution over preferences for large categories. Therefore, a posterior over values for μ^{PL} and σ^{PL} is a posterior over possible population-level distributions of the PL parameter. Figure 5.7 shows the defined hyperprior distributions for μ^{PL} and σ^{PL} (left) as well as some typical population-level distributions sampled from them (right).

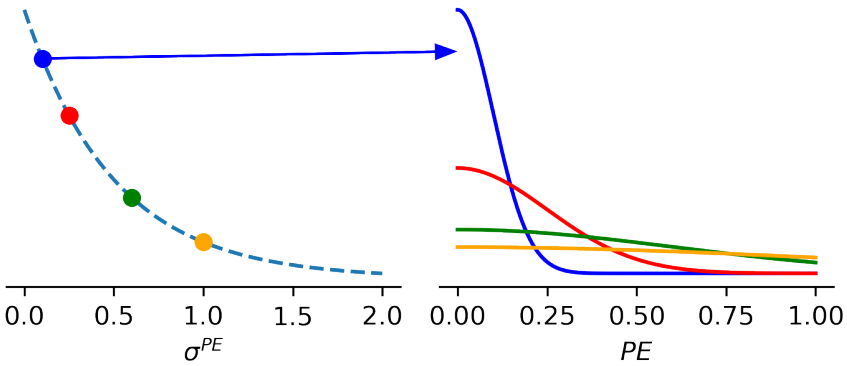


Figure 5.8: Hyperprior parameters for PE. The left plot shows the hyperprior distribution over σ^{PE}. Like in figure 5.7, each combination of sampled values defines a population-level distribution over PE, from which the PE parameter of individual participants is drawn. The left plot shows the distribution corresponding to the color-coded dots sampled in the right plot. This is illustrated in the blue case by an arrow.

Even small variations in PL have a substantial effect on behaviour, e.g. -0.1 and 0.1. Therefore, most of the population-level distributions over PL_i produced by the hyperprior specified above will put a lot of probability mass on implausible values for PL. However, this is not a problem in practice. Since the effect of variations in PL is very characteristic, the data makes extreme values of PL very unlikely. Therefore, even a very wide hyperprior for PL will be overcome by the likelihood. This is confirmed by the posterior distribution for the participants below, which are sharply peaked around 0.

The next distribution to consider is the hyperprior distribution over variances σ^{PE} of population-level distributions of production errors. The variances σ^{PE} have a $\Gamma(1, 2)$ distribution parameterized by α and β. When greater values of σ^{PE} are sampled, the resulting population-level distribution over PE puts more mass on higher levels of PE. The choice of a halfnormal distribution encodes the assumption that the production error should be assumed to be as little as possible, so that explanatory priority is given to a preference for monotonicity or large categories.

The final hyperprior parameters to consider are parameters defining distribution

over means and variances of population-level distributions over PM. The process of obtaining samples of the parameter of interest is in this case slightly more complex. Rather than sampling the parameters of a population-level distribution directly, the parameter is sampled in logit space and then transformed to the relevant parameter space (figure 5.9). This means that the population-level distributions over monotonicity parameters are *logit-normal* distributions. The reason for this choice are explained in details in section 5.2.5. The logit-normal distribution, which has support in $(0, 1)$, is very flexible. Depending on the values of its parameters, logit-normals can be unimodal distributions highly concentrated on any point of the unit interval or with a large spread, corresponding to populations that vary little or that vary a lot with respect to their preference for monotonic hypotheses. When the variance is large enough, logit-normals can also be bimodal with two peaks at the extremes of the support. The bimodal case corresponds to a population where individuals tend to have extreme preferences, but vary in terms of whether they prefer or disprefer monotonic hypotheses.

Importantly, the mean of the logit-normal distribution is not the same as the logit^{-1} of the mean of the normal source. In the experiment's preregistered analysis, we test for a difference between the means of the normal source. However, this might introduce a bias in the test, because differences between means near 0 in the logit space are further away in the bounded space than differences far away from 0.

There is an asymmetry between the estimation of strong preferences and strong dispreferences for monotonicity. Most of the categories are non-monotonic. Therefore, a strong influence of monotone categories leads to behaviour that is influenced by only a small proportion of all categories. Such behaviour can be very characteristic. On the other hand, a strong dispreference for monotone categories will only have a small effect on behaviour, since monotone categories have a small influence to begin with. As an analogy, consider a cloth merchant who sells cloths of very many different colors and who is trying to figure out whether a particular buyer likes red cloth. If the buyer repeatedly buys red cloths, it will be easy for the merchant to determine the buyer's taste. On the other hand, consider the same merchant trying to figure out whether a buyer dislikes red cloths. The buyer repeatedly purchases cloth from the merchant, and never picks red cloth. However, since the merchant offers cloth of many colors, the buyer's behaviour could be the result of chance rather than a dislike of red. It will then be much harder for the merchant to establish the buyer's taste.

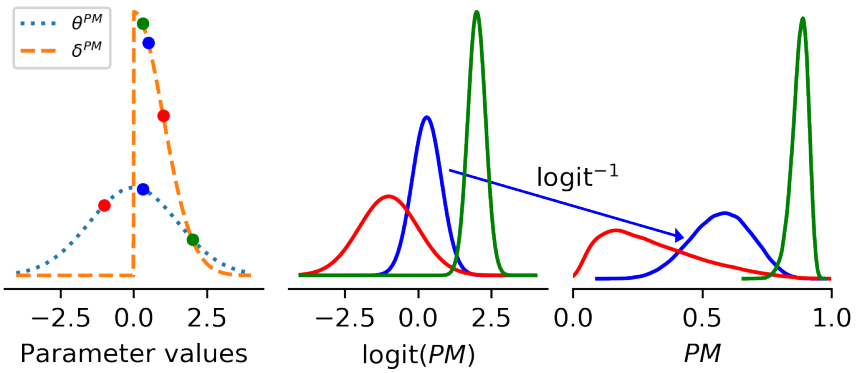


Figure 5.9: Hyperprior parameters for PM. The dots and lines are color coded as in figure 5.7. Two parameters θ^{PM} and δ^{PM} are drawn from the hyperprior distributions (left plot), and determine population-level distributions over the logit of the PM parameter (center plot). Drawing samples from the logit(PM) distribution and then applying the inverse logit, logit^{-1}, results in samples from the population-level distribution over PM (right plot).

Once the hyperprior parameters are specified for all the population-level variables, the joint effect of the chosen hyperparameter values can be observed by repeating the following process. First, sample a parameter for the population-level distribution for all such distributions. Then, sample individual-level parameters from the population-level distributions. Lastly, calculate the behaviour of the sampled individual. In the case of the present model, participants' behaviour depends on the probability of the participant accepting each stimulus conditional on the precategorized observed stimuli. The results of this individual-level prior predictive distributions are shown in figure 5.10. The figure shows that the chosen hyperprior parameter values are compatible with a large variety of participants behaviour. The hyperpriors are compatible both with participants that tend to accept most stimuli and with participants that tend to accept only stimuli very closed to the observed stimuli. As the number of observed stimuli increases (left to right), the same hyperprior distribution predicts a sharper average decrease of acceptance probabilities around the observed stimuli. This follows from the cognitive model as visualized in 5.3. Moreover, the model is compatible with a great variability of production errors, displayed by the varying

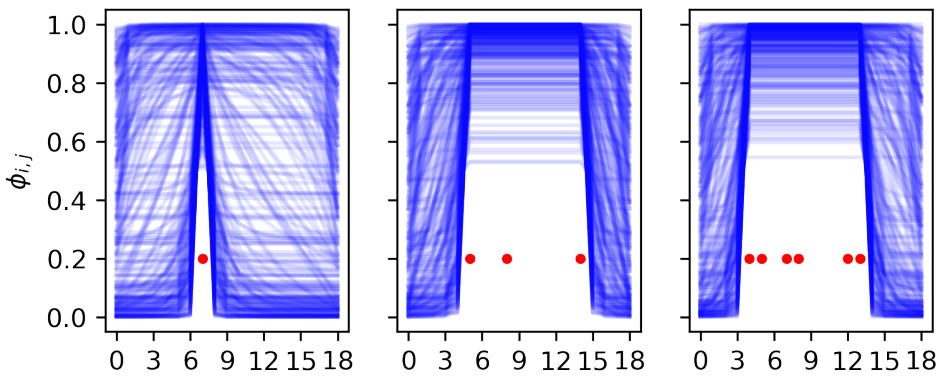


Figure 5.10: Effects of specified hyperprior parameters on individual participants' behaviour. The black dots show the stimuli observed by a participant. Each blue line shows the participant's acceptance probability for a set of parameters sampled from a population-level distribution, whose parameters were itself samples from the specified hyperprior distributions. As the plot shows, the prior is compatible with a great variety of acceptance probabilities, and therefore participant behaviour. While for some combination of parameters the acceptance probabilities decrease rapidly for stimuli away from the observations, for other parameters the acceptance probability is high for all stimuli.

probability of excluding the precategorized stimuli from the category.

5.2.3 The ROPE test

In this section, I discuss the hypothesis-testing strategy that I use for the experiments in the next sections. This strategy is not based on model comparison. Rather, I check if a difference of 0 between conditions is in the set of most probable hypotheses. If it is, the null hypothesis is maintained. If it is not, the null hypothesis is rejected and the alternative hypothesis—that there is a difference between the conditions—accepted. The problem is how to develop this intuitive strategy into a precise algorithm for hypothesis testing. I discuss the ROPE test as explained in Kruschke (2018) as one such algorithm.

The ROPE test consists of a comparison between two intervals in the parameter of interest, the *region of practical equivalence* (ROPE) and the HPD interval discussed

above. The ROPE is the set of points that are practically equivalent to the null hypothesis. I use the 95% HPD interval in the following. When comparing HPD interval and ROPE, three cases are possible. If the ROPE lies completely outside of the HPD interval, the null hypothesis is rejected and the alternative hypothesis accepted, in the way that is consistent with the side of the HPD that ROPE lies on. If the ROPE falls wholly within the HPD interval, the null hypothesis is maintained. If neither is true and there is a partial overlap, judgement is suspended until more data is available. It is clear then that the ROPE test solves the problem of testing a sharp hypothesis by having the null hypothesis be a set of point with non-zero probability mass, by assuming that the hypothesis of interest is not a unique value but rather a (possibly small) interval.

The alternative hypothesis concerns a difference between two population-level means, namely the means of the preference for monotonicity in the similarity and property conditions, which I called $\mu_{PM,0}$ and $\mu_{PM,1}$ respectively. The ROPE and HPD intervals are intervals on the space of differences $\mu_{PM,0} - \mu_{PM,1}$. The sharp null hypothesis is the point 0 on this scale, while the corresponding alternative hypothesis is the space minus the point of difference 0. To obtain posterior samples from the space of differences, the fact that a function of a parameter's posterior samples gives samples from the posterior over the function's parameters is exploited. First, I calculate the posterior distribution over differences between the population-level means over PM in the two conditions across all the S posterior samples:

$$\{\mu^s_{PM,0} - \mu^s_{PM,1}\}^S_{s=1}$$

The 95% HPD interval has to be calculated from posterior samples. Determining the ROPE interval is harder, for two reasons. First, as discussed above, the same difference between PMs causes different amount of changes in behaviour for different parts of the scale. Second, it is unclear what practical equivalence means in the present context.[9] I consider the limit where the ROPE becomes the sharp null hypothesis. This will turn out to not matter for accepting the alternative hypothesis, as in the experiments below the sharp hypothesis will always be contained within the HPD interval. Therefore, depending on the size of the ROPE, either the null hypothesis can be accepted or judgement can be suspended.

[9]This also makes it difficult to embed the hypothesis testing strategy in a Bayesian decision theory framework, in particular because there is no obvious way of specify a loss function.

To work with the bounded space in which the model is formulated, it is important to discuss a complication coming from a parameter transformation in the statistical model on line 5.17, which we already mentioned in section 5.2.2:

$$PM_i = \text{logit}^{-1}(\eta_i)$$

Here, to determine the PM value for participant i we take the inverse logit of a sampled parameter η_i. The test will be based on comparison of intervals that are sensitive to transformations of the space, and therefore we have to be careful. The PM parameter of each participant is not sampled directly from a distribution bounded in $[0, 1]$, but rather first sampled from a normal distribution and then transformed. The individual-level PM parameters are distributed as logit-normal distributions. For hypothesis testing, we are not interested in the difference between the means of the normal sources in the two conditions, but rather in the difference between the means of the logit-normal distributions in the two conditions. As discussed above, the mean of a logit-normal is in general different from the logit of the mean of the relative normal. Moreover, there is no closed way of calculating the relation between them. A high precision approximation can be obtained with the following:

$$\mathbb{E}_{\mathcal{N}(\mu,\sigma^2)}\left[\text{logit}^{-1}(x)\right] = \int_{x=-\infty}^{x=\infty} \text{logit}^{-1}(x) f(x \mid \mu, \sigma^2) dx \tag{5.26}$$

Recall that $f(x) = \frac{d}{dx}\int_{-\infty}^{x} f(x)dx = \frac{d}{dx}\Phi(x)$, and apply substitution $x = \Phi^{-1}_{\mu,\sigma^2}(y)$. Then, $y = \Phi_{\mu,\sigma^2}(x)$ and $\frac{dy}{dx} = f(x \mid \mu, \sigma^2)$, and therefore $dy = f(x \mid \mu, \sigma^2)dx$:

$$= \int_{y=0}^{y=1} \text{logit}^{-1}\left(\Phi^{-1}_{\mu,\sigma^2}(y)\right) dy \tag{5.27}$$

$$\approx \frac{1}{K-1}\sum_{i=1}^{K-1}\left(\text{logit}^{-1}\left(\Phi^{-1}_{\mu,\sigma^2}(i/K)\right)\right) \tag{5.28}$$

The last line is a quasi-Monte Carlo approximation of the preceding integral, where the function is evaluated at a grid of K many point of the unit interval.[10] Φ^{-1} is called the quantile function. Note that equation 5.28 is much more efficient than the

[10]The approximation, without justification, is reported in the wikipedia page "Logit-Normal Distribution" (2019). An explanation of the formula for a slightly simpler case can be found in *Does a Univariate Random Variable's Mean Always Equal the Integral of Its Quantile Function?* (n.d.).

simpler way of estimating the logit-normal mean by drawing samples from the normal source, transforming them, and finally taking their mean. While failing to make the adjustment between the mean of the normal source and the mean of the logit-normal can introduce a bias in principle, this bias is too small to make a difference for the experiments below.

Other approaches beyond the ROPE test could be used in this context. One possibility is to develop models encoding the null and alternative hypotheses, and then to test them in terms of their predictive accuracy for out-of-sample data. I did not use this method for computational reasons as well as the unclear connection between the truth of a model and its predictive accuracy. A second approach is Bayes factor calculation, which I also rejected for computational reasons and its high sensitivity to prior choices. The ROPE method has a computational advantage over the other methods. Pareto-Smoothed Importance Sampling (PSIS-LOO) (Vehtari, Simpson, Gelman, Yao, & Gabry, 2019) requires fitting two models, one encoding the null hypothesis and one encoding the alternative hypothesis. Leave-One-Out Cross Validation (LOO-CV) (Vehtari, Gelman, & Gabry, 2017) requires fitting multiple model, as many as there are units in the data (at the chosen level of the hierarchy). On the other hand, the ROPE test only requires fitting one model, the one corresponding to the alternative hypothesis in the PSIS-LOO method. The computational advantage of the ROPE method makes it especially apt for complex hierarchical Bayesian models, for which a single fit to the data can take hours or days.

5.2.4 Model checks on simulated data

The main task of a statistical Bayesian model is to find a posterior distribution expressing the uncertainty about parameter values after observing experimental data. We cannot determine how good the Bayesian model in combination with the real data is at recovering the true parameters, because the true value of the parameters is unknown. However, we can determine if the model is capable of recovering the true value of the parameters with some accuracy by simulating data with known parameters, and using the model to try to recover the parameter values. This can be done in various ways. If the model is not too computationally expensive, the parameter space can be explored exhaustively with respect to the model's ability to recover the parameters. Unfortunately, the model presented in this chapter is too complex to explore the parameter space. Therefore, I use simulations of real data

to check two qualitative facts about the model. The first is that the model can consistently pick up on a difference between conditions when the true population-level PM means are different, and confirm the alternative hypothesis according to the ROPE test picked above. The second fact about the model I check is that the model can consistently pick up on the lack of difference between conditions when there is no difference.

I simulated datasets with a true difference between conditions and datasets with no true difference between conditions. Figure 5.11 shows the results of this process. Overall, the stimulated check shows that the model fitting plus hypothesis testing procedure we picked are capable of low rates of both type I and type II error. However, this is obviously dependent on the number of participants and true magnitude of the difference between the conditions.

5.2.5 Solving computational problems

The model is coded with PyMC3, a Python implementation of *Hamiltonian Monte Carlo* (HMC) algorithms and related tools for Bayesian statistics. HMC is an algorithm for obtaining sample from a distribution that is too complex to be dealt with analytically or is only known up to a normalization parameter. This is particularly useful in the context of Bayesian modelling, because often the normalization parameter of Bayesian models is not known. Classical Monte Carlo sampling algorithms, such as Metropolis-Hastings, propose new points according to a proposal distribution, and accept them or reject them as a function of their unnormalized probability density. However, the strategy of proposing new points by sampling a proposal distribution becomes less efficient as the dimensionality of the parameter space increases. For a point in a highly dimensional parameter space, the probability density only increases in a very specific direction, meaning that most proposals will be rejected. HMC solves this problem by exploiting geometrical properties of the parameter space and distribution beyond simply comparing its density at pairs of points. More specifically, HMC runs a physical simulation of a particle moving without friction in the parameter space. The height of the particle is proportional to the posterior density, and HMC keeps track of its kinetic and potential energy.[11]

An important method in the specification of hierarchical Bayesian models is *reparameterization*. Reparameterization changes the way that the density is expressed

[11] A much more detailed discussion of HMC can be found in Betancourt (2018).

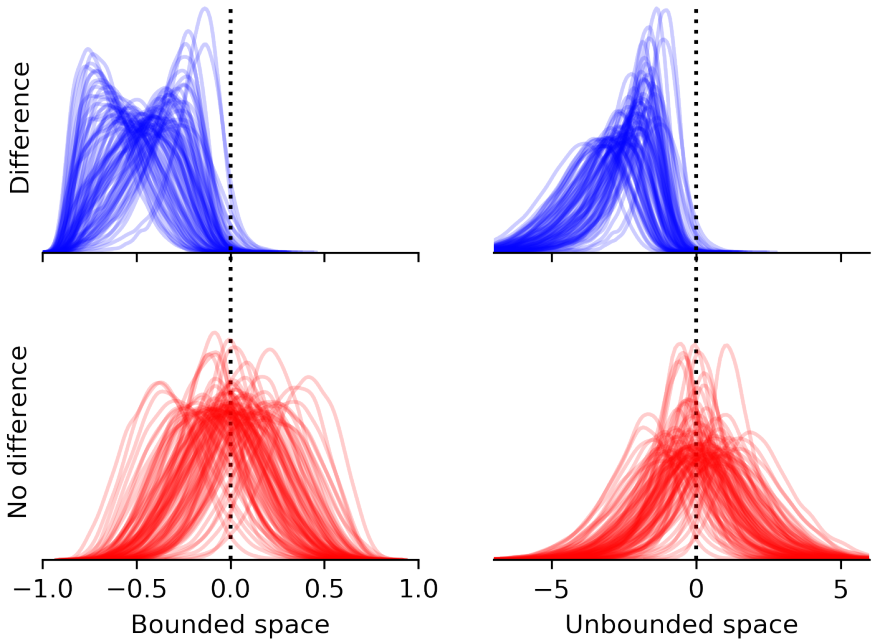


Figure 5.11: The plot shows the posteriors over differences between the population level means over the PM parameter. The posterior difference is displayed for both untransformed and inverse-logit space. The top plots show the posterior of models fitted on simulated data, where the true parameters were as follows: (1) 160 participants, (2) 36 stimuli, (3) $\mu^{PL} = -0.001$, (4) $\sigma^{PL} = 0.005$, (5) $PM_i \sim \mathcal{B}(20, 20)$ for condition 0 and $\sim \mathcal{B}(20, 3)$ for condition 1, (6) $\sigma^{PE} = 0.4$, (7) 20 judgements were simulated for each participant, (8) shown stimuli were at least 9 stimuli away from the border, and (9) participants were shown at most 5 categorized stimuli. The error rate of applying the ROPE test for the top left plot (i.e. failing to accept the null hypothesis) is about 0.06, and about 0.05 for the top right. The bottom plots show the fit to data simulated with the same parameter except that all participants' PM was distributed as $\mathcal{B}(7, 7)$. The error rate for the bottom left simulations (i.e. accepting the alternative hypothesis when the null hypothesis is true) is about 0.02, and it is 0.0 for the bottom right plot. More in general, the plot shows that the model is capable of testing the intended phenomenon.

in terms of input parameters, and can modify the gradient of the posterior density function over the parameters space. I discuss an instance were reparameterization

was needed when developing and testing the model above. Since HMC exploits the curvature of the posterior density function in the parameter space, zones of high curvature can lead to *divergent* samples. This happens when the curvature of the posterior density function is so steep that numerical approximation of the physical system that the HMC algorithm approximates breaks down, returning results that do not have a meaningful physical interpretation.[12] The generated samples therefore have to be rejected. Rejecting sample means that a part of the parameter space is not being explored, causing the Markov Chain to stop being ergodic and introducing a bias in the estimation. A solution to the problem of divergent samples is to reparameterize the model to ensure that the change is nowhere too steep for the HMC to explore.

A related source of divergences requiring reparameterization is extreme correlation between parameters. High correlation can create so-called *funnels* in the parameter space, where the posterior density becomes extremely concentrated in a high-dimensional thin tube which is hard to access and explore. In practice, a standard way to change the geometry of the parameter space by reparameterization when we are dealing with normal distributions is *non-centered* parameterization. Mean and variance are decorrelated when sampling from a normal distribution by exploiting the following equivalence:

$$f(x \mid \mu, \sigma^2) = \mu + \sigma^2 f(x \mid 0, 1) \tag{5.29}$$

where f is the density function of the normal distribution. It follows from equation 5.29 that in the following model X and Y are distributed identically:

$$X \sim \mathcal{N}(\mu, \sigma^2) \tag{5.30}$$

$$Y = \mu + \sigma^2 z \tag{5.31}$$

$$z \sim \mathcal{N}(0, 1) \tag{5.32}$$

When sampling Y, μ, σ^2 and a standard normal distribution can be sampled independently and then combined. While the exact mechanism by which the HMC algorithm is helped by the non-centered parameterization is beyond the scope of the present discussion, it is easy to see how non-centered parameterization might help HMC. The parameter space that HMC has to explore is different when sampling X

[12] More specifically, the total amount of energy (kinetic plus potential) is not conserved.

as opposed to Y. When sampling X, the HMC algorithm has to navigate a space consisting of a mean parameter and a variance parameter. When sampling Y, the space consist of the mean, the variance, and z which has to be sampled itself.

A problem with non-centered parameterization is that it can only be done with some distribution, such as those parameterized by location and scale. Ideally, we would want to apply non-centered parameterization to the population-level mean over PM, which cannot be done directly as PM is bound in the $[0, 1]$ interval. To work around the problem, I transform the bounded space of monotonicity into an unbounded space with the logit transformation. While parameters are always implicitly transformed by PyMC3 so that they have no boundaries, I do it explicitly to code the parameters in a non-centered way.

Before reparameterizing the normal distributions, every fit of the model to the data of the fourth experiment, introduced below, had a few hundred divergences. Reparameterizing solved the problem, leaving at most 5 divergences per model fit. The normal distributions in the model were reparameterized all together. Therefore, it is possible that reparameterizing the PM mean parameter did not help with the convergence issues. However, fitting the model is too computationally expensive to study the effects of reparameterizing only some of the distributions.

One last technical issues that is worth discussing is that of the *Jacobian adjustment*. Jacobian adjustments are much discussed in manuals and forums, with users being often confused about whether a specific model needs it. The present model does not, despite the inverse-logit transformation of the population-level PM parameter. In general terms, assume we have a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$. The Jacobian of f is the determinant of a matrix containing all of f's partial derivatives, i.e. the derivatives of every input component with respect to every output component.[13] Intuitively, the Jacobian encodes the infinitesimal amount of change of each output component in response to changes in each input component. For a scalar valued function of a single parameter, i.e. if $f : \mathbb{R} \to \mathbb{R}$, the Jacobian reduces simply to the absolute value of the derivative of f.

Assume that we want to work with a random variable X defined as a continuous and monotonic $f : \mathbb{R} \to \mathbb{R}$ transformation of another random variable Y, so that

[13]The word "Jacobian" is also used for the matrix itself rather than its determinant. In a Bayesian modelling context however this interpretation is less frequent.

$X = f(Y)$. Moreover, we know the distribution p_Y of variable Y. Then:

$$p_X(x) = p_Y(f^{-1}(x)) \left| \frac{d}{dx} f^{-1}(x) \right| \tag{5.33}$$

The second element of the product is the Jacobian[14] of the inverse transformation f^{-1}. Therefore, the Jacobian adjustment allows us to know the density of a transformed variable at a specific point in the parameter space as a function of the distribution of the untransformed variable. Why is this important for sampling? Consider the following statement:

$$\text{logit}(X) \sim \mathcal{N}(0, 1) \tag{5.34}$$

Here, we are giving a distribution to a transformation of a variable, rather than to a variable directly. A way to evaluate the density of X, which we are interested in, is to use equation 5.34:

$$p_X(x) = f(\text{logit}^{-1}(x) \mid 0, 1) \left| \frac{d}{dx} \text{logit}^{-1}(x) \right| \tag{5.35}$$

where f is the density function of the normal distribution. Note that without the Jacobian adjustment, the function over the space of x would not be integrate to 1 since x is still the parameter of integration, and therefore would not be a density function:

$$\int_{-\infty}^{\infty} f(\text{logit}^{-1}(x) \mid 0, 1) dx \neq \int_{-\infty}^{\infty} f(x \mid 0, 1) dx \tag{5.36}$$

On the other hand, if the sampling happens the following way there is no need for a Jacobian adjustment:

$$X = \text{logit}^{-1}(Y) \tag{5.37}$$

$$Y \sim \mathcal{N}(0, 1) \tag{5.38}$$

In this case, a distribution is not being given to a transformed parameter directly. Rather, a parameter is being sampled, and then the resulting sample is transformed into the new space. The STAN reference manual (Development Team, 2017, p. 291) explains this difference in the following passage:

> A transformation samples a parameter, then transforms it, whereas a

[14] I am using a one-dimensional example for simplicity and because the example in the model presented in this chapter is one-dimensional.

change of variables transforms a parameter, then samples it. Only the latter requires a Jacobian adjustment.

The model above only contains a parameter transformation, but not a change of variable. Therefore, no Jacobian adjustment is needed.

5.3 Experiment 4: Rich categorization data

In the last chapter, I presented three experiments that tried to test whether thinking in terms of a scalar property would induce a preference for monotonic categories. The experiments presented in the last chapter did not find support for this hypothesis. I concluded that, under the assumption that the hypothesis is true, the lack of evidence could be explained by the fact that the chosen combination of experimental design and statistical analysis was incapable of capturing the relevant differences in behaviour. Therefore, I developed a much more sophisticated statistical model to quantify subtler patterns in the data. In the rest of this chapter, I present three experiments that are designed to produce data that can be analysed by the Bayesian model developed above.

5.3.1 Materials and Methods

Participants

The same mode of recruitment, participation, and restrictions apply as in the previous experiments. Participants were be paid £0.75 for the experiment, which lasted about 5 minutes. Data for a total of 77 participants was collected, 37 in the property condition and 40 in the similarity condition.

Materials and Stimuli

Stimuli are identical to the ones in experiment 3 (see figure 4.5).

Procedure

This experiment is similar to the ones presented in the last chapter. Like the previous experiments, it contains two conditions, each featuring a domain of objects. One domain is the appropriate domain for scalar categories: the objects in this domain

are related with respect to an order. The other domain is not structured according to an order, and objects in this second domain are compared to each other using distance to each other in terms of similarity. The experiment tests the hypothesis that participants have a higher preference for monotone categories in a conceptual domain if the domain is mentally structured by a scalar property rather than similarity relations. I predict that participants show a greater preference for monotone categories in the property than in the similarity condition.

The fourth experiment differs from the third experiment in some details of the phrasing of the instruction. Moreover, the plants do not become bigger when hovered over. This is because the plants were large enough to be seen clearly. The main differences are in the testing phase. The participants see between 1 and 4 (inclusive) randomly picked stimuli, surrounded by a red box. Participants are told that the aliens use the word "  " to refer to the boxed plants. " " is used so that the word does not afford an interpretation as an adjective or a noun. While in the third experiment participants were asked to pick between two categories, in this experiment participants are asked to categorize individual stimuli. The testing phase consists of a series of trials where participants are asked whether the aliens would use the word to refer to other stimuli. For more details on the phrasing of the instructions, see flowchart of experiment four in D.4.

Like in the previous experiments, participants filled out a post-test questionnaire, in which they are asked for any feedback they might have about the experiment, and what their criterion was for judging whether a stimulus belonged to the category or not.

5.3.2 Results

The raw data from the fourth experiment is shown in figure 5.12. The data is very rich, so it is difficult to see whether there is an effect just by visualizing the raw data. The data was fitted with the Bayesian model described in the previous section. In this and the following runs of the Bayesian HMC sampler, 5000 samples were taken in one chain for each fit after a burn-in of 5000 samples. Convergence was checked visually from the shape of the traces, which after initial convergence moved in the same range of the parameter space and with constant variance in the step size. Figure 5.13 shows the 95% HPD interval for the population-level monotonicity preference mean in the bounded and unbounded space. In both cases, a difference of 0 is within

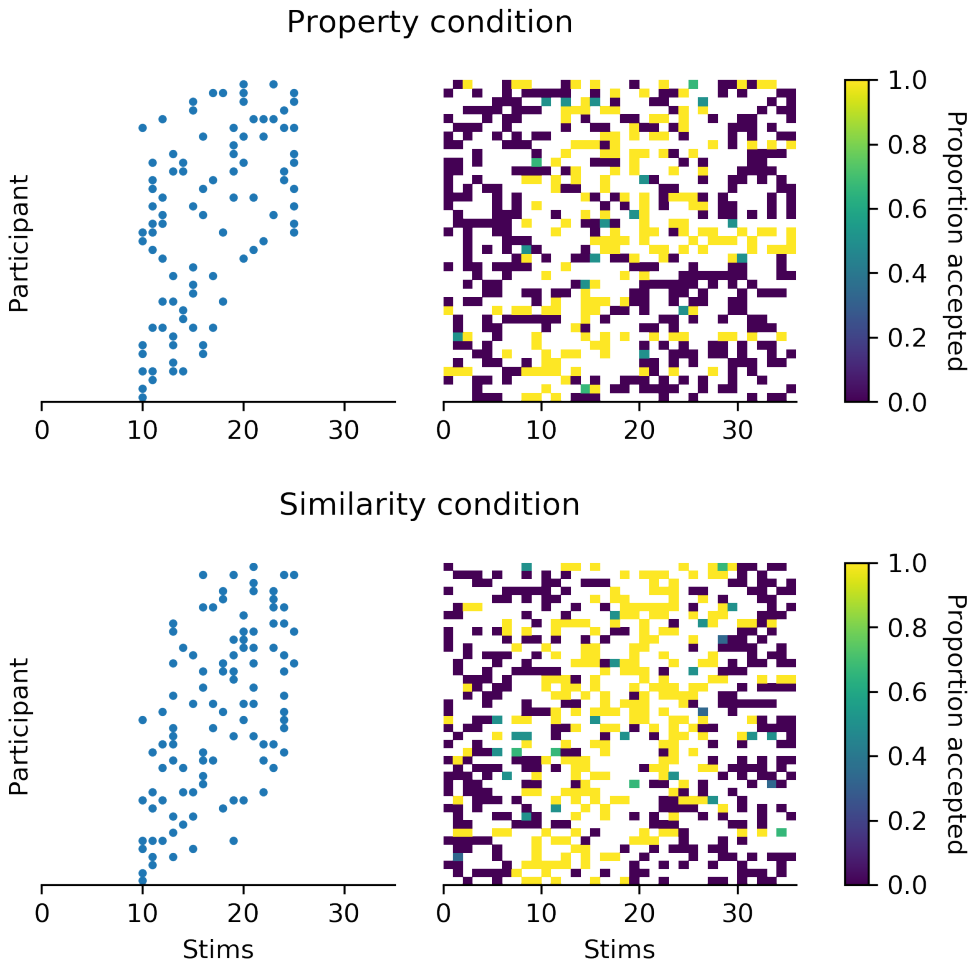


Figure 5.12: Plots visualizing the raw data from the fourth experiment. The plots on the left column show the stimuli that were presented to each participant as belonging to the category, where each row is a single participant. The rows are ordered by the mean of the pre-categorized stimuli: near the bottom are participants that observed stimuli more on the left of the scale. The plots on the right column show the proportion of times that each stimulus was accepted as belonging to the category. The white points correspond to stimuli that a participant was not asked about. Roughly, a comparison between the left and the right plot shows that the accepted categories tend to be close to the pre-categorized stimuli.

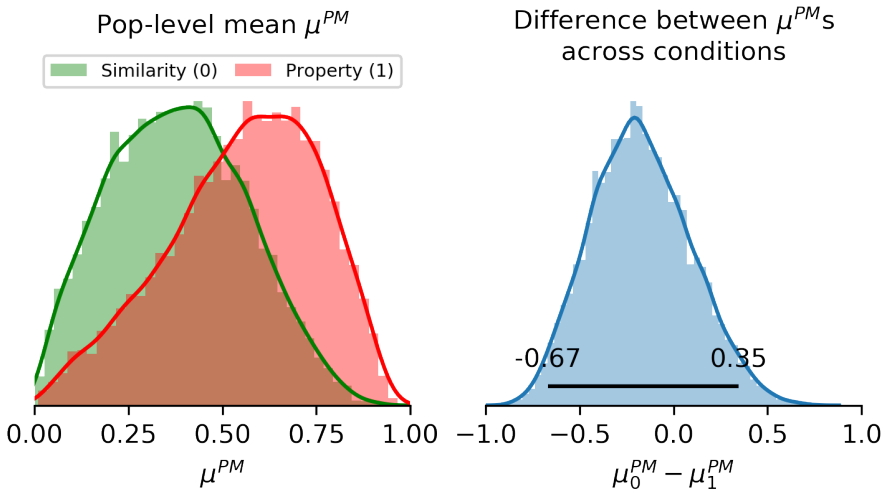


Figure 5.13: Plots visualizing the posterior distribution, given the data from the fourth experiment, over the population-level parameter determining the mean preference for monotonicity in each condition. High level of μ_0^{PM} means for instance that the distribution over monotonicity preferences in the similarity condition has a high mean. Despite a tendency towards the direction of the alternative hypothesis, in the right plot 0 is contained in the 95% HPD interval. Therefore, the ROPE test says either that the null hypothesis should be maintained or that judgement should be suspended, depending on the size of the ROPE. Note that the samples of μ_0^{PM} and μ_1^{PM} are correlated. Depending on the values taken by other parameters, they both tend to be large or small together. This implies that the plot on the left is not sufficient to evaluate whether there a difference between the conditions. The sample-wise differences between conditions have to be considered. The correlation between the μ^{PM} parameters is a consequence of the assumption that while the two conditions differ with respect to preference for monotonicity, they do not with respect to the other parameters. This assumption could be given up, at the cost of more parameters and therefore greater risk of overfitting the data.

the 95% HPD interval and therefore the null hypothesis is maintained.

A last check that is worth performing on the model is its prediction of participant's categorization behaviour for untested stimuli. Figure 5.14 shows the posterior probability of participants accepting each stimulus as belonging to the category. While the way that PM and PL affect behaviour is not easy to see, it is clear that the model is to some extent tracking participants' noisiness in categorizing stimuli

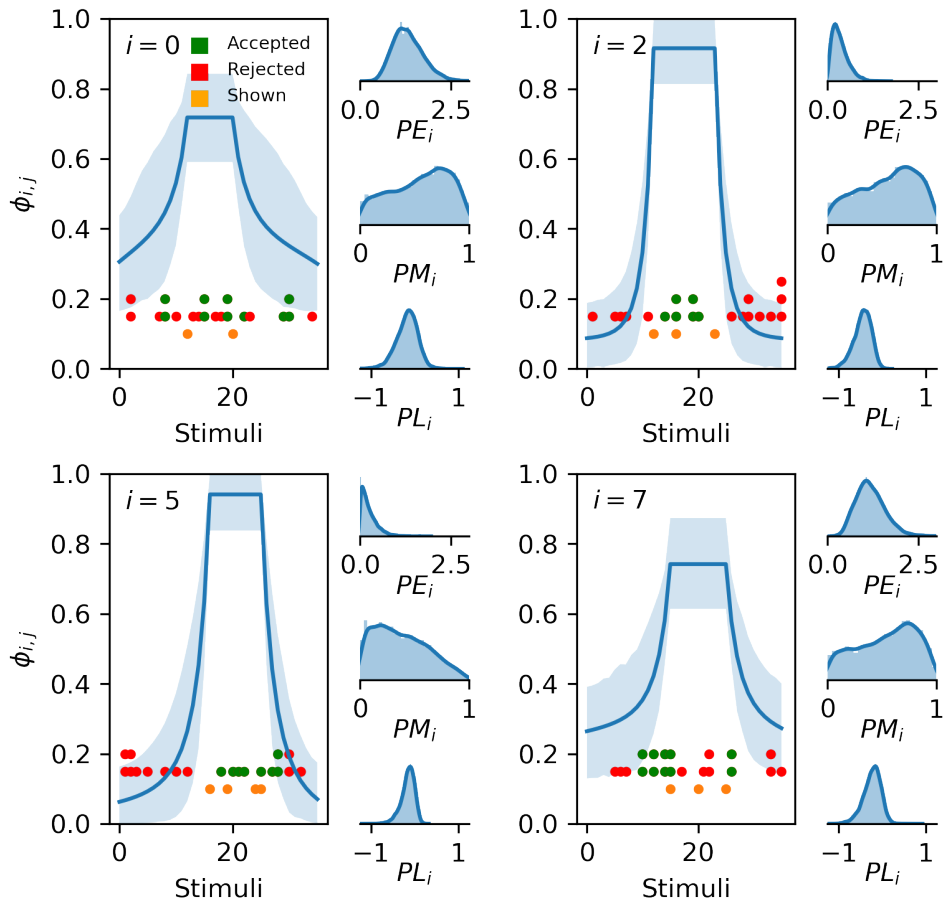


Figure 5.14: The plot show the posterior distribution over the probability of accepting each stimulus as belonging to the category for four participants, along with 95% HPD interval. For each participant, the plot also shows the posterior distribution over their individual-level parameters. The plot also shows which stimuli the participants saw as pre-categorized, which stimuli they were asked to categorize, and whether they accepted them or rejected them. The stimuli that the participants were asked about are stacked vertically.

with the PE parameter.

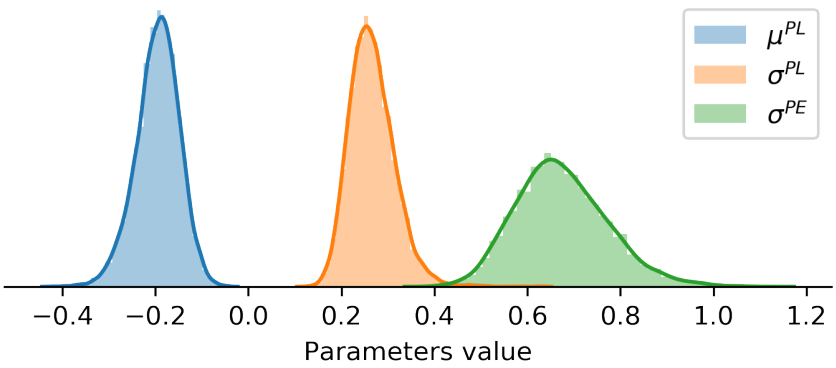


Figure 5.15: The plot shows the posterior densities for the population level parameters that do not concern the preference for monotonicity, given the data from experiment 4.

5.3.3 Discussion

While the posterior density over the difference between conditions shows a slight tendency in the direction of the alternative hypothesis, this tendency is not strong enough to reject the hypothesis according to the chosen inference criterion. Therefore, I do not accept the alternative hypothesis based on the data from the fifth experiment. Beside fitting the parameters relating to monotonicity, the Bayesian model also fit other population-level parameters, which are assumed to not vary across conditions. These are shown in figure 5.15.

One worry about the model that is worth mentioning is that noisy participants, who accept or reject stimuli randomly, might introduce a lot of noise in the estimation of the population-level parameters. However, the Bayesian model is capable of dealing with this. If the acceptance behaviour is hard to explain given the population-level parameters, the model explains it in terms of production error. This plays two different roles in the estimation process. Consider first the view from the top of the hierarchy down. If the estimated production error for a participant is high, the other individual-level parameters cannot influence much the calculated acceptance probabilities, which will tend to be around 0.5 for all values of the other parameters. This means that the estimation of the underlying pre-error acceptance probabilities

are free to vary without becoming much more incompatible with the data, and will be mostly informed by the population-level distributions. The estimated pre-error acceptance probabilities of a clearly error-prone participant will therefore have high variance. If the estimated production error is small, the underlying pre-error acceptance probabilities will have a large impact on the acceptance probabilities when the error is included. The other view is from the bottom of the hierarchy up. If the mean production error of a participant is high, it is more difficult to estimate the value of the other parameters for the participant, and this means in turn that that participant's data has less influence in the inference of the population-level distribution. This mechanism allows the Bayesian model to automatically penalize data coming from noisy participants. The relation between the posterior density over production error and probability of accepting the stimuli are shown in figure 5.16.

5.4 Experiment 5: Adding an affordance for scalar interpretation

Like in the series of experiments presented in the previous chapter, a natural variation of the experimental design is a change in the stimuli. While various versions of the stimuli were tested in the first three experiments, the data cannot be analysed with the Bayesian model except for in an exploratory way. Therefore, in experiment 5 we change the stimuli to increase affordance for scalarity. A possible problem with the stimuli of experiment 3 and 4 (see figure 4.5) is that there is nothing about the stimuli themselves that affords a scalar interpretation. The participants' only pressure towards a scalar way of thinking about the stimuli comes from the story of what causes the stimuli to change, namely the amount of blagardium.

5.4.1 Materials and Methods

Participants

Conditions and modes of recruitment are identical to the previous experiment. Data was collected for a total of 80 participants, 39 in the property condition and 41 in the similarity condition.

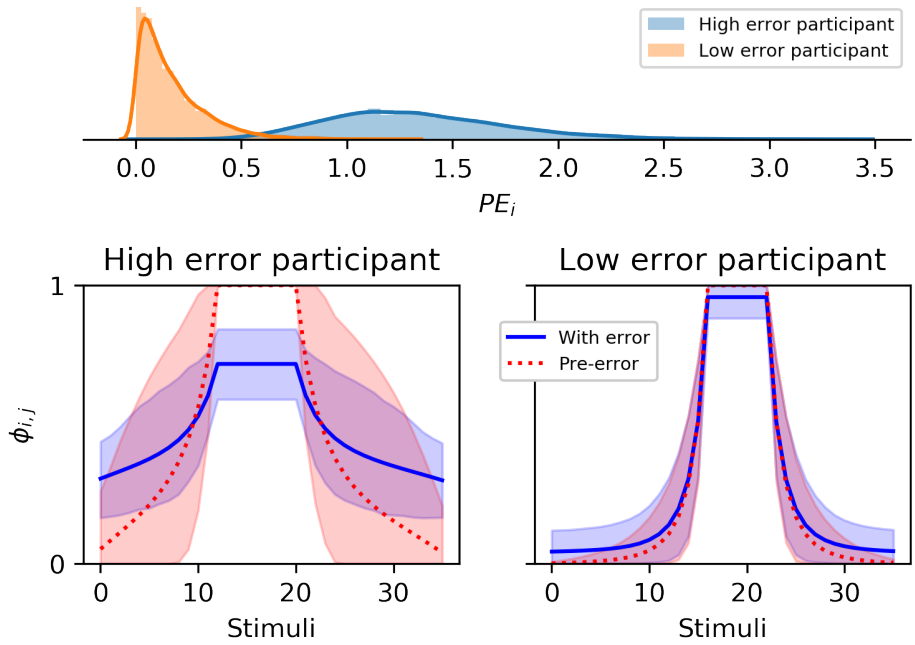


Figure 5.16: The plot shows the interaction between the estimated production error and the estimated probability of accepting a stimulus as belonging to the unknown category with and without error. The data is from two participants from the fourth experiment. The top plot shows the posterior densities over the production error for two participants, one with a high mean error and one with a low mean error. The bottom plots show the posterior means and 95% HPD interval of the two participants accepting each stimulus in the category, before and after applying the production error. Crucially, higher estimates of production error lead to more variance in the estimate of the probability of accepting a stimulus before the error disturbance is applied. This is shown by the fact that the 95% HPD intervals for the pre-error estimates in the left plot are much larger than in the right plot.

Materials and Stimuli

The only different from experiment 4 is the set of stimuli. To create a stronger affordance to interpret the stimuli as organized in a scalar way, we add two scalar features, namely border thickness and fill colour's saturation. The new stimuli are displayed in figure 5.17.

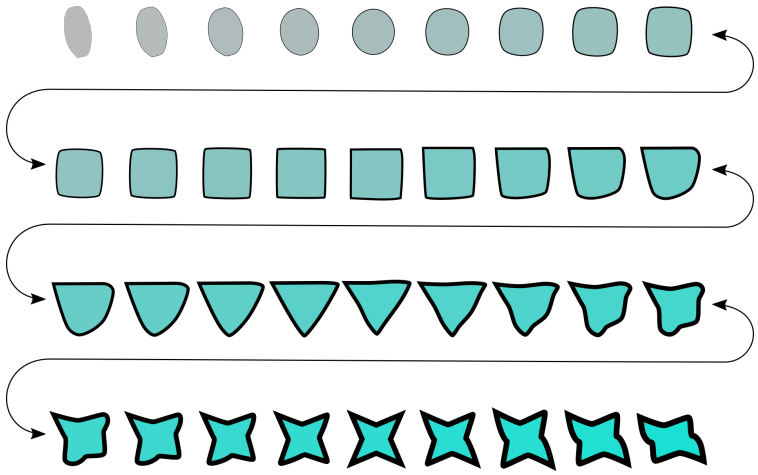


Figure 5.17: Stimuli for the fifth experiment, arranged in four rows for ease of visualization. The stimuli consist of 36 images of alien plants.

Procedure

The training and testing phases are identical to the fourth experiment with respect to structure and instructions. To summarise, the two conditions vary with respect to the framing of the stimuli. In the similarity condition, the attention of the participant is drawn to how similar or different the stimuli are to one another. In the property condition, the attention of the participant is drawn to the position of the stimuli in an order. Besides the noise coming from random sampling, participants vary with respect to (1) the condition they are tested in, (2) which stimuli they observe when they are familiarized with the stimuli, (3) which stimuli they are shown as example of the category to be inferred, (4) which stimuli they are asked to categorize as belonging or not to the category, (5) the order of the stimuli shown on the screen.

The order of the stimuli on the screen is randomized across participants. The 14 sets of stimuli presented in the training phase are picked randomly. In the testing phase, each participant sees a random number between 1 and 4 (included) of stimuli presented as belonging to the category. The categorized stimuli are also picked randomly, with the restriction that they are not within 9 (included) of the most extreme stimuli on the scale. The 20 stimuli that the participants are asked to categorize are also picked randomly. For more details on the instructions and look

of the experiment, see flowchart D.5.

5.4.2 Results

The raw data for the fifth experiment is shown in figure 5.18. The results of the ROPE test are shown in figure 5.19. A difference of 0 is contained in the 95% HPD interval, and therefore according to the chosen test the alternative hypothesis cannot be accepted. Moreover, the posterior does not show a tendency in the direction of the alternative hypothesis. The posterior for the other population-level parameters is displayed in figure 5.20. The posterior distributions for some individual participants are shown in figure 5.21.

5.5 Experiment 6: Different stimuli across conditions

In the experiments presented above, the stimuli were identical across conditions. This was to prevent noise in the way participants' behaviour differed across the two conditions, and particularly noise in the difference between conditions determined by features of the stimuli not involving scalarity. However, using different stimuli across conditions has the advantage that different affordances can be created in the two conditions. Specifically, it is possible to show stimuli without a clear order in the similarity condition and intuitively ordered stimuli in the property conditions. In the sixth experiment, I try to reach a balance between the advantages and disadvantages of using different stimuli in the two conditions. Specifically, I use stimuli that differ minimally across the two conditions while still affording different interpretations.

Participants

The data collection procedure is identical to the previous preregistration, with the only difference that the experiment lasts 2.5 minutes and participants will be paid \$0.40. The data was collected for 163 participants, 92 in the property condition and 71 in the similarity condition.

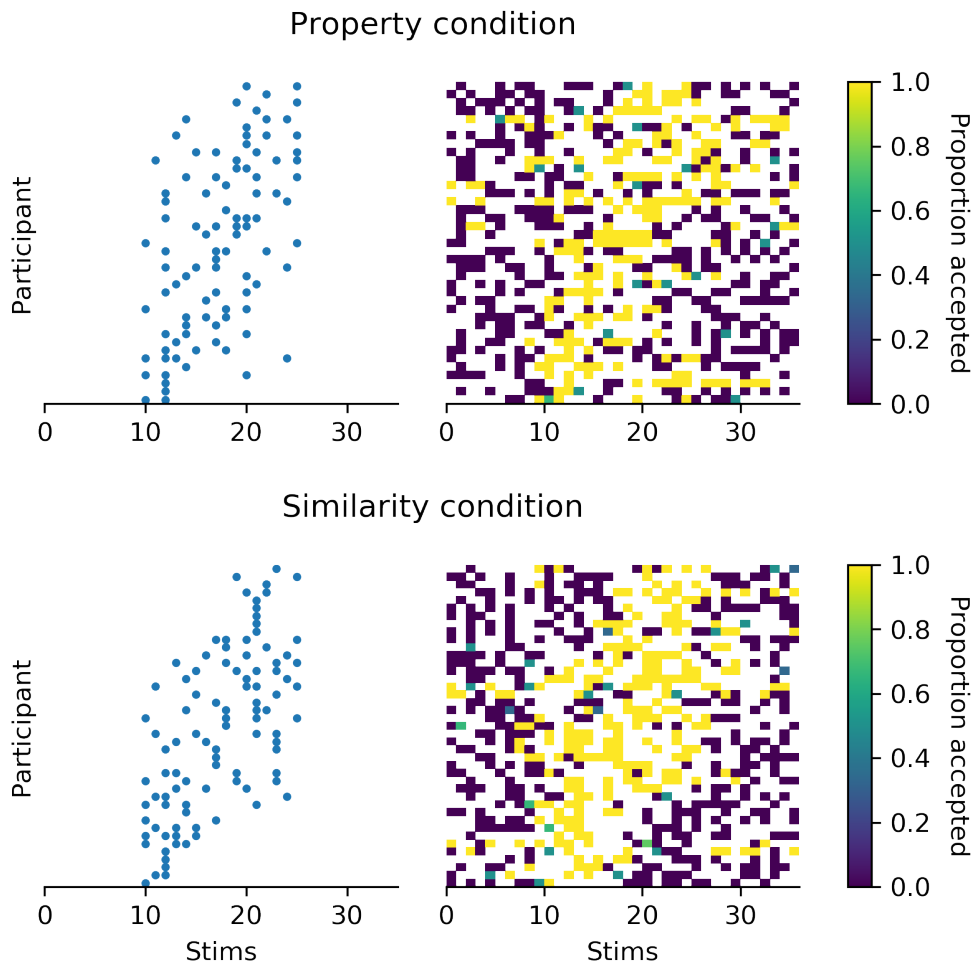


Figure 5.18: Visualization of the raw data of the fifth experiment. For an explanation of the plot, see figure 5.12.

Materials and Stimuli

The set of stimuli is the most substantial difference to the previous experiments. The stimuli are generated procedurally, and are different in the two conditions. In the order condition, there is something about the stimuli themselves that makes them ordered, namely the number of "petals", which varies monotonically, either

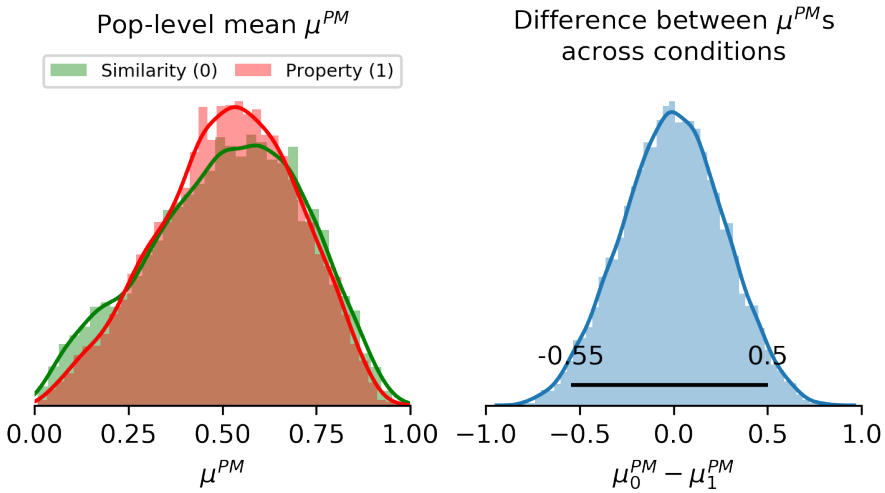


Figure 5.19: Plots visualizing the posteriors over monotonicity, along with 95% HPD interval, for the data from the fifth experiment. For more on how to interpret the plot, see figure 5.13.

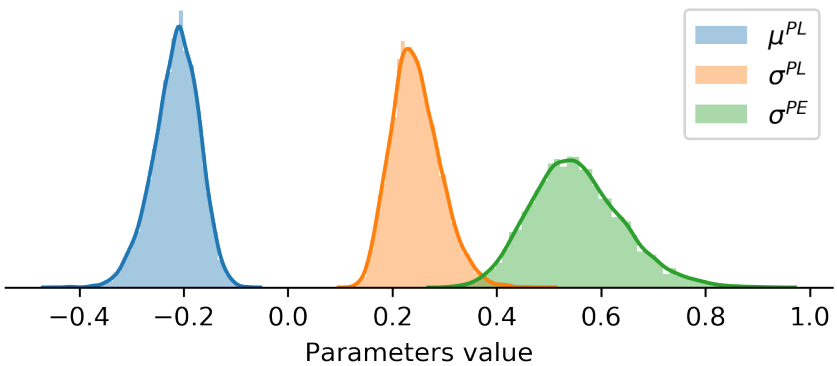


Figure 5.20: Plots visualizing the posterior densities over population-level parameters given the data from the fifth experiment. See figure 5.15 for same plot with data from the fourth experiment.

always increasing or always decreasing. In the distance conditions, the number of petals both increases and decreases when going through the stimuli (see the stimuli in figure 5.22). Both sets of stimuli have pivotal stimuli with 2, 4, 5, and 6 petals.

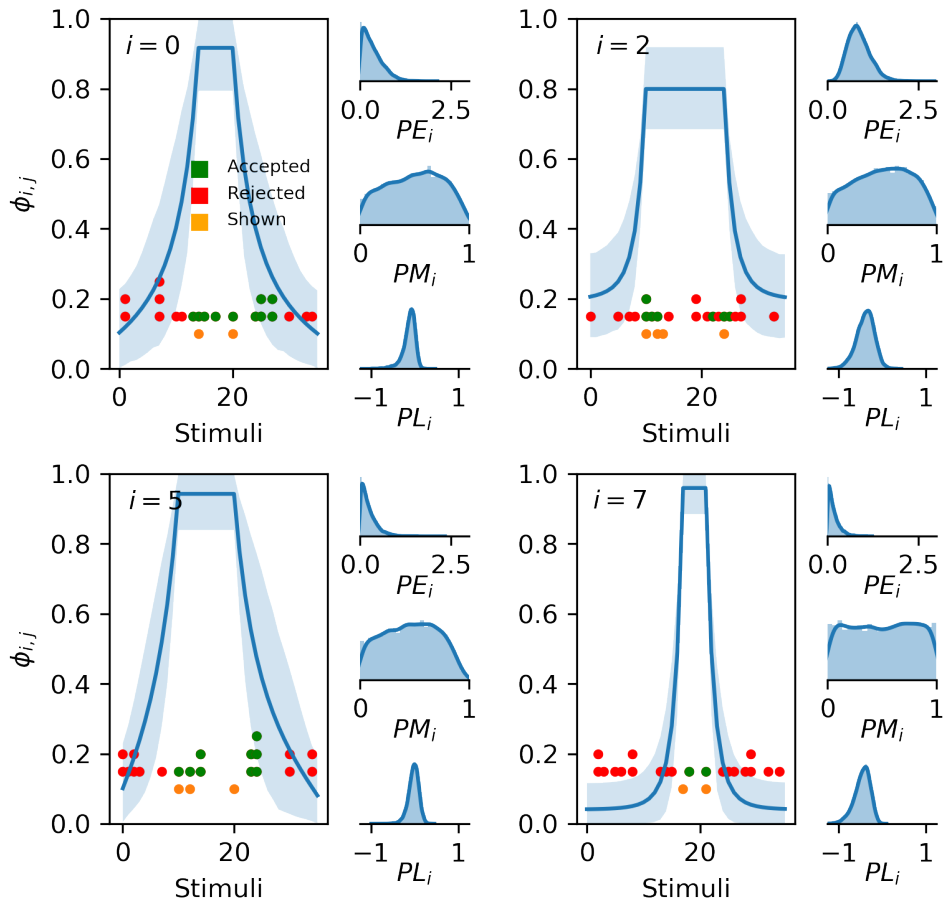


Figure 5.21: For explanation of this plot see figure 5.14.

The number of petals for the pivotal stimuli was chosen so that it was possible for all the stimuli in both conditions to have vertical symmetry. In both conditions, these four pivots are placed in different orders and 5 new stimuli are created between any two adjacent pivots by interpolation. In the order condition, the order is 2, 4, 5, 6. In the distance condition, the order is 2, 5, 4, 6.

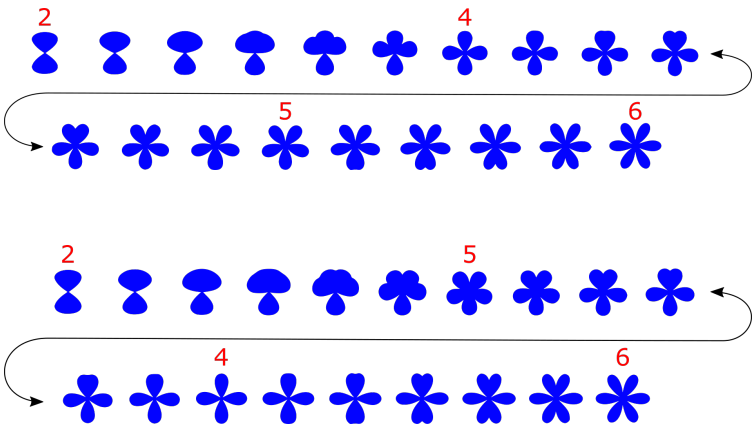


Figure 5.22: Stimuli for property condition (top) and similarity condition (bottom) of the sixth experiment, arranged in two rows for visualization purposes.

Procedure

The study design is similar to the fifth experiment, except for the following differences. First, the training phase is reduced to two screens, the first one giving the context and the second one to familiarize the participants with the stimuli. Second, the set of stimuli is different from the previous experiment, and there are 19 rather than 36 stimuli. Third, the instructions refer to the stimuli as "crystals" rather than "plants". Randomization is the same as previous preregistration, with the only difference that the categorized stimuli are not within 3 (included) stimuli from the border, rather than 9. This change is motivated by the fact that the set of stimuli is smaller than in the previous experiment. For more details on the instructions and look of the experiment, see flowchart D.6.

5.5.1 Results

Raw data for the sixth experiment is visualized in figure 5.23. The results of the ROPE test are shown in figure 5.24. A difference of 0 is included in the 95% HPD interval, and therefore the alternative hypothesis cannot be accepted according to the chosen hypothesis testing method. Figure 5.25 shows the posterior densities for the other population-level parameters. Note that while the posterior densities of the fourth and fifth experiments closely resembled each other, they are quite different

from the posterior densities of the sixth experiment. The likely cause for this is the fact that the set of stimuli in the sixth experiment is different. Note that the value of the parameter for preference for large categories does not, by itself, have a clear interpretation. Instead, its interpretation is relative to a particular set of stimuli, making the difference between posterior densities less strange than it might seem at first. The intuitively reasonable posterior distributions over acceptance probability for some participants shown in figure 5.26 provide further confirmation that the model is working.

5.6 Discussion

The experiments did not confirm the hypothesis that thinking in terms of scales causes a preference for monotonic over other categories. Does this mean that the hypothesis should be rejected? As always, statistical modelling alone cannot answer a theoretical question. There are reasons for thinking that, despite the lack of evidence from the experiments, scalarity causes a preference for monotonicity in learning. First of all, as discussed in section 1.2 the deep ties between monotonicity and scalarity are not manifested only in the semantics of gradable adjectives, but also of quantifiers. Second, previous research indicates that learning has a role to play in the evolution of quantificational monotonicity (section 1.3). Third, the theoretical considerations developed in the introduction give a mechanism for the way monotonicity might be caused by learning. Fourth, the evolutionary model presented in chapter 3 offers an elegant picture of the way a preference for monotonicity in learning might cause monotonicity to spread. Lastly, the models show that even a small bias for monotonicity might suffice to cause monotonicity to spread. Such a bias might be difficult to measure in experimental data. These considerations motivate further experimental research on the relation between scale based thinking and monotonicity.

Various possible variations on the design above might be implemented to explore further the relation between learning and monotonicity. A first possibility is to change the stimuli. If the stimuli are identical in the two conditions, they have to be capable of affording both a scalar and a distance-based interpretation, depending on the framing story. It is hard to strike a balance with respect to the difference affordances, for the following reason. A perceptually simple property cannot be

chosen as the dimension of stimuli variation—e.g. size or brightness—because there will usually be a pre-existing English category that might bias the category guessing behaviour. Therefore, stimuli will have to vary with respect to multiple properties. However, the multidimensional variation cannot be a linear combination of one-dimensional variations, or it will be possible to encode all change with any one of the dimensions. For instance, if the stimuli constantly increase in both height and brightness, any way of dividing the stimuli in convex sets will be encodable in terms of heights alone and in terms of brightness alone. The set of stimuli has to explore a non-linear[15] arc of a multidimensional feature space. However, such complicated patterns are difficult for participants to learn without extensive experience.

A second possible variation of the experiment is to use sets of stimuli that vary across conditions in a substantial way, as opposed to the relatively small variations of the sixth experiment. For instance, a stimulus set that is obviously scalar could be used in the property condition and highly-dimensional stimuli in the similarity condition. The problem with using very different sets of stimuli is the model's assumption that the parameters other than PM have the same distribution for both conditions. Using very different stimuli would invalidate this assumption, particularly for the PL parameter. A possible fix is to fit a population-level distributions for PL for each condition. However, doing so would make it harder for the model to distinguish between the effects of the preference for monotonicity and large categories.

A third possibility is to not show participants the scale throughout the experiment, but rather to let them infer it from observations of the single stimuli. This might force them to encode the stimuli, and this encoding is what is predicted to make a difference in the experiment. I did not do this because it might be very difficult for participants to infer the structure of the scale just by observing the stimuli.

Lastly, the simplest change would be to use a smaller set of stimuli. Since the number of convex categories increases faster than the number of monotonic categories as the number of stimuli increases, a smaller set of stimuli implies a greater proportion of monotonic categories. A greater proportion of monotonic categories would mean that a preference for monotonicity would have a greater effect on behaviour. In principle, simulations on artificial data could be run with different number of stimuli to check what the optimal proportion between monotonic and convex categories is

[15]To be more specific, the arc has to be non-monotonic in every dimension to avoid the possibility that every convex category can be reduced to any of the space's dimensions.

for inferring the true preference for monotonicity.

The Bayesian model could also be extended in exciting directions. In the current state, the model fits a parameters for monotonicity, trying to model to some extent the scale-based way of thinking discussed in the introduction. However, the model disregards some of the complexities of the distance-based way of thinking. In particular, the affordance provoked by salient stimuli—the "perfect" shapes like the circle—to be the prototypes of the category to be guessed is not modelled. Ignoring this aspect of prototype-based categorization strategies might have an impact on estimation. For instance, if a perfect shape exists near the border, it might cause participants to extend a category to include stimuli closer to the border, which the Bayesian model above might interpret as a preference for monotonicity. We tried to avoid this happening by disallowing pre-categorized stimuli too close to the border. An advantage of modelling the affordances given by salient stimuli would thus be the possibility of showing pre-categorized stimuli close to the border. The Bayesian model might be particularly helped in estimating a preference for monotonicity by how close to the border the pre-categorized stimuli have to be before participants guess a fully monotone category.

A second consequence of prototype-based thinking would be its interaction with the size of the guessed categories. In a scale with fewer stimuli that afford a prototype interpretation, the guessed categories will tend on average to be larger in a distance-based way of thinking. This is important because an unmodelled factor that interacts with the size of guessed categories might cause a bias in the estimation of the preference for large categories, which itself interacts with the estimation of the preference for monotonic categories.

The theoretical work presented in the introduction makes further empirical predictions that could be tested, allowing to go beyond minor manipulations to the current design and model. I mention two. First of all, a difference in the acceptability of overlaps between categories is predicted. If prototype-based categories form Voronoi tessellations, overlap between categories should be mostly excluded. On the other hand, if scalar categories are coded in terms of a binary order relation, overlaps should be relatively acceptable. A possible difficulty with testing this prediction comes from pragmatics. For instance, in a director-matcher task the director observes a stimulus and sends a signal representing a category to the matcher, who then picks among a set of possible stimuli. If overlapping categories are available, the director will always tend to pick the smaller ones among the ones that contain

the observed stimulus, calculating a scalar implicature. Therefore, the behaviour for overlapping categories will end up being similar as the behaviour for non-overlapping categories.

A second empirical prediction made by the theory in the introduction is that non-convex double-bounded categories should be equally complex to learn as convex double-bounded categories for scale-based categorization, but they should be harder to learn for prototype-based categorization. This follows from the fact that in scale-based categories a non-convex double-bounded category and a convex double-bounded category both require memorizing two thresholds plus one bit of information. On the other hand, in prototype-based categorization a convex double-bounded category requires two prototypes, while a non-convex double-bounded category requires three. A difficulty with this approach is the exponential increase in the number of possible categories as the number of bounds increases. This increase, barring other simplifying assumptions, might make it impossible to fit a Bayesian model to the data.

This chapter presented a computational way to detect preferences for monotonicity in category learning from rich categorization data. While the results did not confirm the hypothesis, tests on simulated data shows that the model is capable of recovering the true parameters. The approach presented above can be extended in multiple exciting directions which I leave for future work. In the next chapter, I move to a different universal of scalar semantics, namely extremeness.

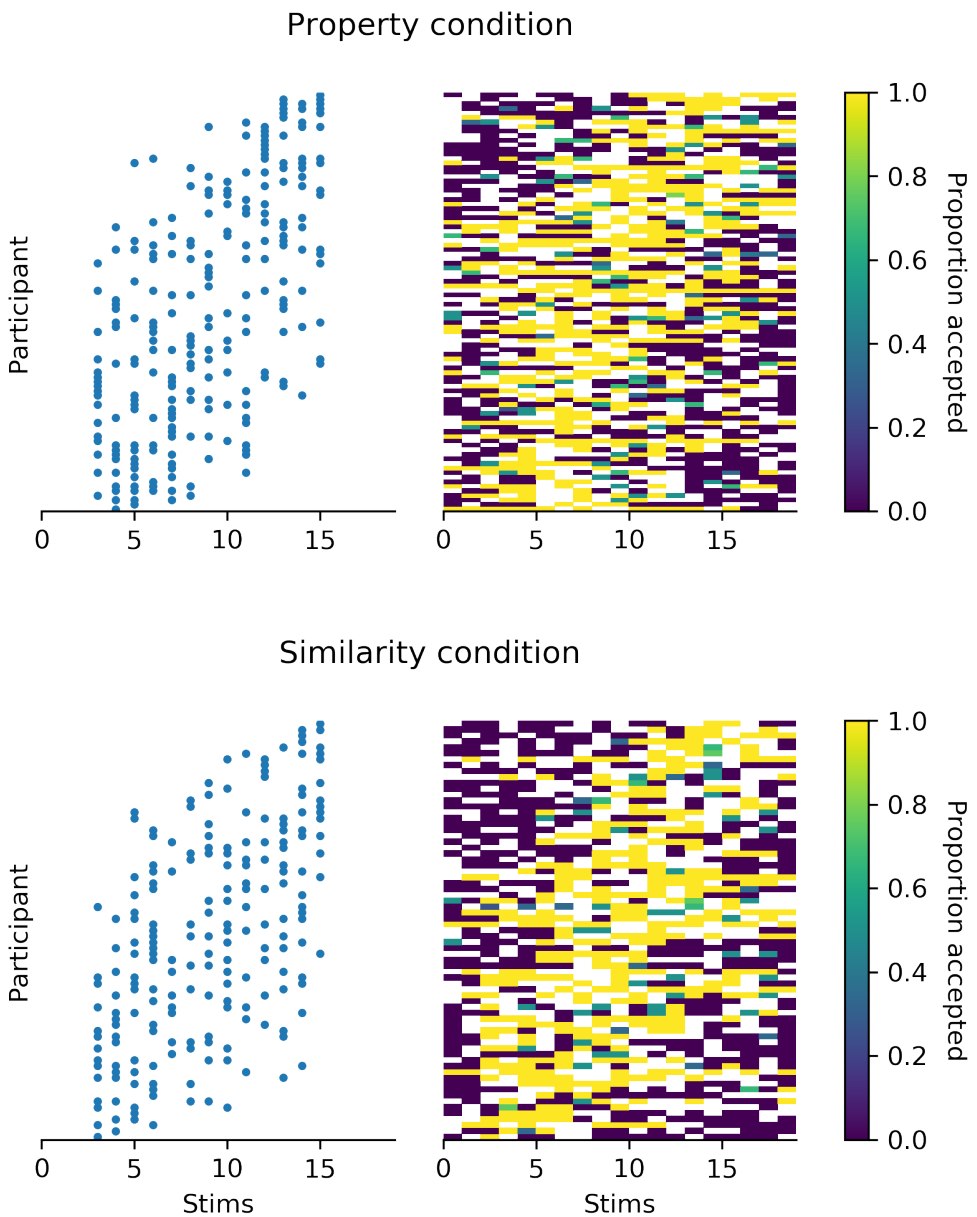


Figure 5.23: Visualization of the raw data of the sixth experiment. For an explanation of the plot, see figure 5.12.

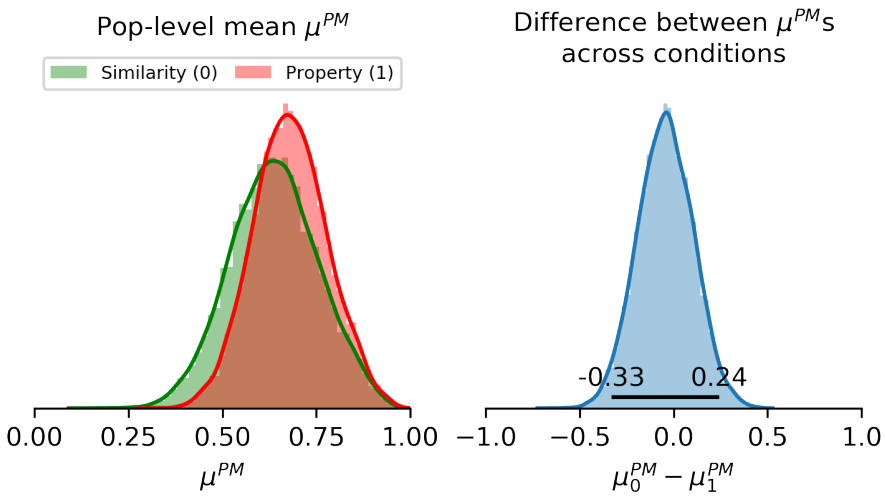


Figure 5.24: Plots visualizing the posteriors over monotonicity, along with 95% HPD interval, for the data from the sixth experiment. For more on how to interpret the plot, see figure 5.13.

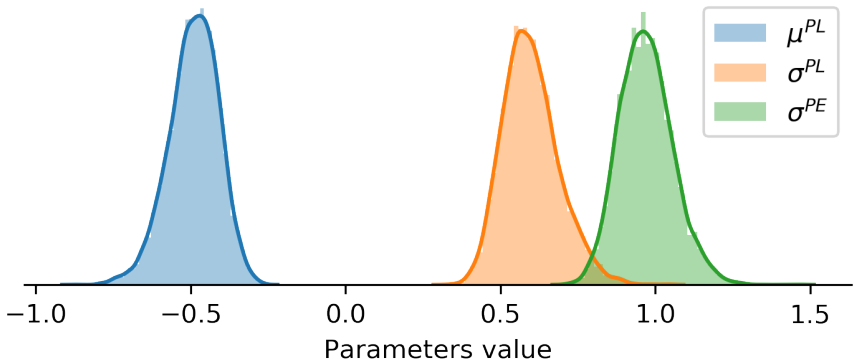


Figure 5.25: Plots visualizing the posterior densities over population-level parameters given the data from the sixth experiment. See figure 5.15 and 5.20 for same plot with data from previous experiments.

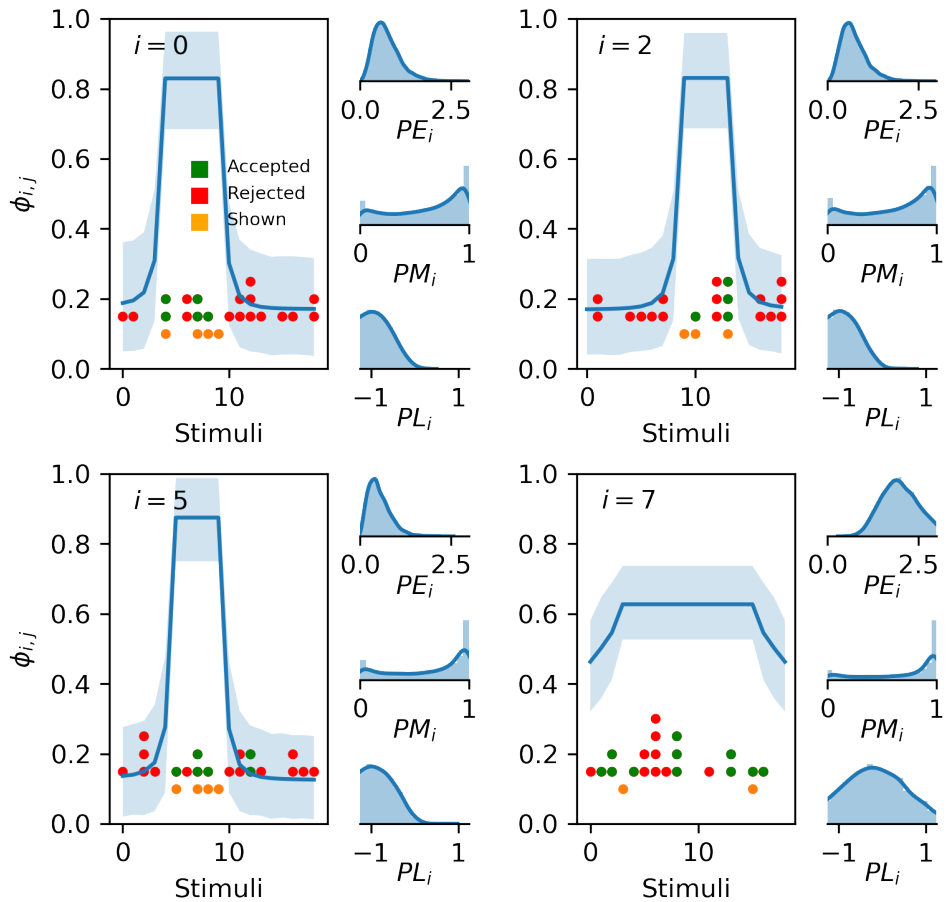


Figure 5.26: For explanation of this plot see figure 5.14.

Chapter 6

Modelling the evolution of absolute thresholds

In chapter 1, I presented previous work on the evolution of extreme categories, and identified two gaps. First, most previous work focussed on the role of communication in the evolution of extremeness. Second, most previous work considered categories that were close to extreme—in the sense of extremeness discussed in this thesis—rather than truly extreme. In this chapter, I examine cultural transmission as a possible cause for the evolution of extreme transitions.

To support this picture of the evolution of extreme transitions, I develop a computational model of cultural transmission and use of gradable adjectives based on previous work on pragmatic communication. Similarly to chapter 3, I combine an Iterated Learning (IL) model of cultural evolution (Kirby et al., 2014, 2015) with a Rational Speech Act (RSA) model of pragmatic communication (Goodman & Frank, 2016). The conclusion of the model in this chapter is that extreme categories can emerge often from a pure pressure from cultural transmission. I start in section 6.1 by presenting the model of communication and learning. Then, in section 6.2 I present the results.

6.1 Evolutionary model

6.1.1 Model of pragmatic communication

As I argued above in section 1.3.2, an explicit model of *pragmatic slack* is needed to model the evolution of adjectival semantics. Pragmatic slack, which I explain in more detail below, is the usage of categories to refer to individuals outside of their extension, as long as the literally inappropriate usage does not make difference for practical purposes. I implement such a model in the RSA framework. Similarly to the RSA models in chapter 3, a pragmatic receiver L_1 receives an utterance $s \in S$ with a known meaning and finds a posterior distribution over world states, $p_{L_1}(w|s, ...)$ (where "..." represents any further semantic or contextual parameter that might influence the interpretation of the signal). L_1 calculates the posterior assuming that the signal has been produced by a rational, cooperating sender S_1. More specifically, L_1 imagines a scenario where S_1 has perceived the real world state w and picked a signal that tends to be as helpful as possible to a literal receiver L_0. Helpfulness is quantified by a utility function $U_{L_0}(s, w, ...)$, and S_1's choice of signal is modelled by a softmax choice function:

$$p_{S_1}(s|w, ...) = \frac{e^{\alpha U_{L_0}(s,w,...)}}{\sum_{s_i \in S} e^{\alpha U_{L_0}(s_i,w,...)}} \tag{6.1}$$

The intuition behind this function is that the more useful an utterance is, the more likely it is that S_1 will produce it. The strength of this tendency is controlled by α; the greater the α, the stronger the tendency. In the limit of $\alpha = 0$, the choice is uniform across signals. For $\alpha = \infty$, S_1 always picks the most useful signal. Finally, L_0 is a literal receiver, who calculates a posterior over world states assuming simply that the signal is true.

The language of the agents in the model has three signals: s_σ (silence), s_+ (a positive polarity adjective), and s_- (a negative polarity adjective). Each adjective conveys that the real observed degree d_o falls in a certain part of the relevant scale. s_σ leaves the position unspecified, and is compatible with the whole scale. s_+ conveys that d_o is greater on the scale than a value θ_+. s_- conveys that d_o is lower on the scale than a value θ_-. Assume that the literal receiver and the pragmatic receiver have accurate priors about the distribution of degrees in the world, and that this distribution is $f(d)$.

L_0's posterior over degrees after receiving a signal is calculated as follows:

$$p_{L_0}(d|s_\sigma, \vec{\theta}) = f(d) \tag{6.2}$$

$$p_{L_0}(d|s_+, \vec{\theta}) = \begin{cases} \frac{f(d)}{\int_{\theta_+}^{\infty} f(x)dx} & \text{if } d \geq \theta_+ \\ 0 & \text{otherwise} \end{cases} \tag{6.3}$$

$$p_{L_0}(d|s_-, \vec{\theta}) = \begin{cases} \frac{f(d)}{\int_{-\infty}^{\theta_-} f(x)dx} & \text{if } d \leq \theta_- \\ 0 & \text{otherwise} \end{cases} \tag{6.4}$$

where $\vec{\theta} = \begin{bmatrix} \theta_+ & \theta_- \end{bmatrix}'$. In practice, L_0 renormalizes its prior over the portion of the scale compatible with the received signal, i.e. the whole scale for silence, degrees above θ_+ for s_+, and degrees below θ_- for s_-.

The rational sender calculates the utility of a signal after observing a degree based on L_0's posterior. Following previous literature (Franke, 2014), the utility of a signal is calculated as a function of the expected distance between the degree observed by S_1 and the degree that L_0 guesses:

$$U_{L_0}(s_i|d_o, \vec{\theta}, \vec{\lambda}) = \int_{-\infty}^{\infty} e^{-\lambda_i(d_o - x)^2} p_{L_0}(x|s_i, \vec{\theta})dx \tag{6.5}$$

where $\vec{\lambda} = \begin{bmatrix} \lambda_\sigma & \lambda_+ & \lambda_- \end{bmatrix}'$. The intuitive meaning of this formula is that the smaller the expected distance between S_1's observation and L_0's guess after sending a signal s, the greater the utility of s. How much greater? This is regulated by the λ parameters. For each signal, as its corresponding λ parameter increases, the importance of minimizing the expected distance also increases, and therefore the pragmatic slack decreases.[1] S_1 picks a signal according to eq. 6.1.

Finally, the pragmatic receiver L_1 simply does Bayesian inference to define a posterior over degrees given its prior over degrees and S_1's pragmatic behaviour:

$$p_{L_1}(d|s_i, \vec{\theta}, \vec{\lambda}, \alpha) \propto p_{S_1}(s_i|d, \vec{\theta}, \vec{\lambda}, \alpha) f(d) \tag{6.6}$$

In the models below, f is a uniform distribution in the $[0, 1]$ interval. This assumption allows simplifications that make the model computationally much less expensive, but

[1] Note that $\lim_{\theta_+ \to 1} U_{L_0}(s_+|d_o, \vec{\theta}, \vec{\lambda}) = e^{-\lambda_+(d_o - 1)^2}$. This is intuitively right because when the threshold of s_+ is at 1, the receiver always guesses the scale's maximum. For similar reasons, $\lim_{\theta_- \to 0} U_{L_0}(s_-|d_o, \vec{\theta}, \vec{\lambda}) = e^{-\lambda_- d_o^2}$.

could be lifted in future research. α is set to 4.

This models inherits from Franke (2014) the prediction that signals can be used to refer to degrees outside of the literal extension, a phenomenon known as *pragmatic slack* (Lasersohn, 1999). A real example of pragmatic slack is the literally false but pragmatically apt usage of "full" for recipients that are less than completely full (Fig 6.1, top row). Pragmatic slack is a crucial aspect of the phenomenon at hand. Short of considerations about the finite precision of the human perceptual system, absolute adjectives refer to single points of a continuous scale, which is a set of measure zero for a continuous distribution. Therefore, a degree sampled from the scale will (almost surely) fall outside of the absolute adjective's extension. Literal language users, who only produce adjectives that contain the observed degree, would (almost surely) never have an occasion to use an absolute adjective. Absolute adjectives would perform poorly in communication, and language learners would not have any occasion to acquire them. Pragmatic slack is therefore a prerequisite for absolute adjectives to stand a chance to evolve.

A second phenomenon that is captured by the RSA model is the fact that the amount of pragmatic slack varies in different contexts. For instance, the same metal rod might count as straight in a construction site but not in a clocks factory. The amount of pragmatic slack is regulated by the parameter λ_i; the higher λ_i, the smaller the slack, the closer a degree outside of s_i's extension has to be to the θ_i for senders to use s_i (Fig 6.1, bottom row). Therefore, by changing the value of λ_i, different pragmatic slacks can be obtained.

In the model, I assume that extreme thresholds are fixed at the semantic level and are in that respect different from relative adjectives (Qing & Franke, 2014a).[2] This is in contrast with a view according to which the extreme threshold of absolute adjectives is determined by the context (Lassiter & Goodman, 2013). I assume this picture and leave implications of the model under a different picture to future work.

6.1.2 Learning from pragmatic data

A learner in the model observes data produced by a rational sender and tries to infer the unobserved sender's semantic parameters $\theta_+$and θ_-. The data D consists of a set of n tuples $\langle o_i, s_i \rangle$, where o_i is the i^{th} degree observed by the sender and s_i is the signal sent by the sender to refer to o_i.

[2]See Aparicio et al. (2016) for an overview.

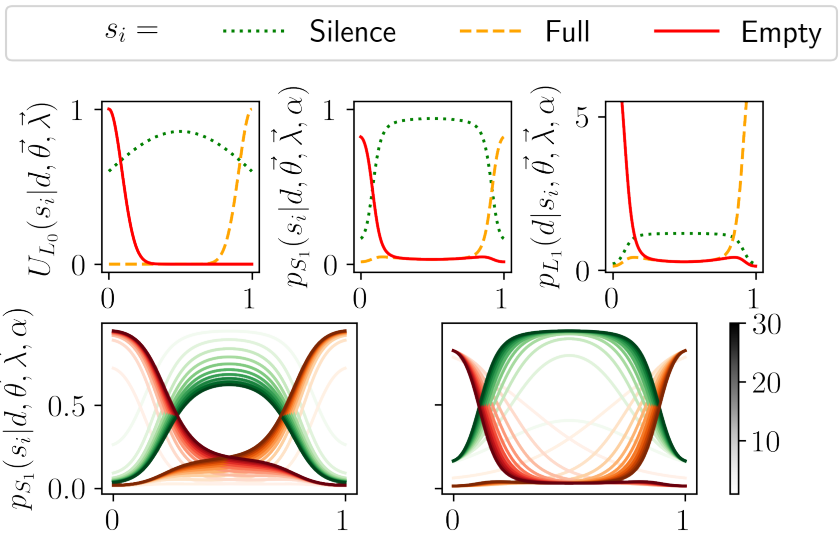


Figure 6.1: The Rational Speech Act model of pragmatic communication with "full" and "empty", i.e. with $\theta_+ = 1$ and $\theta_- = 0$. All x-axes are the scale's degrees. Top row: Utilities, rational sender and rational receiver with $\alpha = 4, \lambda_\sigma = 2, \lambda_+ = \lambda_- = 60$. Bottom row: Effect of variation of λ parameters on production behaviour. The left plot shows the effect of changing λ_0 as $\lambda_+ = \lambda_- = 15$. As λ_+ increases, silence is used less and less even for central degrees. The right plot shows variations of λ_+, λ_- with $\lambda_\sigma = 2$. As λ_+, λ_- increase, less pragmatic slack for "full" and "empty" is allowed. The colorbar shows the relation between color lightness and value for all colors.

The learners in the model are Maximum A Posteriori (MAP) agents, meaning that they pick the combination of parameters that has the greatest posterior probability given the observed data. There are two reasons for using MAP agents. First, sampling from the posterior distribution would be computationally more expensive than finding the MAP. Since the model is already computationally intensive, using sample agents would make the model prohibitively slow. Second, and more importantly, sample agents will (almost surely) not sample an extreme category. This is because extreme categories cover a part of the parameter space with measure 0, namely the four individual points where the transitions are at 0 or 1. Since extreme

categories are often observed in natural language, we can infer that learners cannot be samplers in acquiring them. MAP is a natural alternative option that allows for extremeness to occur.

Since the MAP cannot be found analytically and the posterior over language parameters can be multimodal, I used the sampling-based *basin-hopping* algorithm (Wales & Doye, 1997) to find the language parameters that maximise the probability of the learner's observations. Basin-hopping is an algorithm to find the global minimum of a function by drawing successive samples. The algorithm is particularly useful when the function to optimize is multimodal, and the point of interest is the global maximum. The algorithm is based on an iterative process of finding a local minimum, and then introducing noise to attempt to enter the basin of attraction of another local minimum. The algorithm is displayed in algorithm 1.

Algorithm 1 Basin-hopping

```
function BASIN-HOPPING
    x_1 ← RANDOM()                                  ▷ Random initialization
    i ← 1
    while STOP is false do                          ▷ STOP is some stopping condition
        y ← PERTURB(x_i)    ▷ Find a point by perturbing the present minimum
        x_{i+1} ← LOCALOPTIM(y)                     ▷ Perform local optimization from y
        if f(x_{i+1}) < f(x_i) then
            i ← i + 1
        end if
    end while
    return x_i
end function
```

In order to check that the model is behaving as intended, it is instructive to observe that the more datapoints are observed by the learner, the more accurate the guess of the parent's language becomes. This is visualized in figure 6.2 for a variety of parameter combinations.

6.2 Results

I simulate chains of IL for various combinations of parameters. Each chain consists of 10000 generations of single agents. As the results in figure 6.3 show, signals with extreme thresholds evolve more often than would happen by chance.

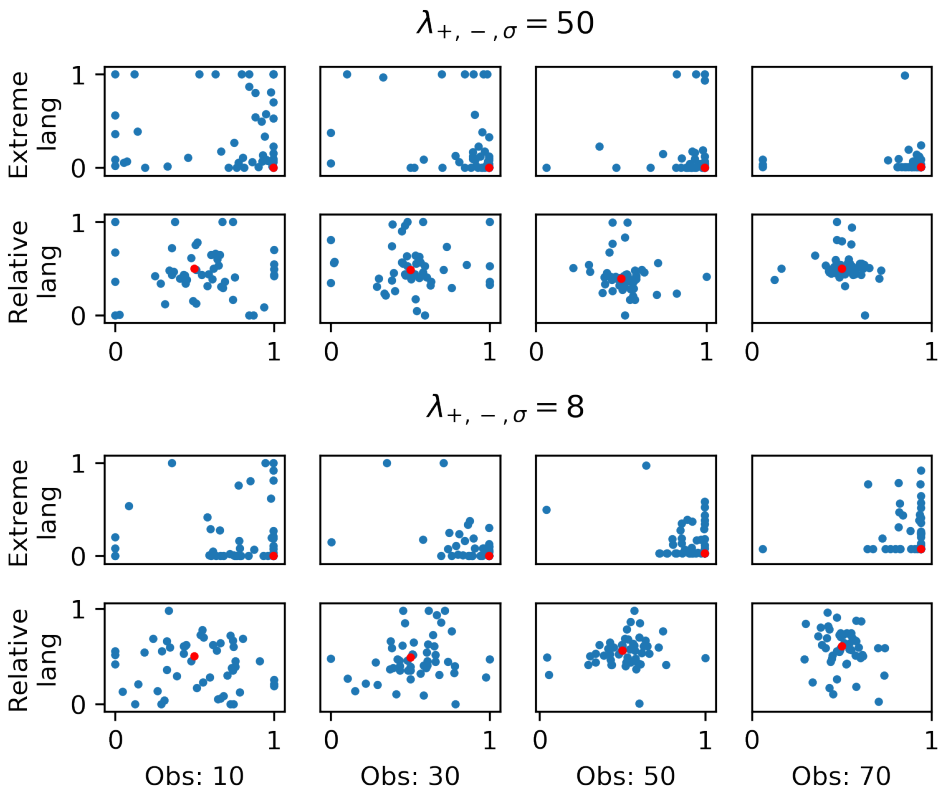


Figure 6.2: Languages learned by agents (blue dots) after various numbers of observations and combinations of parameters given the true language (red dots). λ determines the amount of pragmatic slack for the different signals. Relative languages are the ones where the thresholds do not lie on an extreme of the scale. In the plot relative languages have $\sigma_+ = \sigma_- = 0.5$. The transitions of extreme languages, on the other hand, are at the scale's extrema. The main result from the plot is that, as expected, learners that observe more production data on average guess the teacher's language more accurately.

Two observations can be drawn from the results about how parameter values influence the tendency of extreme languages to emerge. Firstly, more observations lead to fewer extreme meanings evolving. Secondly, greater values of all λ parameters, i.e. less pragmatic slack, lead to fewer extreme meanings evolving.

More information about the behaviour of the IL chains can be obtained by comparing the parent (figure 6.4) and children (figure 6.5) languages of extreme languages

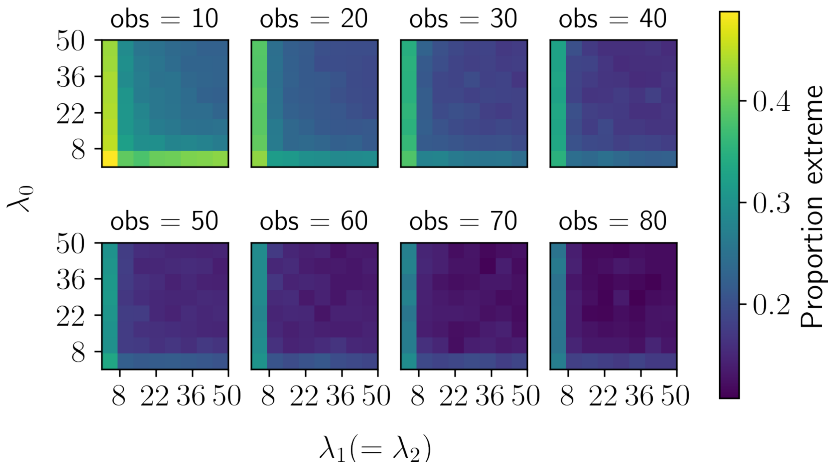


Figure 6.3: Results of IL model for various combinations of numbers of learner observations and values of the λ parameters. The color indicates the proportion of all learned meanings that are extreme, where extreme means that the transition happens at 1 or 0. When the λ parameters take smaller values, i.e. when there is a high level of pragmatic slack, up to half of all meanings are extreme.

for combinations of parameters that produces a high or a low proportion of extreme languages. This comparison shows that combinations of parameters that develop a high proportion of extreme meanings are the ones where the estimation of the language is noisier.

6.3 Conclusions

At this stage, it is unclear why extreme meanings evolve often in a pure Iterated Learning condition. A prima facie possible reason for the stability of extreme thresholds in the cultural evolution of language is that they produce very characteristic data. In other words, a learner observing data produced by extreme thresholds will have an easier time learning it accurately than if the language had been produced by non-extreme thresholds. Once a population stumbles upon an extreme language, the language will be transmitted with high fidelity and will tend to persist in the population over time. However, this hypothesis is contradicted by the observation

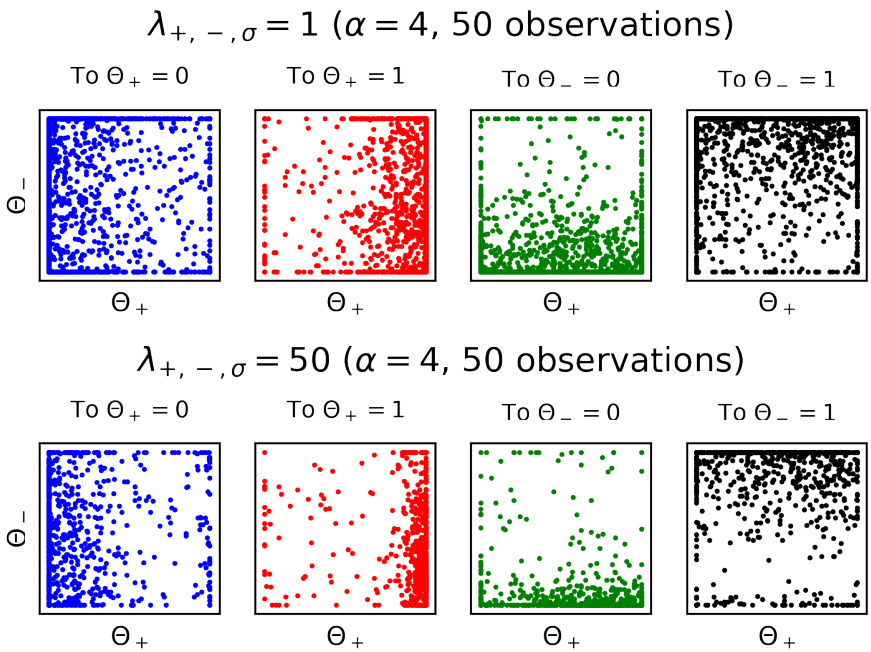


Figure 6.4: In all plots, the x-axis is the position of the threshold for the positive signal, and the y-axis is the position of the threshold for the negative signal. Top row: languages of the teachers whose learners acquired at least one extreme meaning, for a combination of λ parameters that produced a higher proportion of extreme languages. Bottom row: languages of the teachers whose learners acquired at least one extreme meaning, for a combination of parameters that produced a lower proportion of extreme languages. Each column of plots (indicated by color) corresponds to a different extreme meaning acquired by the learner. For instance, the first column (blue) shows the languages of teachers whose learners acquired a positive signal with a threshold at the scale's minimum ($\Theta_+ = 0$). The figure shows that extreme languages do not tend to have almost-extreme parents more often for the combination of parameters that produced a high proportion of extreme languages.

that learners with extreme-language teachers do not end up closer to their teacher's extreme language in combinations of parameters that produce a high proportion of extreme languages. Further, the average distance between two successive languages for a parameter setting that produced a relatively high proportion of extreme languages ($\lambda_{+,-,\sigma} = 1$, with $\alpha = 4$ and 50 observations) is greater (≈ 0.334) than the

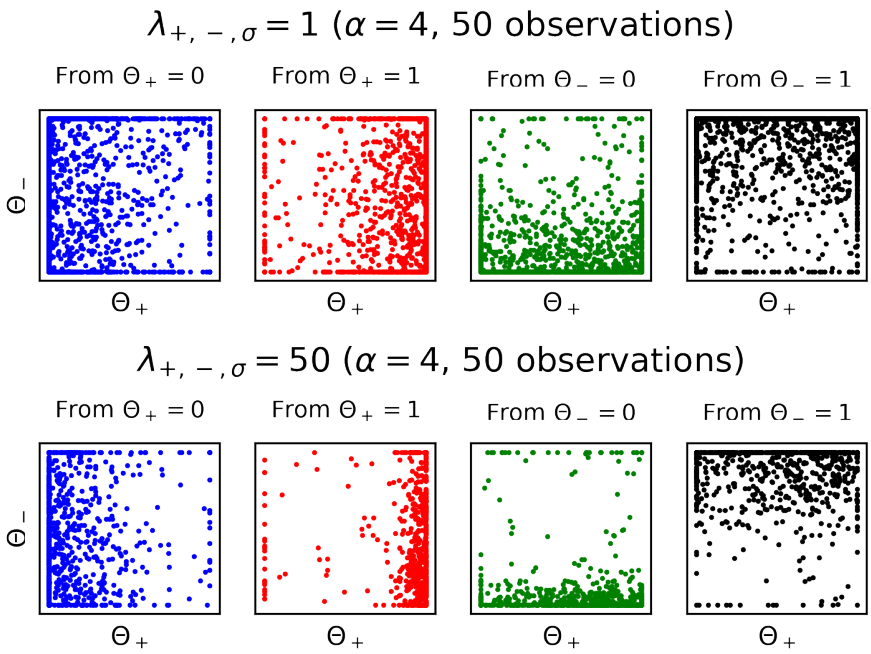


Figure 6.5: Top row: languages of the learners whose teachers acquired at least one extreme meaning, for a combination of λ parameters that produced a higher proportion of extreme languages. Bottom row: languages of the learners whose teachers acquired at least one extreme meaning, for a combination of parameters that produced a lower proportion of extreme languages. See figure 6.4 for more information.

average distance between successive languages with a combination of parameter values that produced few extreme languages ($\lambda_{0,1,2} = 50$, with an average distance of ≈ 0.18). Understanding the reason for the tendency towards extremeness is made more difficult by the complex inference of RSA agents. I leave further exploration of the model to future work.

The model presented in this chapter could be extended in various ways. First, direct selection for communicative accuracy could be implemented. Communicative accuracy might have the effect of keeping the languages more stable across generations, and pragmatic agents could use extreme languages without losing in terms of communicative accuracy. However, selection for communicative accuracy would

require a population of multiple agents, which would be computationally more expensive.

A second possible extension of the model considers more complex languages, especially in combination with a pressure for communicative success, to study whether patterns observed in real languages emerge. A language in the present model consists of two signals of opposite polarity, plus silence. However, more signals could be introduced, therefore having multiple signals of the same polarity.

Another extension would be to consider non-uniform prior over degrees. In the present model, the uniform prior over the unit segment corresponds to an open scale, i.e. a distribution with significant probability mass at two extremes. Such distributions generally give rise to two extreme antonyms. Other distributions could be implemented that correspond to half-open (significant probability mass on only one of the extremes) and open (little mass at the extremes) scales.

Finally, one the inaccuracies of the model is that thresholds, even when non-extreme, are crisp. Vagueness could be introduced in the model via uncertainty of the agents about the distribution of the property in the population. This could be implemented as a hyperprior over the parameters of the prior distribution over property, similarly to (Zhao & Cremers, n.d.).

In this chapter, I made two contributions to the literature. First, I built on previous work on pragmatic communication to provide a computational model where the cultural evolution of crisp extreme thresholds can be studied, because extreme categories are sometimes used to refer to non-extreme observations. Second, I showed that even without an explicit pressure from a cognitive bias, extremeness can evolve more often than it would by chance. This advantage might combine with cognitive biases for extreme thresholds coming from their salience and other previously explored mechanisms.

Declaration

The following chapter was written collaboratively with Shane Steinert-Threlkeld and Jakub Szymanik, at the Institute for Logic, Language, and Computation (ILLC) at the University of Amsterdam. The ideas were developed and the results mostly produced during my visit to the ILLC in Amsterdam between the 20th and the 28th of November 2018. Roughly, I developed the part of the project concerning cultural evolution and IL, while Shane Steinert-Threlkeld and Jakub Szymanik developed the parts concerning the neural networks and agents model, based on the previous work in Steinert-Threlkeld and Szymanik (2019). The work below was accepted for an oral presentation at CogSci in 2019 (Carcassi, Steinert-Threlkeld, & Szymanik, 2019), and appears here unmodified.

Chapter 7

Modelling the evolution of quantificational monotonicity

7.1 Introduction

While natural languages show great variability, there are features that they all appear to share. Linguists call these features linguistic *universals*. Universals have been found at all levels of linguistic structure: phonological, syntactic, and semantic.[1] Some universals might follow from constraints on what humans are physically capable of doing. For instance, there is no language whose prosody requires the production of ultrasounds. The reasons for other universals are harder to understand, leading to multiple proposed explanations.

One well-supported claim is that at least some universals are to be explained in terms of *learnability*. [2] More precisely, it is easier to learn a language that satisfies the universal than it is to learn a language that does not satisfy the universal, and this difference in the complexity of acquisition causes languages that satisfy universals to spread. In the case of universals of lexical semantics such as the one we focus on below, the learnability explanation says that lexical entries whose meaning satisfies the universal are easier to learn, and therefore more likely to be lexicalized. Complicated meanings can be obtained through complex grammatical constructions and compositional interpretation thereof.

The learnability explanation is an empirical, causal claim about the origins of

[1] For some examples see, respectively, Hyman (2008), Newmeyer (2008), and Barwise and Cooper (1981).

[2] See, e.g., Steinert-Threlkeld (2019), Piantadosi et al. (2012), and Peters and Westerståhl (2008).

linguistic universals. One way to support the learnability explanation for a specific universal is to provide a model of learning that is cognitively realistic and on which expressions that satisfy the universal are indeed easier to learn.

Finding an appropriate model of learning can however only partially explain a linguistic universal. Learnability is a fact about individual cognition, while a universal is a feature of a whole language. A second challenge consists in connecting these two levels, showing the effects of learnability on emerging language structure. This is the so-called problem of *linkage*.[3]

Iterated learning (IL) is a method that addresses the problem of linkage. In IL, parents teach children their language, who teach the next generation their language, and so on and so forth. The crucial insight of IL is that learning is not an inert process in cultural evolution, since the languages of a cultural child and its cultural parent are generally slightly different. The changes caused by learning are not random, but rather tend to be guided by the learner's cognitive biases. As a consequence, over time languages adapt better and better to the agents' cognitive biases. Learnability can then affect the frequency of different traits.[4]

Previous work has addressed the learnability challenge by showing that quantifiers, responsive predicates, and color terms that satisfy certain semantic universals are easier to learn for neural networks.[5] In this paper, we address the problem of linkage by building an iterated learning model of the evolution of the semantic structure of quantifiers. In particular, we will use neural networks as our agents and standard gradient descent as the learning method inside the context of iterated learning. The next section briefly reviews the theory of generalized quantification and the universal of *monotonicity*. After that, the following section presents the model of cognition and the iterated learning model, as well as an information-theoretic measure of the *degree of monotonicity* of a quantifier. Experiments with this model and their results are presented in the following section. Results are discussed in the final section, along with possible future directions.

[3]The problem of linkage was introduced in Kirby (1999).

[4]See, e.g., Tamariz and Kirby (2016); Culbertson and Kirby (2016); Kirby et al. (2008) for discussions of the way individual cognition is reflected in language structure through IL and experimental evidence supporting the connection.

[5]See, respectively, Steinert-Threlkeld and Szymanik (2018); Steinert-Threlkeld (2019); Steinert-Threlkeld and Szymanik (2020).

7.2 Quantifiers and monotonicity

Determiners are expressions that take a common noun as an argument and return a Noun Phrase. Determiners can be grammatically simple—e.g. *some*, *few*, *most*—or complex—e.g. *fewer than three* or *at most five*.[6] Determiners express generalized quantifiers.[7] (Monadic) Generalized quantifiers are properties of sets of subsets of a domain of discourse. The generalized quantifiers expressed by natural language determiners are of type $\langle 1, 1\rangle$, i.e. properties of exactly two sets. Equivalently, a quantifier of type $\langle 1, 1\rangle$ takes (the characteristic function of) a set A and returns a function from (the characteristic function of) a set B to truth values. A is the *left argument* and B the *right argument* of the quantifier. For instance, the sentence "most As are B" is true if and only if the number of As that are B (cardinality of the intersection of A and B, i.e., $|A \cap B|$) is greater than the number of As that are not Bs (i.e., $|A - B|$), i.e.:

$$[\![\text{most}]\!] = \{(A, B) : |A \cap B| > |A \setminus B|\}$$

Various universals have been proposed about which generalized quantifiers are expressed by simple determiners. In the following, we focus on the *monotonicity* universal proposed by Barwise and Cooper (1981). This says that all simple determiners across all languages express monotone quantifiers. A quantifier is monotone iff it is *upward* monotone or *downward* monotone. A quantifier Q is upward monotone [downward monotone] iff for any three sets A, B and B', if $Q(A)(B)$ and $B \subseteq B'$ [$B' \subseteq B$] then $Q(A)(B')$. As an example, consider the upward monotone quantifier $[\![\text{most}]\!]$. Assume that the sentence "Most cats sleep" is true and that everything that sleeps is alive, i.e. $[\![\textit{sleep}]\!] \subseteq [\![\textit{alive}]\!]$. The monotonicity of $[\![\text{most}]\!]$ ensures then that "Most cats are alive" is true.

Monotonicity is an interesting universal because it is easy to imagine non-monotone quantifiers. Examples of non-monotone quantifiers abound among the meanings of complex determiners: "an even/odd number of" or "exactly 2", etc. The commonness of non-monotonicity among complex quantifiers makes the lack of simple non-monotone quantifiers especially puzzling and in need of an explanation. Previ-

[6] Exactly how to draw the distinction between simple and complex and whether, for instance, *most* is simple or complex, do not matter for present purposes.

[7] For more information on generalized quantifier theory from linguistic, computational, and cognitive perspectives, see also Peters and Westerståhl (2008) and Szymanik (2016).

ous work proposed to explain the universal of monotonicity in terms of the greater learnability of monotone quantifiers.

Steinert-Threlkeld and Szymanik (2018) propose to use neural networks in this context. A neural network is a computational device that can learn to approximate functions by observing tuples of inputs and relevant outputs, and progressively minimizing a suitably defined distance between the true output and the network's own prediction. In the case of a quantifier, the input is a structure where the relevant sets are specified and the output is 1 iff the structure verifies the quantifier and 0 otherwise. In practice, given a structure the neural network outputs a probability that can be interpreted as confidence that the structure verifies the quantifier.

Data about how fast neural networks learn different kinds of quantifiers was produced with the following algorithms. First, two quantifiers are picked such that one satisfies the universal and the other does not. Then, the two quantifiers are taught to a neural network until it has accurately learned them. The crucial information is how long on average it takes neural networks to accurately learn quantifiers that satisfy the universal compared to ones that do not. Various universals were tested in this way. In the case of monotonicity, the data was produced both for a downward monotone and for an upward monotone quantifier. The neural networks were strikingly faster at learning monotone compared to non-monotone quantifiers. Figure 7.1 shows an example.

As discussed above, knowing that meanings with certain features can be learned more easily only goes some of the way in explaining the features' universality across various languages. A full explanation also needs to show that the structure can and eventually will be reached by processes of cultural evolution. In the rest of this paper, we develop an iterated learning model of the cultural evolution of quantifiers that embeds the learning model of neural networks, and show that monotonicity reliably emerges.

7.3 Methods

7.3.1 Iterated learning

IL models start with two groups of randomly initialized agents, the first and second generations. Each agent in the first generation—the *cultural parent*—is associated with one agent in the second generation—the *cultural child*. A set of linguistic

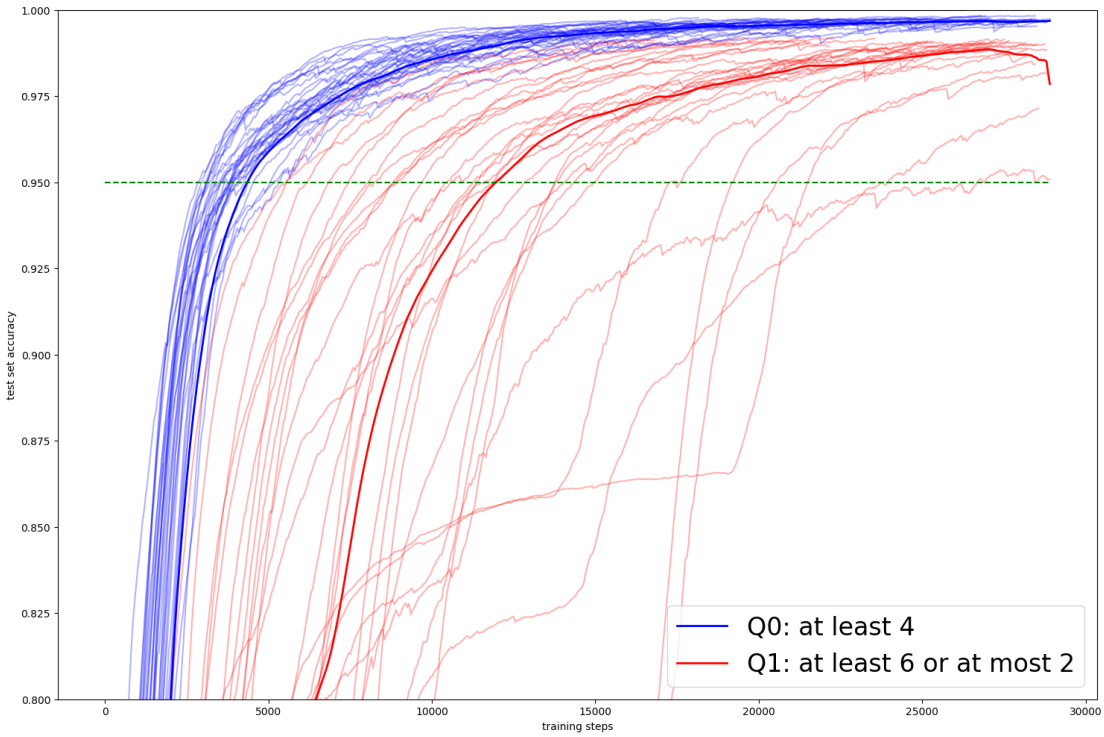


Figure 7.1: Learning curves on a neural network for the monotone *at least 4* (blue) versus *at least 6 or at most 2* (red). The x-axis is number of training steps; the y-axis is accuracy (percentage correct) on a test set of examples the network has not yet seen. This was Figure 4 in Steinert-Threlkeld and Szymanik (2019).

production data is generated for each cultural parent and used as input for the cultural child. The cultural child tries to approximate its cultural parent's language. In the following step, the process is repeated with agents in the second generation as cultural parents and the new agents in a third generation as cultural children. The cultural transmission process is iterated for some number of generations. Each cultural family line is called a *chain* of IL.

Crucially, the agents do not learn their parent's language perfectly. There can be various reasons for this. First, there can be a bottleneck in learning. This happens when the learner does not observe everything that is needed to perfectly reconstruct the language, and therefore has to guess some aspects of it. The number of data points given to the learners is fixed for all generations and agents and is called the *bottleneck size*. A second reason is that the agent might not have perfect memory or perfect reasoning abilities, and might therefore learn a language that does not perfectly conform to the given data. In this case, the more rational the agent, the closer the learned language will be to the teacher's language. A third reason is that the cultural parents might produce language in a way that is stochastic rather than deterministic. This can make the language harder to approximate and impossible to learn perfectly. For instance, a cultural parent might pick among the signals compatible with a certain observation according to a categorical distribution. The cultural child would need to infer the parameters of the distribution, a task which cannot in general be accomplished perfectly with a finite number of observations.

The changes introduced by each learner accumulate over generations. Since these changes are not completely random, but rather tend to be consistent across agents, the languages tend to change in the same way over time in different chains. In sum, IL is a way to study how the cognitive system of the learners determine which languages one should expect to see spoken in a population of such agents. The crucial individual level components of an IL model are the set of possible languages, and the way the agents learn them. We now explain these two components in turn.

7.3.2 Model of models, quantifiers, and language

Since the focus is on the evolution of monotonicity, we simplify the language model by assuming that the quantifiers are conservative and extensional.[8] This amounts

[8]These, next to monotonicity, are two prominent semantic universals distinguishing natural language quantifiers from all logically possible quantifiers. Extensionality means that extending or

to saying that the truth value of each quantifier only depends on the elements in A and $A \cap B$, and not on $\overline{A \cup B}$ or $B \setminus A$. Therefore, the truth of any quantifier depends only on which of the elements of A are also elements of B, and which are not. Assuming conservativity and extensionality both reduces the number of possible quantifiers that agents can speak and simplifies the model of each quantifier, since only A and $A \cap B$ need to be encoded. Moreover, we assume that the left argument of the quantifiers is fixed to some set A with cardinality n.

Assuming conservativity/extensionality and a fixed set A, we can represent the part of the world—called a *model*—that is relevant to determining the truth value of a quantifier as a bit vector of a fixed length n. Each element of the model represents an object in A. Each element has value 1 iff the object corresponding to that bit is also an element of B, and 0 otherwise. For instance, the vector $[\,0, 1, 1\,]$ would model a situation where $A = \{o_1, o_2, o_3\}$ and o_2, $o_3 \in B$. The set of models is the set of all binary strings of length n, i.e. the set of possible relations between a fixed A and any possible B. We call M' a *submodel* of a model M iff M' is 0 everywhere where M is 0. For instance, $[\,0, 1, 1, 0, 0\,]$ is a submodel of $[\,0, 1, 1, 1, 1\,]$. Note that each model is a submodel of itself.

We represent a *quantifier* as a function from models to single bits. An example of a quantifier is $Q(x) = 1$ if $\sum_{i=1}^{n} x_i > 2$ otherwise 0, meaning "more than two". Since for A of size n there are 2^n different models, each quantifier is a 2^n-sized bit vector. Each element of the quantifier vector corresponds to a model and has value 1 iff the model verifies the quantifier and 0 otherwise.

To see how this works in practice, consider a set A of size 3. There are 8 possible ways in which any other set B can overlap with A. Each of these is modelled as a bit vector of size 3. For instance, $[\,0, 1, 1\,]$ says that the second and third object of A are also elements of B, but the first is not. The English expression "all As are B" is modelled by a bit vector of size 8 that has value 1 at the index corresponding to the model $[\,1, 1, 1\,]$ and 0 otherwise. If the models are ordered lexicographically[9] and the last model is therefore $[\,1, 1, 1\,]$, then the quantifier corresponds to the vector

shrinking the universe of discourse has no effect on the truth-value of the quantifier sentence as long as the left and right arguments are unchanged. Conservativity means that only the part of B that is common to A matters for the truth-value of the sentences. In other words, the elements in $B \setminus A$ can be safely ignored when determining the truth-value. See Peters and Westerståhl (2008) for definitions.

[9]In that case, lexicographic order is the dictionary order over sequences of letters from the alphabet $\{0, 1)$ with 0 preceding 1 in the order.

$[\,0, 0, 0, 0, 0, 0, 0, 1\,]$. We call a quantifier *degenerate* if and only if it corresponds to a vector of identical elements, 0s or 1s. A degenerate quantifier corresponds intuitively to a quantifier that is true or false of every model.

Each agent encodes a single quantifier. Agents do not encode the quantifiers directly. Rather, given a model they produce a truth value by using a neural network. The next two sections clarify the connection between the neural networks and the agent's behaviour.

7.3.3 Neural Networks

Because of the aforementioned learnability results of Steinert-Threlkeld and Szymanik (2018), the agents that make up the generations in our iterated learning setup are *neural networks*. Each network has n input neurons (one for each bit of a vector corresponding to a model) and one output neuron (how probable it thinks that the true output is a 1), with two hidden layers of 16 neurons each. We made this choice so that the networks had enough expressive power to represent many quantifiers, including complex ones. Future work will analyze the effect of architecture choices on the results presented below. The networks and learning, which will be described in the next section, were implemented in PyTorch.[10]

Such a network learns from input/output pairs using a fancier version of gradient descent called Adam (Kingma & Ba, 2017). The network receives a number of true input/output pairs, which it iterates over in small batches. For each batch, it guesses the correct outputs for the inputs, and then updates its parameters (weights and biases connecting the neurons) in such a way that its future outputs are guaranteed to be closer to the truth.[11] Because this style of learning is fairly gradual, we introduce one more parameter to our simulations, namely *number of epochs*: this is how many times the network processes its training set in each generation. In other words, the network sees a portion of its parent's language, as determined by *bottleneck size*, but gets to learn from that portion number-of-epochs times.[12]

[10] `http://pytorch.org`

[11] For general introductions, see Nielsen (2015); Goodfellow, Bengio, and Courville (2016).

[12] In some experimental literature — for example, Carr, Smith, Culbertson, and Simon (2018) — this is also referred to as *exposures*.

7.3.4 Model of the agents

Each agent plays two roles in an IL simulation. The first role is to learn a language given data from the previous generation. The second role is to produce data used to teach to the following generation. To produce this data, the agent is prompted with randomly chosen models.

In the learning phase, each agent receives learning data consisting of a set of tuples ⟨model, judgment⟩. The judgment is a single bit expressing whether the quantifier used by the agent is compatible or not with the model. This data is used to train the agent's neural network as described in the previous subsection.

Production works as follows. The agent feeds an observed model to its neural network. The neural network returns a number in the $[0, 1]$ interval. Then, the agent rounds the number and returns it. The returned number expresses whether the agent's quantifier is compatible or not with the model that the agent observed. The production behaviour is deterministic, since an agent always produces the same bit given the same model.

Prompted with a string of 1s and 0s, agents produce a 1 or 0. The former models a state of the world, the latter models the compatibility of the agent's quantifier with the world state. However, nothing in the simulation implies that neural networks are interpreting 1 and 0 as True and False respectively in their input and output. Therefore, the output of an agent under-determines which quantifier the agent speaks, even when the output for all models is known. For instance, an agent that returns 1 for input $[\,0, 0, 1, 1\,]$ can be interpreted as accepting the model where $B = \{o_3, o_4\}$ (if 1 is interpreted as True in the model and in the quantifier), as rejecting the model where $B = \{o_3, o_4\}$ (if 1 is interpreted as False in the quantifier and True in the models), as accepting the model where $B = \{o_1, o_2\}$ (if 1 is interpreted as True in the quantifier and False in the models), or as rejecting the model where $B = \{o_1, o_2\}$ (if 1 is interpreted as False in the quantifier and the models). Crucially, the interpretation of the bits has to be consistent across the models and across the quantifier judgments. Therefore, each agent can be interpreted as speaking four quantifiers, depending on whether 1 and 0 are interpreted as meaning true or false in the models and in the agent's output. We discuss below how we deal with underdeterminacy when it might make a difference to the interpretation of the results.

7.3.5 Measures of monotonicity

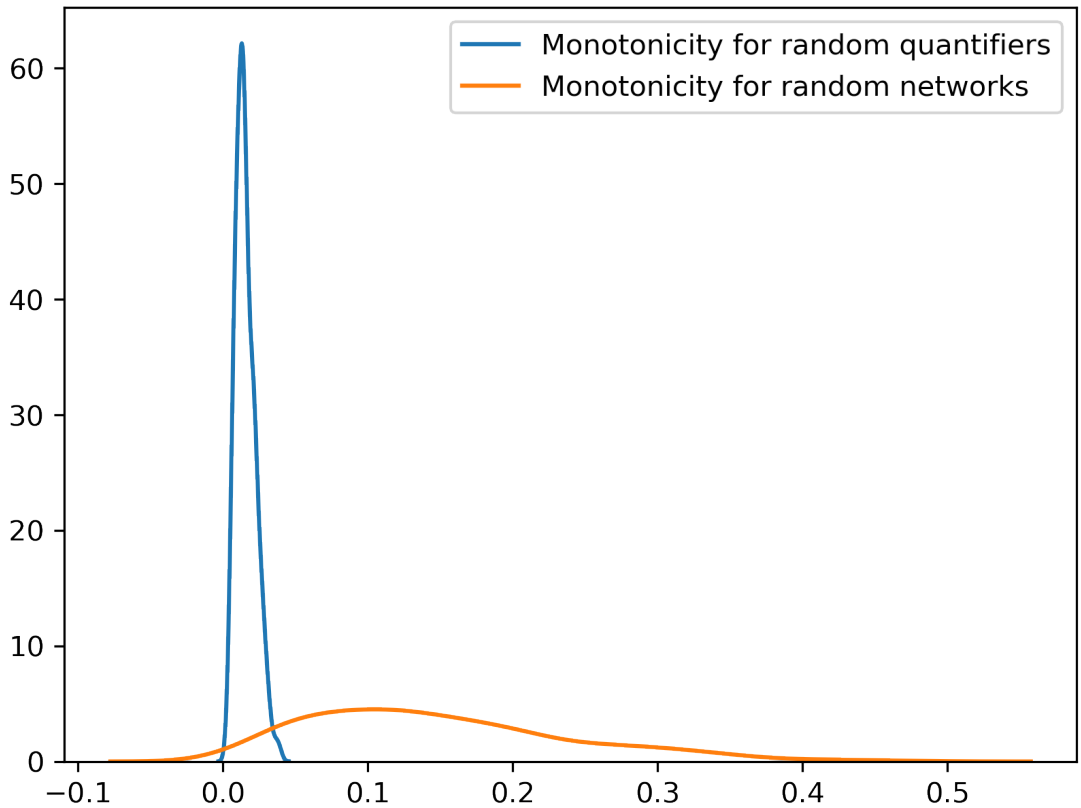


Figure 7.2: Kernel Density Estimation of the distribution of degrees of monotonicity from a sample of 300 completely random quantifiers and 300 random neural network agents. The x-axis is the measure of monotonicity we describe in the main text.

According to the standard definition, monotonicity is a binary property. A possible way of analyzing the results would be to find the proportion of monotone languages in every generation. However, some quantifiers are intuitively more monotone than other quantifiers. For instance, consider the three quantifiers "some", "between 3 and 5" and "an even number of". While "some" is monotone and the other two quantifiers are not, intuitively "an even number of" is the least monotone of the three. To track finer changes in monotonicity level over time, we define a graded measure of monotonicity.

We measure monotonicity in information-theoretic terms as the proportion of uncertainty in the output of a quantifier that is removed after knowing that there is

a submodel where the quantifier is true (i.e. a 1). For a perfectly (upward) monotone quantifier Q, if a model M has a submodel to which the quantifier assigns 1 then Q will assign 1 to M. Therefore, for a monotone quantifier all the uncertainty is removed and the measure has value 1.

More formally, first define the random variables $\mathbb{1}_Q$ and $\mathbb{1}_Q^{\subset}$ on the space of possible models as follows. $\mathbb{1}_Q$ is the value that Q assigns to the model M. $\mathbb{1}_Q^{\subset}$ is whether a model has a submodel that the quantifier considers true (assigns 1 to). The entropy of $\mathbb{1}_Q$, $H(\mathbb{1}_Q)$, quantifies the uncertainty about what truth value Q will assign to a model. The conditional entropy $H(\mathbb{1}_Q \mid \mathbb{1}_Q^{\subset})$ quantifies the uncertainty about what Q will assign to a model, given that one knows whether the model has a submodel that Q considers true (assigns 1 to). $H(\mathbb{1}_Q \mid \mathbb{1}_Q^{\subset})$ is minimized (attains value 0) for a perfectly monotone quantifier: if you know that a model has a true submodel, and the quantifier is upward monotone, you know the truth value of that model. The difference between the entropy and the conditional entropy between these variables is known as the mutual information:

$$I(\mathbb{1}_Q; \mathbb{1}_Q^{\subset}) := H(\mathbb{1}_Q) - H(\mathbb{1}_Q|\mathbb{1}_Q^{\subset})$$

This measures how much information $\mathbb{1}_Q^{\subset}$ provides about $\mathbb{1}_Q$. For a perfectly monotone quantifier, $H(\mathbb{1}_Q|\mathbb{1}_Q^{\subset}) = 0$, and so $I(\mathbb{1}_Q; \mathbb{1}_Q^{\subset}) = H(\mathbb{1}_Q)$. In other words: for a monotone quantifier, knowing which models have a true sub-model provides as much information as knowing the entire quantifier.

While this roughly captures what we want from a measure of monotonicity, it needs to be normalized to form a degree that applies across quantifiers, since $0 \leq I(\mathbb{1}_Q; \mathbb{1}_Q^{\subset}) \leq H(\mathbb{1}_Q)$. We do this by dividing by $H(\mathbb{1}_Q)$, moving the upper bound to 1. In total then, we measure monotonicity as

$$\begin{aligned}
\mathsf{mon}(Q) &:= \frac{I(\mathbb{1}_Q; \mathbb{1}_Q^{\subset})}{H(\mathbb{1}_Q)} \\
&= \frac{H(\mathbb{1}_Q) - H(\mathbb{1}_Q|\mathbb{1}_Q^{\subset})}{H(\mathbb{1}_Q)} \\
&= 1 - \frac{H(\mathbb{1}_Q \mid \mathbb{1}_Q^{\subset})}{H(\mathbb{1}_Q)}
\end{aligned}$$

To see how this measure tracks intuitions, consider the previous mentioned quantifiers "some", "between 3 and 5" and "an even number of". "Some" gets mono-

tonicity 1.0, because knowing whether a model has a submodel that verifies "some" eliminates all uncertainty about the truth of the model. Recall that each agent can be interpreted as instantiating any of four quantifiers, which can be monotone to different degrees. This raises the question of which of the four degrees of monotonicity should be considered in the analysis of the results. The monotonicity of an agent's language is the highest among the degrees of the quantifiers compatible with the agent's language. For instance, an agent whose quantifier is "between 3 and 5" has degree 0.7517 and one with "an even number of" has degree 0.001.

We compare the results of the simulation to the distribution of the measure in randomly generated quantifiers. There are two different random distributions of quantifiers. On the one hand, there are the quantifiers instantiated by randomly initialized agents. On the other hand, there are the quantifiers sampled uniformly from the space of possible quantifiers. These two distributions are depicted in Figure 7.2. While the completely random quantifiers have a narrower distribution, both types of random distribution are very skewed towards low degree of monotonicity. This makes sense: monotonicity is a relatively rare property, and so should not be expected to appear randomly. We now turn to the results, showing that higher degrees do emerge via iterated learning.

7.3.6 Materials

For our experiments, we used a fixed model size of 10 (which, recall, is also the size of the input to the agents), with 10 agents in each generation, and varied the bottleneck size (200, 512, 715, 1024) and number of epochs (4 and 8). For each setting of those two parameters, we ran 20 trials.

The code, data, and instructions for running experiments may be found at `https://github.com/thelogicalgrammar/NeuralNetIteratedQuantifiers`.

7.4 Results

The first result is that monotone quantifiers evolve consistently and rapidly for some values of the simulation parameters. More specifically, the evolution of monotonicity depends on the bottleneck size and the number of epochs, i.e. how much of the parent's language is observed by the cultural child. See Figure 7.3 for the results. If the networks get too much input, they learn the quantifier accurately and change is

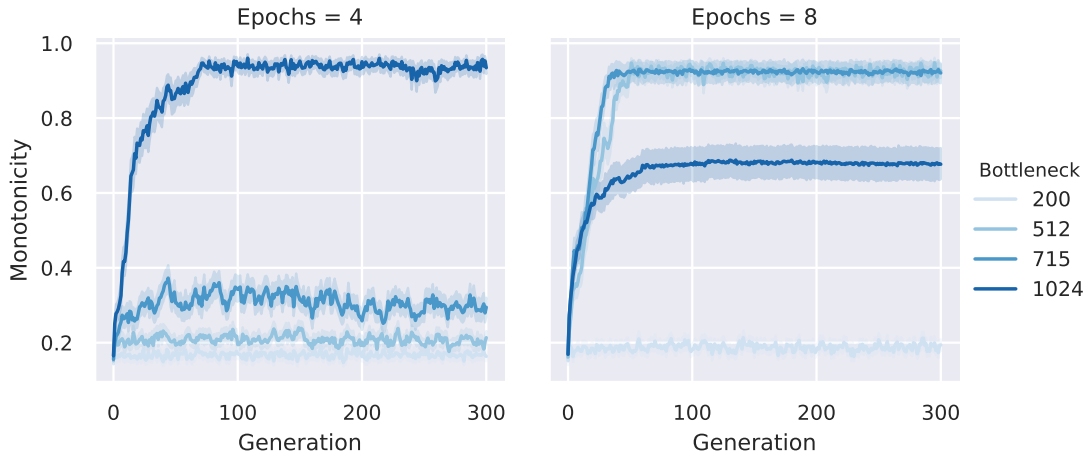


Figure 7.3: The simulation was ran 20 times for each combination of bottleneck size and number of epochs in a population of 10 agents and a maximum model size of 10. The plot shows how the average monotonicity level across all languages changes over 300 generations. Convergence to monotonicity depends on how much the learners' neural networks are trained, which itself depends on the number of epochs and the bottleneck size. With small bottleneck and few epochs, monotonicity does not evolve. With a bigger bottleneck size and more training epochs, monotone languages become widespread. However, increasing the training data further tends to impede the development of monotone languages.

very slow. If the networks get too little input, the learning has little effect and no pattern emerges. If languages are somewhat stable across generations, but enough variation is allowed by not over-training the cultural children, monotonicity evolves.

A second result is that the monotone quantifiers that emerge are in large part non degenerate. In Bayesian models that include a prior for simplicity, degenerate languages become widespread under pure IL (Kirby et al., 2015). Here, however, degenerate quantifiers are a small minority (about 0.005% of all quantifiers).

The third result is that most non-degenerate monotone quantifiers fall in one of a few types. About 79% of the perfectly monotone quantifiers show the following pattern: there is some index i such that the quantifier—call it Q_i—assigns 1 to a model iff the model is 1 at i (or an equivalent pattern obtained by switching 0 and 1 uniformly in the models and/or in the quantifier). Q_i is true iff o_i, the object

represented by index i, belongs to the set B.[13]. Therefore $Q_i(A)$ functions like a proper noun for o_i. Just like "Anna is human" is true iff Anna belongs to the set of humans, "$Q_i(A)$ is B" is true iff o_i belongs to the set B.

For other monotone quantifiers $Q_{\{j,k\}}$, there are two indices j, k (with $j \neq k$) such that $Q_{\{j,k\}}$ assigns 1 to a model iff the model has value 1 at both j and k (or, again, an equivalent patterns obtained by switching 0 and 1 in the models and/or in the quantifier). $Q_{\{j,k\}}$ is true iff B contains two specific elements of A, and false otherwise.[14] It can be interpreted as the conjunction of two proper nouns. Like "Anna and Rob are human" is true iff Anna is human and Rob is human, "$Q_{\{j,k\}}(A)$ is B" is true iff o_j is B and o_k is B.

7.5 Discussion

The results we presented support the learnability account of the origins of semantic universals of quantification. While previous work compared quantifiers satisfying semantic universals to quantifiers that do not, we have presented a model where the former are selected out of all of the possible quantifiers by a process of cultural evolution. Moreover, the preference for monotone quantifiers is not a consequence of an explicitly coded bias for simplicity, but rather of an independently motivated, biologically plausible model of learning. The results therefore suggest that not only are monotone quantifiers easier to learn, but they are also widespread in language *because* of their learnability.

This model can be straightforwardly extended in various ways. The agents judged their quantifier compatible with a given model simply by rounding the output of their neural network. An alternative to this is for the agents to accept a model with a probability proportional to the network's output. Such so-called sample agents do not straightforwardly instantiate a quantifier, since they can produce inconsistent output when repeatedly prompted with the same model. However, preliminary results have shown that neural networks are capable of doing *statistical learning*: given enough data, they approximate not just whether their parents tend to reject or accept a

[13]In set-theoretic terms, Q_i is a *principal ultrafilter* If U is a finite non-empty set, a set F is a principal ultrafilter on U if there is an $a \in U$ such that $F = \{B \in \mathcal{P}(U) | a \in B\}$. In the present model, Q_i is (the characteristic function of) a principal ultrafilter on B because it it contains every subset of B that contains i.

[14]These are called in set-theoretic terms *principal filters*. They are not principal ultrafilters because their truth depends on more than one element.

model, but also the probability of acceptance.

While the quantifiers that emerge from our experiment are monotone, they are unnatural in certain respects. For instance, the proper-name-like quantifiers that emerge are not *quantitative*, i.e. their truth value depends not simply on the number of 1s and 0s, but on the identity of particular elements.[15]

To try and explain the emergence of quantifiers which are both monotone and quantitative, it might be necessary to make it more difficult for the networks to rely on the identity of particular objects by, for instance, shuffling the order of models in the parent and the teacher's inputs. Another pressure that might contribute to shape the meaning of quantifiers comes from communication (Kirby et al., 2015). While some semantic universals of quantification might have an advantage in cultural evolution because they conform well with learning biases, other universals might evolve because they lead to more successful communication. Therefore, combining iterated learning with a pressure for accurate communication might help more natural quantifiers emerge. We leave all of these exciting possibilities to future work.

[15]See Steinert-Threlkeld and Szymanik (2018) for the definition of and motivation for quantity, which generalizes the isomorphism/permutation constraint in generalized quantifier theory as discussed, for instance, in Peters and Westerståhl (2008).

Chapter 8

Conclusion

In caminu s'acconzat su garrigu[1]

Sardinian saying

In this thesis, I have tried to combine different sources of evidence to study the evolution of two universals of scalar semantics, monotonicity and extremeness. One problem with combining theoretical, experimental, and modelling work is that the different approaches might push in different directions. This is to some extent what happened: while the theoretical work in chapter 2 and the modelling work in chapter 3 pointed in the same direction, i.e. transitions-based encoding as the basis for monotonicity, the experimental results of chapters 4 and 5 did not provide support for the same picture. The picture was then made more complex by the results in chapter 7, which showed that a general model of learning without a pressure for communicative accuracy does indeed result in monotonic quantifiers evolving, but not necessarily degenerate ones. Finally, chapter 6 was an initial exploration of how cultural transmission might influence the evolution of extreme categories, and it is not yet clear how this interacts with the mechanisms that create a pressure for extremeness discussed in section 1.3.

Despite the complexity of the emerging picture, some general lessons can be drawn from the work presented in this thesis. Chapter 1 identified two universals of the semantics of scalar language, namely monotonicity and extremeness, and reviewed previous discussion of them in the literature. The conclusion of the discussion is that there is no unified account of monotonicity across different grammatical classes

[1] One balances the weight along the way.

(e.g. gradable adjectives and quantifiers), and previous accounts of the evolution of extremeness mainly focus on approximate rather than true extremeness and communication rather than learning. These gaps in the literature provided questions for the rest of the thesis.

Assuming that language evolution is guided in part by pressures coming from cognition, in chapter 2 I developed a simple account of the cognitive roots of scalar categorization based on conceptual spaces theory. I argued that a classical picture of categorization in conceptual spaces, based on prototypes, is unsuitable for modelling scalar categories. I proposed an alternative picture based on transitions which addresses the issues that arose with a prototype-bases picture. The picture of scalar categorization based on transitions offers a natural measure of the complexity of different categories as the number of transitions needed to encode them.

In chapter 3, I developed further the implications of the theory developed in chapter 2. I proposed, based on modelling results, that the evolution of monotonicity follows from a combination of two pressures, namely a pressure for simplicity coming from iterated cultural transmission and a pressure for communicative accuracy coming from usage. Crucially, in order for these two pressures to combine in such a way that monotonic categories evolve, the language users have to be pragmatically skilful.

In chapter 4, I presented results from three category-learning experiments that attempted to test the measure of complexity developed in chapter 2 and used in the models in chapter 3. The results were inconclusive, and I identified a possible problem in the fact that the design could not capture small enough differences in the behaviour of participants across conditions.

In order to make the analysis more sensitive to differences in behavioural patterns, I run three more experiments that produced much richer behavioural data. The experiments are presented in chapter 5. In order to analyse the more complex data, I developed a hierarchical Bayesian model wrapped around a cognitive model of categorization that was designed to capture the participants' preferences for monotonic categories. The results of the latter three experiments were also inconclusive.

In chapter 6, I turned to the other universal of scalar semantics that was identified and discussed in chapter 1, namely extremeness. I show that even without assuming a cognitive bias in favour of extreme categories, extremeness can evolve purely due to cultural transmission. This work provides a basis for future analyses of the role of learning in the evolution of extremeness.

Finally, in chapter 7, written in collaboration with Shane Steinert-Threlkeld and Jakub Szymanik, we show for the more complex semantic space of quantifiers that monotonicity can emerge from IL and the general purpose learning model of neural networks.

Much work remains to be done to understand the origins of the universals of scalarity, and disentangle the contribution of different pressures on their evolution. In this thesis, I have focussed on gradable adjectives, which constitute the simplest example of scalar categories, and on quantifiers. However, other grammatical classes express scalar categories. One promising approach for instance would be to apply the work presented in this thesis to modal semantics. Modal expressions have been argued to require a scalar semantics (Lassiter, 2016; Santorio & Romoli, 2017; Klecha, 2014), or to be implicitly quantificational (Kratzer, 2012). The complexity of modal semantics remains an obstacle to developing accounts of its evolution.

Another interesting way to develop the work presented in this thesis further is to collect evidence to assess the commonness of the two discussed universals of monotonicity and extremeness cross-linguistically. While much evidence exists for the monotonicity of quantifiers cross-linguistically, more evidence is needed for the monotonicity of gradable adjectives. Moreover, more data is needed on the proportion of extreme categories cross-linguistically.

A third direction for future research on the evolution of scalar universals is to develop new experiments. The picture that I argued for in chapter 2 implies that scalar categories expressed by gradable adjectives have a particularly simple mental representation, which lends itself to cognitive Bayesian modelling in the style of chapter 5. As opposed to adjectival scales, quantifiers use scales with a complex internal structure, which is reflected in their complex mental representations. However, it is crucial to distinguish between the simplicity biases that emerge once a scale such as the scale of proportions is given (e.g. a bias for monotonicity) and the biases that are involved in constituting the scale in the first place (e.g. bias for independence of truth conditions from non-argument sets).

Appendix A

Proofs

Proof for lemma 1:

Proof. Consider the case where $x < y$, $f(x) = \text{true}$, and $f(y) = \text{false}$. The proof for the other cases are similar. Call a the greatest element in the domain such that f is true for all q with $x < q < a$. a has to be lower than y, otherwise y would be a point $x < y < a$ where f is false, which contradicts the definition of a. a can be true or false. If a is false, then a counts as a false transition from true to false. If a is true, by definition of a there have to be points greater than a where f is false. Call b a point such that f is false for all q with $a < q < b$. a is then a true transition from true to false. □

Proof for lemma 2:

Proof. Assume by contradiction that they are both true transitions from true to false, and assume t_1 is the smaller of the two. The proofs for the other cases are similar. By definition 2, there is some $y > t_1$ such that all points a with $t_1 < a \leq y$ evaluate to false, and there is some $x \leq t_2$ such that all points b with $x \leq b \leq t_2$ evaluate to true. Since x is false, y is true, and $x < y$, by lemma 1 there has to be a transition between x and y. This contradicts the assumption that t_1 and t_2 are adjacent. □

Proof for lemma 3:

Proof. First show that if f is monotonic, then it has one of fewer transitions. Assume by contradiction that f has exactly two distinct transitions t_1 and t_2 but is monotonic. A proof for a function with more than two transitions can be obtained by simply

considering any two adjacent transitions. Since we assumed that the domain is totally ordered and there are countably many transitions, we can call t_1 the lowest of the two. Assume that t_1 is a transition from true to false, with the proof for the other cases being similar. Since t_2 is the only other transition, t_1 and t_2 are adjacent. By lemma 2, transition t_2 has to be from false to true. Call a any of the points $\leq t_1$ that evaluates to true, b any point between t_1 and t_2 that is false, and c any point $t_2 \geq$ that is true. The existence of a, b, and c is guaranteed by the definition of transition. Note that $a < b$ and $f(a) < f(b)$, implying that f cannot be monotonic decreasing. Moreover, $b < c$ and $f(b) > f(c)$, implying that the function cannot be monotone increasing. Therefore, f cannot be monotonic, contradicting the assumption.

Next show that if f is non-monotonic, then f has at least two transitions. Since f is non-monotonic, it is neither monotone increasing nor monotone decreasing. Therefore, there are some x, y in the domain such that $x \leq y$ and $f(x) > f(y)$, and some z, w such that $z \leq w$ and $f(z) < f(w)$.[1] Since $f(x) > f(y)$, the truth value at x is different from the truth value at y. By lemma 1, there has to be a transition between x and y. Therefore, f cannot have zero transitions. A similar argument shows that there has to a transition between z and w. Therefore, f has to have at least two transitions. □

Proof for lemma 4:

Proof. Index the transitions with integers, so that they can be called $t_1, \ldots, t_n$, such that consecutive integers correspond to adjacent transitions. This is possible because we assumed there are countably many transitions and that their order is known. By lemma 2, for any index j the types of transition t_{j+1} and of transition t_{j-1}, if they exist, can be inferred from the type of transition t_j. By assumption, the type of some transition t_i is known. Therefore, the transition type can be inferred for all indices. □

Proof for lemma 5:

Proof. Assume that f and g have both a transition from false to true. The proof for the other type is similar. Moreover, assume that the transition of f, t_f, is lower than or equal to the transition of g, t_g. The case where $t_g \leq t_f$ is similar. We want to show that f includes g. Assume by contradiction that there is a point x where f

[1] From a logical point of view, since the definition of monotonicity says that for all x, y, $P(x, y) \rightarrow Q(x, q)$, its negation says that for some x, y, $P(x, y) \wedge \neg Q(x, y)$.

is false but g true, which is the negation of the assertion that f includes g. f has to be true everywhere above t_f, otherwise by lemma 1 f would have two have two transitions, contradicting the assumption. Therefore, $x \leq t_f$. By the definition of transition from false to true, there are points y such that $y \leq t_g$ and $g(y) = \text{false}$. In sum, $x < y \leq t_g$. Since $g(x) \neq g(y)$, g has a transition between x and y by lemma 1. But this contradicts the assumption that g has a single transition. □

Appendix B

Category selection algorithms

Algorithm 2 Category selection in experiment 2 in pseudo-code

function SELECTCATEGORIES
  $maxIndex \leftarrow 35$ ▷ There are 36 stims but indexing starts at 0
  $distBorders \leftarrow \text{UNIF}(6, 12)$ ▷ Uniform excludes max
  $pickedId \leftarrow \text{UNIF}(distBorders, maxIndex + 1 - DistBorders)$
  if $pickedId < maxIndex - pickedId$ **then**
    $pickedExtreme \leftarrow 0$ ▷ Picked stim is closer to min than max of scale
  else if $pickedId > maxIndex - pickedId$ **then**
    $pickedExtreme \leftarrow maxIndex$
  else
    $pickedExtreme \leftarrow \text{RANDCHOICE}([0, maxIndex])$
  end if
  $distCloseBorder \leftarrow \text{UNIF}(0, 3)$ ▷ Distance of internal border from $pickedId$
  if $pickedExtreme = 0$ **then**
    $monList \leftarrow \text{RANGE}(0, pickedId + 1 + distCloseBorder)$
  else
    $monList \leftarrow \text{RANGE}(pickedId - distCloseBorder, maxIndex + 1)$
  end if ▷ At this point a monotone list is defined
  $center \leftarrow pickedId + \text{UNIF}(-1, 2)$
  $size \leftarrow \text{LEN}(monList)$
  $startNonMon \leftarrow center - (size/2)$ ▷ Position of internal border
  if $size$ is even **then**
    if RANDBOOL() **then**
      $startNonMon \leftarrow \text{ROUND}(startNonMon - 1)$
    else
      $startNonMon \leftarrow \text{ROUND}(startNonMon)$
    end if
  end if
  $nonMonList \leftarrow \text{RANGELENGTH}(startNonMon, size + 1)$
  return $[monList, nonMonList, pickedId]$
end function

Algorithm 3 Category selection in experiment 3 in pseudo-code

function SELECTCATEGORIES
    $cat \leftarrow \text{RANDCHOICE}([B, C])$
    if $cat = B$ **then**
        $prototypeId \leftarrow 13$
        $pickedId \leftarrow \text{UNIF}(10, 14)$
        $monInternalBorder \leftarrow selectedId + 2$
        $monList \leftarrow \text{RANGE}(0, monInternalBorder + 1)$
        $nonMonList \leftarrow \text{RANGE}(9, 18)$
    else
        $prototypeId \leftarrow 22$
        $pickedId \leftarrow \text{UNIF}(22, 26)$
        $monInternalBorder \leftarrow selectedId - 2$
        $monList \leftarrow \text{RANGE}(monInternalBorder, 36)$
        $nonMonList \leftarrow \text{RANGE}(18, 27)$
    end if
    return $[monList, nonMonList, pickedId]$
end function

Appendix C

Efficient calculation of the utility

Assuming that the prior over degrees p is uniform allows a simplification of the utility function that significantly improves performance. First, consider the utility of silence (u_0) for the literal listener L_0 given that the rational speaker S_1 has observed degree s_0, with thresholds $\vec{\Theta}$ and pragmatic slack parameters $\vec{\lambda}$:

$$U_{L_0}(u_0|s_o, \Theta_1, \Theta_2, \lambda) = \int_{s=0}^{s=1} e^{-\lambda(s_o-s)^2} ds \tag{C.1}$$

$$= \int_{s=0}^{s=1} e^{-\left(\sqrt{\lambda}(s_o-s)\right)^2} ds \tag{C.2}$$

Variable substitution $x = \sqrt{\lambda}(s_o - s)$, $ds = -\frac{1}{\sqrt{\lambda}}dx$:

$$= -\frac{1}{\sqrt{\lambda}} \int_{x=\sqrt{\lambda}s_o}^{x=\sqrt{\lambda}(s_o-1)} e^{-x^2} dx \tag{C.3}$$

Manipulate into the error function:

$$= -\frac{\sqrt{\pi}}{2\sqrt{\lambda}} \frac{2}{\sqrt{\pi}} \int_{x=\sqrt{\lambda}s_o}^{x=\sqrt{\lambda}(s_o-1)} e^{-x^2} dx \tag{C.4}$$

Use the fact that $\int_a^c f(x)dx = \int_a^b f(x)dx + \int_b^c f(x)dx$:

$$= -\frac{\sqrt{\pi}}{2\sqrt{\lambda}} \left(\frac{2}{\sqrt{\pi}} \int_{x=\sqrt{\lambda}s_o}^{x=0} e^{-x^2} dx + \frac{2}{\sqrt{\pi}} \int_{x=0}^{x=\sqrt{\lambda}(s_o-1)} e^{-x^2} dx \right) \tag{C.5}$$

Use the fact that $\int_a^b f(x)dx = -\int_b^a f(x)dx$

$$= \frac{\sqrt{\pi}}{2\sqrt{\lambda}} \left(\frac{2}{\sqrt{\pi}} \int_{x=0}^{x=\sqrt{\lambda}s_o} e^{-x^2} dx - \frac{2}{\sqrt{\pi}} \int_{x=0}^{x=\sqrt{\lambda}(s_o-1)} e^{-x^2} dx \right) \tag{C.6}$$

Finally, note that $\frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt = \text{erf}(x)$:

$$= \frac{\sqrt{\pi}}{2\sqrt{\lambda}} \left(\text{erf}\left(\sqrt{\lambda}(s_o - 1)\right) - \text{erf}\left(\sqrt{\lambda}s_o\right)\right) \tag{C.7}$$

This expression can be calculated much more efficiently than the original expression, because there are algorithms to approximate the error function erf. Similar reasoning gives the utility of the positive polarity signal s_1 with the same setup:

$$U_{L_0}(u_1|s_o, \Theta_1, \Theta_2, \lambda) = \frac{\sqrt{\pi}}{2\sqrt{\lambda}(1-\Theta_1)} \left(\text{erf}\left(\sqrt{\lambda}(1 - s_o)\right) - \text{erf}\left(\sqrt{\lambda}(\Theta_1 - s_o)\right)\right)$$

Finally, the utility of the negative polarity item s_2 with the same setup:

$$U_{L_0}(u_2|s_o, \Theta_1, \Theta_2, \lambda) = \frac{\sqrt{\pi}}{2\sqrt{\lambda}\Theta_2} \left(\text{erf}\left(\sqrt{\lambda}s_o)\right) - \text{erf}\left(\sqrt{\lambda}(s_o - \Theta_2)\right)\right)$$

Appendix D

Experiments Flowcharts

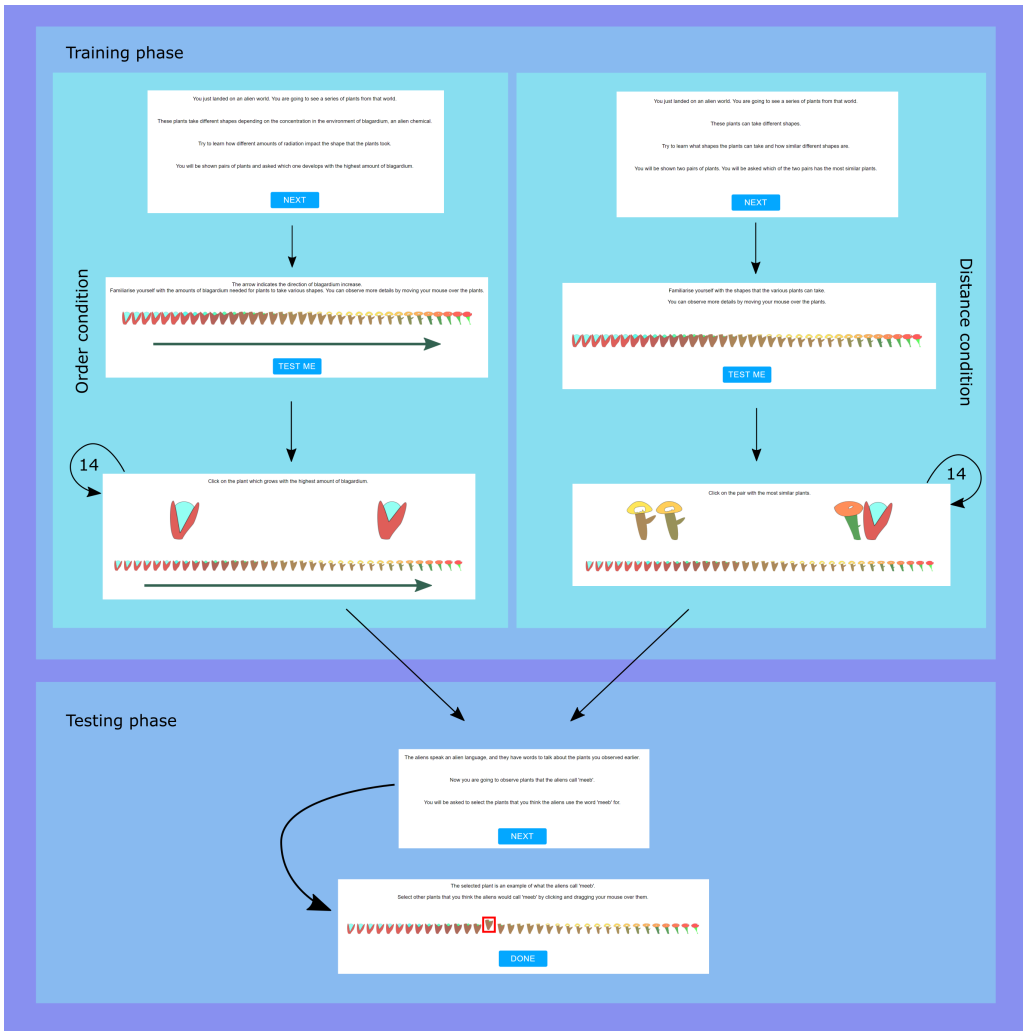


Figure D.1: Flowchart for the first experiment.

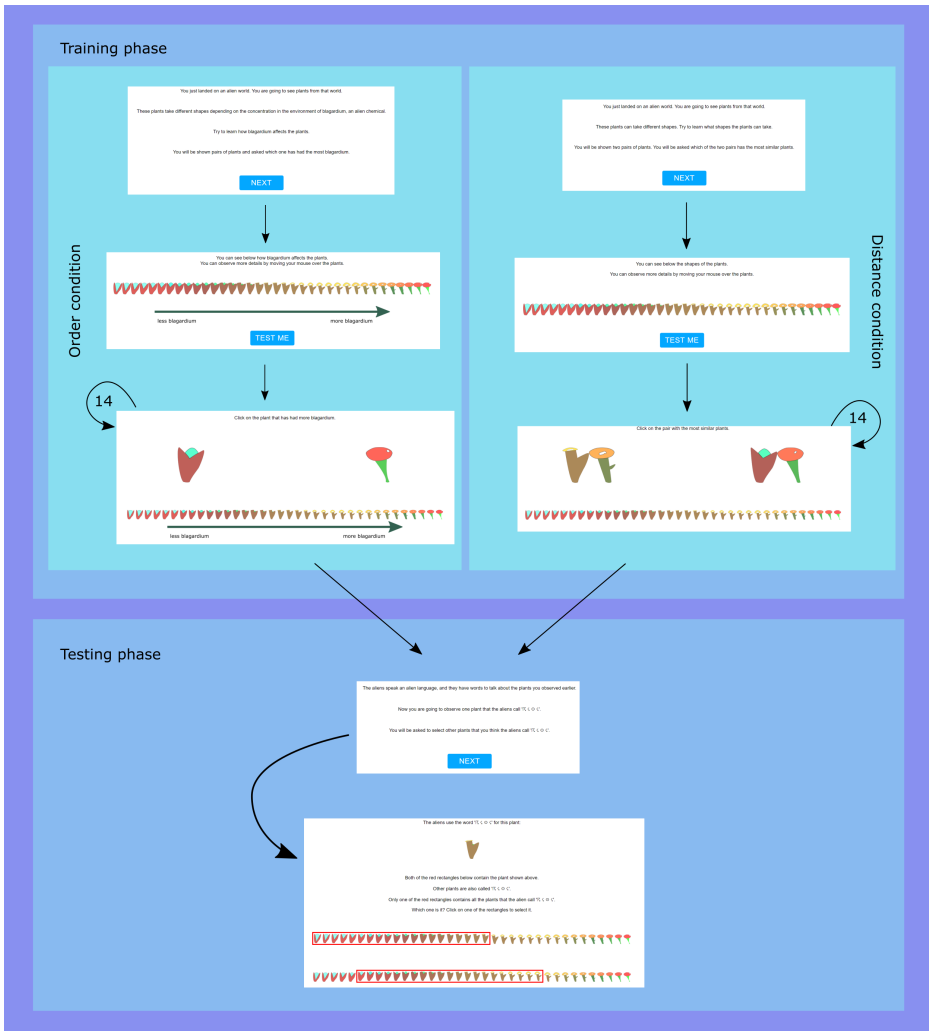


Figure D.2: Flowchart for the second experiment.

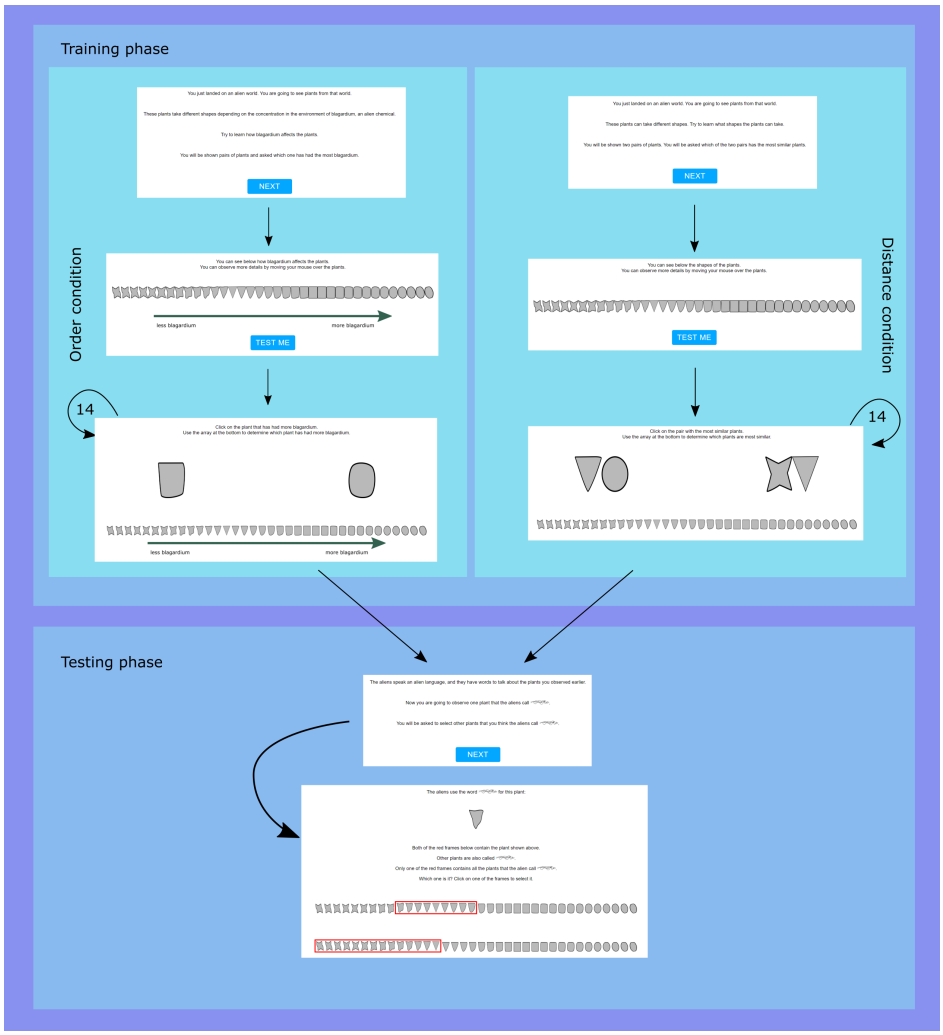


Figure D.3: Flowchart for the third experiment.

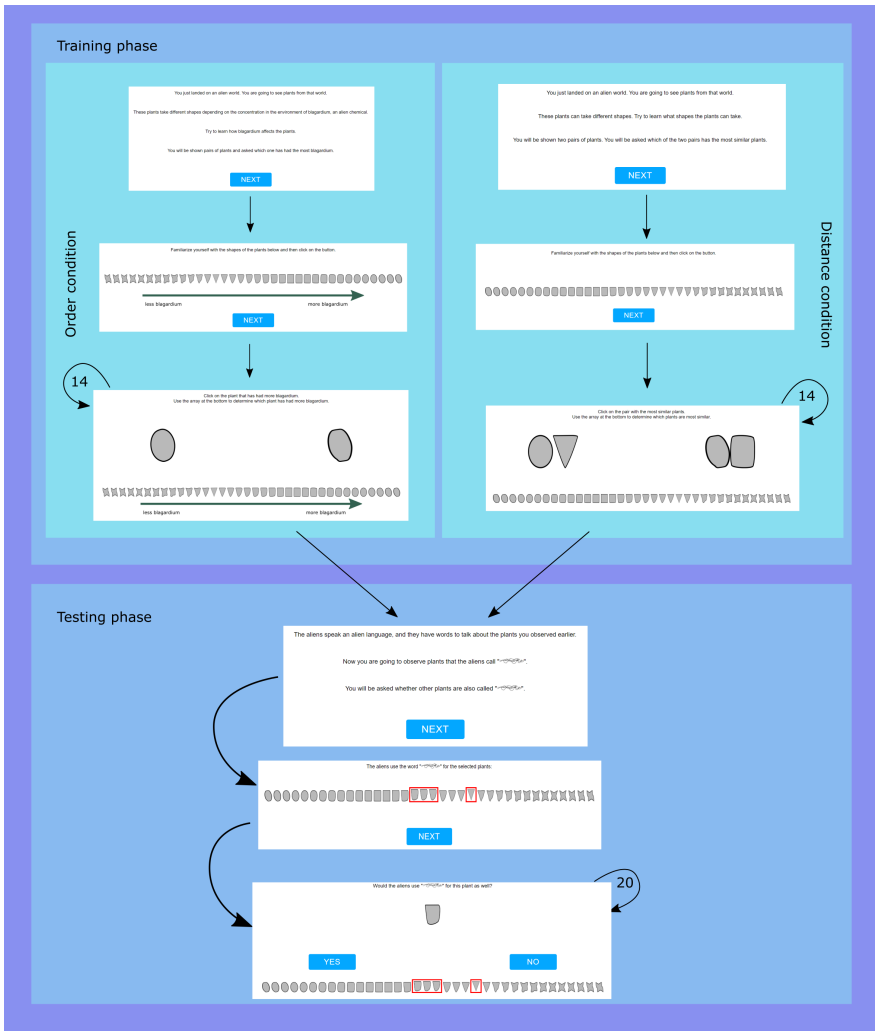


Figure D.4: Flowchart for the fourth experiment.

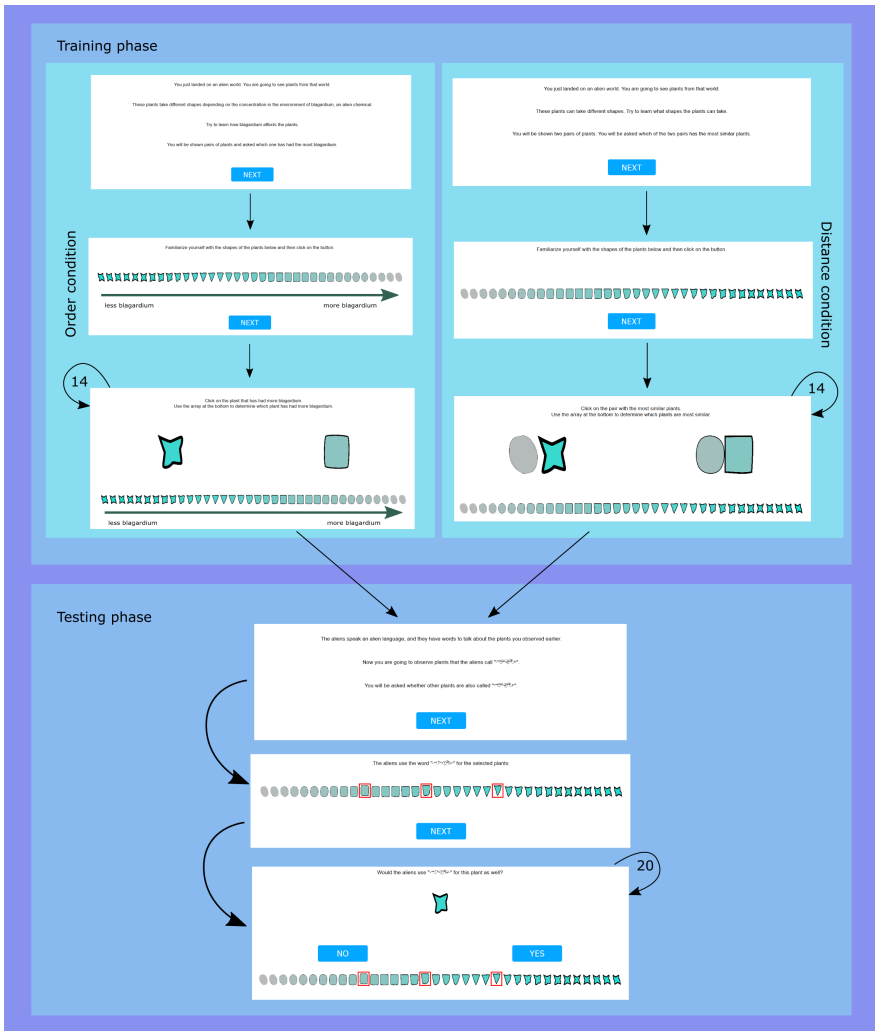


Figure D.5: Flowchart for the fifth experiment.

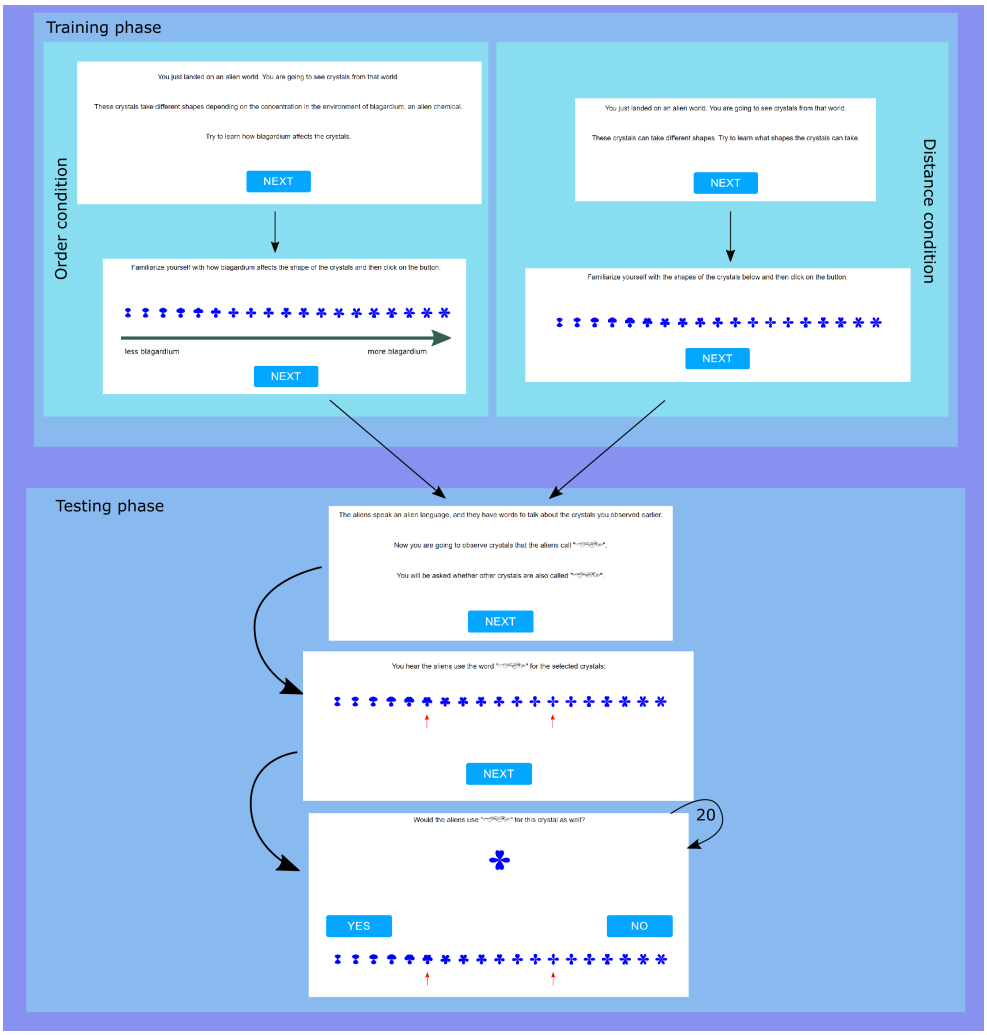


Figure D.6: Flowchart for the sixth experiment.

References

Aparicio, H., Xiang, M., & Kennedy, C. (2016, January). Processing gradable adjectives in context: A Visual World study. *Semantics and Linguistic Theory*, *25*, 413. doi: 10.3765/salt.v25i0.3128

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53. doi: 10.1037/0278-7393.14.1.33

Ashby, F. G., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 598–612. doi: 10.1037/0096-1523.16.3.598

Ashby, F. G., & Maddox, W. T. (1993, September). Relations between Prototype, Exemplar, and Decision Bound Models of Categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400. doi: 10.1006/jmps.1993.1023

Barker, C. (2002). The Dynamics of Vagueness. *Linguistics and Philosophy*, *25*, 36.

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

Betancourt, M. (2018, July). A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat]*.

Bogal-Allbritten, E. (2013, September). Decomposing notions of adjectival transitivity in Navajo. *Natural Language Semantics*, *21*(3), 277–314. doi: 10.1007/s11050-012-9093-2

Brochhagen, T., Franke, M., & van Rooij, R. (2016). Learning biases may prevent lexicalization of pragmatic inferences: A case study combining iterated (Bayesian) learning and functional selection. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 6.

Brochhagen, T., Franke, M., & van Rooij, R. (2018, November). Coevolution of

Lexical Meaning and Pragmatic Use. *Cognitive Science*, *42*(8), 2757–2789. doi: 10.1111/cogs.12681

Burnett, H. (2014, February). A Delineation solution to the puzzles of absolute adjectives. *Linguistics and Philosophy*, *37*(1), 1–39. doi: 10.1007/s10988-014-9145-9

Burnett, H. (2016). *Gradability in Natural Language: Logical and Grammatical Foundations*. Oxford, New York: Oxford University Press.

Carcassi, F. (2020, January). *Thelogicalgrammar/evoModelsAdjectivesReimplementation.*

Carcassi, F., Steinert-Threlkeld, S., & Szymanik, J. (2019). The emergence of monotone quantifiers via iterated learning. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 190–196). doi: 10.31234/osf.io/8swtd

Carr, J. W., Smith, K., Culbertson, J., & Simon, K. (2018, July). *Simplicity and informativeness in semantic category systems* (Preprint). PsyArXiv. doi: 10.31234/osf.io/jkfyx

Carstairs-McCarthy, A. (2010). *The evolution of morphology*. Oxford: Oxford University Press. (OCLC: ocn437305846)

Chater, N., & Christiansen, M. H. (2007, August). Two views of simplicity in linguistic theory: Which connects better with cognitive science? *Trends in Cognitive Sciences*, *11*(8), 324–326. doi: 10.1016/j.tics.2007.06.006

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in cognitive sciences*, *7*(1), 19–22.

Chemla, E., Buccola, B., & Dautriche, I. (2019, June). Connecting Content and Logical Words. *Journal of Semantics*, *36*(3), 531–547. doi: 10.1093/jos/ffz001

Christiansen, M. H., Collins, C., & Edelman, S. (Eds.). (2009). *Language Universals*. Oxford University Press. doi: 10.1093/acprof:oso/9780195305432.001.0001

Clark, E. V. (1993). *The Lexicon in Acquisition* (No. 65). Cambridge: Cambridge University Press.

Culbertson, J., & Kirby, S. (2016, January). Simplicity and Specificity in Language: Domain-General Biases Have Domain-Specific Effects. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.01964

Culbertson, J., Kirby, S., & Schouwstra, M. (2016). Word Order Universals Reflect Cognitive Biases: Evidence From Silent Gesture. In *The Evolution of Language: Proceedings of the 11th International Conference* (pp. 391–393).

Decock, L., Dietz, R., & Douven, I. (2013). Modelling Comparative Concepts in Conceptual Spaces. In Y. Motomura, A. Butler, & D. Bekki (Eds.), *New Frontiers in Artificial Intelligence* (pp. 69–86). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-39931-2_6

Decock, L., & Douven, I. (2014). What Is Graded Membership? *Noûs*, *48*(4), 653–682. doi: 10.1111/nous.12003

Development Team, S. (2017). *Stan Modeling Language Users Guide and Reference Manual, Version 2.17.0.*

Does a univariate random variable's mean always equal the integral of its quantile function? (n.d.). https://stats.stackexchange.com/questions/18438/does-a-univariate-random-variables-mean-always-equal-the-integral-of-its-quanti.

Doetjes, J., Constantinescu, C., & Součková, K. (2009, September). A Neo-Kleinian Approach to Comparatives. *Semantics and Linguistic Theory*, *19*(0), 124–141. doi: 10.3765/salt.v19i0.2544

Doran, R., Baker, R., McNabb, Y., Larson, M., & Ward, G. (2009). On the Non-Unified Nature of Scalar Implicature: An Empirical Investigation. *International Review of Pragmatics*, *1*(2), 211–248. doi: 10.1163/187730909X12538045489854

Douven, I. (2016, June). Vagueness, graded membership, and conceptual spaces. *Cognition*, *151*, 80–95. doi: 10.1016/j.cognition.2016.03.007

Feldman, J. (2016, September). The simplicity principle in perception and cognition. *Wiley Interdisciplinary Reviews. Cognitive Science*, *7*(5), 330–340. doi: 10.1002/wcs.1406

Fine, K. (1975, September). Vagueness, truth and logic. *Synthese*, *30*(3), 265–300. doi: 10.1007/BF00485047

Franke, M. (2012a). On Scales, Salience and Referential Language Use. In M. Aloni, V. Kimmelman, F. Roelofsen, G. W. Sassoon, K. Schulz, & M. Westera (Eds.), *Logic, Language and Meaning* (pp. 311–320). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-31482-7_32

Franke, M. (2012b). Scales, Salience, and Referential Safety: The benefit of communicating the extreme. In *The Evolution of Language: Proceedings of the 9th International Conference.* Kyoto. doi: 10.1142/9789814401500_0016

Franke, M. (2014). Typical use of quantifiers: A probabilistic speaker model. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36st Annual Conference of the Cognitive Science Society* (Vol. 36, pp. 487–492).

Austin, TX.

Gaio, S. (2009). A Granular Account for Gradable Adjectives. *Logique et Analyse*, *52*(208), 407–422.

Gärdenfors, P. (2004). *Conceptual Spaces: The Geometry of Thought* (New Ed edition ed.). Cambridge, Mass.: MIT Press.

Gärdenfors, P. (2011). Semantics Based on Conceptual Spaces. In M. Banerjee & A. Seth (Eds.), *Logic and Its Applications* (pp. 1–11). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-18026-2_1

Gärdenfors, P. (2014). A Semantic Theory of Word Classes. *Croatian Journal of Philosophy*, *14*(41), 16.

Gärdenfors, P. (2017). *The Geometry of Meaning: Semantics Based on Conceptual Spaces* (Reprint edition ed.). MIT Press.

Gärdenfors, P. (2019, June). Convexity Is an Empirical Law in the Theory of Conceptual Spaces: Reply to Hernández-Conde. In M. Kaipainen, F. Zenker, A. Hautamäki, & P. Gardenfors (Eds.), *Conceptual Spaces: Elaborations and Applications* (pp. 77–80). Springer International Publishing. doi: 10.1007/978-3-030-12800-5_5

Gauker, C. (2007, September). A Critique of the Similarity Space Theory of Concepts. *Mind & Language*, *22*(4), 317–345. doi: 10.1111/j.1468-0017.2007.00311.x

Geurts, B., & Van Der Slik, F. (2005, February). Monotonicity and Processing Load. *Journal of Semantics*, *22*(1), 97–117. doi: 10.1093/jos/ffh018

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Goodman, N. D., & Frank, M. C. (2016, November). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. doi: 10.1016/j.tics.2016.08.005

Grano, T., & Davis, S. (2018, February). Universal markedness in gradable adjectives revisited: The morpho-semantics of the positive form in Arabic. *Natural Language & Linguistic Theory*, *36*(1), 131–147. doi: 10.1007/s11049-017-9365-0

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge: Harvard University Press.

Griffiths, T. L., & Kalish, M. L. (2007, May). Language Evolution by Iterated Learning With Bayesian Agents. *Cognitive Science*, *31*(3), 441–480. doi: 10.1080/15326900701326576

Hampton, J. A. (2007). Typicality, Graded Membership, and Vagueness. *Cognitive*

Science, *31*(3), 355–384. doi: 10.1080/15326900701326402

Heim, I. (2006). Little. In *Proceedings of SALT 16*. Ithaca.

Heim, I. (2008). Decomposing Antonyms? Oslo: University of Oslo, Department of literature, Area studies and European languages. (OCLC: 929799760)

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar* (No. 13). Malden, MA: Blackwell.

Hernández-Conde, J. V. (2017, October). A case against convexity in conceptual spaces. *Synthese*, *194*(10), 4011–4037. doi: 10.1007/s11229-016-1123-z

Horn, L. (1972). *On the Semantic Properties of Logical Operators in English* (Unpublished doctoral dissertation). University of California, Los Angeles.

Hyman, L. M. (2008). Universals in phonology. *The Linguistic Review*, *25*(1-2), 83–137. doi: 10.1515/TLIR.2008.003

Ibbotson, P., & Tomasello, M. (2009, March). Prototype constructions in early language acquisition. *Language and Cognition*, *1*(1), 59–85. doi: 10.1515/LANGCOG.2009.004

Kalish, M. L., & Kruschke, J. K. (1997, November). Decision boundaries in one-dimensional categorization. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *23*(6), 1362–1377. doi: 10.1037//0278-7393.23.6.1362

Kamp, H. (1995, November). Prototype theory and compositionality. *Cognition*, *57*(2), 129–191. doi: 10.1016/0010-0277(94)00659-9

Kamp, H. (2013, January). Two Theories about Adjectives. *Meaning and the Dynamics of Interpretation*, 225–261. doi: 10.1163/9789004252882_011

Keenan, E. L., & Westerståhl, D. (2011). Generalized Quantifiers in Linguistics and Logic. In *Handbook of Logic and Language* (pp. 859–910). Elsevier. doi: 10.1016/B978-0-444-53726-3.00019-0

Kennedy, C. (2001). On the monotonicity of polar adjectives. In J. Hoeksema, H. Rullmann, V. Sánchez-Valencia, & T. van derWouden (Eds.), *Linguistik Aktuell/Linguistics Today* (Vol. 40, pp. 201–221). Amsterdam: John Benjamins Publishing Company. doi: 10.1075/la.40.09ken

Kennedy, C. (2007, March). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*(1), 1–45. doi: 10.1007/s10988-006-9008-0

Kennedy, C. (2013). *Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison* (1st ed.). Routledge. doi: 10.4324/9780203055458

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

Kingma, D. P., & Ba, J. (2017, January). Adam: A Method for Stochastic Optimization. In *International Confer-ence of Learning Representations (ICLR)*.

Kirby, S. (1999). *Selection and Innateness: The Emergence of Language Universals* (Unpublished doctoral dissertation). University of Edinburgh, Edinburgh.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007, March). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241–5245. doi: 10.1073/pnas.0608222104

Kirby, S., Griffiths, T., & Smith, K. (2014, October). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114. doi: 10.1016/j.conb.2014.07.014

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015, August). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. doi: 10.1016/j.cognition.2015.03.016

Klecha, P. (2014). *Bridging the divide: Scalarity and modality*. University of Chicago, Division of the Humanities, Department of Linguistics.

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and philosophy*, *4*(1), 1–45.

Koptjevskaja-Tamm, M. (Ed.). (2015). *The Linguistics of Temperature*. John Benjamins Publishing Company.

Kranich, S. (2016). *Contrastive Pragmatics and Translation: Evaluation, epistemic modality and communicative styles in English and German* (Vol. 261). John Benjamins Publishing Company.

Kratzer, A. (2012). *Modals and conditionals*. Oxford, Oxford ; New York: Oxford University Press.

Kruschke, J. K. (2018, June). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. doi: 10.1177/2515245918771304

Langacker, R. W. (2003). One Any. *The Association For Korean Linguistics*, *1*(18), 42.

Lasersohn, P. (1999). Pragmatic Halos. *Language*, *75*(3), 522–551. doi: 10.2307/417059

Lassiter, D. (2015). Adjectival modification and gradation. *Handbook of contemporary semantic theory*, 143–167.

Lassiter, D. (2016). *Graded modality: Qualitative and quantitative perspectives.* Oxford: Oxford University Press (to appear).

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and linguistic theory* (Vol. 23, pp. 587–610). Santa Cruz. doi: http://dx.doi.org/10.3765/salt.v23i0.2658

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, *194*(10), 3801–3836. doi: https://doi.org/10.1007/s11229-015-0786-1

Leffel, T., Xiang, M., & Kennedy, C. (2017). *Interpreting Gradable Adjectives in Context: Domain Distribution vs. Scalar Representation.*

Lewis, D. (1970). General Semantics. *Synthese*, *22*(1/2), 18–67.

Lewis, D. (1975, December). Adverbs of Quantification. In *Formal Semantics of Natural Language* (1st ed., pp. 3–15). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511897696

Li, M., & Vitányi, P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications* (3rd ed.). New York: Springer-Verlag. doi: 10.1007/978-0-387-49820-1

Liu, C.-S. L. (2010, April). The positive morpheme in Chinese and the adjectival structure. *Lingua*, *120*(4), 1010–1056. doi: 10.1016/j.lingua.2009.06.001

Logit-normal distribution. (2019, December). *Wikipedia*. (Page Version ID: 929485835)

Maddox, W. T., & Ashby, F. G. (1993, January). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*(1), 49–70. doi: 10.3758/BF03211715

Magri, G. (2015). Universals on natural language determiners from a PAC-learnability perspective. In *CogSci.*

Markman, E. (1989). *Categorization and Naming in Children.* Cambridge, MA: MIT Press.

Marr, D., Ullman, S., & Poggio, T. (2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Cam-

bridge, Mass: MIT Press.

Morzycki, M. (2009). Degree modification of extreme adjectives. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society* (Vol. 45, pp. 471–485). Chicago Linguistic Society.

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012, March). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(2), 187–223. doi: 10.1111/j.1551-6709.2011.01212.x

Newmeyer, F. J. (2008, January). Universals in syntax. *The Linguistic Review*, *25*(1-2). doi: 10.1515/TLIR.2008.002

Nielsen, M. A. (2015). *Neural Networks and Deep Learning*.

Parsons, T. (2014). *Articulating medieval logic*. Oxford: Oxford University Press. (OCLC: ocn852235311)

Peters, S., & Westerståhl, D. (2008). *Quantifiers in Language and Logic*. Oxford University Press.

Peters, S., Westerstahl, D., & Westerståhl, D. (2006). *Quantifiers in Language and Logic*. Clarendon Press.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). *Modeling the acquisition of quantifier semantics : A case study in function word learnability.*

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016, July). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392–424. doi: 10.1037/a0039980

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

Qing, C., & Franke, M. (2014a). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In *Semantics and linguistic theory* (Vol. 24, pp. 23–41). New York. doi: http://dx.doi.org/10.3765/salt.v24i0.2412

Qing, C., & Franke, M. (2014b). Meaning and Use of Gradable Adjectives: Formal Modeling Meets Empirical Data. In *CogSci* (Vol. 36). Quebec City.

Rett, J. (2008). *Degree Modification in Natural Language* (Unpublished doctoral dissertation). Rutgers University.

Rett, J. (2014, May). The polysemy of measurement. *Lingua*, *143*, 242–266. doi: 10.1016/j.lingua.2014.02.001

Rett, J. (2015, December). Antonymy In Space And Other Strictly Ordered Domains. *Baltic International Yearbook of Cognition, Logic and Communication*,

10(1). doi: 10.4148/1944-3676.1095

Rett, J. (2018, January). The semantics of many, much, few, and little. *Language and Linguistics Compass*, *12*(1). doi: 10.1111/lnc3.12269

Rosch, E., & Lloyd, B. B. (Eds.). (1978). *Cognition and categorization.* Oxford, England: Lawrence Erlbaum.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976, July). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. doi: 10.1016/0010-0285(76)90013-X

Rosch, E. H. (1973, May). Natural categories. *Cognitive Psychology*, *4*(3), 328–350. doi: 10.1016/0010-0285(73)90017-0

Santorio, P., & Romoli, J. (2017, May). Probability and implicatures: A unified account of distributive and free choice inferences under epistemic modals.

Sapir, E. (1944). Grading, A Study in Semantics. *Philosophy of Science*, *11*(2), 93–116.

Sassoon, G. W. (2013, August). A Typology of Multidimensional Adjectives. *Journal of Semantics*, *30*(3), 335–380. doi: 10.1093/jos/ffs012

Sawada, O., & Grano, T. (2011, June). Scale structure, coercion, and the interpretation of measure phrases in Japanese. *Natural Language Semantics*, *19*(2), 191–226. doi: 10.1007/s11050-011-9070-1

Shramko, Y., & Wansing, H. (2018). Truth Values. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018 ed.). Metaphysics Research Lab, Stanford University.

Spike, M., Stadler, K., Kirby, S., & Smith, K. (2017). Minimal Requirements for the Emergence of Learned Signaling. *Cognitive Science*, *41*(3), 623–658. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12351) doi: 10.1111/cogs.12351

Stechow, A. V. (1984, January). Comparing Semantic Theories of Comparison. *Journal of Semantics*, *3*(1-2), 1–77. doi: 10.1093/jos/3.1-2.1

Steinert-Threlkeld, S. (2019). An Explanation of the Veridical Uniformity Universal. , 16.

Steinert-Threlkeld, S., & Szymanik, J. (2018). *Learnability and Semantic Universals.*

Steinert-Threlkeld, S., & Szymanik, J. (2019, November). Learnability and semantic universals. *Semantics and Pragmatics*, *12*(4), 1. doi: 10.3765/sp.12.4

Steinert-Threlkeld, S., & Szymanik, J. (2020). Ease of Learning Explains Semantic Universals. *Cognition*, *195*, 21.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*(2684), 677–680. doi: 10.1126/science.103.2684.677

Svenonius, P., & Kennedy, C. (2006, August). Northern Norwegian degree questions and the syntax of measurement. In H. van Riemsdijk, H. van der Hulst, J. Koster, & M. Frascarelli (Eds.), *Phases of Interpretation* (Vol. 91, pp. 133–162). Berlin, New York: Mouton de Gruyter. doi: 10.1515/9783110197723.3.133

Szymanik, J. (2016). *Quantifiers and Cognition: Logical and Computational Perspectives* (Vol. 96). Cham: Springer International Publishing.

Szymanik, J., & Thorne, C. (2017, March). Exploring the relation between semantic complexity and quantifier distribution in large corpora. *Language Sciences*, *60*, 80–93. doi: 10.1016/j.langsci.2017.01.006

Szymanik, J., & Zajenkowski, M. (2009). Understanding Quantifiers in Language. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

Tamariz, M., & Kirby, S. (2016, April). The cultural evolution of language. *Current Opinion in Psychology*, *8*, 37–43. doi: 10.1016/j.copsyc.2015.09.003

Tenenbaum, J. B., & Griffiths, T. L. (2001, August). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. doi: 10.1017/S0140525X01000061

van de Pol, I., Steinert-Threlkeld, S., & Szymanik, J. (2019, May). *Complexity and learnability in the explanation of semantic universals of quantifiers* (Preprint). PsyArXiv. doi: 10.31234/osf.io/f8dbp

van Fraassen, B. C. (1966). The Completeness of Free Logic. *Mathematical Logic Quarterly*, *12*(1), 219–234. doi: 10.1002/malq.19660120117

Van Tiel, B., Van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016, February). Scalar Diversity. *Journal of Semantics*, *33*(1), 137–175. doi: 10.1093/jos/ffu017

Vehtari, A., Gelman, A., & Gabry, J. (2017, September). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. doi: 10.1007/s11222-016-9696-4

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2019, July). Pareto Smoothed Importance Sampling. *arXiv:1507.02646 [stat]*.

Vennemann, T. (1972). *Semantic structures: A study in the relation between semantics and syntax.* Athenäum Verlag.

Verheyen, S., & Égré, P. (2018). Typicality and Graded Membership in Dimensional Adjectives. *Cognitive Science*, *42*(7), 2250–2286. doi: 10.1111/cogs.12649

von Heusinger, K., Maienborn, C., & Portner, P. (2011). *Semantics*. Walter de Gruyter.

Wales, D., & Doye, J. (1997, July). Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *The Journal of Physical Chemistry A*, *101*(28), 5111–5116. doi: 10.1021/jp970984n

Zhao, Z. (2019). Interpreting Intensifiers for Relative Adjectives: Comparing Models and Theories. In J. Sikos & E. Pacuit (Eds.), *At the Intersection of Language, Logic, and Information* (pp. 213–224). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-662-59620-3_14

Zhao, Z., & Cremers, A. (n.d.). *A Prior-Uncertainty Model for Gradable Adjectives*. Amsterdam.